OPIM 5511

Survival Analysis with SAS

Project Report

**Team 3**

Kumari Juhi

Arpita Ghosh

Manishi Kumari

Sourav Sethia

Contents

[Project Overview 2](#_Toc17394064)

[Business Case 2](#_Toc17394065)

[Data Preprocessing 3](#_Toc17394066)

[Data Exploration 4](#_Toc17394067)

[Hazards for all Event Types 15](#_Toc17394068)

[Model Building 19](#_Toc17394069)

[COX Proportional Hazards Modelling Technique 19](#_Toc17394070)

[Tied Dataset 20](#_Toc17394071)

[Initial Assumptions for the Model 20](#_Toc17394072)

[Variable Handling 20](#_Toc17394073)

[Time-Dependent Variable 20](#_Toc17394074)

[Non-Proportionality Identification 20](#_Toc17394075)

[Martingale Residuals: 20](#_Toc17394076)

[Schoenfeld Residuals: 22](#_Toc17394077)

[Conclusion 32](#_Toc17394078)

[Recommendations 33](#_Toc17394079)

# Project Overview

Among all business-related problems, employee attrition is one of the key issues in today's scenario. Many of the employees tend to leave the job for various undisclosed factors, such as lack of job security, lack of career advancement, desire for change to new opportunities, anticipation of higher pay, problems with supervisors and few other personal reasons.

Larry Hansen, COO of a FermaLogis, a pharmaceutical company is dealing with attrition in his company. He is investing more on various employee development programs like professional training for new recruits to boost their productivity, executive development programs for experienced employees, and is trying to get benefits of employee retention out of all these investment programs. As employees learn valuable skills and become more competent each year, they are hired by the rival companies with higher salary.

This study helps in understanding the reasons why FermaLogis is experiencing heavy employee attrition and suggests some measures in retaining employees. Survival analysis with SAS is the best tool to analyze the current issue of attrition of Fermalogis with the available information like demographics and other factors.

# Business Case

Companies are poaching the employees of Fermalogis once they attain skills that are in demand in the Market. The company needs to identify the employee sub-groups that are most prone to attrition in order to understand the underlying reasons for the exits. The company is planning to solve this by seeking the help of an analytics company proficient in survival analysis of data to investigate and probe the key determinants of attrition and develop a report with the compelling discoveries and recommendations to Fermalogis.

# Data Preprocessing

Preprocessing of data is the first step towards building any machine learning model. Individual feature/attribute study would help to understand the data better. The FermaLogis dataset has details of 1470 employees aged between 18 and 60, with 76 variables which help in analyzing attrition. The table below has key attribute details which highly contribute towards identifying the reason and rate of attrition.

|  |  |
| --- | --- |
| **Features** | **Description** |
| BusinessTravel | Describes if the employee had opportunity to travel |
| EnvironmentSatisfaction | Rates the level of work environment satisfaction |
| Department | States the department in which the employee works |
| Education | Indicates the highest education degree of the employee in the organization |
| OverTime | Denotes whether the employee works overtime or not |
| YearsAtCompany | Specifies the number of years at FermaLogis |

Figure 1: Data Description

Further detailed study of features indicates that most of the variables are categorical, like attrition, gender marital status, etc. and we tried to recode them as numeric for ease of modeling.

A new variable **Turnover Type** has been created as follows:

* If type= “0” then Turnover Type = No turnover
* If type= “1” then Turnover Type = Retirement
* If type= “2” then Turnover Type = Voluntary Resignation
* If type= “3” then Turnover Type = Involuntary Resignation (Health Problem, Family problems)
* If type= “4” then Turnover Type = Job termination

(Note: If Type = 0 then no employee is leaving the organization)

There are few variables which have been recoded as follows:

**Business Travel:**

0 = Non-Travel

1 = Travel rarely

2 = Travel frequently

**Over Time:**

0 = not worked overtime

1 = worked overtime

**Gender:**

0 = Female

1 = Male

**Marital Status:**

0 = Single

1 = Married

2 = Divorced

**Department:**

0 = Human Resources

1 = Research & Development

2 = Sales

**Education Field has been recoded based on <=3 and >3**

0 = educ field from 0 - 3

1 = High educ from 4 - 5

**Job Satisfaction has been recoded as follows:**

0 = Not satisfied

1 = Satisfied

There are few redundant variables that do not give any significant information about employee attrition, like Employee number, standard hours over 18, etc.; hence, these variables were removed from our analysis.

# Data Exploration

Data exploration is one of the key steps in data analysis and typically involves an unstructured way to draw initial patterns, characteristics, and points of interest. It helps to understand the data and the relationship between the independent variables and its impact on the target variable.

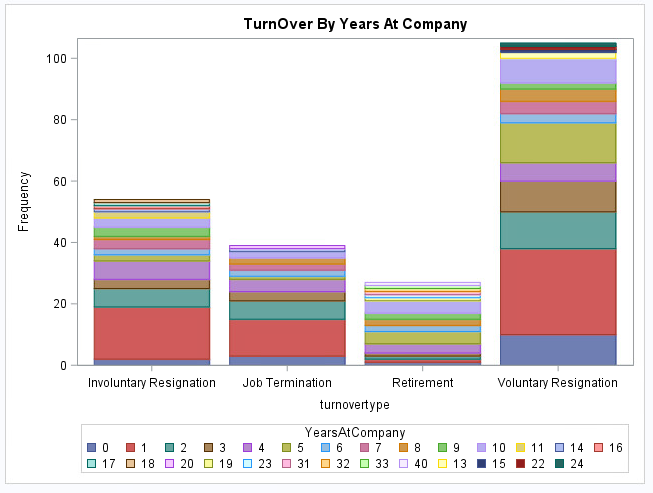
****

Figure 2: Distribution of TurnOver with YearsAtCompany

We can see a very wide range of employee experience within the company. Most of the employees have been with the company for less than or equal to 10 years. There are very few employees with FermaLogis for more than 10 years. It can be observed from the plot that employees who are there in the company for less than 5 years are more likely to leave the company than other employees, but the employees who are there with the company for 1-year are the ones whose attrition rates are the highest.

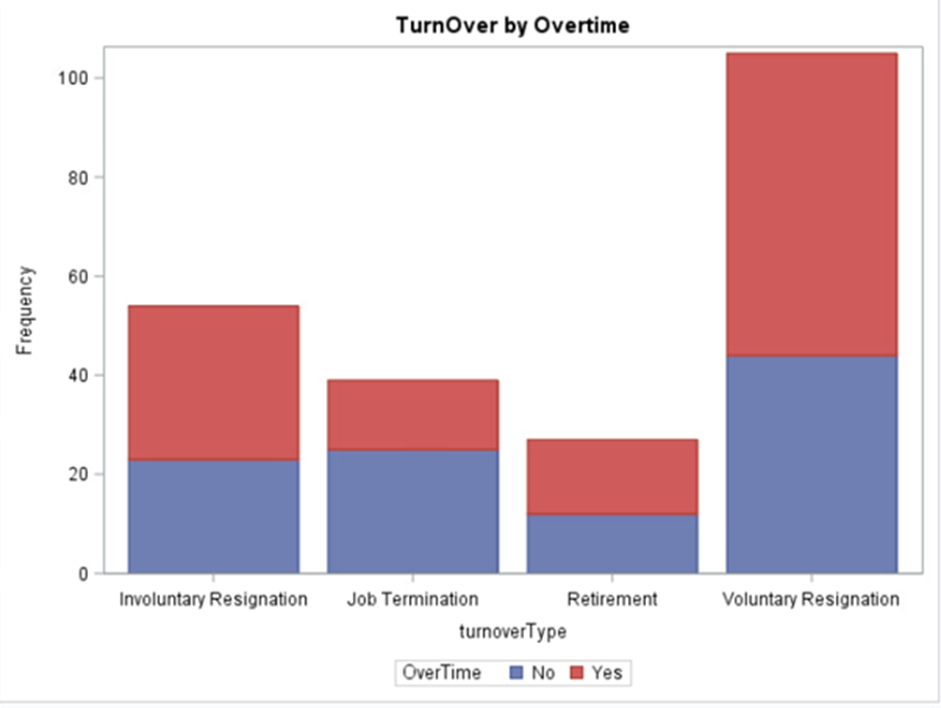


Figure 3: Distribution of TurnOver with OverTime

It can be observed from the plot above that the employees who work overtime are more likely to go for Voluntary Resignation. The company is not interested in employees who are not willing to work overtime, and hence terminate those employees.

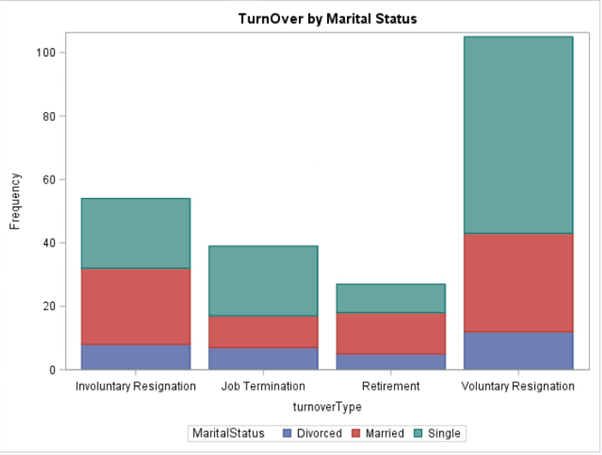


Figure 4: Distribution of TurnOver with MaritalStatus

It can be observed from the above plot that the employees whose marital status is single are more likely to go for Voluntary Resignation. It may be inferred that the employees who are single have less responsibility as compared to people who are married, so married people are prone to stay more with the company; hence, the company would like to have married employees instead of single ones.

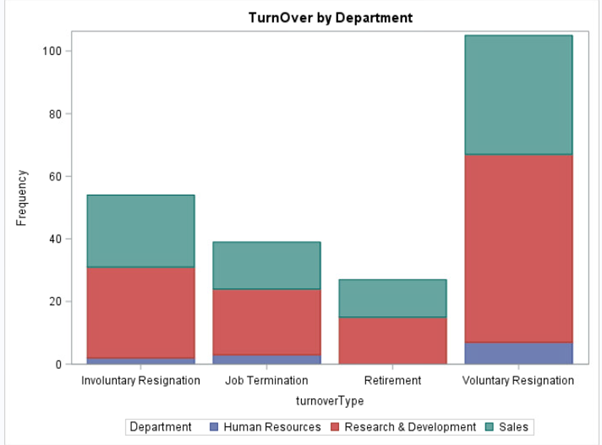


Figure 5: Distribution of TurnOver with Department

It can be observed from the above plot that the employees who work with Research & Development department are more likely to go for Voluntary Resignation. We can even say that employees working with Research & Development department are more likely to be terminated and the reason could be that the company did not like their performance. The employees who are working with Human Resources department are really doing good in the company at all levels of turnover types.

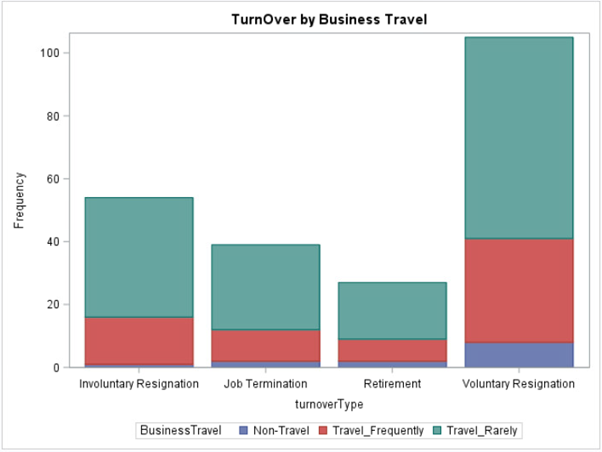


Figure 5: Distribution of TurnOver with BusinessTravel

The relation between the variable turnoverType and Business Travel indicates that the employees who travel frequently are more likely to leave the company, and the employees who travel rarely are also more likely to leave the company. So, altogether we can say that the employees who are expected to travel for any job purpose either frequently or rarely, leave the company and the reason could be that they do not like to travel at all.

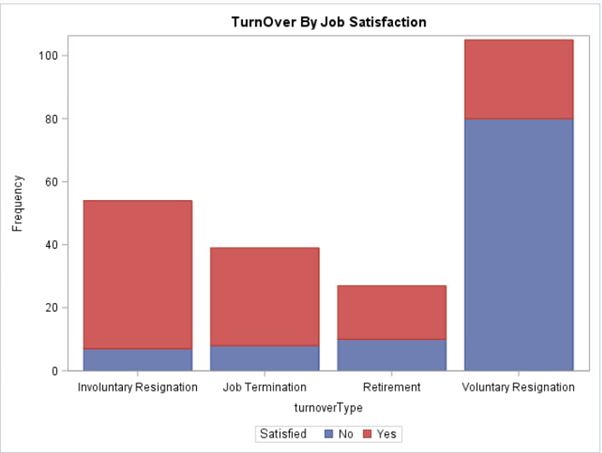


Figure 6: Distribution of TurnOver with JobSatisfaction

It can be observed from the above plot that the employees who are not satisfied with their job profiles are more likely to go for Voluntary Resignation. However, the employees who are highly satisfied with their jobs are being terminated more. The reason could be that they are not given much work and the company is not getting much benefit out of those employees.

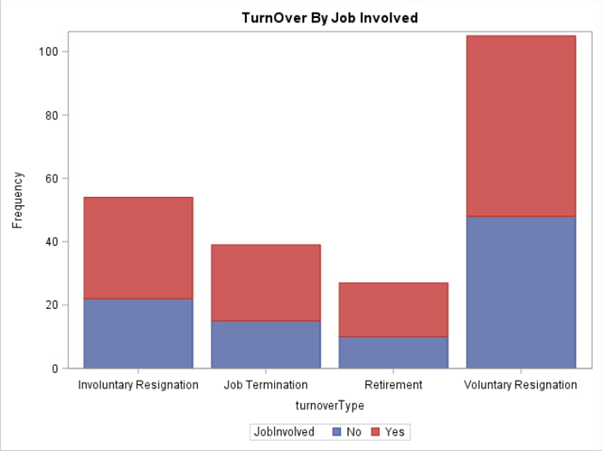


Figure 7: Distribution of TurnOver with JobInvolved

It can be observed from the above plot that the employees who are highly involved in their jobs, which means that they are given lots of work than they expect, maybe they are overloaded, are more likely to go for Voluntary Resignation.

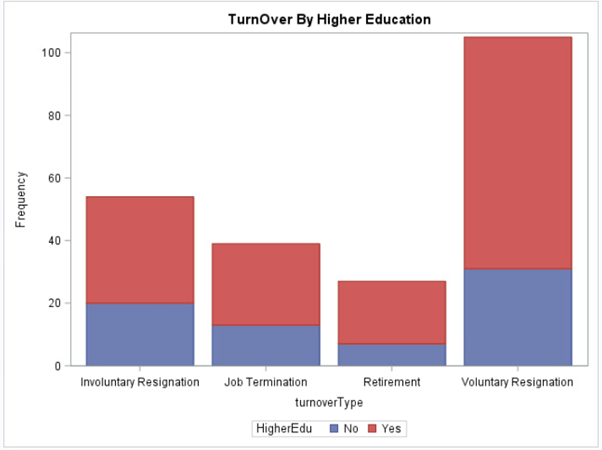


Figure 8: Distribution of TurnOver with HigherEduc

It can be observed from the above plot that the employees who are highly educated are more likely to go for Voluntary Resignation. Highly educated employees are more qualified with respect to less educated employees; hence, they have high chances of getting any new high-paying jobs as per their level of studies. We can even see that the employees with high education are terminated more as the company must pay more to these employees, but they might not have enough jobs at their level.

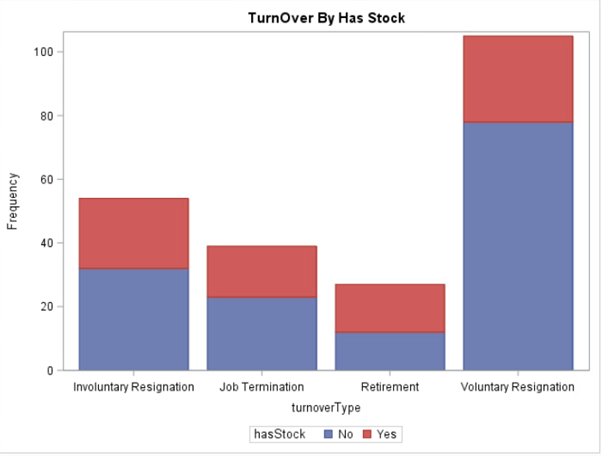


Figure 9: Distribution of TurnOver with hasStock

It can be observed from the above plot that the employees who have no stock are more likely to go for Voluntary Resignation.

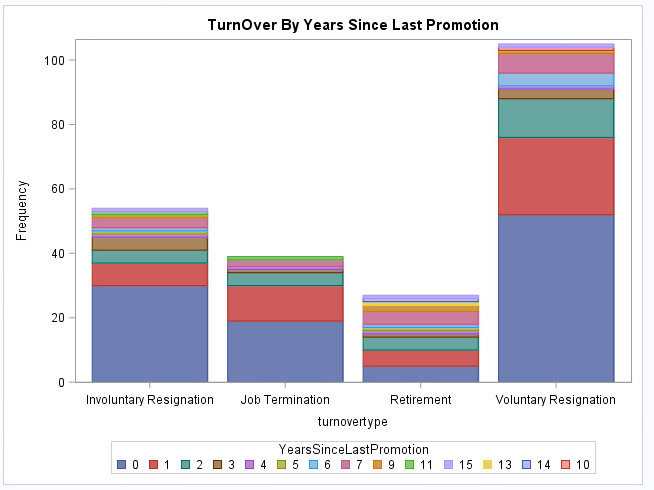


Figure 10: Distribution of TurnOver with YearsSinceLastPromotion

The above plot gives a very interesting insight about the relationship between TurnOverType and YearSinceLastPromotion. Employees are more likely to leave the company as soon as they get promoted and the reason could be that they were the right candidates for promotion but for some internal political issues they had not been promoted earlier.

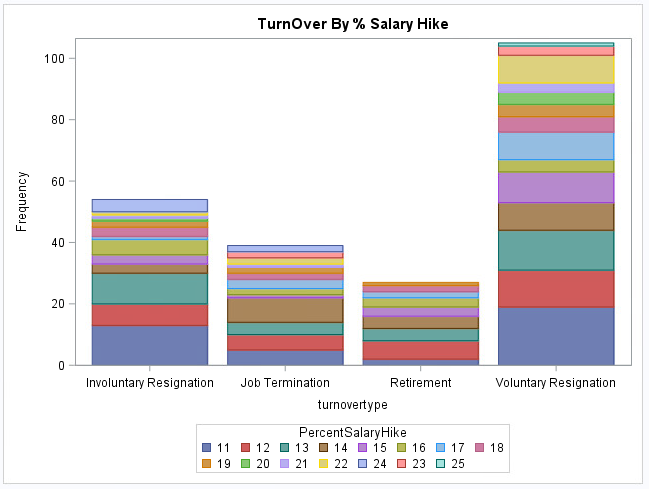


Figure 11: Distribution of TurnOver with PercentSalaryHike

It can be observed from the above plot that the employees whose SalaryHike is 11% or less are more likely to go for Voluntary Resignation. With the increasing SalaryHike the attrition decreases as the employees are very satisfied with the Salary.

## Hazards for all Event Types

The dataset with multiple event types can be analyzed and interpreted in different ways which are required to decide whether all the events should be considered together or analyzed separately. So, while computing the risks in survival analysis for different event types, other types of events would be considered as censored.

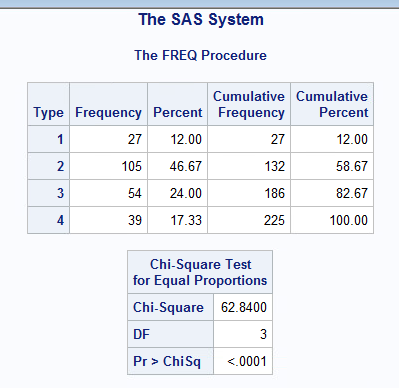


Figure 12: Frequency plot of all event type

As seen from the above frequency results, all events have different number of observations. The remaining 1,245 observations out of 1,470 belong to event type = 0. There 225 employees out of 1,470 who are leaving the company because of Retirement (Type = 1), Voluntary Resignation (Type = 2), Involuntary Resignation (Type = 3, personal and health reasons) and Job Termination (Type = 4). Out of these, 47% employees (105/225) are leaving the company voluntarily as compared to other types, which is the main concern for the company.

To test the linearity of hazards with all other event types, an event variable has been created which has censored No TurnOver type and combined all the other event types. Log-Log Survival plot (LLS) has been used for checking the linearity. LLS is used for Weibull distribution, which helps in interpreting the shape of the hazard function.

“Diff = All” in PROC LIFETEST is used to compare the Strata of different turnover types.

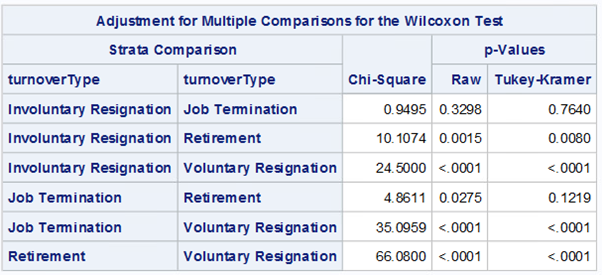


Figure 13: Strata Comparison Table/ p-value comparison between different event types

As seen from the above Strata Comparison table, all other combinations of turnover types are statistically significant based on p-values, except Involuntary Resignation and Job Termination.

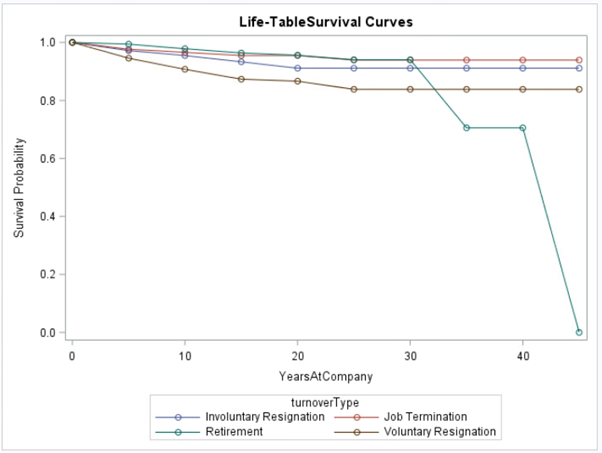


Figure 14: Survival estimate plot for all TurnOver Type

To prove the previous conclusion, survival estimate plot has been made, which shows that the number of employees who are resigning involuntarily is linearly proportional to those who are terminated. Hence, these two event types can be clubbed together.

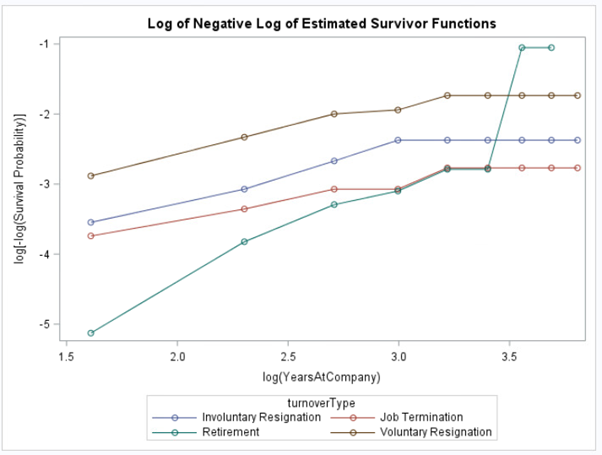


Figure 15: LLS plot of all event types

Adding to the above point, LLS graph of all event types plotted above shows that the ω values for Involuntary and Terminated employees are almost same. To check the possibility of using both event types together, PHREG model for each event types separately and combined are built.

**Modeling using PHREG:**

*Nested event type Retirement Event Voluntary Resignation Involuntary Resignation*

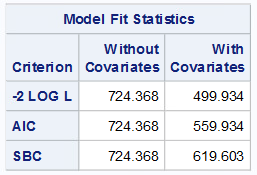
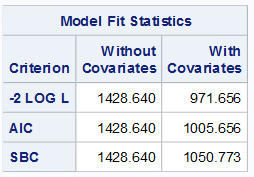
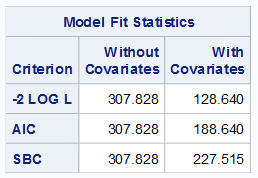
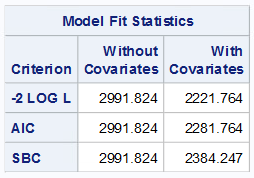


Figure 16: Nested and Individual event type

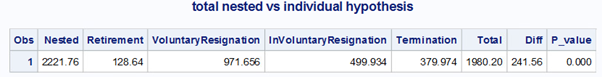


Figure 17: Total nested vs individual hypothesis

The null hypothesis for our model states that – with all the event types combined, the effect on turnover of employees is the same as the individual event types. Given that the p-value of the individual models and nested model is 0.00, the null hypothesis is rejected and hence, separate models need to be built for each event type.

*Nested (Involuntary resignation + Termination) model -*

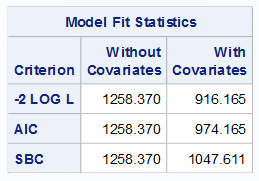


Figure 18: Nested model (Involuntary resignation + Job Termination)

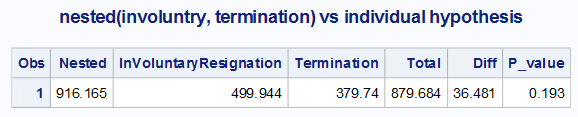


Figure 19: Nested (Involuntary resignation + Jon Termination) vs individual hypothesis

As shown above, the p-value for nested and individual models is 0.193, which is insignificant. Hence, the null hypothesis will be accepted in this case, and nested model can be built for Involuntary resignation and Termination event types.

# Model Building

## COX Proportional Hazards Modelling Technique

Since this model does not require knowing the distribution beforehand, it can be used for calculating the hazard ratio of Turnover to No Turnover employees. This model enables us to know the hazard ratio between two groups considering the covariates.

### Tied Dataset

When two records in a dataset have the same time to event, it is called Tied data. This has been handled by using Efron method in this analysis. Efron method has been used here because the Cox proportional hazard model is sensitive to tied datasets.

### Initial Assumptions for the Model

All the four types of turnovers have been considered – Retirement, Voluntary Resignation, Involuntary Resignation and Termination. Separate models have been built for all the four types of event. As a result, separate models have shown better performance as well as helped in explaining different turnover types as compared to nested model. All these steps have been explained in detail later.

The primary focus of building these models was to identify the important covariates affecting different turnover types significantly. This will help Fermalogis to focus on these attributes and reduce the current attrition rate.

### Variable Handling

For nested model, 0 has been considered as censored and all other event types as events. Step-wise regression approach has been used to find the significant time-dependent covariates by eliminating the non-significant covariates. The previously created variables have also been used in the model.

### Time-Dependent Variable

Bonus has been identified as the time-dependent variable, and has been handled by taking the cumulative sum of all the bonuses received by the employee during his/her tenure of up to forty years. Hence, forty new variables have been created to calculate the cumulative bonus: cum1-cum40. In the model, bonus variable has been passed with respect to “YearsAtCompany” which provides the tenure information of the employees.

## Non-Proportionality Identification

For identifying non-proportionality in the data, residual test like Martingale and Schoenfeld tests have been used to identify the time-dependent variables.

### Martingale Residuals:

Martingale residuals can be calculated using ASSESS statement in SAS. If a covariate is time-dependent, then, the pattern of covariate will be heavily biased towards the created set of simulations. The covariate’s pattern will be significantly different from the generated simulations.

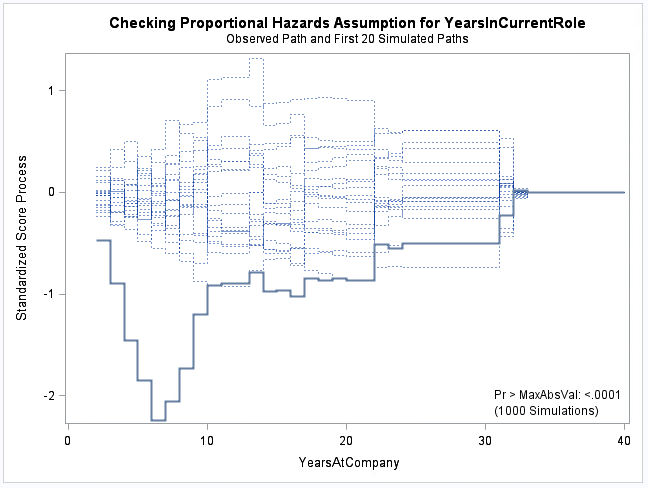


Figure 20: Proportional Hazard assumption for Years in Current role

The above figure shows the thick blue line for YearsInCurentRole variable heavily deviating from the simulations having significant p-value. Hence, this is a time-dependent covariate.

The initial item for proportional hazard assumption is that variables whose pattern deviates significantly from the simulated patterns with significant p-value are non-proportional covariates.

Since, Years in current role is a time-dependent covariate based on the above graph, it means that years in current role is dependent on the YearsAtCompany.

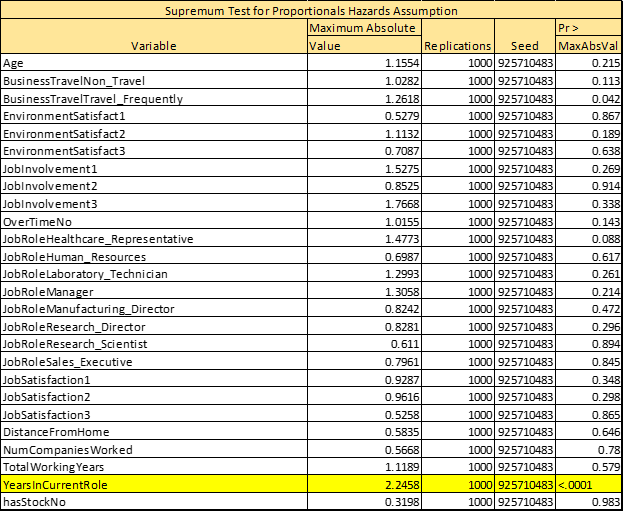


Figure 21: Null Hypothesis test for finding time dependent covariate

### Schoenfeld Residuals:

This is another way of testing the non-proportionality of the covariates. It tests the independence between residuals and time. This test has been used in the analysis to check if the residuals of significant covariates are correlated to employee’s tenure with the company. If it is significant, it is a time dependent covariate, otherwise, it is an independent covariate.

To check this, only numerical values have been considered to find the linearity. Years at company, log (years at company) and quadratic term for this variable have been used to check for linearity and non-linearity.

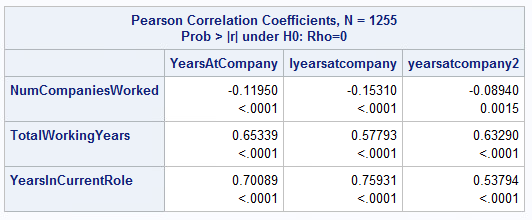


Figure 22: Pearson Correlation Coefficient for Time Dependent (Non-Proportional) Covariate

As shown from the Pearson Correlation Coefficients, Number of companies worked, Total Working Years and Years in Current Role have significant p-value. Hence, they are significant covariates.

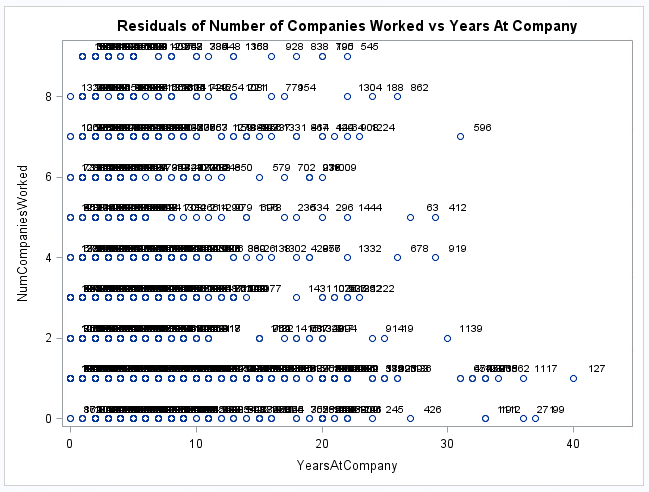


Figure 23: Schoenfeld residual of Number of companies worked vs Years at Company

On further investigating, the negative correlation between Years at company and Number of companies worked holds true, hence these two variables are non-proportional or time-dependent variables. These inputs could be included for further modeling.

Total Working Years, which is a non-proportional covariate, increases linearly with the increase in YearsAtCompany. This is expected as the number of years in the current company adds to the total years of experience. This is also evident from the figure below with the inference that total working years is a time-dependent covariate.

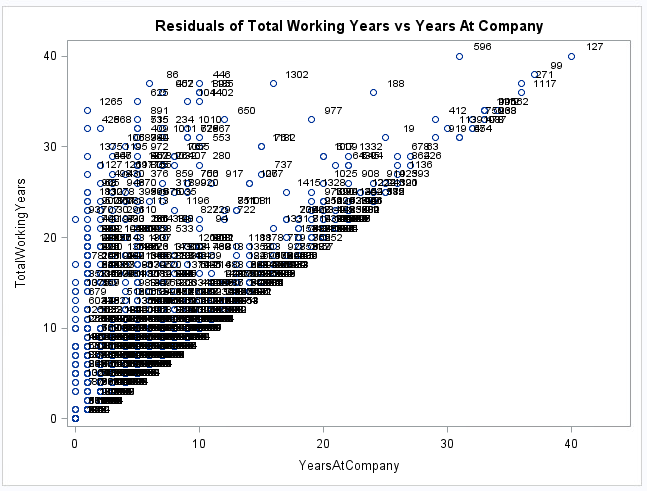


Figure 24: Schoenfeld residual of Total Working Years vs Years at Company

Non-proportionality has been observed in “Years in current role”. For promotion, the trend is that employees with less years of experience move to the next level in shorter time span, and with the increase in employees’ experience, the chances of getting promoted to higher levels also increase. Thus, with the increase in the “Years at company” of an employee, her/his “Years in current role” also increases, as seen in the figure below.

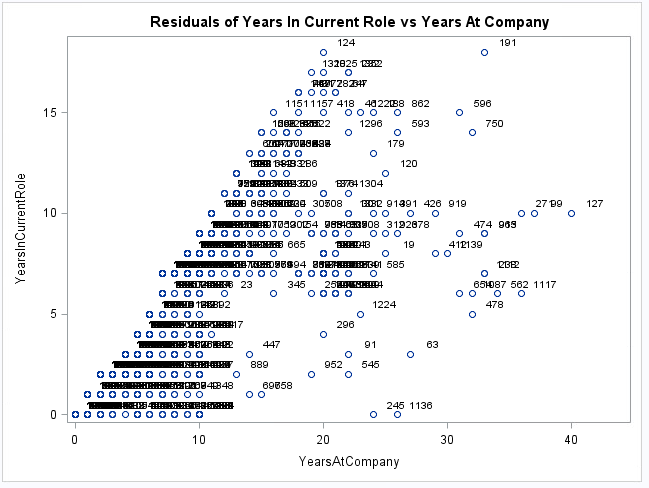


Figure 25: Schoenfeld residual of Years In currentRole vs Years at Company

**Handling Non-Proportionality using Interactions:**

Non-proportionality could be best handled with the use of interaction terms with Years at company. For each observation in the data, a new variable has been created by adding an interaction term between all the non-proportional variables and Years at Company. Now these new variables will be used for building models.



Sum of coefficients of the original variables and the interaction term will be the actual value at any given time.

**Attributes increase/decrease the hazard rates for certain event types:**

After selecting all the significant variables and coding the interaction terms, the next step is to identify the important variables which have a high impact on each of the Turnover types.

After modeling all the four independent hypotheses and the nested hypothesis, it could be concluded that covariates and hazard rates are better explained by independent models than a nested one. A detailed analysis on significant covariates and hazard rates on individual event types can be seen in the figure below.

**Note**: Significant covariates are marked. When hazard rate >1, hazard is increasing with a unit increase in the covariates, whereas hazard rate <1 shows that hazard is decreasing with a unit increase in the covariates value.

**Retirement Event Type:**

Here Retirement type is modeled as the event and all other types are considered as censored, and this model is significantly different and even better than the Null mode. F-statistics is quite different for models with covariates and those without covariates.

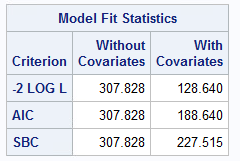


Figure 26: Model result of event type Retirement

**Important variables observed for Retirement:**

Below are the most important factors that are affecting employees’ retirements throughout the company - all these variable coefficients are significantly different from zero (which is explained by P<0.05). However, if these coefficients have positive value it increases the hazard rate whereas if the coefficient is negative, it decreases the hazard rate. Few important covariates affecting retirement are Age, Business Travel, Job involvement, Years in current role, Employee Bonus, Number of companies worked with respect to time, Year in Current Role with respect to time.

**Important negative factors which should be considered by management** **are**:

1. A one unit increase in age would increase the hazard rate by 54%, which says that older employees are more likely to retire from the company.
2. People travelling frequently are retiring almost 8 times higher than the people who are travelling less.
3. Old people who are less involved in the job are retiring from the company. However, we only have 27 events of this type to believe in this and we need more of this event type to train the model.
4. People who are working overtime are more likely to retire.
5. Employees who do not get bonuses in their experience are likely to retire.
6. The interactions terms “Years in current role” and “number of companies” worked are increasing hazard rate.

**Important Positive factors that should be considered by management are:**

1. Employees who are not working overtime are less likely to retire, or the hazard decreases for them.
2. A unit increase in the current role decreases the hazard rate by almost 40%.

**Impact of Bonus on Employee Turnover:**

Retirement and Bonus given to employees are highly correlated; the reason could be that with high experience the salaries are also usually high, and now the employees expect some other benefits from the company. Since the number of events of type Retiring is less in number, so it a judgement call for the management. We cannot strongly say that bonus affects retirement.

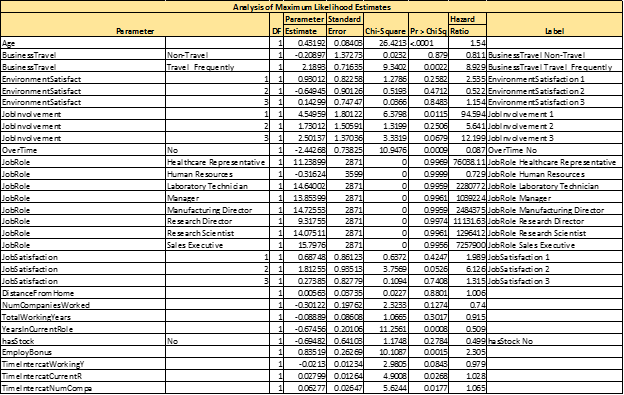


Figure 27: Maximum likelihood estimates of event type Retirement

**Combined model for involuntary resignation and job termination:**

As stated earlier, nested models with Involuntary Resignation and Job Termination are not significantly different from each other, hence we combine them for better insights than modeling them individually.

From the below Model Fit Statistics, it is clearly evident that modeling involuntary resignation and termination as the event, keeping all other event types as censored, shows significant difference in F-Statistics values. The model with covariates (full model) is better than the null model or one without covariates.

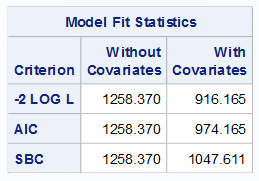


Figure 28: model result of combined event type of Involuntary resignation and Job termination

**Based on the Maximum Likelihood Estimates we can analyze important negative features which is primary reason of concern for the Management: -**

1. Employees who are less satisfied with environment (having “EnvironmentSatisfaction” with value 1) is twice as prone to leave the company.
2. Employees having no stock are leaving more by 75% in comparison to people who have stocks.
3. Considering the interaction terms, our model shows years in current role and number of companies statistically significant and it is increasing the hazard ratio with respect to time.
4. In terms of job satisfaction, we have three levels – 1, 2 & 3. Assuming level 1 & 2 employees are least satisfied, the attrition level is quiet high among them.

**Important Positive Features that should be taken into consideration by the managerial authority: -**

1. Employees who are not frequently traveling are less likely to resign the company. It is almost five times less than people who frequently travel for business reasons.
2. Employees who do not work overtime have less hazard ratio. In other words, probability of leaving the company is quiet less so as an employer, should give additional benefits and incentive to employees who work overtime, that would control the attrition rate.
3. The Interactive term for total years of experience, shows that more experienced employees are less likely to leave the company and at the same time chances of being terminated also is quiet low. Hence, in order to retain them FermaLogis should pay attention and give timely promotions.
4. With increase in years in current role, the hazard rate decreases. Thus, it concludes that experienced people are less likely to be terminated.

Interaction Terms are significant for Employee Bonus but not for employees leaving from company due to Involuntary and Termination types.

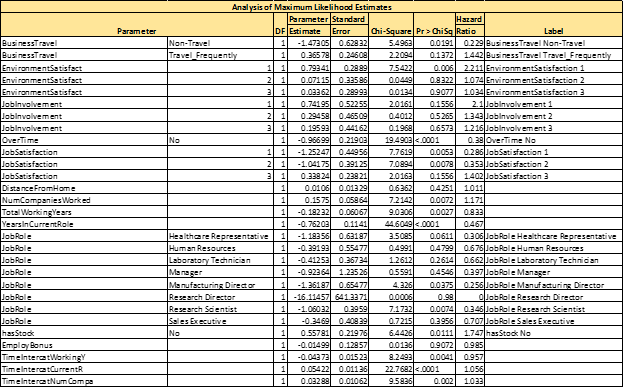


Figure 29: Maximum likelihood estimates for combined event type of Involuntary resignation and Job termination

**Voluntary Resignation:**

While building the model for Voluntary Retirement turnover as the event type, keeping all other event type censored, we can see that the full model (with covariates) performs better than Null Model (without covariates) in terms of -2 Log L value, AIC and BIC. For all the specified criteria, smaller the value, better it is in terms of the model’s performance. The fit statistics clearly showcase the difference in criteria for with and without covariates.

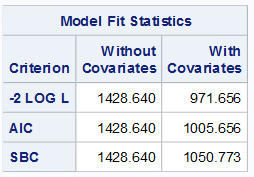


Figure 30: Model result of Voluntary resignation

**Important negative features which are the primary reason for concern for Management: -**

* Since the parameter “hasStock” with value “No”, is statistically significant with a p-value less than the default significance value of 0.05, hence we can conclude that people who hold no stock are leaving as parameter estimate has a positive magnitude. Based on the hazard ratio it is almost 4 times as much as people who have stocks.
* Employees who are traveling frequently have a higher attrition rate. This variable is statistically significant and its positive magnitude of the estimate show that with increase in one unit of BussinessTravel (feature), the voluntary resignation event is increasing. From the hazard ratio, we can see that it is almost 8 times more than people who travel less (where BusinessTravel equals “Non-Travel”).
* Employees who are not at all satisfied with the environment (having value of 1), are quitting the company more often. The statistics show that they are twice as likely to leave than others.
* Employees who are less involved in their job (where “JobInvolvement” equals “No”), either due to lack of interest in current job role or their skillset being different from what is required, often has higher attrition rate than others.
* Employees who are less satisfied with their job roles and responsibility often leave the company; hence, the feature shows statistical significance. It is one of the important factors for higher attrition rate, and the management needs to talk to their employees to understand their concerns. It is almost 6 times more than people who are satisfied with their job role.
* For employees who are lateral hires having prior experience, as represented by the feature “NumCompaniesWorked”, almost 23% resign voluntarily.

Interaction Term for working in the current role for number of years “TimeInteractCurrentR” is increasing the hazard rate by 9.2%.

**Important Positive Features that should be taken into consideration by the managerial authority: -**

* People who are not traveling (BusinessTravel with value of “Non-Travel”), is less likely to leave the company and the total percentage is almost 50% compared to people who are frequently traveling.
* Employees who have higher work experience are less prone to leave the company. From the negative magnitude of parameter estimate for “TotalWorkingYears”, we can derive the relationship. Hence, FermaLogis, as an employer, should be more concerned about employees who have less experience.
* Employees who do not work overtime are less likely to leave the company. FermaLogis can hire new employees to balance out.

Employee Bonus feature (“EmployeeBonus”) is not statistically significant for increasing the attrition rate, which is clearly shown from the below Maximum Likelihood Estimates table.

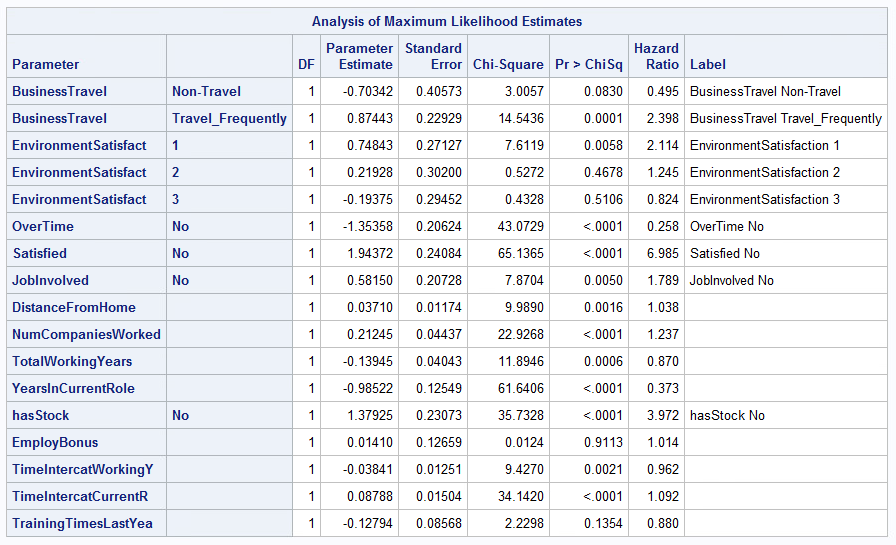


Figure 31: maximum likelihood estimates of event type Voluntary resignation

# Conclusion

* From Figure 12, it is clear that frequency of each individual TurnOver type is significantly different from one another in terms of attrition rate; hence, each TurnOver type is different. Linearity between TurnOver Types could be observed from LLS plot in Figure 15, where Involuntary resignation and Job Termination are linearly related. The same point could be concluded from Proportional hazard plot in Figure 20 and Wilcoxon test in Figure 13 as well. P-value of these tests are not significant, thus Involuntary Resignation and Job Termination are modeled together.
* P-value of all event types combined is 0 which could be seen from the Log Ratio test in Figure 17 which says that all event types are different and needs to be handled individually. P-value for combination of Involuntary Resignation and Job Termination is 0.193, which is not significant, hence nested model would be best.
* For Retirement – Age, Job Involvement, Business Travel, Bonus and Interaction Variable - Years in current role are significant variables.
* For Voluntary Resignation- hasStocks, BusinessTravel, Environment Satisfaction, Job satisfaction, Job Involvement, interaction term - number of companies worked and Years in current role are significant variables.
* For Involuntary Resignation and Job termination – Environment Satisfaction, Business Travel, Overtime, Job satisfaction, hasStocks and Interaction variable - Years in current role and number of companies worked are significant.
* Employee Bonus has a significant impact on employees who are retiring from the company but has no significant impact on the other event types.
* NumCompaniesWorked, TotalWorkingYears and YearsInCurrentRole are the non-proportional variables with respect to hazard.

# Recommendations

There are few variables that have high impact on both Voluntary and Involuntary Resignation:

* Management could reduce overtime of the employees by hiring few temporary employees and shifting the extra work.
* Management should try reducing the Business travel associated with any employee by hiring new employees at required locations, training other employees on the same skillset to reduce dependency on particular employees, giving some incentives to those employees who travel frequently.
* There is a strong impact of hasstocks on employee turnover, so Management should promote or give additional benefits or lucrative offers like issuing stocks at subsidized rates with higher returns on investment for in-house employees who are interested in buying stock.
* By motivating employees with good leadership and encouraging them with their work can lead to a better job satisfaction and involvement levels. This might reduce the turnover rate associated with these factors.
* Environment satisfaction is one of the major factors of attrition, so Management can arrange for some fun events or family events every quarter to boost up the morale and productivity of the employees.

***Code for the analysis:***

