Factors Affecting Employment Duration in the Food Retail Industry

BEATA SUNDLING & LOVA HÖFT
Abstract

Measuring and tracking the employee turnover rate is a crucial part when evaluating a company’s performance. An important part of this is measuring the employment duration within an organization. The purpose of this report is to investigate if employment duration in a food retail company can be explained by predetermined variables using multiple linear regression. Data from five years ago until today has been collected and processed to analyze and fit the best choice of the linear model. Gender, employment rate, industry experience and age are the predictors used for conducting the analysis. The result shows that a low linear correlation can be seen between employment duration and the explanatory variables: gender, employment rate, industry experience and age. In the discussion, the results are analyzed as well as potential problems and improvements of the regression.
Sammanfattning

4.7 Bootstrapping ......................................................... 21

5 Analysis & Discussion .................................................. 21
  5.1 Analysis of Methodology ........................................... 21
  5.2 Analysis of Original Model ........................................ 22
  5.3 Analysis of Final Model ........................................... 23
  5.4 Future Research .................................................. 23

6 Conclusion .......................................................... 24
Acknowledgements

This bachelor thesis was written by Beata Sundling and Lova Höft as part of a five-year M.Sc program in Industrial Engineering and Management at KTH Royal Institute of Technology in Stockholm, Sweden.

First, we want to express our gratitude towards our collaboration company for trusting us with this thesis. We want to thank our supervisor at the collaboration company for her help, support and valuable knowledge about the industry. Second, we would like to thank our mathematical supervisor at KTH, Isaac Ren, for giving advice and guidance when it comes to the mathematical aspects of the thesis. Finally, we want to thank our Industrial Engineering and Management supervisor, Hans Lööf.
1 Introduction

1.1 Background

1.1.1 Employee Turnover Rate and Employment Duration

The employee turnover rate is a metric that shows “the percentage of employees that leave your organization during a given time period” [10]. Often the employee turnover rate is measured annually, but it can also be measured quarterly or monthly. In equation 1, the formula for calculating the employee turnover rate is stated.

\[
\text{Annual turnover } \% = \frac{\text{Employees leaving}}{\text{Average number of employees}} \cdot 100
\]  

(1)

Employee turnover rate is an important metric for every company to track since it can have significant effects on a company’s success. For example, a high turnover rate can increase costs and have a negative impact on the company’s culture and reputation. Not only are recruiting, hiring and training costly, but a high turnover comes with indirect costs such as decreased productivity. The reputation might also be affected if potential applicants interpret the high turnover rate as a sign of poor leadership or a bad work environment. In addition, frequent turnover can lead to decreased engagement among employees and loss of talent. However, a turnover that is too low should also be avoided since it can limit diversity and innovation. New employees bring new ideas and perspectives, which enhance innovation and challenge old ideas and routines within the organization. [6] Ultimately, the turnover should not be too high or too low. Hence it is of great interest for organizations to understand what factors drive the turnover rate and be able to predict how long the employees will stay with the company.

The employee turnover rate cannot be calculated on an individual basis and is often used as an organisational Key Performance Indicator (KPI). Therefore, a measurement that can be used when looking at individual employees is employment duration, which can be described as the length of employment. If the employment duration is long for the majority of employees, the employee turnover will decrease and vice versa if the employment duration is short. Hence, tracking employment duration can be a useful tool to monitor employee retention and engagement within an organization.

1.1.2 Trade and Retail Industry

The trade industry is the largest industrial branch in Sweden and in 2020 about 11% of all Swedish employed worked in the industry. The trade industry consists of the retail industry, the wholesale industry and the car trade. The retail industry is the branch where most people are employed out of the three [4]. In 2020 the retail industry employed about 220 000 people with 25% being young adults aged 16 to 24. In the last 20 years, the
number of employed in the retail industry has grown with a compounded annual growth rate of about 0.8% with a slight dip in 2020, potentially due to the COVID-19 pandemic [4]. Retail is defined as “the activity of selling goods to the public, usually in shops” and includes, for example, grocery stores, clothing stores and drug stores [3]. Furthermore, the retail industry is known for having a high employee turnover rate. In Sweden, the employee turnover rate was 32% in 2016 with plenty of people considering to change company or industry. Even though a high turnover rate is associated with higher costs and sometimes inadequate working conditions, the high turnover rate can be seen as more natural when it comes to the retail industry [5]. The reason for this might be that the industry is characterized by having different types of employment and diversity among its employees. There are full-time and part-time employments, permanent and temporary employments, different seniority levels, several departments to work at as well as several age groups working in the industry. For example, students and younger adults might work part-time during evenings and weekends and have the intention to work in the industry for a shorter period of time [5]. Store managers and more senior employees that usually have more expertise and experience might on the other hand work at the company for a longer period as their primary job and profession.

1.1.3 Food Retail Industry

Grocery stores are one type of retail store where different types of food items are sold, as well as some other necessary products such as hygiene products. The Swedish food retail industry is dominated by five main actors: ICA, COOP, Axfood (brands such as Hemköp and Willys), LIDL and Bergendahls (brands such as City Gross). About 110 000 people in Sweden are employed in grocery stores and over 3000 grocery stores are located in Sweden. Additionally, the annual revenue in the grocery store industry in 2021 was about 340 billion SEK. A large share, about 32%, of people working in grocery stores are young adults aged 16 to 24. [2]

1.2 Purpose

The purpose of this report is to examine what predetermined factors drive employee turnover in the food retail industry and if it is possible to predict how long the employees will stay with the company based on these factors. This will be investigated by performing multiple linear regression with employment duration as the response variable and four predictors in the initial model - employment rate, industry experience, age and gender. The company in focus is active in the food retail industry and runs grocery stores across the country. The company will throughout this report be referred to as “The Company”.

8
1.3 Research Question

The research question that will be investigated in this report is stated below.

- To what extent is employment duration in the food retail industry driven by employment rate, industry experience, age, and gender?

1.4 Delimitations and Scope

There are some limitations that should be considered to scope a manageable and relevant project. First, the data that has been provided by The Company spans from January 2018 until March 2023. Data before January 2018 will not be considered as well as data after March 2023. This means that the data is censored, which might result in biased and skewed results and conclusions. Second, only people employed in retail will be included in the analysis, meaning people working at other departments, such as the finance department, R&D, etc., are excluded from the analysis. Third, the predictors have been chosen based on discussions with The Company, company data available and hypotheses on which factors might affect the turnover rate. Even so, other factors, not included in or available for our analysis, might have an impact on the response as well.

1.5 Earlier Research

Plenty of research has been done when it comes to factors, both qualitative and quantitative, that affect the employee turnover rate in the retail industry. For example, Simon Booth and Kristian Hamer wrote a research paper on “Labour turnover in the retail industry” in the UK and concluded, by doing a stepwise regression, that environmental factors and company culture factors had a significant influence on the turnover rate [1]. Several papers have also been written on employee turnover in specific markets, for example, the Indian market [8], Chinese market [11] and Lebanese market [9] where the data was collected from self-completed questionnaires.

Common findings in these papers were relationships between a high turnover and low job satisfaction, high level of work-related stress, absence of motivation and poor organisational commitment. Also, all of these studies have been conducted using questionnaires that rely on employees’ thoughts. In addition, all studies are limited to particular geographical markets. However, this paper aims to investigate how selected numeric variables, which are automatically collected for all employees, might affect the employee turnover in the Swedish food retail industry.
2 Mathematical Definitions

2.1 Regression Analysis

To investigate the relationship between the dependent and the independent variable, a multiple regression analysis is conducted. Below a multiple linear regression model is stated, where $i$ is the index of the datapoint. \[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \] (2)

2.2 Residual Analysis

The residual can be viewed as the deviation between the data and the fit. Plotting the residuals can be an effective way to check how well the data fit the model, as well as check the assumptions that should apply. \[ \] [7]

2.3 Outliers and Influential Points

Outliers are defined as datapoints that differ considerably from the rest of the data. Influential points are observations in a dataset that have a significant impact on the result. \[ \] [7]

2.4 Multicollinearity

Multicollinearity is defined as the phenomenon where two or more regressors are highly correlated. It is important to be aware of this since it can affect the result and be difficult to distinguish the effect of each variable on the dependent variable. \[ \] [7]

2.5 Variable Selection

Variable selection is one of the most effective techniques when dealing with multicollinearity. It does not guarantee the elimination of multicollinearity but it helps justify or eliminate regressors in the final model. \[ \] [7]

2.6 Cross Validation

Cross-validation is used to evaluate the performance of the regression model. It involves partitioning the available data into subsets, using one subset to train the model and the other subset to test its performance. \[ \] [7]
2.7 Bootstrapping

Bootstrapping is a technique that involves resampling the available data to obtain multiple subsets and estimating the model parameters on each subset, allowing for the assessment of the model’s stability and accuracy. [7]

3 Methodology

3.1 Data Collection

The dataset provided by The Company includes information about every employee that has worked at The Company between 2018-01-01 and 2023-03-01. A total of 17309 employees were initially included in the dataset before processing the data, giving us 17309 initial datapoints. The dataset was anonymized by giving every employee a unique identification number. Also, the dataset was structured by giving each employee one row for each month working at The Company with corresponding monthly information. One employee could have up to 63 rows in the dataset depending on how many months they have worked. When an employee quits, the dataset continues with the next employee for a total of about 500000 rows. The data was processed in order to retrieve a dataset where one employee equals one datapoint.

3.2 Data Processing

3.2.1 Response Variable

The regression analysis was conducted with employment duration as the response variable. This variable was created by counting the number of rows for each employee in the dataset and could range from one month to 63 months.

3.2.2 Predictors

In the initial model, a total of four predictors were used in the regression analysis; gender, industry experience, employment rate and age.

Gender was included in the analysis to investigate whether it affects an employee’s employment duration. In the dataset, each employee was marked as either a man, woman or empty. When no gender was provided, it could have been the result of the employee not wanting to provide their gender, being non-binary or The Company not having collected the information. In total, three employees did not have the gender man or woman and these were excluded from the analysis. Since gender is a binary variable all men were given the value 0 and all women were given the value 1.
Moreover, industry experience was a relevant predictor for the analysis. Industry experience shows how many days the employee has worked in the food retail industry, measured from the last time the dataset was updated, 2023-03-01. Consequently, a person that has been working in the industry for a longer period has more days than a person that has been working for a shorter period of time. Some employees did not have data provided for industry experience. The reason for this could have been that The Company did not collect it from the employee when starting at The Company or that this variable was not calculated when The Company extracted the dataset. All employees that had an empty industry experience were deleted from the dataset used in the regression. In total 4154 datapoints were erased.

The next predictor included in the regression was employment rate. The employment rate shows, in percentage, how much an employee works. An employment rate of 100% means a working week of 40 hours. However, the employment rate range between 2% and 100% amongst the employees. This variable is investigated since it is interesting to see if there is a relationship between how long an employee stays at the company and how much the employee works. It is worth noting that the employment rate of each employee can fluctuate on a monthly basis. To address this issue, approximately 50% of all datapoints were removed by excluding employees who had a significant fluctuation in their employment rate. Even though individuals were deleted, the dataset still included over 7000 datapoints, more precisely 7151 employees which was the final number of employees included in the regression.

Lastly, age was used as a predictor in the regression. The age variable was calculated as an average over all months the employee worked at The Company. A hypothesis was that age could correlate with industry experience since an older person has had a longer time to collect experience within the industry, however, this is a generalization and might not be the case all the time.

### 3.2.3 Type of Variables

The variables and their value types are summarized below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Response/Regressor</th>
<th>Type of variable</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment duration</td>
<td>Response</td>
<td>Quantitative</td>
<td>Months</td>
</tr>
<tr>
<td>Gender</td>
<td>Regressor</td>
<td>Binary</td>
<td>0 or 1</td>
</tr>
<tr>
<td>Employment rate</td>
<td>Regressor</td>
<td>Proportion</td>
<td>%</td>
</tr>
<tr>
<td>Industry experience</td>
<td>Regressor</td>
<td>Quantitative</td>
<td>Days</td>
</tr>
<tr>
<td>Age</td>
<td>Regressor</td>
<td>Quantitative</td>
<td>Years</td>
</tr>
</tbody>
</table>

Table 1: Variables including their type and unit
3.3 Variable Selection and Model Creation

The initial model included all predictors described in table 2 and was inserted in the model presented below.

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \text{ where } i = 1, \ldots, 7151 \]  

\[ (3) \]

| \( x_{i1} \) | age       |
| \( x_{i2} \) | gender   |
| \( x_{i3} \) | industry experience |
| \( x_{i4} \) | employment rate |

Table 2: Predictor variables \( x_{ik} \)

The model was first evaluated based on \( R^2 \) value and adjusted \( R^2 \) value to see how well the model fitted the data. Thereafter the linearity assumption was checked by doing a residual analysis using both unscaled residuals and the square root of the standardized residuals. The normality assumption was checked by doing a Q-Q plot. Thereafter, Cooks’ distance was used to find and eventually eliminate outliers and influential points. Later, multicollinearity was examined through Variance Inflation Factor’s (VIF’s) and when finding variables with a VIF above five, these variables were excluded from the model. Variable selection was performed based on four different evaluation criteria; \( R^2 \), adjusted \( R^2 \), BIC and Mallow’s \( C_P \). Both backward elimination and forward selection were used to find the best fitted model based on our data. Cross validation was performed and two models, one with four regressors and one with three regressors, were compared and evaluated. Finally, the initial model parameters were estimated and its accuracy was checked using bootstrapping confidence intervals.

3.4 Limitations

Limitations regarding the methodology can be related to the data processing, see section 3.2.

4 Results

4.1 Initial Model

To recall, the following equation shows the initial model with four explanatory variables.

\[ y_i = \beta_0 + \beta_1 (\text{Gender}) + \beta_2 (\text{EmploymentRate}) \]
\[ + \beta_3 (\text{IndustryExperience}) + \beta_4 (\text{Age}) + \epsilon_i \text{ where } i \]
\[ = 1, \ldots, 7151 \]

(4)
In table 3 one can observe the goodness of fit of the initial model. The adjusted $R^2$ is 0.5529 which can be interpreted as if 55.29% of the variance of the response variable is explained by the explanatory variables.

<table>
<thead>
<tr>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Degrees of freedom</th>
<th>Residual standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5532</td>
<td>0.5529</td>
<td>7146</td>
<td>15.78</td>
</tr>
</tbody>
</table>

Table 3: Goodness of fit, initial model

In table 4 the p-values for each regressor can be observed. Since the p-values for gender, employment rate and industry experience are less than 0.05, it implies that there is a significant relationship between employment duration and these variables.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.00144</td>
</tr>
<tr>
<td>Employment rate</td>
<td>$&lt; 2e-16$</td>
</tr>
<tr>
<td>Industry Experience</td>
<td>$&lt; 2e-16$</td>
</tr>
<tr>
<td>Age</td>
<td>0.18539</td>
</tr>
</tbody>
</table>

Table 4: P-values, initial model

4.2 Residual Analysis

Initially, a residual analysis was conducted to verify the linearity assumption, ensuring that the connection between our response variable and explanatory variables exhibits linearity. Figure 1 shows the initial residual analysis made on all datapoints and with no data transformation. One can observe that there is a pattern in the plot. The dots are laying in a horizontal band that is slightly decreasing, meaning that the linearity condition is violated.

The normal Q-Q plot should be approximated as a straight line, which is the case in figure 1. There is a tail in the bottom left part of the figure which can be an indication that the residuals are not normally distributed [7]. However, the tail is not that heavy meaning it can assume that the residuals are roughly normally distributed which is preferred.

Furthermore, the scale-location plot in figure 1 shows a V-shaped pattern where the tip of the V is pointing at 63, the maximum value of months worked. Also, one can see a more heavy density around values 5-40 and thereafter a funnel shape which is a sign of heteroscedasticity, that the variance of the residuals increases as the fitted values increase.

The residual vs leverage plot in the bottom right corner of figure 1 indicates that there is no influential point in our data set.
To improve the relationship between the response variable and the regressors, the response variable was processed in different ways. First, all employees that had worked for 63 months were removed from the dataset since many employees had worked for the whole period giving the distinct upper line in the residual vs fitted plot in figure 1. Also, the response variable was transformed using a logistic (logit) transformation. This was done by converting the number of months to a probability, $p$, where a probability of 0 means an employee worked at The Company for zero months and a probability of 1 means an employee worked at The Company for 63 months. Then the logit transformation was used, shown in equation 6.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

A logit transformation is used when the response variable is restricted to an absolute
interval, in this case [1,63]. Lastly, 40 datapoints were found with an age variable of 120 or above and these were removed. Figure 2 shows plots using this new model.

In figure 2, the residual vs fitted plot shows a more random pattern than in figure 1 and the red line is more or less centralized around zero and horizontal even though it is following the same pattern as in figure 1. This indicates that the linear model fits better after transforming and processing the data further.

The normal Q-Q plot shows that the normality condition is fulfilled. When it comes to the scale-location plot one can observe that the strange V-shape is gone in comparison with figure 1. However, the density is still high on the left side of the scale-location plot with some pattern of a funnel, indicating non-homoscedasticity. The residual vs leverage plot shows no sign of high influence points.
Table 4 shows the goodness of fit for the transformed model. The adjusted $R^2$ of 0.2167 indicate that about 22% of the variance in employment duration can be explained by the predictors. Since the adjusted $R^2$ was worse for this transformed model, the regression analysis proceeded with the initial model.

<table>
<thead>
<tr>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Degrees of freedom</th>
<th>Residual standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2172</td>
<td>0.2167</td>
<td>5720</td>
<td>0.616</td>
</tr>
</tbody>
</table>

Table 5: Goodness of fit, transformed model

4.3 Outliers and Influential Points

Cook’s distance was used to detect high influence datapoints and in figure 3, a Cook’s distance plot can be observed. There were three datapoints that were pointed out as high influential points: datapoint 163, 204 and 325. These were manually analyzed to detect why they differ from the rest and in order to decide if they should be excluded or not. Datapoint 163 differed because there was a 65 years old person with 45 years of industry experience that only worked at The Company for one month. Even though this is odd, it might still be an accurate datapoint if the person in question started to work in the industry when being a young adult. Datapoint 204 and 325 were considered influential, for the same reason. The Cook’s cutoff value [7] was calculated to 0.87. Since none of the potentially influential datapoints exceeded that value at the same time as it did not seem to be incorrect data collection when it comes to these points, no point was excluded from the model. This lies in line with the Residual vs Leverage plot in figure 1 and where no datapoint exceeds Cook’s distance.

![Figure 3: Cook’s distance](image)

Figure 3: Cook’s distance
4.4 Multicollinearity

4.4.1 Covariance Matrix

The covariance matrix shows the variance of each variable on the diagonal and the covariance between each pair of variables in the off-diagonal elements. When two or more predictor variables exhibit a strong linear relationship, their correlation coefficient absolute value approaches one. As a result, if the values are not near one, it is possible to detect non-multicollinearity. This relationship can be visualized by plotting the regressors against one another. Age and industry experience have the highest absolute value of their correlation coefficient. Hence, these predictors are the most correlated. The reason for this could also be explained by the fact that younger people naturally have less industry experience compared to older people.

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Employment Rate</th>
<th>Industry Experience</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.000000</td>
<td>-0.090867</td>
<td>-0.028713</td>
<td>-0.036357</td>
</tr>
<tr>
<td>Employment Rate</td>
<td>-0.090867</td>
<td>1.000000</td>
<td>0.498256</td>
<td>0.400308</td>
</tr>
<tr>
<td>Industry Experience</td>
<td>-0.028713</td>
<td>0.498256</td>
<td>1.000000</td>
<td>0.592052</td>
</tr>
<tr>
<td>Age</td>
<td>-0.036357</td>
<td>0.400308</td>
<td>0.592052</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

Table 6: Covariance matrix

4.4.2 VIF

To analyze multicollinearity, another approach is to utilize Variance Inflation Factors (VIFs). When a predictor variable has a high VIF value, it indicates a significant correlation with other predictor variables. If the VIF value exceeds 5 or 10, it suggests the presence of multicollinearity. Since non of the calculated VIF values are greater than 5 or 10, the detected multicollinearity is considered low.

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Employment Rate</th>
<th>Industry Experience</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.008816</td>
<td>1.371244</td>
<td>1.760813</td>
<td>1.575685</td>
</tr>
</tbody>
</table>

Table 7: VIF Values

4.5 Variable Selection

Not all predictors used are necessarily important. To find the most important subset of predictors, variable selection is used. This includes forward selection and backward elimination. To evaluate the performance of the predictors, $R^2$, Adjusted $R^2$, BIC and Mallow’s $C_P$ are used as measurements.
4.5.1 Forward Selection

According to forward selection, the model will perform best with three or four predictors. \( R^2 \) and Adjusted \( R^2 \) suggest that the model should include all four predictors, while BIC and Mallow’s \( C_P \) suggest that age should be excluded.

![Figure 4: Forward selection](image)

<table>
<thead>
<tr>
<th>Regressors</th>
<th>Gender</th>
<th>Employment Rate</th>
<th>Industry Experience</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 8: Forward selection regressors

<table>
<thead>
<tr>
<th></th>
<th>Max ( R^2 )</th>
<th>Max Adjusted ( R^2 )</th>
<th>Min BIC</th>
<th>Min Mallow’s ( C_P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.5531637</td>
<td>0.5529136</td>
<td>4.754187</td>
<td>-5723.325</td>
</tr>
<tr>
<td>Regressors</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 9: Forward selection
4.5.2 Backward elimination

<table>
<thead>
<tr>
<th>Regressors</th>
<th>Gender</th>
<th>Employment Rate</th>
<th>Industry Experience</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>YES</td>
<td></td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 10: Backward elimination regressors

<table>
<thead>
<tr>
<th>Max $R^2$</th>
<th>Max Adjusted $R^2$</th>
<th>Min BIC</th>
<th>Min Mallow’s $C_P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.5531637</td>
<td>0.5529136</td>
<td>4.754187</td>
</tr>
<tr>
<td>Regressors</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 11: Backward elimination

Similar to forward selection, backward elimination suggests that the model will perform best with three or four predictors. In addition, $R^2$ and Adjusted $R^2$ also suggest that the model should include all four predictors, while BIC and Mallow’s $C_P$ suggest that age should be excluded.
4.6 Cross Validation

Cross validation is used to compare two models. One model consists of all original predictors, while the other one has excluded age. The results show that the models perform similarly in regards to Max Adjusted $R^2$ and Mean Squared Error. However, there is a slight improvement to the model when age is excluded.

<table>
<thead>
<tr>
<th>Max Adjusted $R^2$</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5536</td>
<td>249.0174</td>
</tr>
</tbody>
</table>

Table 12: Model with four regressors

<table>
<thead>
<tr>
<th>Max Adjusted $R^2$</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5537</td>
<td>248.9939</td>
</tr>
</tbody>
</table>

Table 13: Model with three regressors

4.7 Bootstrapping

By using bootstrapping, it is possible to estimate the confidence intervals of the model. The confidence intervals for Months Employed are compared to the estimated coefficient. The confidence interval contains the estimated coefficient but the interval is still quite large. This implies that the model parameters vary.

<table>
<thead>
<tr>
<th>Confidence interval</th>
<th>Estimated coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 3.511, 5.932 )</td>
<td>4.752</td>
</tr>
</tbody>
</table>

Table 14: Confidence intervals and estimated coefficient

5 Analysis & Discussion

5.1 Analysis of Methodology

When it comes to the methodology, there is room for improvement and further development of data processing techniques in order to obtain more accurate results and conclusions. Firstly, individuals lacking industry experience were excluded from the dataset. Ideally, it would have been desirable to include these individuals in the regression analysis. However, due to the unavailability of the necessary data from The Company, there was no feasible workaround for this issue.

Secondly, employees with significant fluctuations in their employment rates were eliminated, which accounted for approximately half of the data points. The approach taken to
address this problem was the removal of these specific data points. An alternative strategy could have involved categorizing employment rates into different ranges, such as 0-25%, 25-50%, etc., thereby avoiding the need for data point deletion. Nevertheless, this alternative approach would not have fully resolved the issue since some employees exhibited varying employment rates, such as 10% in one month and 90% in another, unless using a category called "other".

5.2 Analysis of Original Model

The original model had an adjusted $R^2$ value of about 0.55 indicating that there is a low correlation between employment duration and gender, employment rate, industry experience and age. A good adjusted $R^2$ value depends on what relationship is considered, but in this case, a higher value would mean a stronger correlation between the variables investigated, which is preferable. Based on this it is hard to draw a conclusion that it is significant that employment duration would depend on gender, employment rate, industry experience and age.

Moreover, the linearity condition was violated in the initial model. The plot showed signs of patterns that should not be present if a linear relationship is present between the variables investigated. Some problems that might have caused this were first that our response variable was a discrete variable only considering distinct values between 1 and 63 months. Second, the response variable was a finite variable that could not exceed 63 and could not be lower than one. Employment rate was also a finite variable ranging from zero to 100 which could also cause problems in the model. To get around these problems, several transformations were tested as well as removing datapoints. For example, a logistic transformation was used for employment duration. Additionally, all employees that had worked at The Company for 63 months were removed since it was a significant amount of people that had worked for the whole duration that the dataset covered. Even though testing different transformations and removing datapoints the residual vs fitted plot still showed signs of non-linearity, however, the transformed model plot looked more accurate than the original model. One could also say that the normal Q-Q plot was better for the transformed model since its tails did not fluctuate much.

On one hand, when comparing the original model with the transformed model based on the residual vs fitted plot, normal Q-Q-plot, scale-location plot and residual vs leverage plot the transformed model seemed to be a better fit. On the other hand, when comparing the $R^2$ and adjusted $R^2$ value the initial model showed a better fit. The adjusted $R^2$ value for the transformed model was about 0.22 which tells us that 22% of the variance in the employment duration can be explained by the regressors meaning that there is a very low correlation between the response and the regressors. Hence, the regression analysis proceeded with the initial model.
When identifying and evaluating influential points, Cook’s distance was used and since the Cook’s cutoff was higher than Cook’s distance for every outlier, no point was removed from the model. This might indicate good data quality and that most datapoints were correctly collected and measured.

Furthermore, the original model showed low dependence between the regressors which is preferable. As mentioned in section 4.4, all Variance Inflation Factors (VIFs) had a value lower than five and hence there was no clear dependence between any of the regressors. Anyway, following logical reasoning, industry experience and age were the two regressors that correlated the most with a covariance coefficient of about 0.6.

Lastly, when performing variable selection with forward selection and backward elimination, both methods concluded that all four regressors should be used in the final model when looking at $R^2$ and adjusted $R^2$ values. However, both BIC and Mallow’s $C_P$ concluded that the best model would be to exclude the age variable and keep the other three. This model will be analysed further in the next paragraph.

### 5.3 Analysis of Final Model

The final model consists of three regressors: gender, industry experience and employment rate. Although it is important to note that this model only performed slightly better than the model containing all four predictors according to cross validation. The reason for this might be that age and industry experience are correlated. If age is to be eliminated, it will not have a large impact on the final result. However, when looking at variable selection, BIC and Mallow’s $C_P$ suggested that age should be removed while $R^2$ and adjusted $R^2$ suggested that all the predictors should be kept in the model. Therefore, we can conclude that it does really matter if age is included in the final model or not. The value of adjusted $R^2$ according to cross validation in the final model was 0.5537. This number suggests a low correlation.

### 5.4 Future Research

Further investigation is required to discover a suitable model that accurately describes the relationship and aligns with the data in order to draw a reliable conclusion. The results indicate that the linear model fails to meet the data’s requirements, as demonstrated by the residual analysis. Therefore, alternative nonlinear models should be explored. Moreover, since the data is censored, it is advisable to consider other censored regression models. Additionally, the inclusion of additional predictors such as salary and education level could potentially enhance the model and identify other factors that may influence the relationship. To ensure a reliable conclusion, it is imperative to examine additional data from an earlier time period.
6 Conclusion

To conclude, the aim of this thesis was to investigate whether there is a significant linear relationship between employment duration and gender, employment rate, industry experience and age. The research question was stated as follows:

- To what extent is employment duration in the food retail industry driven by employment rate, industry experience, age and gender?

Our final results show that employment duration in the food retail industry is to some extent driven by employment rate, industry experience, age and gender. The p-values calculated show that there is a significant relationship between employment duration and gender, industry experience and employment rate. However, the adjusted $R^2$ value shows that our model is not a perfect fit for the data. The residual analysis concludes that the linearity condition is violated. To find a better model, other models have to be considered. Based on the analysis, a conclusion that can be drawn is that the regressor age can be excluded from the model. The age parameter does not add any information to the relationship and an easier model was found excluding age. What can be said, based on the analysis and models tested, is that more research and analyses need to be conducted to find an appropriate model that explains the relationship in a great way.

References


