**Problem Statement:**

Attrition is a major problem faced by many organizations and there is strong need to develop strategies to mitigate it. Our objective is to use the value of survival analytics in enterprise decision making, specifically in predicting the attrition rate with the goal of reducing it and to retain high-performing employees.

**Business Case:**

The COO of Fermalogis a pharmaceutical company that is facing the issue of employee attrition has written to us to understand the vital factors leading to employee dissatisfaction. Our initial analysis of the data (Project 1) answered some key questions identifying who are leaving the company, why are they leaving the company and when is the danger higher for them to leave the company.

However, a discrepancy in the data provided to us is identified and sent back for analysis, addressing the problem as ‘Turnover’ rather than ‘Attrition’. In addition to this we are provided with the Event Type that caused the Turnover and they are numbered from 0-4 as:

Coding For Event Type:

0 - No turnover

1 - Retirement

2 - Voluntary Resignation

3 - Involuntary Resignation (Health problems, family matters etc.)

4 - Job Termination, Employee is Fired

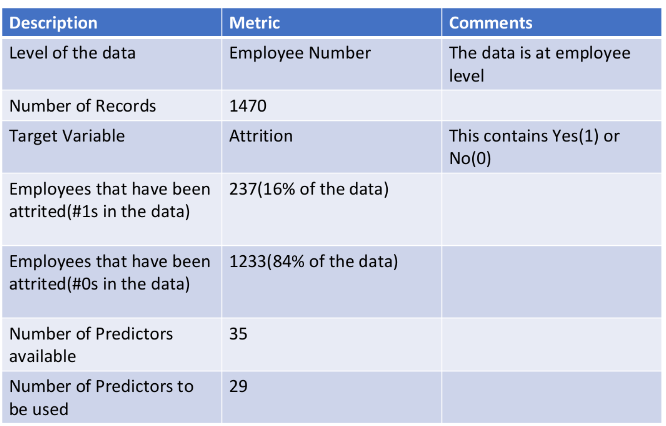
In addition to this, they would also want us to look for a relationship between the Turnover type and the Bonus variables, while also verifying if the impact caused by each of the covariates varies with Turnover event types.

Larry also tasks us with potentially answering the following questions as well:

* 1. Can I combine different event types together? Or do all need to be handled separately?
  2. What attributes increase/decrease the hazard rates for certain event types?
  3. Does bonus affect employee turnover? If yes, how?
  4. Are there any variables which affect hazards non-proportionally?

**Data Description:**

Summary Statistics for the data are as follows-



This excludes the 40 Bonus variables. A detailed description of what was done with them is provided below.

**Data Exploration:**

As a part of the Data Exploration, two types of analysis were considered-

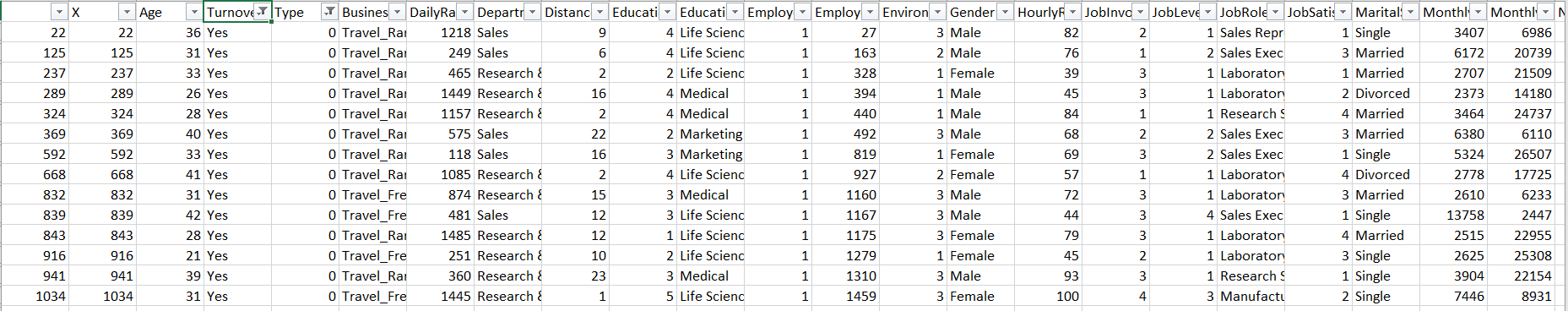
1. Univariate Analysis
2. Bi Variate Analysis

We analyzed the data for outliers and missing values, and found none in the predictors, and excluded NA’s in the Bonus variables.

In addition to the above-mentioned variables, we are also provided with a variable ‘Turnover’ – in the place of attrition and ‘Type’ describing the Turnover event type.

**Data Discrepancies:**

On performing a data quality check, we identify that there are 14 observations where a ‘Yes’ is marked against Turnover and a ‘0’ against Type, which we found to be contradictory. A screenshot of the same is provided below. Given that imputing these values based on other correlated parameters is bound to create biased data, we fix it using: IF Type = 0 THEN Turn\_dum=0; else Turn\_dum=1;



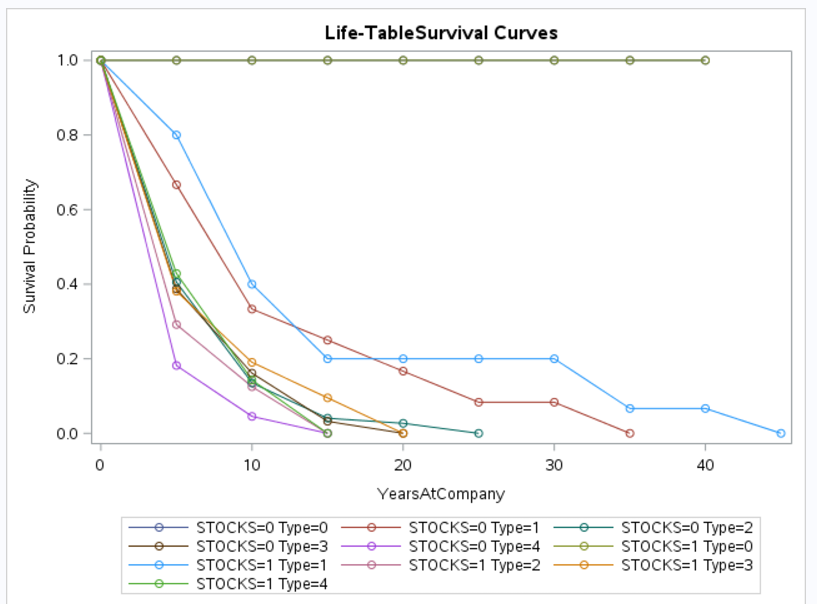
We change our Turnover data and base it on the ‘Event type’ rather than Turnover itself, because, for any given employee, identifying if they undergo ‘turnover’ based on event type (1-4) can be challenging, especially with a limited sample size.

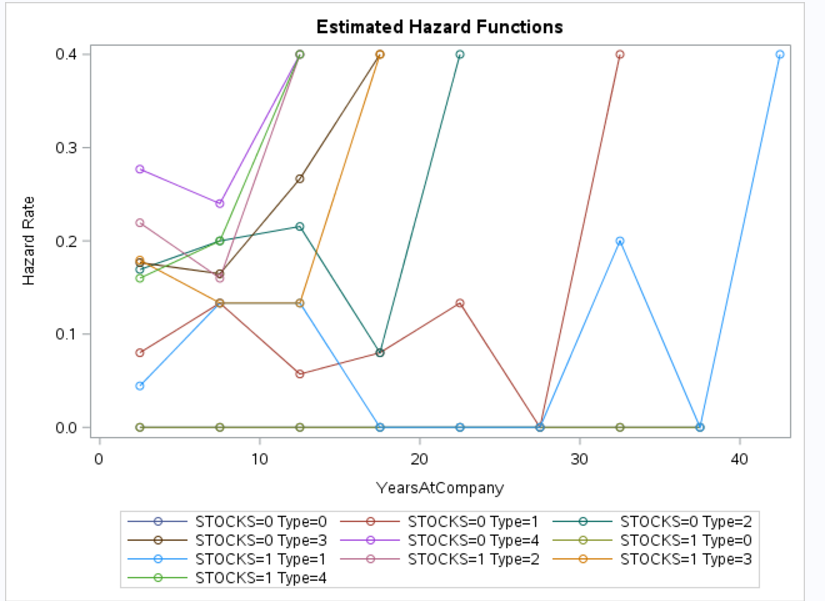
To answer one of the questions we got as feedback from Larry (professor) for Project 1:

**Do the employees who have the stock option not quit or the employees who do not quit get the stock option?**

We create a new variable for Stock Option Level as Stocks such that we can group those with Stock>0 as ‘1’ and Stock=0 as ‘0’. We essentially classify employees based on whether Stock option has been given to them or not.

We then use it along with Type as Strata and evaluate the Survival and Hazard curves for each combination.

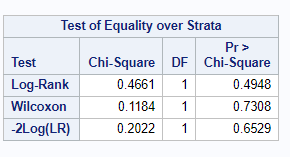


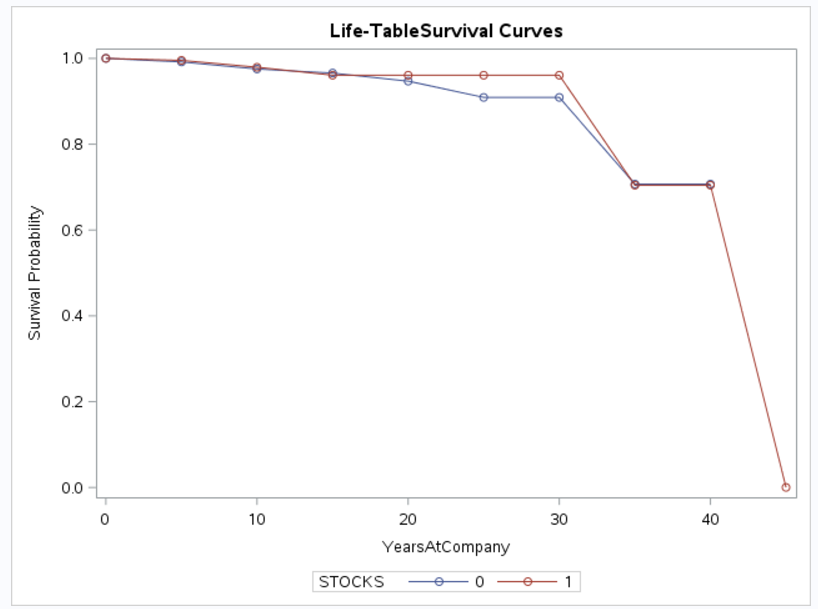


The above graphs suggest that employees who have no stocks but encounter turnover due to job termination/ are fired are bound to have high hazard rate initially. This makes sense because those who are likely to get terminated are perhaps those who haven’t been performing well over a time, and therefore those who haven’t been given stock options. It is interesting to note that those who have a likelihood of being fired are those whose tenure at the company has been less than 10 years. Similarly, we observe maximum survival rates for those who seek retirement and relatively low hazard rate initially and a sudden rise in the hazard rate as their tenure exceeds 40.

But, to answer Larry’s question, we dig a little deeper and using PROC LIFETEST plot individual survival plots against each turnover type, and the results are as below:

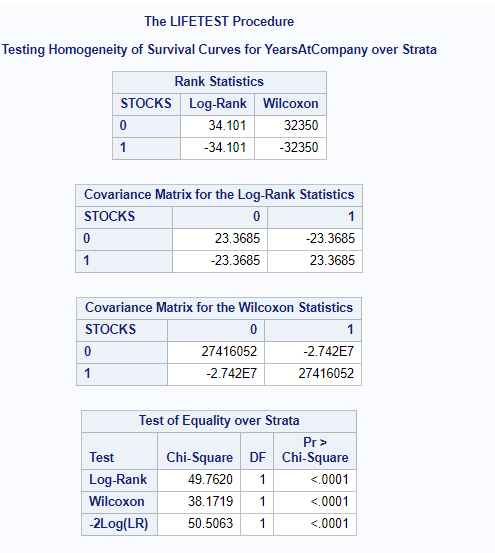
* Retirement and Stocks:



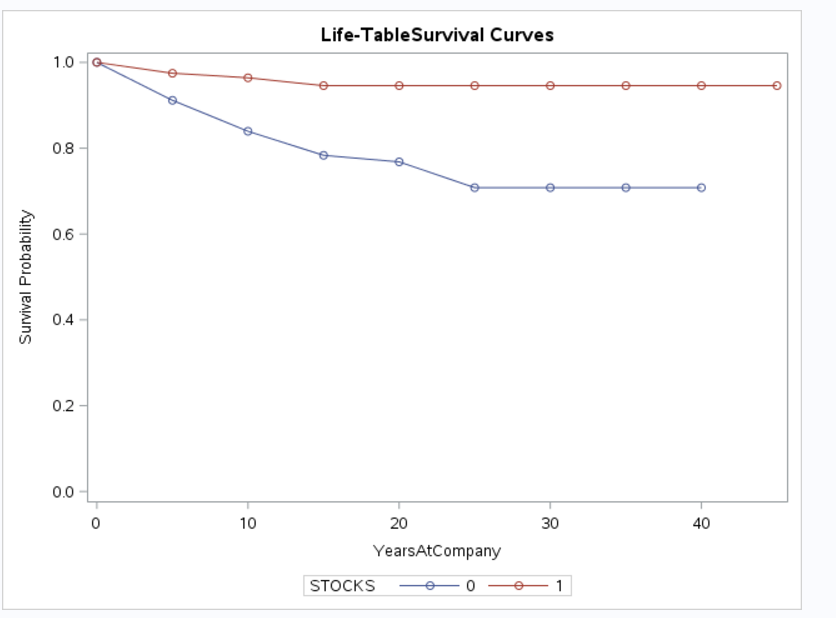


This says that there is no significance difference between the retirement employees with respect to having stocks.

* Voluntary Resignation vs Stocks:



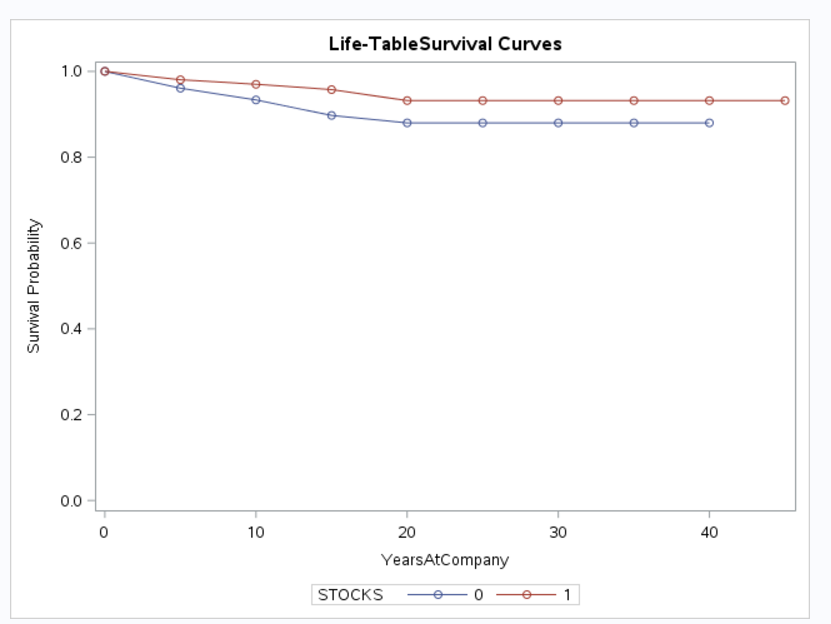
This indicates a significant difference between those who own stocks and those who don’t with respect to Turnover.

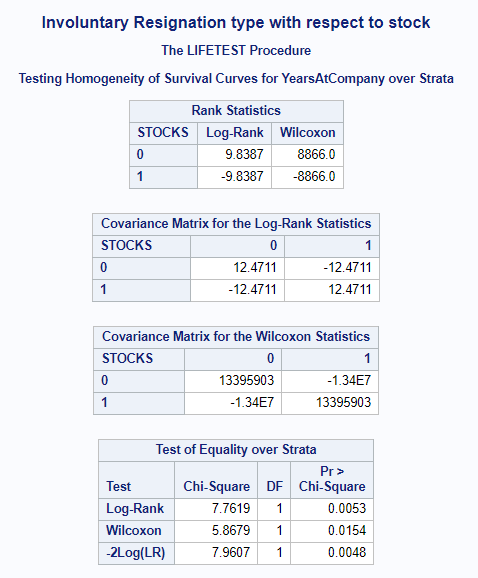


It is evident that employees who hold stocks in the company have a flat survival curve implying that they are not the ones who are leaving the company. However, people without stocks have a decreasing survival rate leading to turnover. This might be of interest to Larry, who can discuss this further to mitigate the risk of turnover being caused by this event.

* Involuntary Resignation:

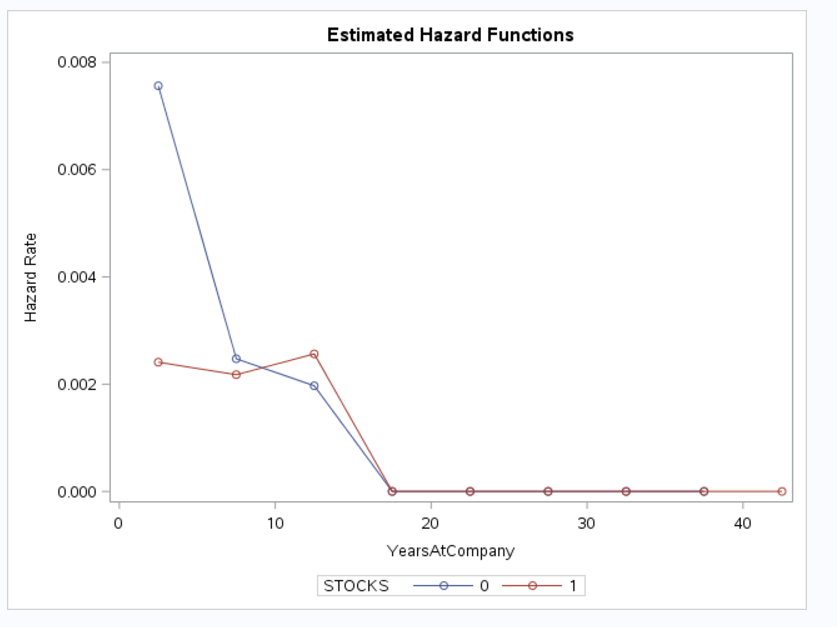
Significant difference between those who have stock option versus those who don’t is implied in the statistical tests. It is possible that those that own stocks may take a leave temporarily for personal reasons but have sufficient motivation to get back once things settle down.

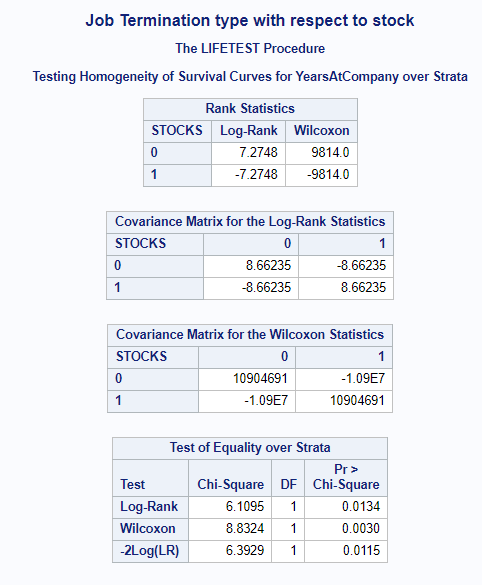
* 



* Job Termination:

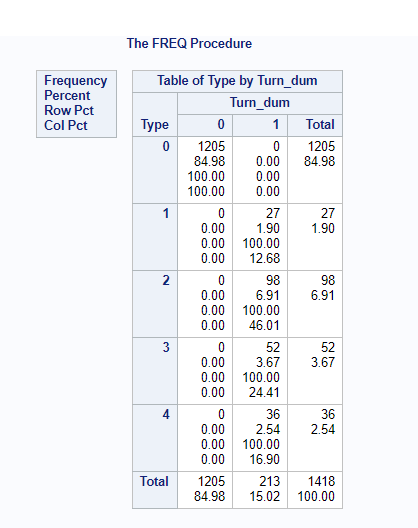
It is interesting to note that the Hazard rate for employees who have a tenure of less than 15 years show varying behavior with and without stocks. Those with stocks have a considerably low termination rate when compared to those who don’t have stocks. Additionally, post the 15-year tenure, we see no significant difference between employees with and without stocks.

* 



**Methodology:**

Since we are now tasked with observing ‘Turnover’ over ‘Attrition’, let us look at the number of the employees from the sample size that belong each of the turnover types.



**Increasing Scope:**

When Larry got back to us, we identified the need for a better model. What changed? The PROC LIFEREG used in project 1 limited the scope of our project by not being robust with respect to the distributions it took, by being parametric and by making the task of accounting for time dependent variables very exhaustive. This leads us to the PROC PHREG, more commonly known as the COX’s Regression Model. What makes it better?

* Semi-parametric nature that doesn’t need it to fit to a distribution, makes it robust in nature
* Ease of incorporation of time-varying covariates
* Allows for stratification
* Quick, when compared to other methods
* Can consider both – Proportional and Non-proportional hazards

It differs from the previous models since it considers that the hazard for any individual is a fixed proportion of the hazard for any other individual and varies as a function of β using partial likelihood estimates in the place of maximum likelihood estimates which in our case indicates the ratio of hazard rate between Turnover and No Turnover, the hazard being the probability of employees leaving the company(Turnover) to the employees who are not leaving the company at a given point of time.

Is our model proportional in nature? And, if not, what factors cause non-proportionality? This will be answered in the later sections. This also leads us to Can I combine different event types together? Or do all need to be handled separately?

Before dwelling into other details, we felt the need to identify the time dependent covariate in this dataset. We found that to be the Bonus variables.

To add a time varying element to Bonus variables, we tried several models and identified that the cumulative bonus option gave us the most optimal result.

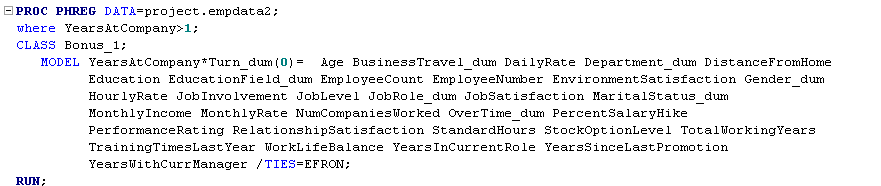
Let us look at the influence the Bonus variables had on employee turnover.

**Model Trails: (part c)**

To detect the impact of the bonus variable on the employee turn over rate, there were several models built and Model fit statistics was used to validate the model. The models can be classified broadly under three different cases.

**Case 1**: **The bonus variables(Bonus\_1 to Bonus\_40) were considered as a class individually to check if they played a significant role for the people who worked during the respective periods.**

Sample Code-



Results of the models built around the variables -Bonus\_1, Bonus\_2 and Bonus\_3 are illustrated in the following screenshots.

Bonus\_1 as class variable -

|  |  |
| --- | --- |
| With Class Variable | Without Class Variable |
|  |  |

Bonus\_2 as class variable -

|  |  |
| --- | --- |
| With Class Variable | Without Class Variable |
|  |  |

Bonus\_3 as class variable -

|  |  |
| --- | --- |
| With Class variable | Without class variable |
|  |  |

Interpretation of the Results –

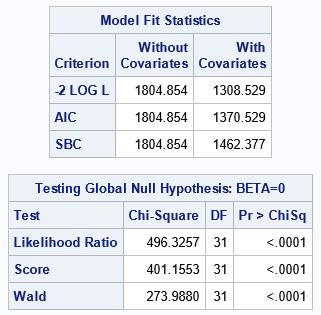
A common trend which was absorbed was the difference in -2 LOG L values between the models built -with and without class. It was almost zero on an average This implies that bonus variables do not play a significant role when treated as a class individually

**Case2: Time dependent covariates measured at regular intervals**

All the bonus variables (Bonus\_0 to Bonus\_40) were included in the model and it was observed that the Bonus\_var was not significant (p-value > 0.05)



The -2 LOG L value is very high and hence this model is also not significant.



**Case 3: The cumulative bonus is calculated and used as a single variable to find out if it plays a significant role and two models with two different types of approximation methods were built to find out the better method between the two**.

Cumulative bonus variable with TIES = Efron, Discrete

|  |  |
| --- | --- |
| TIES = Efron | TIES = Discrete |
|  |  |

The -2 LOG L values(With Covariates) are relatively smaller when compared to the other models and also there is a huge difference between them when the approximation method is changed from Efron to Discrete. Since the smaller the values of -2 LOG L imply that the model is more accurate, the DISCRETE approximation method was the ideal one for this data.

While statistical evidence tells us that we should pick TIES=Discrete, is this the right way to go? Decisions, Decisions.

**Digging deeper to understand the concept of Tied data:**

Partial likelihood is valid only for data in which no two events occur at the same time. It’s quite common for data to contain tied event times, however, we need an alternative formula to handle those situations. Most partial likelihood programs use a technique called Breslow’s approximation, which works well when ties are relatively few. But when data are heavily tied, the approximation can be quite poor.

We anticipated that the EFRON method would work the best. The EFRON method is supposed to be quick and works well with small datasets having heavily tied data. Efron’s approximation is not bad for drawing qualitative conclusions, but there is an appreciable loss of accuracy in estimating the magnitudes of the coefficients.

However, we also thought about the best fit model we can generate in this scenario. Then, why not choose the EXACT method? Unlike the EXACT model, which assumes that ties are merely the result of imprecise measurement of time, the DISCRETE model assumes that time is discrete. When two or more events appear to happen at the same time, there is no underlying ordering—they really happen at the same time.

And, this is true in this case. If 2 employees appear to leave in that year, they are identified as people leaving in that year, the exact time in terms of Month or Day remain unknown.

For our analysis, we could use either of the above-mentioned(EFRON and DISCRETE) models.

TIES = Efron -



TIES = Discrete -



The CumBonus\_var which is the cumulative total of the bonus, seems to be insignificant variable in this model as well, with a negative relationship with the hazard ratio of the employee turnover rate.

Inference:

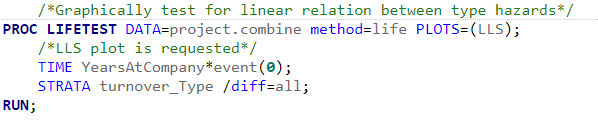
The cumulative total of the bonus does not play a significant role in the employee turnover rate either individually or in cumulative format.

**Hazards/Event Types: (part A)**

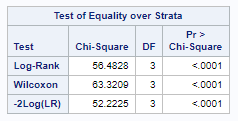
Whenever a dataset is characterized by several event types, a decision needs to be made on whether these event types must be handled individually or together. For this, we decided to go with competing risks in survival analysis which distinguishes between different types of events and when that type of event is considered, the other types are accepted as censored.

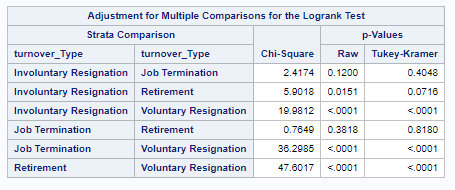
We see that the employees that have undergone Turnover constitute about 15.02 % of the data, of which about 46% belong to a Turnover type of 2 – Voluntary Resignation.

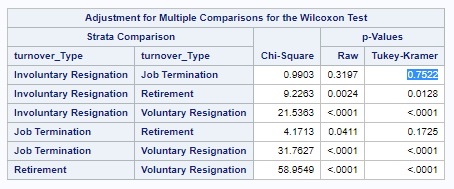
Approach: We have created individual datasets by filtering out the observations corresponding to each type category (1, 2, 3, and 4). Then, we have given a new variable Turnover type and later combined all the data sets together. We performed a proc life test on the aggregated dataset with method as life and plots as LLS with strata variable as Turnover type and got the following results:



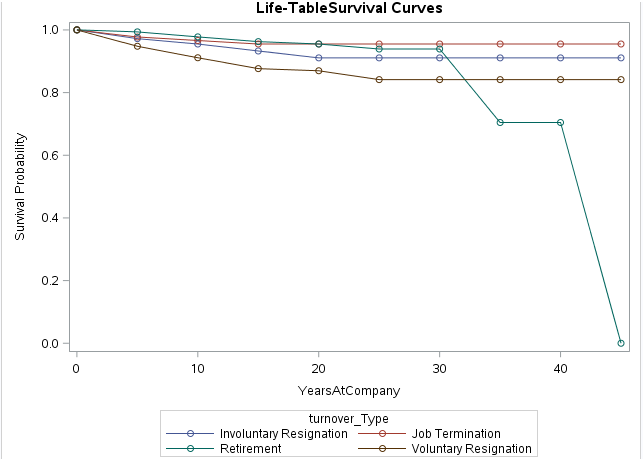
We used the above run for this analysis. Our observations from the screenshot below indicates that this model is significant.



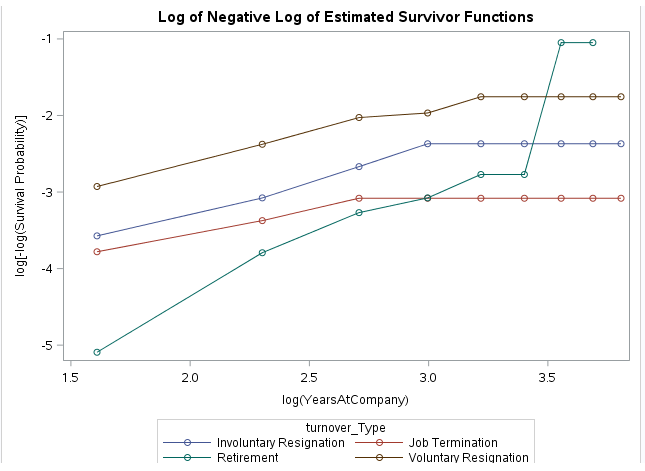




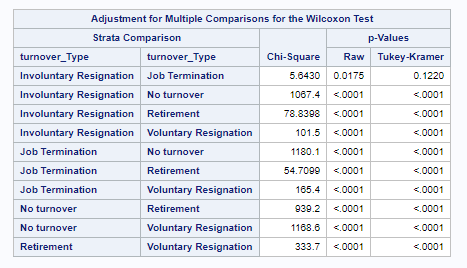
From the strata comparison of variables using the p-values from Tukey-Kramer method proves that the categories involuntary resignation and Job termination are same. The p-value is 0.7522 which is greater than 0.05. So, it is evident that these two categories can be combined.

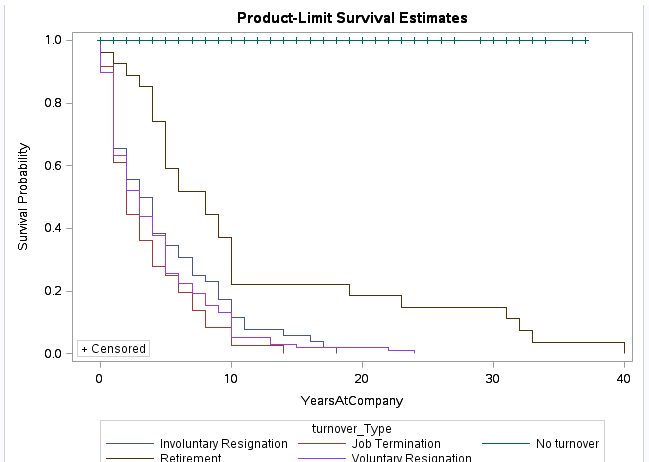


From the life table survival curves, we see that the lines corresponding to involuntary resignation and Job termination are parallel. They do not coincide at any point. From the negative log of estimated survivor functions, the w values are same for involuntary resignation and job termination employees.

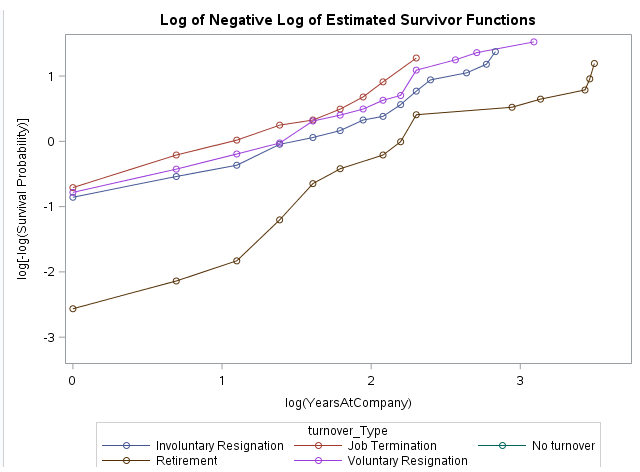


In a different approach, we used the provided dataset and performed strata comparison on the turnover type variable. Checking the results and finding the comparisons of each Turnover type with respect to the p-values to find if the event types are significant. This test helps us to understand if both the Turnover types are different or same. From the results, we understand that “Involuntary Resignation” and “Job Termination” are not different based on the test of p-value < 0.05. So, we choose to combine these two Turnover types.



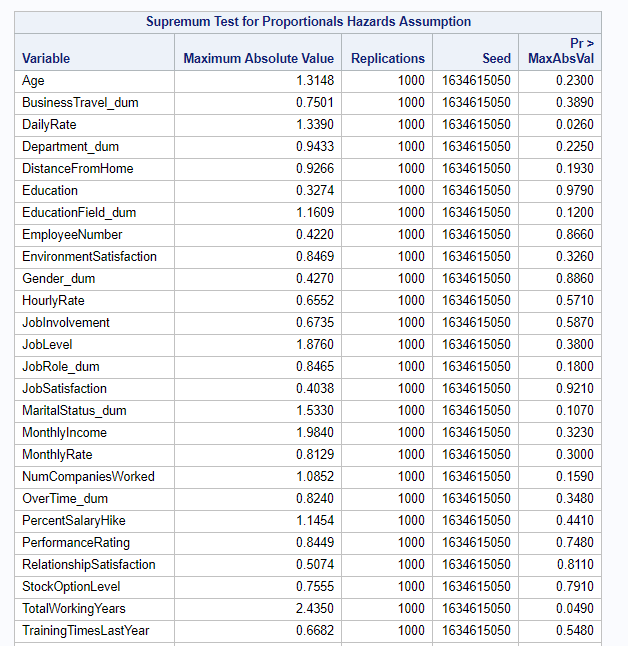


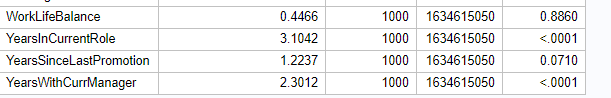
From the product limit survival estimates, we understand that the survival probability of turnover type ‘Job termination’ is the least up to 15 years at company. Below graph shows the log of negative log of estimated survivor functions (LLS plot).



**Testing for nonproportional hazards using residuals: (Part D)**

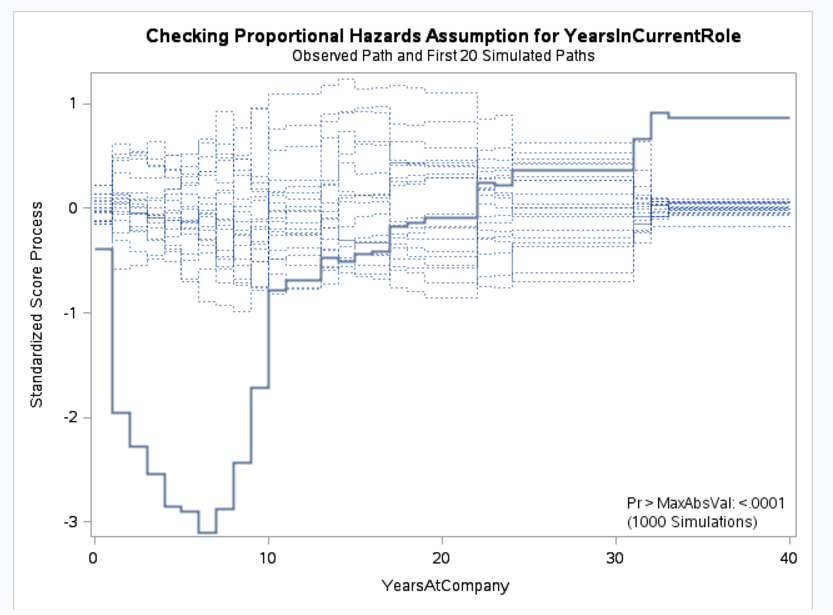
Violations of the PH assumption are equivalent to interactions between one or more covariates and time. That is, the PH model assumes that the effect of each covariate is the same at all points in time. If the effect of a variable varies with time, the PH assumption is violated for that variable. We use the Martingale residuals method to test for nonproportionality. This method has been incorporated into the ASSESS statement in PROC PHREG by excluding time dependent variables to avoid bias. The below simulation graph shows a thick blue line indicating that the actual pattern heavily deviates from the simulated possibilities with an insignificant p-value proving that it is a time dependent covariate. The results of the Supremum tests are as below and the graphs of time dependent covariates have been depicted as well:

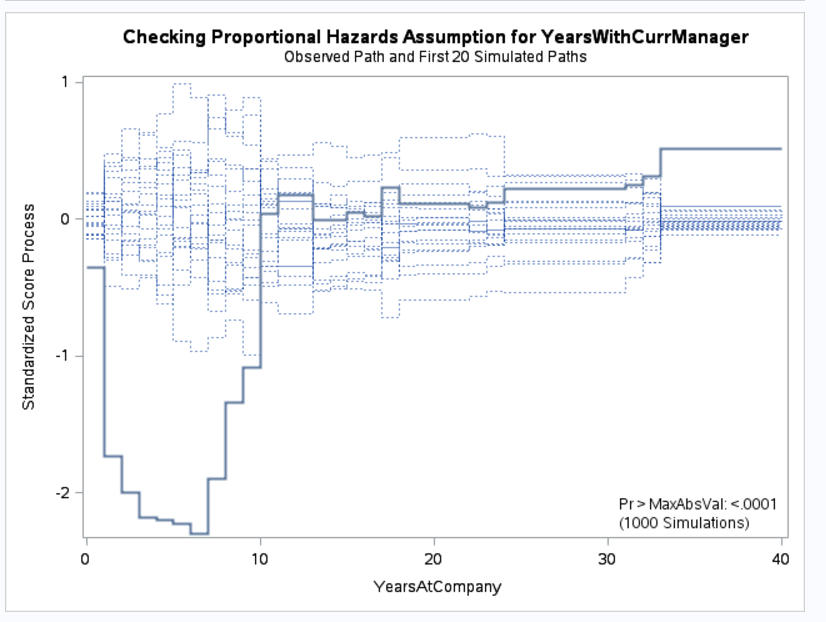




The observed process deviates markedly from the simulated processes, generating evidence against the

proportional hazards assumption.Time dependent covariates can be identified as : YearsInCurrentRole and YearsWithCurrManager





**Handling Non-Proportionality:**

We introduce interactions with respect to the years at company. For each given observation, we create new variables by multiplying the non-proportional variables with years at company. It is interesting to see that the variable ‘Yearswithcurrmanager’ which is insignificant otherwise has a hidden effect in the form of interaction terms.

Timeinteract1 = YearsAtCompany\*YearsinCurrentRole

Timeinteract2=YearsAtCompany\*YearswithCurrManager

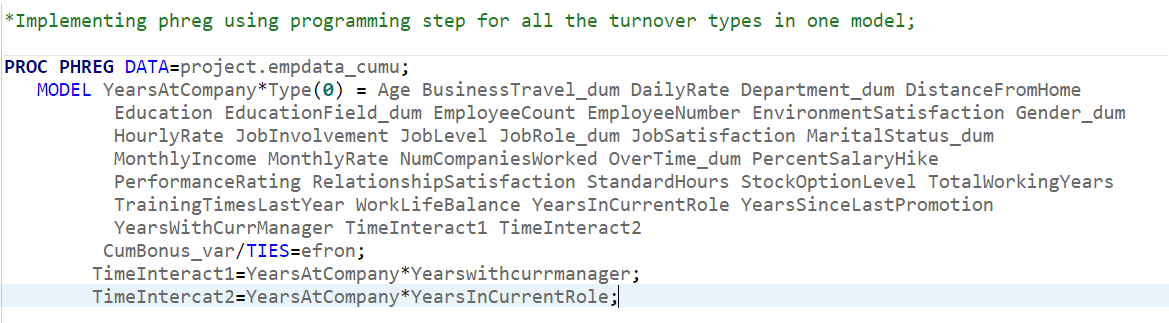
**Evidence: (also comes under part d)**

**Case 1:**

To check the possibility of combining event types, our team has decided to apply PHREG model for each event type and the nested event type.

We have modelled using PHREG for each event type:

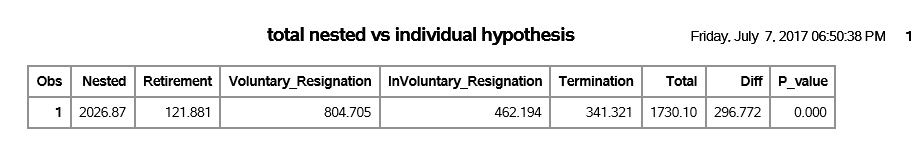
Sample Code:



We build similar models for each turnover type.

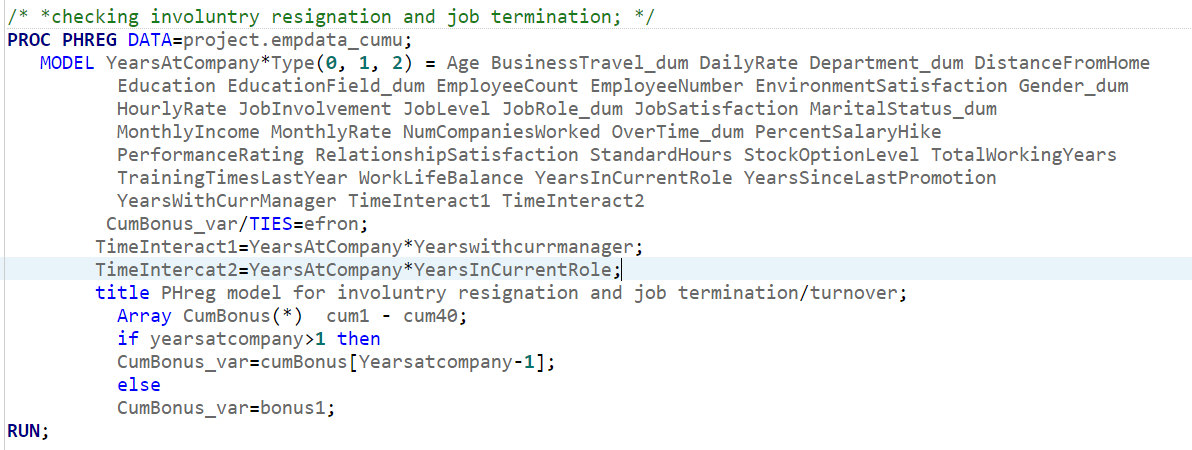
Essentially what we are doing here is performing Hypothesis Testing, with our Null Hypothesis being:

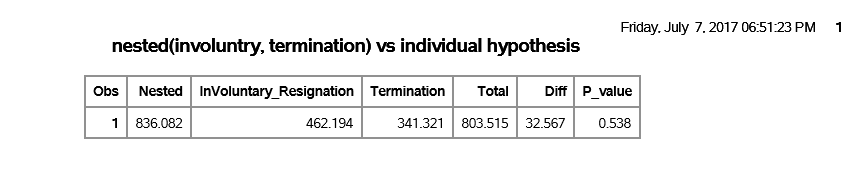
Ho: Nested event types have the same effect on turnover of employees as any Individual event type. We look at the p-value, which in this case is 0, and reject the null hypothesis. This proves the fact that we need to build an independent model for each event type.



**Case 2:**

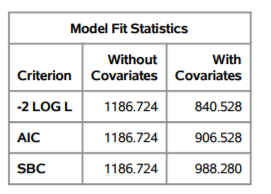
From the Tukey-Kramer test, we identify that Involuntary resignation and Job termination are linearly related so we built a model comprising of them both as event types.

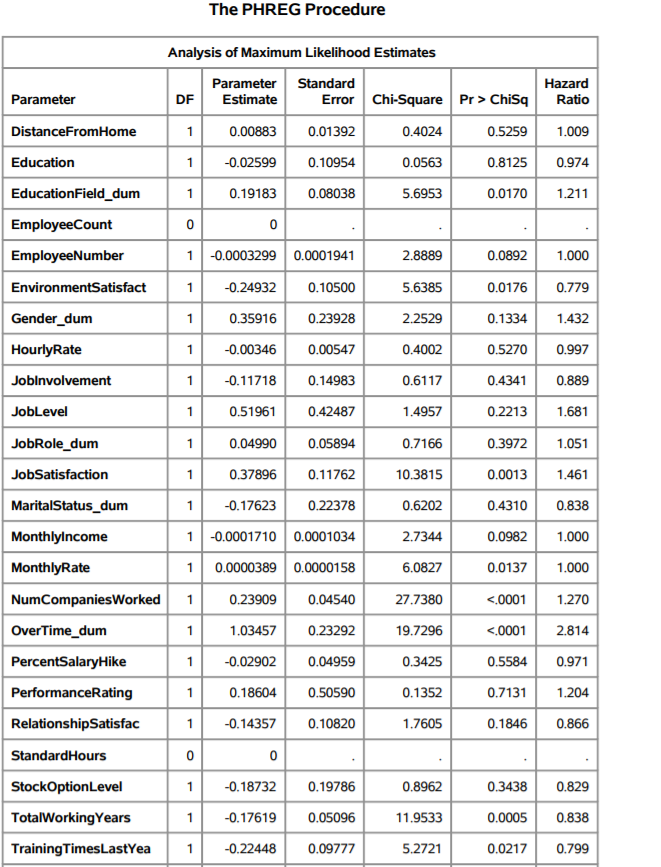


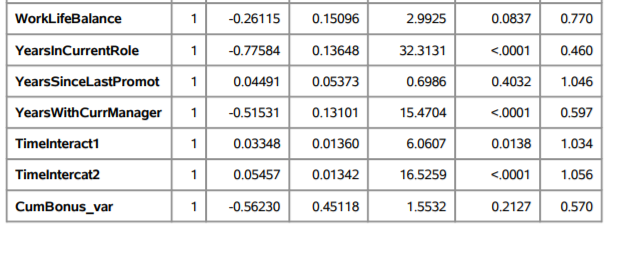


**Conclusion:**

Here, p>0.05, implying that we fail to reject Ho, which tells us that we can indeed consider a nested model. Generating a model with involuntary Resignation and termination as the event type and all other types as censored gives us a model which is significantly different and better than the Null model (Case1: All types)



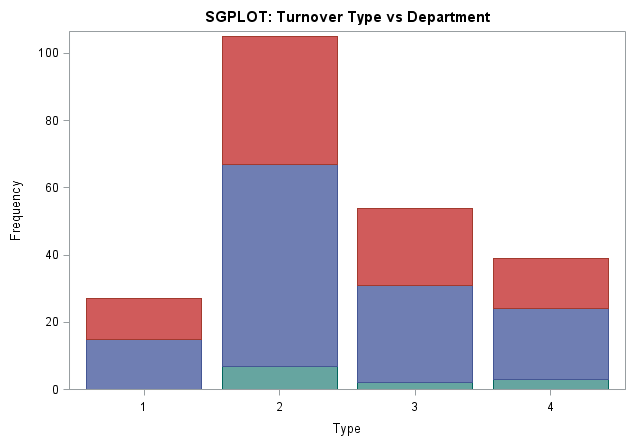
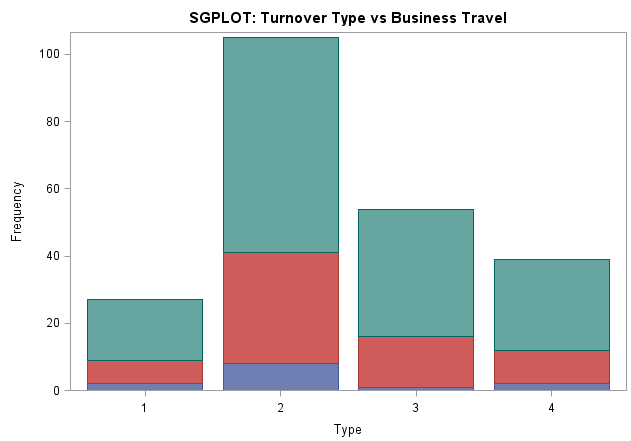
**Significant Variables in Nested Model:** 

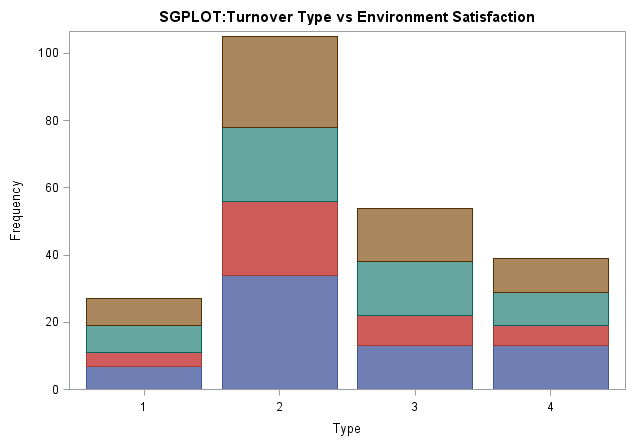
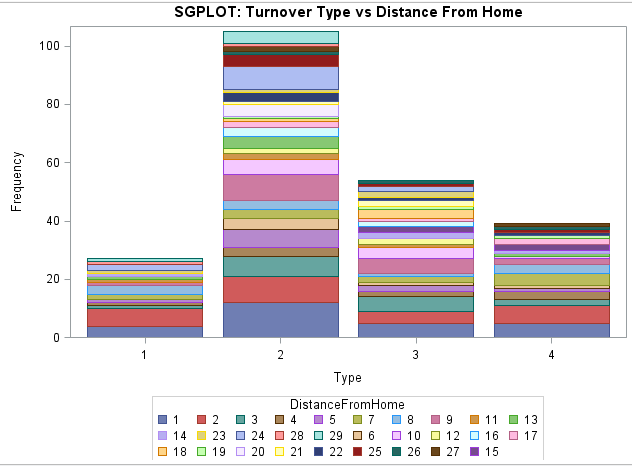


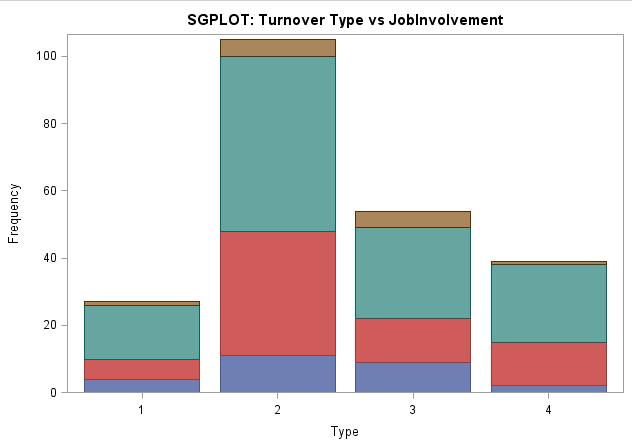
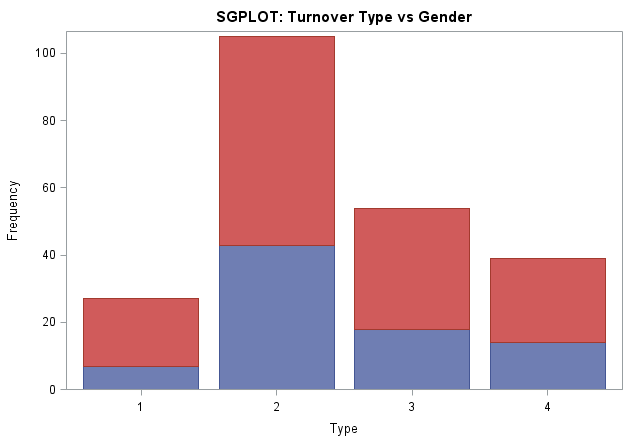
**Impact of attributes that increase/decrease the hazard rates for certain event types: (part B)**

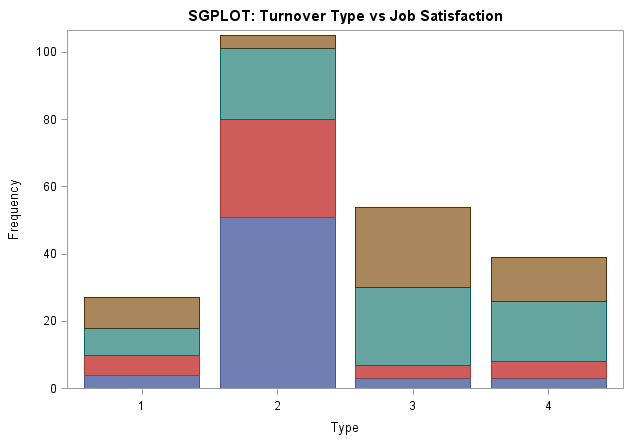
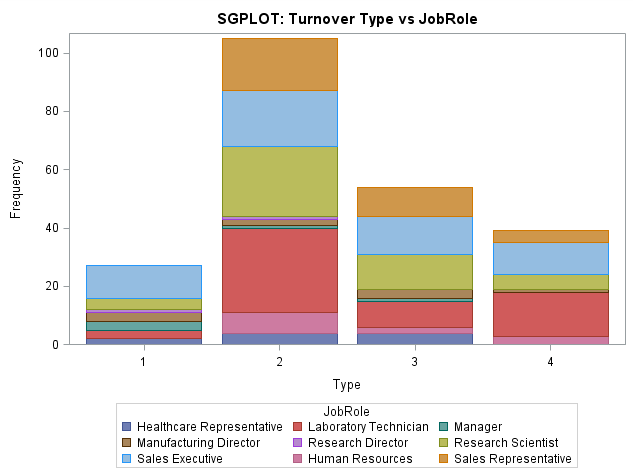
While the above-mentioned model remains our best model, we also looked beyond just finding the correct model. We also looked at how the significant variables in each case impact employee turnover. The reason we didn’t base our model based on just significant variables was because, we didn’t see much variation in the Log likelihood without insignificant variables.

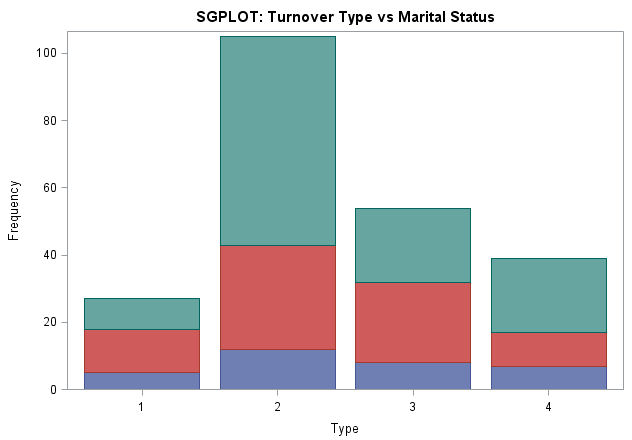
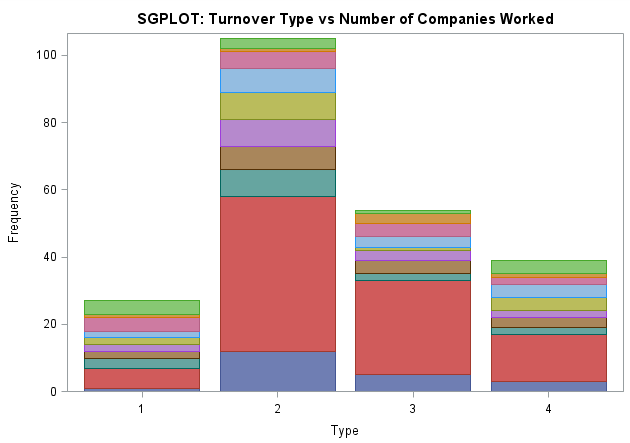
The plots below give us a general idea of how certain factors seem to be impacting Employee Turnover and these are self-explanatory:

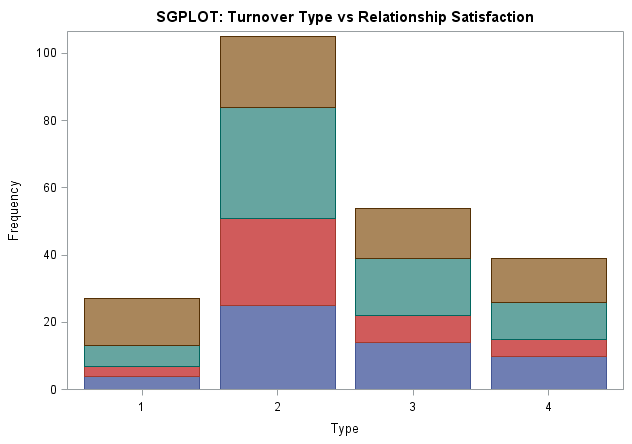
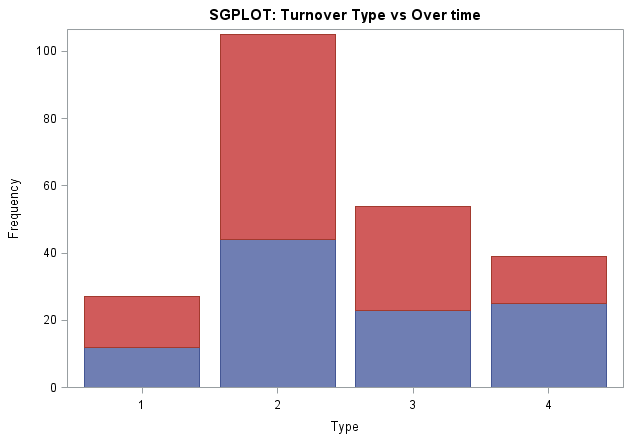


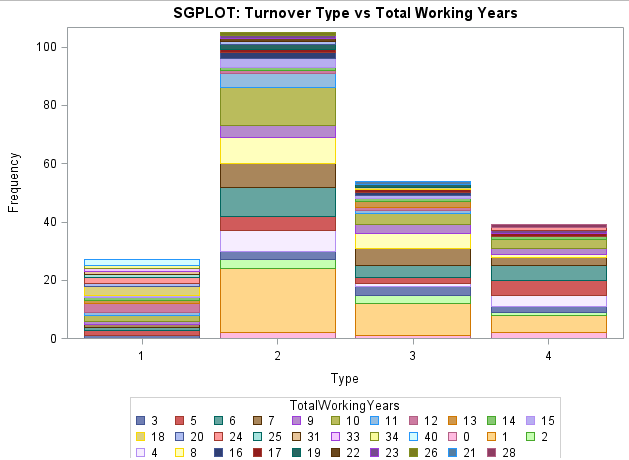


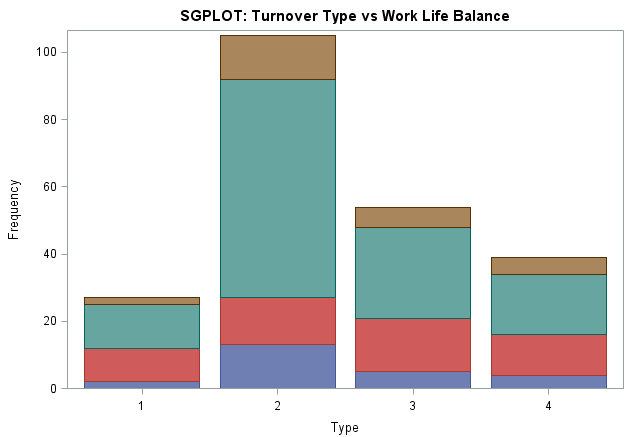
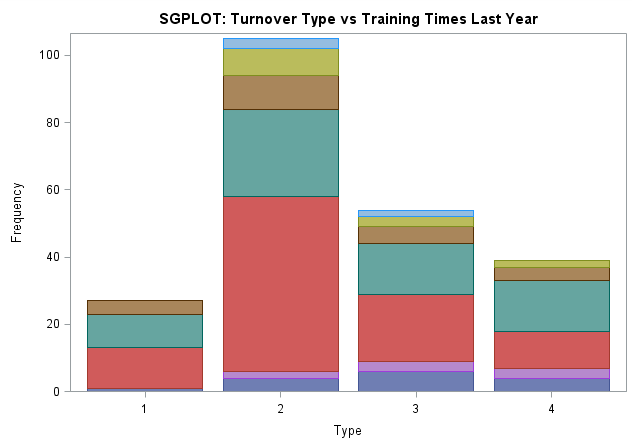


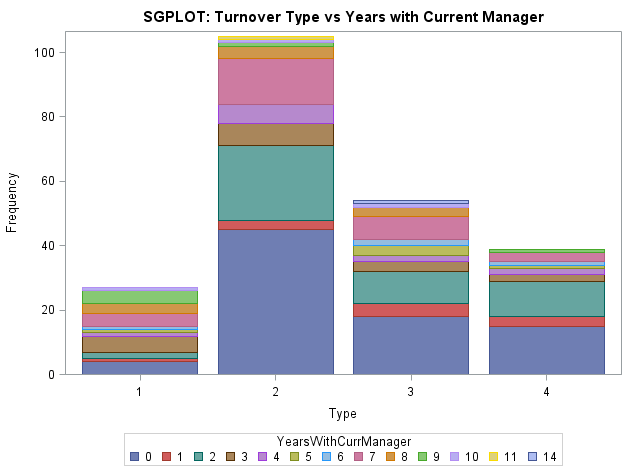
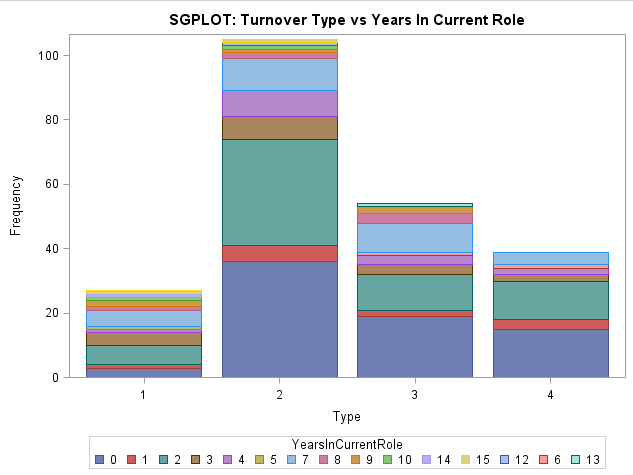












Looking at just the Hazard rates, we identify the below results:

**For Type2 (Voluntary Resignation):**

Interpretation of the effect of significant variables on the hazard rate of employee turnover

|  |  |  |
| --- | --- | --- |
| Significant variable | Positive/Negative relationship | Magnitude of effect on the hazard rate of employee turnover with increase in 1 unit of the significant variable |
| NumCompaniesWorked | Positive | Hazard rate of employee turnover increases by a ratio of  1.346 |
| Overtime Dum | Positive | Hazard rate of employee turnover increases by a ratio of  0.391 |
| Training Time Last year | Positive | Hazard rate of employee turnover increases by a ratio of  0.583 |
| YearsInCurrentRole | Positive | Hazard rate of employee turnover decreases by a ratio of  0.383 |
| YearsWithCurrManager | Negative | Hazard rate of employee turnover decreases by a ratio of  0.379 |

**Type 3 : Involuntary Resignation**

Interpretation of the effect of significant variables on the hazard rate of employee turnover

|  |  |  |
| --- | --- | --- |
| Significant  variable | Positive/Negative  relationship | Magnitude  of effect on the hazard rate of employee turnover with increase in 1 unit of  the significant variable |
| Environment  Satisfaction | Positive | Hazard  rate of employee turnover Increases by a ratio of 0.331 |
| Job Level | Positive | Hazard  rate of employee turnover Increases by a ratio of 0.042 |
| Job  Satisfaction | Positive | Hazard  rate of employee turnover Increases by a ratio of 0.206 |
| YearsInCurrentRole | Positive | Hazard  rate of employee turnover Increases by a ratio of 0.042 |

We also identify that For Type0,1 and Type 4, the increase /decrease of parameters is not significantly affected by parameters.

# RECOMMENDATIONS:

# These are some recommendations that hold good for all types of Turnover types:

**Job and Environment Satisfaction:**

* The company must assess job satisfaction in regular intervals, and this can be done by:
* Ensuring that employees participate in surveys.
* Employee and manager discussions could also help gauge the job satisfaction of employees. In case of low satisfaction levels, reasons should be identified and if found to be relevant, measures must be taken to improve satisfaction.

**Promotion:**

* Given the employee attrition, the company must make necessary changes to their Appraisal process and offer other perks, otherwise.

**Overtime:**

* Company should ensure minimal extended hours for employees with appropriate compensation.
* b. To ensure that this doesn’t hit the company’s budget, checks must be done to verify if it is necessary for them to work over-time. This should be communicated to their respective managers and policies must be implemented to ensure that employees have approvals of the managers while doing so.

**Business Travel:**

* The company should revise their travel policies and consider distributing business travels across different employees to curb travel frequency of one employee. They should also consider using options such as WebEx, wherever applicable.

**Monthly Income:**

* The company should look at the demographics (cost of living) , the tax rate across the areas in which the employees are based, in addition to their performance while revising their salaries.

**StockOption Level:**

* The company must encourage employees to purchase stock or make it available to them at a lower price so ensure longer tenure.