Survival Analysis of Employee Attrition at FermLogis

OPIM5894 Survival Analysis

Team 6

Janeet Arora

Swati Arora

Raozhong Chen

Namratha Kasineni

Jagbir Kaur

Supreet Mangat

Ritika Sinha

Shiqi Xu

Contents

[Executive Summary 2](#_Toc458975388)

[Background 2](#_Toc458975389)

[Exploratory Data Analysis 3](#_Toc458975390)

[Methodology 7](#_Toc458975391)

[Competing Risk 7](#_Toc458975392)

[Time dependent covariates 10](#_Toc458975393)

[Non Proportionality Hazard – Test the proportionality assumption 11](#_Toc458975394)

[Interactions 17](#_Toc458975395)

[Final Model 18](#_Toc458975396)

[Conclusions 20](#_Toc458975397)

[Recommendations 20](#_Toc458975398)

**The work presented is our team’s work and our team’s work alone**

# **Executive Summary**

Larry Hansen, COO of a very famous pharmaceutics company, FermaLogis, wants to apprehend the employee attrition rate by using advanced survival analytic techniques as he is facing a very crucial attrition problem so that he can take steps to prevent the employees from leaving FermaLogis. They trained their employees and made them more competent. Therefore, the company benefits a lot from their well-trained employees. However, recently, their competitors are recruiting their employees with high transfer fees and higher salaries which prevents the company from gaining returns from all the investments it has made to train the employees. Thus, he decided to use the survival analytics technique to find out who are leaving and why they are leaving. By this way, Fermalogis may know about their employees better and help them take some measures to prevent their employees from leaving.

In order to help Fermalogis understand the reasons behind attrition rate, our team conducted in depth analysis to figure out the reasons which led to employees leave the organization. To start with, we performed pre-data exploration to check if all the employee attrition types have the same hazard rate or not using PROC FREQ. In our analysis, we came up with few crucial features which played deciding factors of whether the employee could leave the company or not. The following factors like salary, stock options, years in current role, job satisfaction, years with current manager, years with current manager, marital status, years in current role, age, number of companies worked, work life balance, department, distance from home, education and years in current role came out to be important deciding factors. Moreover, the COO of FermaLogis was also interested in knowing how the bonus payments of the company were affecting the turnover, therefore our team conducted further analysis using PHREG procedure. We laid emphasis on voluntary resignations by dividing distinct types of employee turnover and to see what impact does bonus have in their decisions to leave the job. Final leg of the analysis involved understanding those variables which affected hazards non-proportionally or identifying those variables which had non-proportional hazards varying with time variable.

In the following sections, we analysed how each attribute, including bonus, affects the attrition of different employee groups such as for all employees, young employees, and experienced employees. This report shows the methodology used to determine if the attrition of several types were significantly different and what all are the factors that affect the hazard rates. Based on the results of the analysis performed, the report outlines the recommendations for reducing the attrition rate in FermaLogis.

# **Exploratory Data Analysis**

**Data Exploration:**

We have been provided with the data of 1470 employees and following is the description of the variables present in the csv.

|  |  |  |
| --- | --- | --- |
| **No.** | **Variables** | **Description** |
| **1.** | **Age** | This variable describes the Age of the employee |
| **2.** | **Turnover** | It describes if employee left the company or not |
| **3.** | **Type** | It takes the following values: 0 - No turnover  1 - Retirement  2 - Voluntary Resignation  3 - Involuntary Resignation (Health problems, family matters etc.)  4 - Job Termination, Employee is Fired |
| **4.** | **BusinessTravel** | It shows how much a travel employee makes. It has three values:  a. Travel Rarely  b. Travel Frequently  c. Non-Travel |
| **5.** | **DailyRate** | This variable describes the daily compensation of employee before any cuts and taxes |
| **6.** | **Department** | It describes the department of the employee when this dataset was created |
| **7.** | **DistanceFromHome** | This variable describes the commute distance for the employee in miles |
| **8.** | **Education**: | There are following levels of education of employees  1 - 'Below College' 2- 'College' 3- 'Bachelor' 4- 'Master' 5- 'Doctor' |
| **9.** | **EducationField** | It describes the field of education of an employee |
| **10.** | **EmployeeCount** | This variable contains aggregated calculations |
| **11.** | **EmloyeeNumber** | It contains the ID of the employee |
| **12.** | **EnvironmentSatisfaction** | It represents the score of how much the employee is satisfied with the provided company's facilities  1 'Low'  2 'Medium'  3 'High'  4 'Very High' |
| **13.** | **Gender** | It represents the Gender of the employee |
| **14.** | **HourlyRate** | It contains the hourly compensation of employee before any cuts/taxes deduction |
| **15.** | **JobInvolvement** | It represents the score given to the employee by supervisors describing how much the employee is involved in company's operations which are as follows:  1- 'Low' 2- 'Medium' 3- 'High' 4- 'Very High' |
| **16.** | **JobRole**Job role of the employee in the company | This variable describes the job role of the employee in the organization. |
| **17** | **JobLevel** | This variable describes the management level of the employee |
| **18** | **JobSatisfaction** | This variable shows the last survey result of the employee about his\her job satisfaction |
| **19** | **MaritalStatus** | It shows the marital status |
| **20** | **MonthlyIncome** | It represents the Monthly income of the employee |
| **21** | **MontlyRate** | This variable represents the monthly compensation of employee |
| **22** | **NumCompaniesWorked** | It represents the number of companies the employee has worked before joining the organization. |
| **23** | **Over18** | It represents the employee is over 18 years old or not |
| **24** | **OverTime** | This variable represents if the employee works overtime more than 10 hours a week or not |
| **25** | **PercentSalaryHike** | It represents the percent of agreed yearly salary rise |
| **26** | **PerformanceRating** | It represents the score given to the employee by supervisors on the basis of how good the performance of the employee was for last year. It contains the following values:  1 'Low'  2 'Good'  3 'Excellent'  4 'Outstanding' |
| **27** | **RelationshipSatisfaction** | This variable contains the last survey results of the employee about his\her satisfaction with other employees in the company which contains the following values:  1 'Low'  2 'Medium'  3 'High'  4 'Very High' |
| **28** | **StandardHours** | This variable represents the number of hour’s employee works for one payroll period (two weeks) |
| **29** | **StockOptionLevel**: | It represents the stock option for the employee. |
| **30** | **TotalWorkingYears** | It represents the time which employee worked as a professional in any organization |
| **31** | **TrainingTimesLastYear** | It represents the no of training programs attended by the employee last year |
| **32** | **WorkLifeBalance** | This variable represents the employee satisfaction of the work load (4 is the highest satisfaction level)  1 'Bad' 2 'Good' 3 'Better' 4 'Best' |
| **33** | **YearsAtCompany** | This variable represents the time spent at the company |
| **34** | **YearsInCurrentRole** | It represents the number of years employee has worked in the current position at the company |
| **35** | **YearsSinceLastPromotion** | It represents the number of years passed since the last promotion |
| **36** | **YearsWithCurrentManager** | This variable shows the no of years with the current supervisor. |
| **37** | **bonus\_1-40** | It represents whether the employee received bonus payments since past 40 years. |
| **38** | **Turnover** | This variable represents the rate at which employees leave a company. It has got two values: 1. Yes 2. No |
| **39** | **Attrition Type** | This variable contains the code for the following event types:  0- No turnover  1- Retirement  2-Voluntary Resignation  3-Involuntary Resignation  4- Job Termination, Employee is fired |

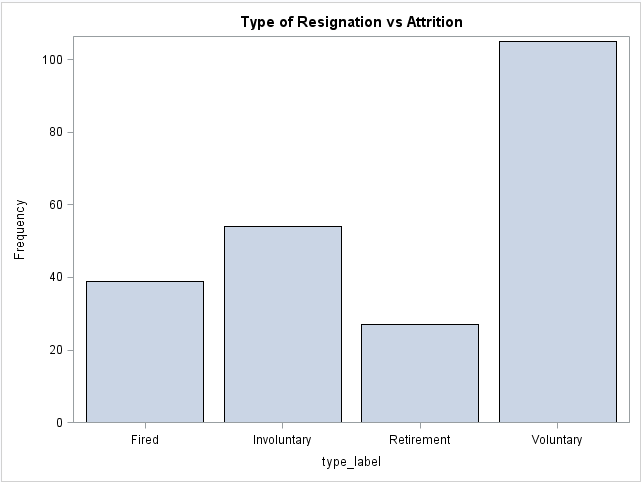
**Data Pre-processing:**

To Analysis the data properly our team decided to create few variables in the dataset.

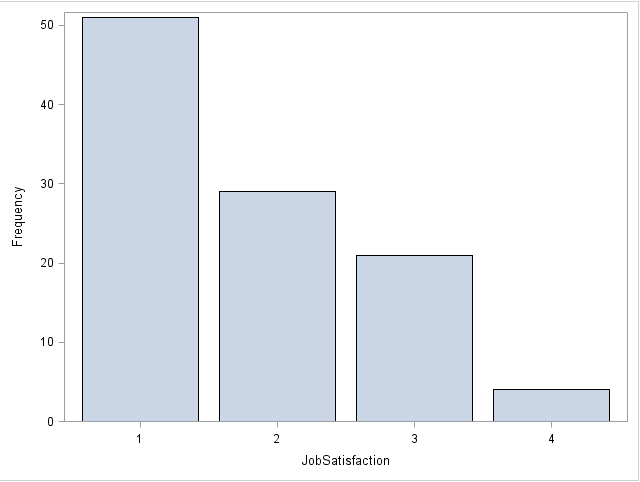
|  |  |  |
| --- | --- | --- |
| 1 | Turnover\_Binary | 0/1 (For Turnover Yes/No) |
| 2 | Type\_Label | Character Variable to convert attrition type from numeric to corresponding categories  No turnover -0  Retirement-1  Voluntary Resignation - 2  Involuntary Resignation - 3  Job Termination, Employee is fired -4 |
| 3 | Bonus | Updated the NA to “.” for all the 40 Bonus variables. |

**Key Observations:**

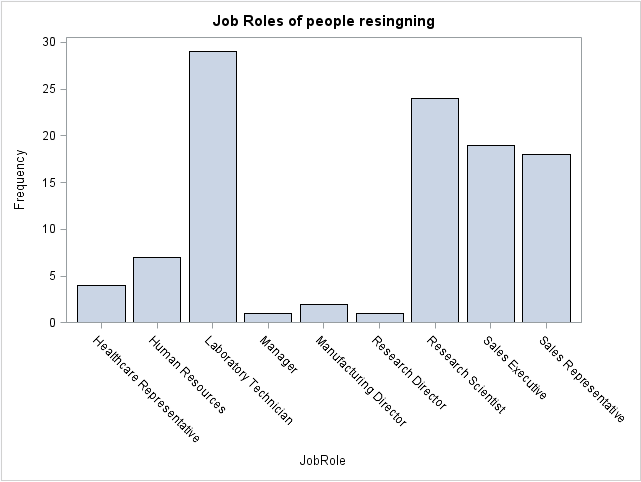
1. We can observe from the following plot that the attrition is highest for category “Voluntary” type among all the others. Also, retirement has the least frequency in attrition which describes that there are few employees who are faithful to the company.



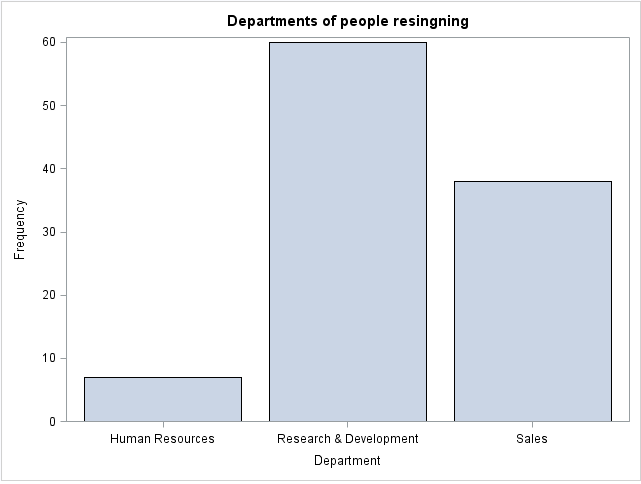
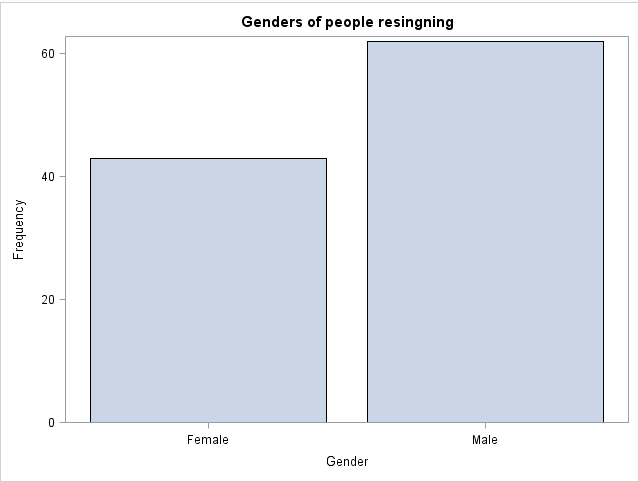
1. The other plot represents the frequency of people corresponding to the Job satisfaction level. It can be observed that people with rating 1 i.e. having least job satisfaction are the ones who have high attrition rate. As the job satisfaction increases the attrition is decreasing drastically.



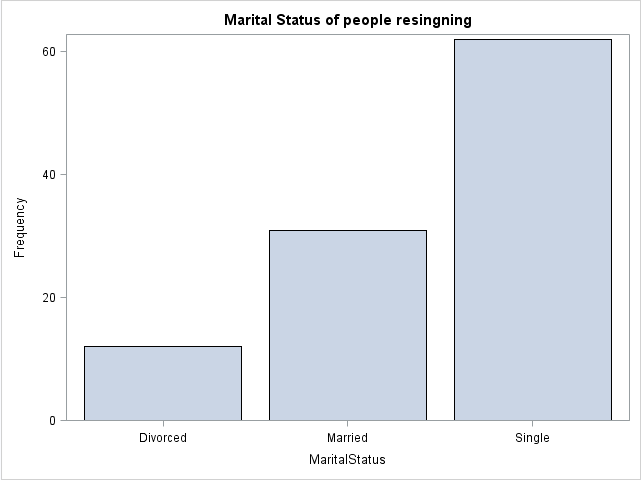
1. The following plot represents the attrition based on Job Role corresponding to the frequency of people. It can be observed that most of the employees with job role as Laboratory Technicians and Research Scientists are leaving the organization. As already explained in previous report, the attrition for managers, directors and HR is very low as compared to others.



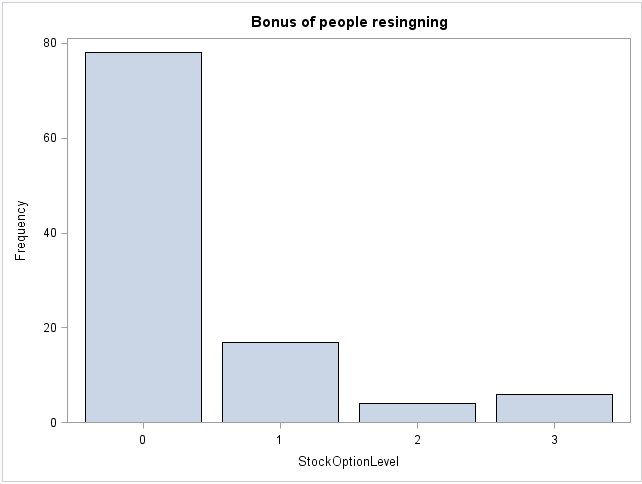
1. The following visual plot shows that higher number of male employees have high rate of attrition. In the second plot between departments of people resigning and number of employees, it can be observed that people lying in the Research and Development, are having higher rate of attrition as compared to other departments.



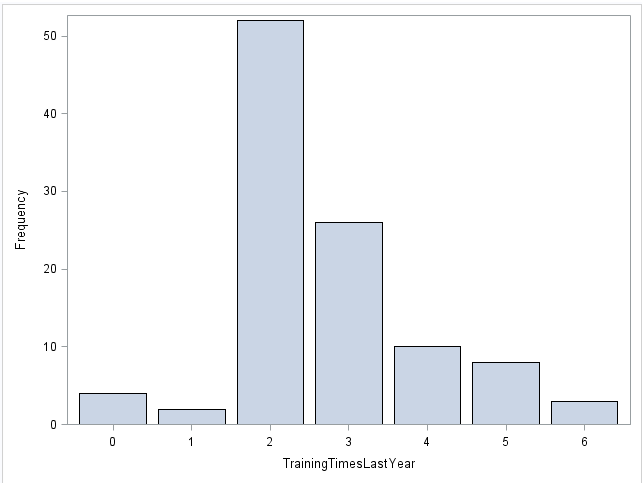
1. The following plot depicts that the count of employees who are single are more likely to leave the company.



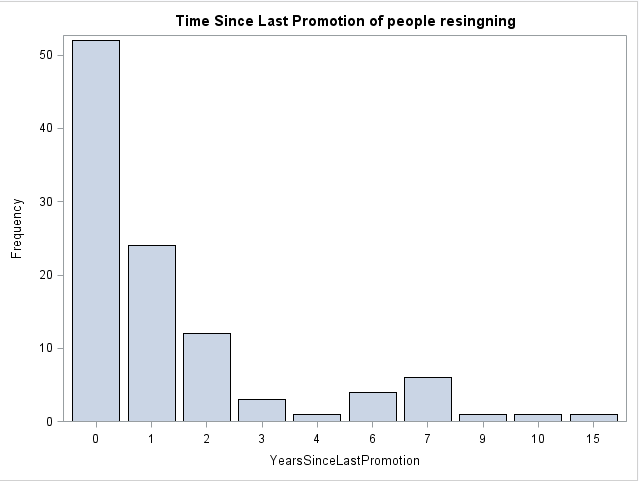
1. The second plot shows that the employees who get the least bonus, are more likely to leave the company.



1. The plot between TrainingTimesLastYear and the count of employees depict that people who went through the trainings twice are very highly likely to leave the company resulting in higher attrition rate.



1. The below plot shows that more than 50% of people resigning are the ones who got promoted within last one year. That means employees start looking for the job as soon as they get promotion which is quiet surprising.



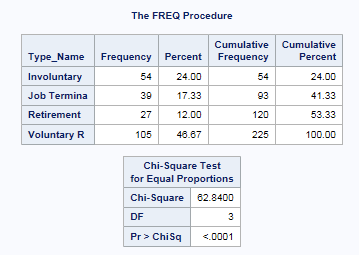
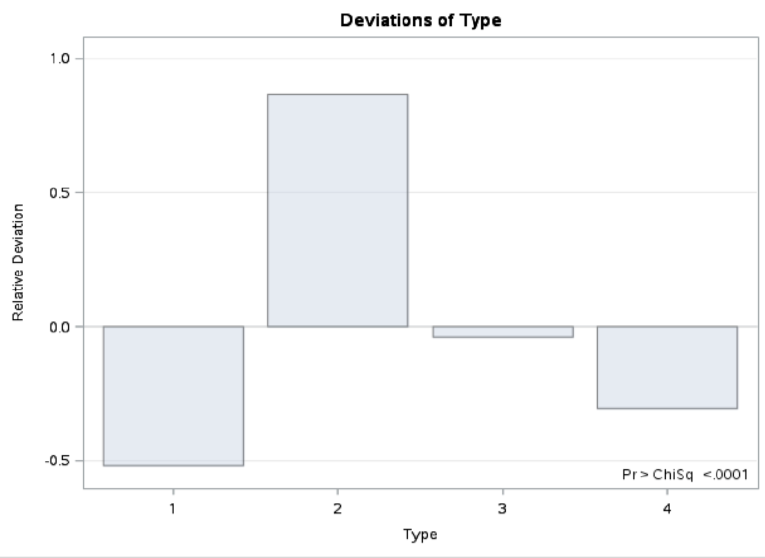
# **METHODOLOGY**

The exploratory analysis shows that the 4 different turnovers might have different dynamics. This was the reason we wanted to go ahead with the Competing Risk (Type Specific Hazard). We also used Cox PH Technique to analyse the effect of bonus (time dependent covariate) on Attrition. We also verified the assumption of proportionality for covariates and if it is violated then we remodelled using interaction terms.

**Competing Risk:**

As already mentioned that different turnovers can involve different covariate, So we are applying competing risk to understand whether different event types needs to be combined or we will model them individually with respect to turnover. Thus this can conclude whether all event types have equal hazards or not.

1. Firstly, we used **FREQUENCY PROCEDURE** to check whether type specific hazard function is similar for all the event type.

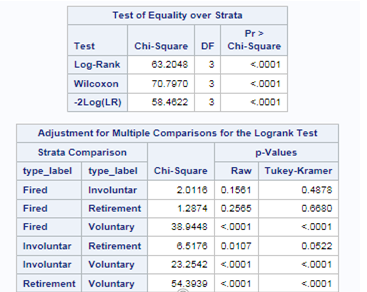
The above screenshot captures the analysis that each turnover type has different frequency value thus we can say that hazard function is different for all turnover type.

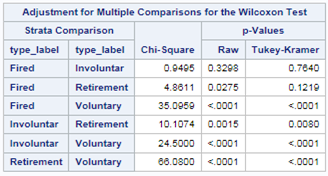
Here we can see that the frequency of Type 2(Voluntary Resignation) is higher than other three types. Thus, we need to handle type 2(Voluntary resignation separately)

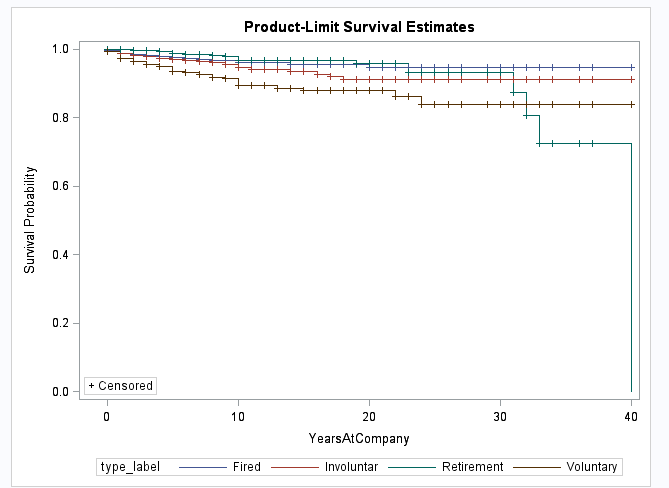
1. Using **LIFETEST PROCEDURE**, we analyzed the linear relation between type hazards without considering the covariates.

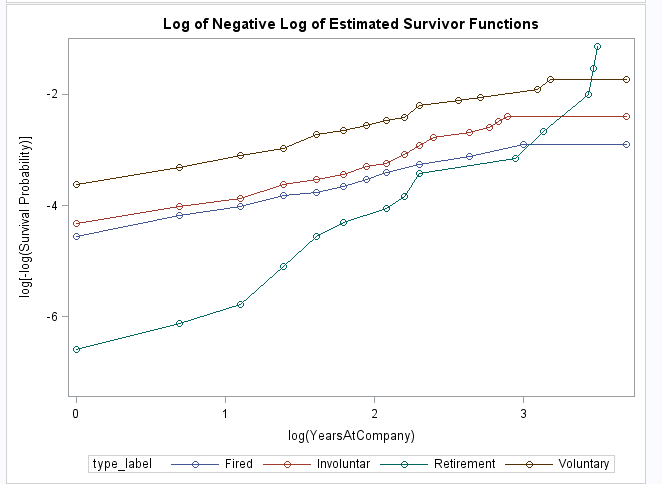
After creating four separate dataset with different turnover type and considering one turnover type at a time while censoring the other turnover type, we ran LifeTest procedure to check if event type needs to be modelled separately or we need to combine the event type.

In the below screenshot for Multiple comparison for the Wilcoxon Test, the p value for strata comparison shows significant value for four pair thus we can reject the null hypothesis and ca say that all event types are not same.









Further from the above graph, we can easily interpret that event types are different.

**Impact of Covariates for different turnover types.**

We used Likelihood Ratio test to check if the coefficient for each turnover type is same as the coefficient for found for the model when there was no distinction made between the event types.

**Separate models and complete model**. From the below screenshot, we see that p value is 0 which is insignificant. Thus, we need to model separately.



We also ran the above test for different combinations of event type to see if any of the combination of event types shows significant p value. But after running all the possible combinations we concluded that none of the combinations are significant. Thus, we came to the conclusion that we need to model the event type separately.

Below are the screenshots for all the possible combinations ran:















Since Type 2 turnover Voluntary resignation is impacting Larry’s business most and as per Larry’s requirement, analysing the voluntary resignation can answer well to the Larry’s question of why the employees were leaving the company.

**Larry’s Next Concern: DOES BONUS AFFECT EMPLOYEE TURNOVER**

Bonus is a Time dependent covariate. We used Cox proportional model to analyse the impact of bonus covariate on employee turnover (basically we are considering Voluntary Resignation as Larry’s concern is related to employees leaving voluntarily)

Thus, analysing the impact of bonus (lag1 and lag2) on voluntary resignation.

Lag1:



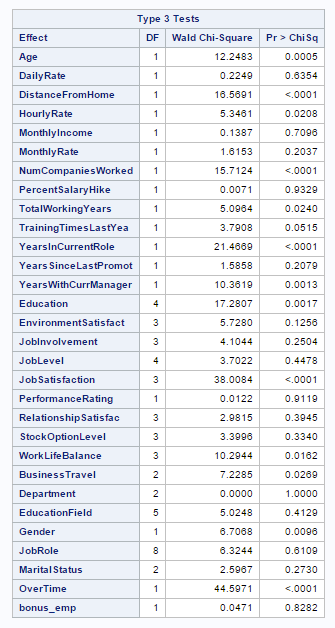
Lag1 for bonus is slightly significant which shows that hazard function of employee for type 2(Voluntary resignation) for last year is slightly impacted by the bonus.

Lag 2:



For last 2 year (lag1), bonus is insignificant thus bonus doesn’t impact the hazard function for employee turnover type2 (voluntary resignation)

We also analysed the effect on hazard ratio due to bonus variable cumulatively.



We can say that bonus variable cumulatively do not affect the hazard ratio significantly.

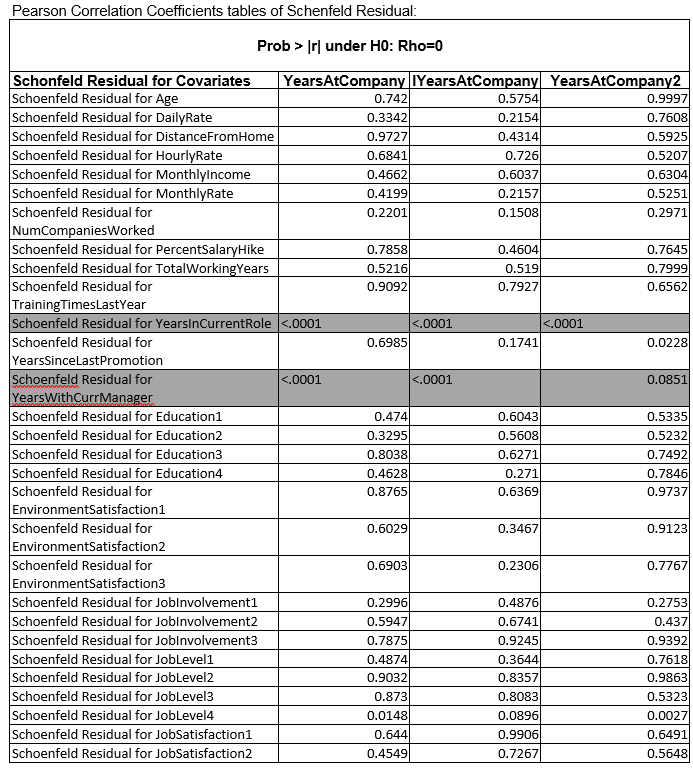
# **Verifying proportionality assumption for covariates**

From the COX model, we now do the test for non-proportional hazards. We are aware that test is dependent on accurate specification of the model.

We omitted ‘Bonus’ covariate here, because we found it has dependency on time and we are dealing with it using linear programing, also as per Shoenfeld residuals it does not seem to be significant.

We used ASSESS statement to check for the Martingale residual as well as the Schoenfeld residual to verify non-proportionality. Any values exceeding 5% critical values show nonproportional hazards assumptions has been violated

1. **Schoenfeld residuals Test:**



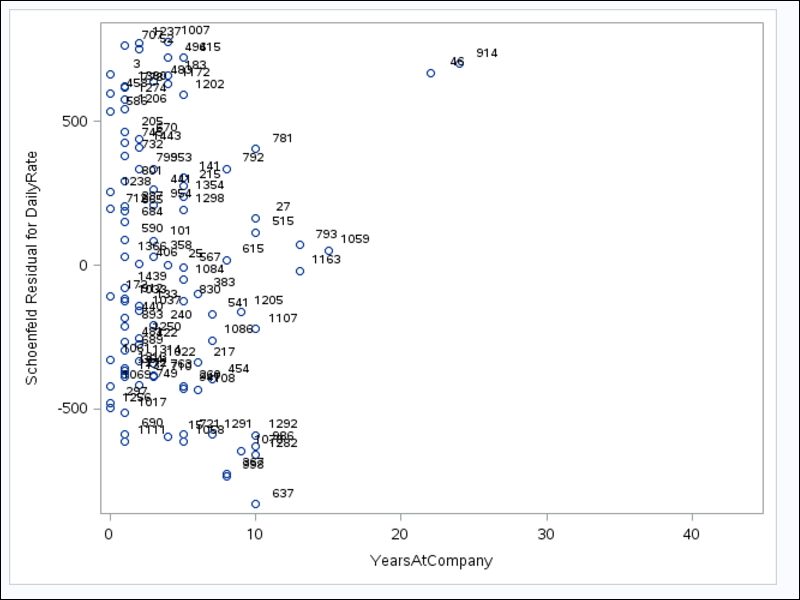
1. **Martingale residuals Test:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Supremum Test for Proportionals Hazards Assumption** | | | |
| **Variable** | **Pr > MaxAbsVal** | **Variable** | **Pr > MaxAbsVal** |
|
| Age | 0.219 | JobSatisfaction1 | 0.955 |
| **DailyRate** | **0.007** | JobSatisfaction2 | 0.476 |
| **DistanceFromHome** | **0.037** | JobSatisfaction3 | 0.041 |
| HourlyRate | 0.544 | PerformanceRating3 | 0.738 |
| **MonthlyIncome** | **0.03** | RelationshipSatisfaction1 | 0.661 |
| MonthlyRate | 0.902 | RelationshipSatisfaction2 | 0.408 |
| NumCompaniesWorked | 0.738 | RelationshipSatisfaction3 | 0.703 |
| PercentSalaryHike | 0.688 | StockOptionLevel0 | 0.615 |
| **TotalWorkingYears** | **0.023** | StockOptionLevel1 | 0.659 |
| TrainingTimesLastYear | 0.85 | StockOptionLevel2 | 0.245 |
| **YearsInCurrentRole** | **<.0001** | WorkLifeBalance1 | 0.424 |
| YearsSinceLastPromotion | 0.101 | WorkLifeBalance2 | 0.657 |
| **YearsWithCurrManager** | **<.0001** | WorkLifeBalance3 | 0.925 |
| Education1 | 0.857 | Business Travel Non\_Travel | 0.064 |
| Education2 | 0.225 | Business Travel Travel\_Frequently | 0.641 |
| Education3 | 0.525 | DepartmentHuman\_Resources | 0.973 |
| Education4 | 0.916 | DepartmentResearch\_\_\_Development | 0.449 |
| EnvironmentSatisfaction1 | 0.891 | EducationFieldHuman\_Resources | 0.521 |
| EnvironmentSatisfaction2 | 0.133 | EducationFieldLife\_Sciences | 0.147 |
| EnvironmentSatisfaction3 | 0.205 | EducationFieldMarketing | 0.346 |
| JobInvolvement1 | 0.14 | EducationFieldMedical | 0.031 |
| JobInvolvement2 | 0.995 | EducationFieldOther | 0.791 |
| JobInvolvement3 | 0.465 | GenderFemale | 0.755 |
| JobLevel1 | 0.046 | JobRoleHealthcare\_Representative | 0.325 |
| JobLevel2 | 0.247 | JobRoleHuman\_Resources | 0.973 |
| JobLevel3 | 0.816 | JobRoleLaboratory\_Technician | 0.983 |
| JobLevel4 | 0.43 | JobRoleManager | 0.148 |
| MaritalStatusDivorced | 0.041 | JobRoleManufacturing\_Director | 0.575 |
| MaritalStatusMarried | 0.736 | JobRoleResearch\_Director | 0.605 |
| OverTimeNo | 0.111 | JobRoleResearch\_Scientist | 0.946 |
|  |  | JobRoleSales\_Executive | 0.372 |

**Residual Plots for :**

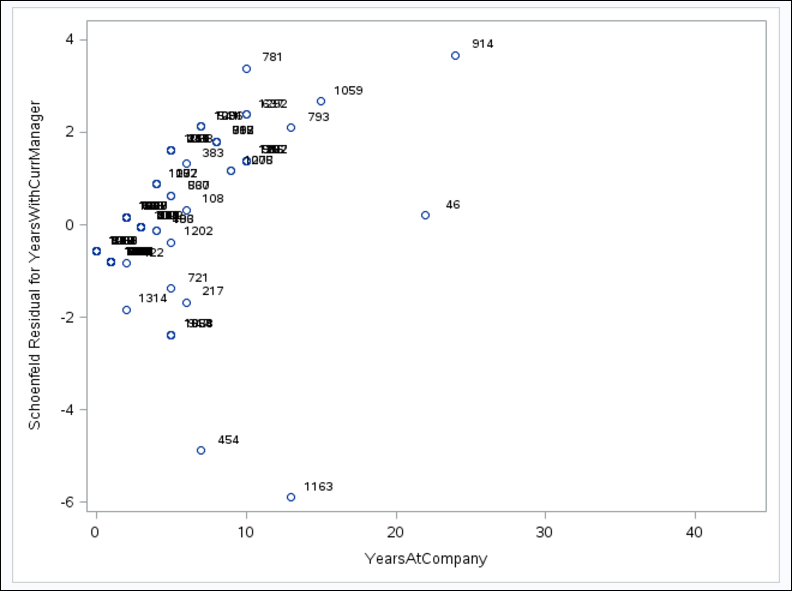
1. **Daily Rate:**

The Schoenfeld Residual graph for Daily Rate of the employees varies from -1000 to 1000.



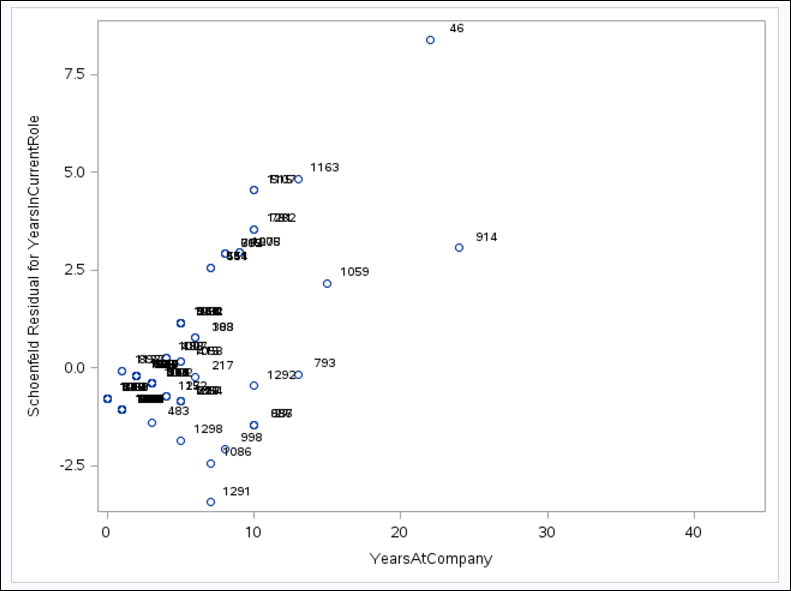
1. **Years with Current Manager:**

The scatterplot ranges between -6 to 4 for Years with current manager.



1. **Years in Current Role:**

The scatterplot ranges between -2.5 to 7 for employees with Years in current Role.



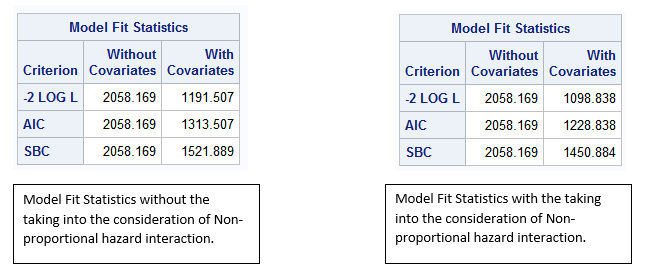
1. **Interpretation:**

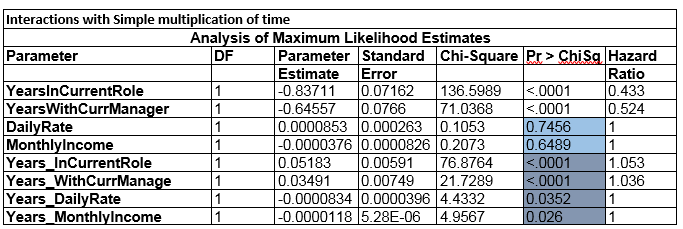
From both the tests we can see that the YearsInCurrentRole and YearsWithCurrManager are coming as non-proportional, whereas Daily Rate, Distance from home, Monthly Income and Total working Years are coming non-proportional with only Martingale residual test. We further checked time interaction of all of the above covariates and we figured that for variables **Distance from home** and **Total working years** p-value deteriorated. Therefore, we took out those two covariates. Thus, for futher analysis we considered only the following covariates

* **YearsInCurrentRole**
* **YearsWithCurrManager**
* **Daily Rate**
* **Monthly Income**

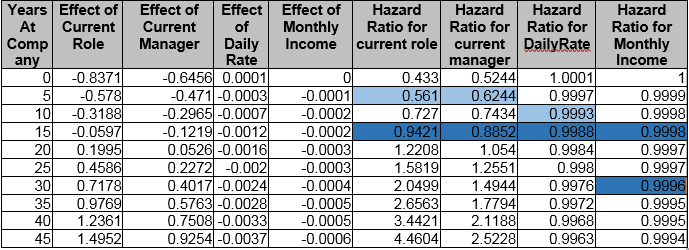
When a covariate does not pass the test, we can include an interaction between the covariate and some function of time, and here we did it using Simple multiplication of time

Simple multiplication of time.

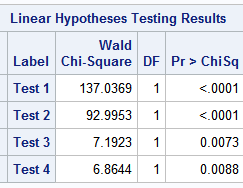
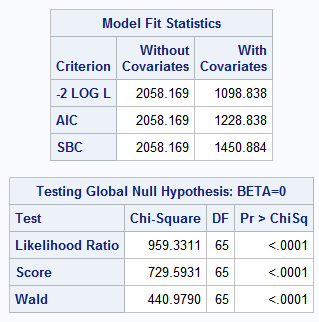




Since, Hazard ratios needs proportional hazards model whose assumptions include linearity and homogeneous standard error. To establish that we did the test whose results can be seen below:

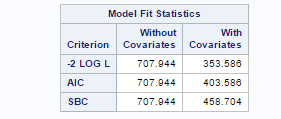


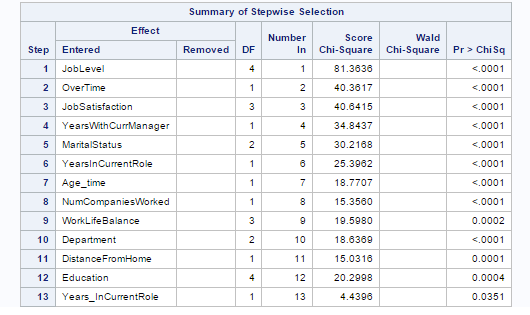
In order to find the impact of years with current role in the company, years spent with the current manager, Daily rate and Monthly we performed another separate test whose are below:

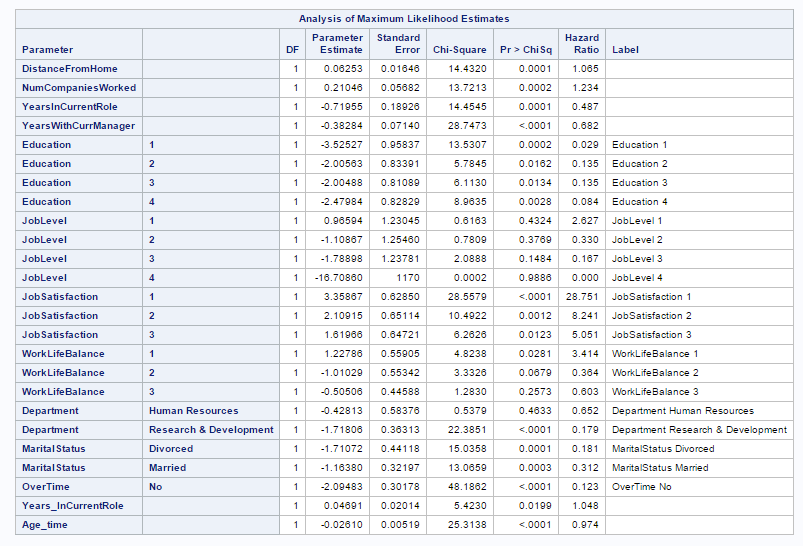


**Final Model for Voluntary Resignation sub group of employees**

To reach our final model, we used bonus (lag1) covariate and interaction terms for non proportional covariates apart from the other covariates. In order to select the significant variables we used the option – stepwise as can be seen in the below table:





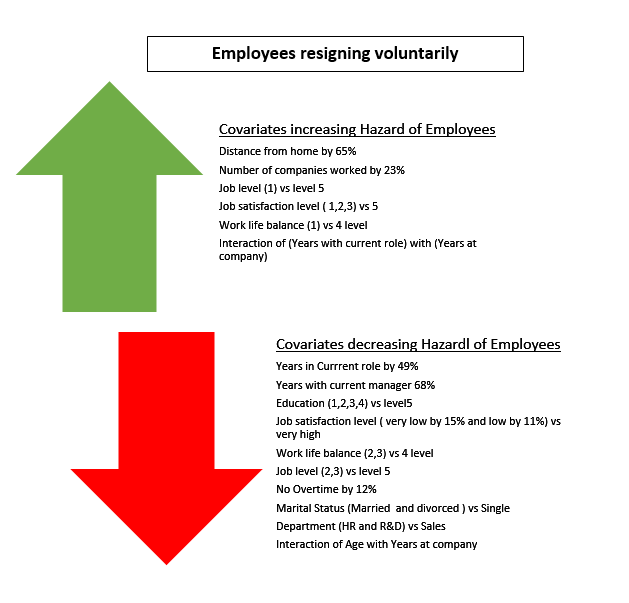


The above tables shows the list of the significant variables along with their impact on hazard ratio. It can be seen that some non-proportional covariates like age\_time and Years\_InCurrentRole are impacting the hazard ratio significantly.

Since Bonus is not significantly influencing the hazard we have not shown it here. Also we did not show any other insignificant covariates.

**Now to answer What attributes increase/decrease the hazard rates for certain event types?**

As per our entire analysis we can summarise our findings in the table below:-



# **Conclusions**

This section discusses key interesting findings from our survival models which also helps in answering the business questions in context.

1. After analysing the subset, it can be observed that it is more meaningful for us to focus on the employees who voluntarily resigned. After running the analysis with and without covariates, it can be observed that several types of events cannot be combined by the company resulting in handling of these events separately.

2. The Bonus variable does not seem to have considerable influence on hazard ratio.

3. We can also observe that employees who are younger in the company are more likely to quit the job. Therefore, it implies that Age factor has impact on the hazard ratio. Hence, when the age increases, the employees are more likely to be stable in the job and less likely to leave the organization.

5. Another interesting factor was the marital status of the employee impacting the hazard ratio. It can be observed that the employees who are married or divorced, they are more loyal for the company and are more likely to stay in the company.

6. Distance to work also plays a key role in affecting the hazard ratio. When the distance from home increases, the hazard ratio increases at the same time.

7. We also observed after the analysis that those employees who had ‘bad’ Work-Life balance are more likely to leave the company resulting in extremely high hazard ratio.

8. The other important feature affecting the hazard ratio was time spent by the employee for their respective job. We observed that the employee who does not do overtime, they have the lower probability to leave the company resulting in lower hazard ratio.

# **Recommendation**

After thorough analysis for the subset of employees containing only voluntary resignation employees, our team would like to suggest Fermalogis to take the following steps into consideration for reducing the attrition rate in the organization

1. Fermalogis should focus on Job satisfaction feature i.e., they should lay emphasis on providing basic facilities like providing the following features:

* dormitory rooms for staff
* Medical facilities
* Provide promotions to deserving and loyal employees
* Improve employees’ working environment by giving them good eating joints and sports complex to relax
* Yoga rooms and Gym

2. In order to retain the young blood, Fermalogis should assign a contract with the new employee to make them work for the company for at least 4 years. This would greatly impact the attrition rate.

3. They should focus on providing such kind of environment to the employees so that their Work-life balance is maintained which can be done as follows:

* Arrange entertainment parties for the employees on regular basis
* Avoid promoting the concept of staying late at office and by cutting down overtime activity so that the employees can comfortably manage their personal as well as professional life efficiently.

4. After identifying the employees who has the tendency to leave, Fermalogis should make an effort to redeem them on regular basis. Moreover, they should also work on finding the appropriate replacement for those employees who are definitely about to leave.

5. Fermalogis should also focus on endorsing on following factors that can go a long way to decrease the attrition rate of the company such as:

* Improving employee engagement mechanism
* Provide employees with more opportunities to address their thoughts
* Encouraging them to make the decision
* Propaganda out of box thinking

6. Set up a mechanism to award loyal employees quaterly and improve their welfare level.