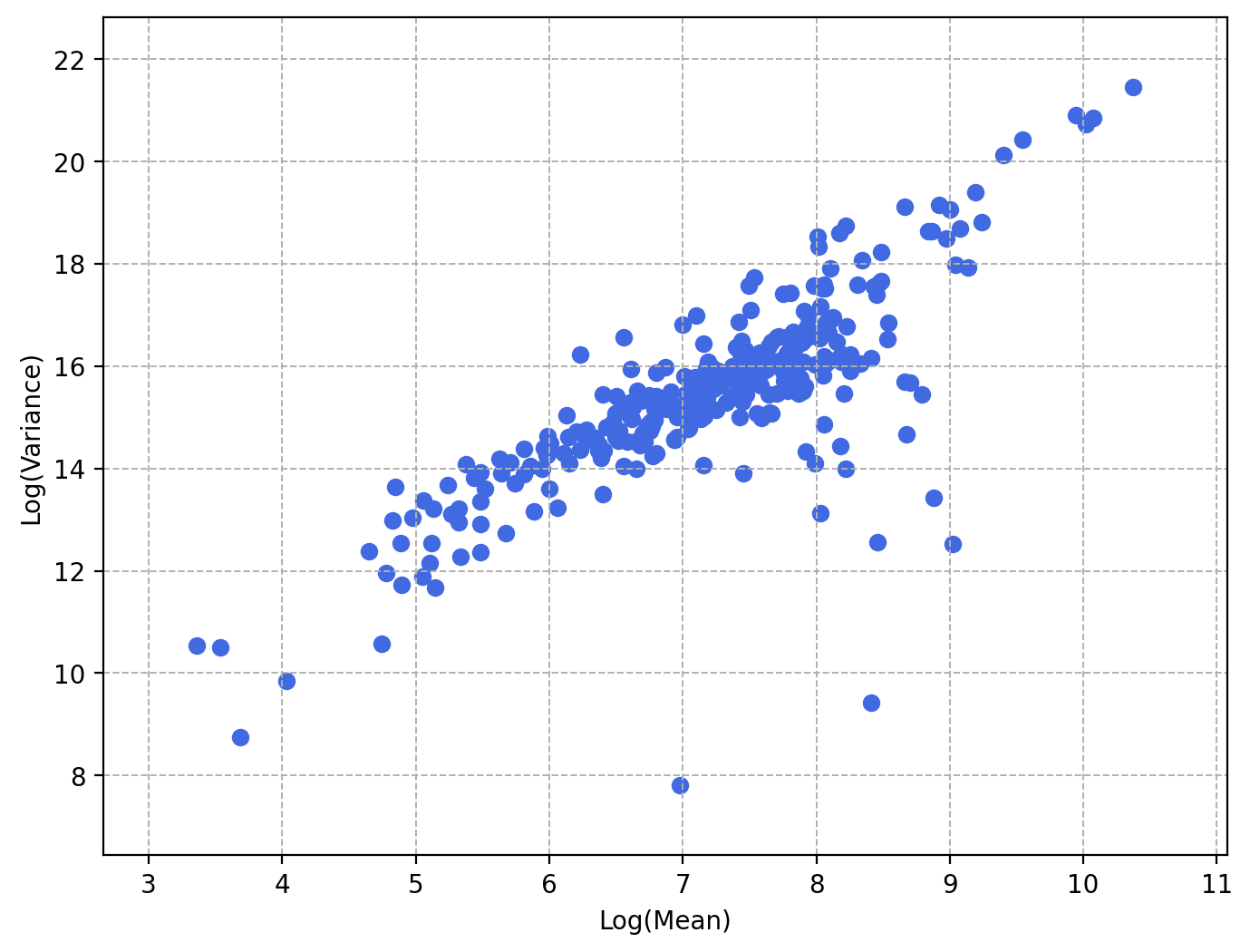
**ASSIGNMENT 5**

**QUESTION 1**

1. **(10 points). We will first estimate the Tweedie distribution’s Power parameter 𝑝 and Scale parameter 𝜙. To this end, we will calculate the sample means and the sample variances of the claim amount for each value combination of the categorical predictors. Then, we will train a linear regression model to help us estimate the two parameters. What are their values? Please provide us with your appropriate chart.**



The Power parameter p = 1.2840608802234919

The Scale parameter 𝜙 = 495.39032035257253

1. **(10 points). We will use the Forward Selection method to enter predictors into our model. Our entry threshold is 0.05. Please provide a summary report of the Forward Selection in a table. The report should include (1) the Step Number, (2) the Predictor Entered, (3) the Model Degree of Freedom (i.e., the number of non-aliased parameters), (4) the Quasi-Loglikelihood value, (5) the Deviance Chi-squares statistic between the current and the previous models, (6) the corresponding Deviance Degree of Freedom, and (7) the corresponding Chi-square significance.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Predictor** | **N Non-Aliased Parameters** | **Quasi Log-Likelihood** | **Deviance ChiSquare** | **Deviance DF** | **Deviance Sig.** |
| 0 | Intercept | 1 | -2217255.2459542 | NaN | NaN | NaN |
| 1 | URBANICITY | 2 | -2118973.9898024 | 506.5534053 | 1 | 0.0000000 |
| 2 | EDUCATION | 6 | -2057057.2451762 | 333.8694706 | 4 | 0.0000000 |
| 3 | CAR\_TYPE | 11 | -1999000.2771733 | 322.2537740 | 5 | 0.0000000 |
| 4 | PARENT1 | 12 | -1953492.1239025 | 259.7083244 | 1 | 0.0000000 |
| 5 | MVR\_PTS | 13 | -1918087.9738579 | 206.7168485 | 1 | 0.0000000 |
| 6 | TRAVTIME | 14 | -1902663.0152115 | 91.7091556 | 1 | 0.0000000 |
| 7 | CAR\_USE | 15 | -1888044.9607430 | 87.6009339 | 1 | 0.0000000 |
| 8 | REVOKED | 16 | -1873857.9187160 | 85.6612435 | 1 | 0.0000000 |
| 9 | KIDSDRIV | 17 | -1860340.5734043 | 82.2211227 | 1 | 0.0000000 |
| 10 | TIF | 18 | -1848415.9277319 | 73.0475183 | 1 | 0.0000000 |
| 11 | INCOME | 19 | -1836307.0066460 | 74.6417756 | 1 | 0.0000000 |
| 12 | MSTATUS | 20 | -1828104.4856810 | 50.8864359 | 1 | 0.0000000 |
| 13 | CAR\_AGE | 21 | -1823989.5801422 | 25.6379077 | 1 | 0.0000004 |
| 14 | YOJ | 22 | -1820045.7816882 | 24.6229350 | 1 | 0.0000007 |
| 15 | HOMEKIDS | 23 | -1818924.8912574 | 7.0121634 | 1 | 0.0080958 |
| 16 | GENDER | 24 | -1818178.6998704 | 4.6701449 | 1 | 0.0306913 |
| 17 | RED\_CAR | 25 | -1817282.3284297 | 5.6113845 | 1 | 0.0178442 |

1. **(10 points). We will calculate the Root Mean Squared Error, the Relative Error, the Pearson correlation, and the Distance correlation between the observed and the predicted claim amounts of your final model. Please comment on their values.**

The Root Mean Squared Error = 4116.064009419276

The Relative Error = 1.0078985075249889

The Pearson Correlation = 0.18768098705727357

The Distance Correlation = 0.27019055675583475

The RMSE value represents the average difference between the predicted and actual amounts which being so high isn't great.

The relative error suggests that the model's predictions are, on average, about 1% higher than the actual payout amounts.

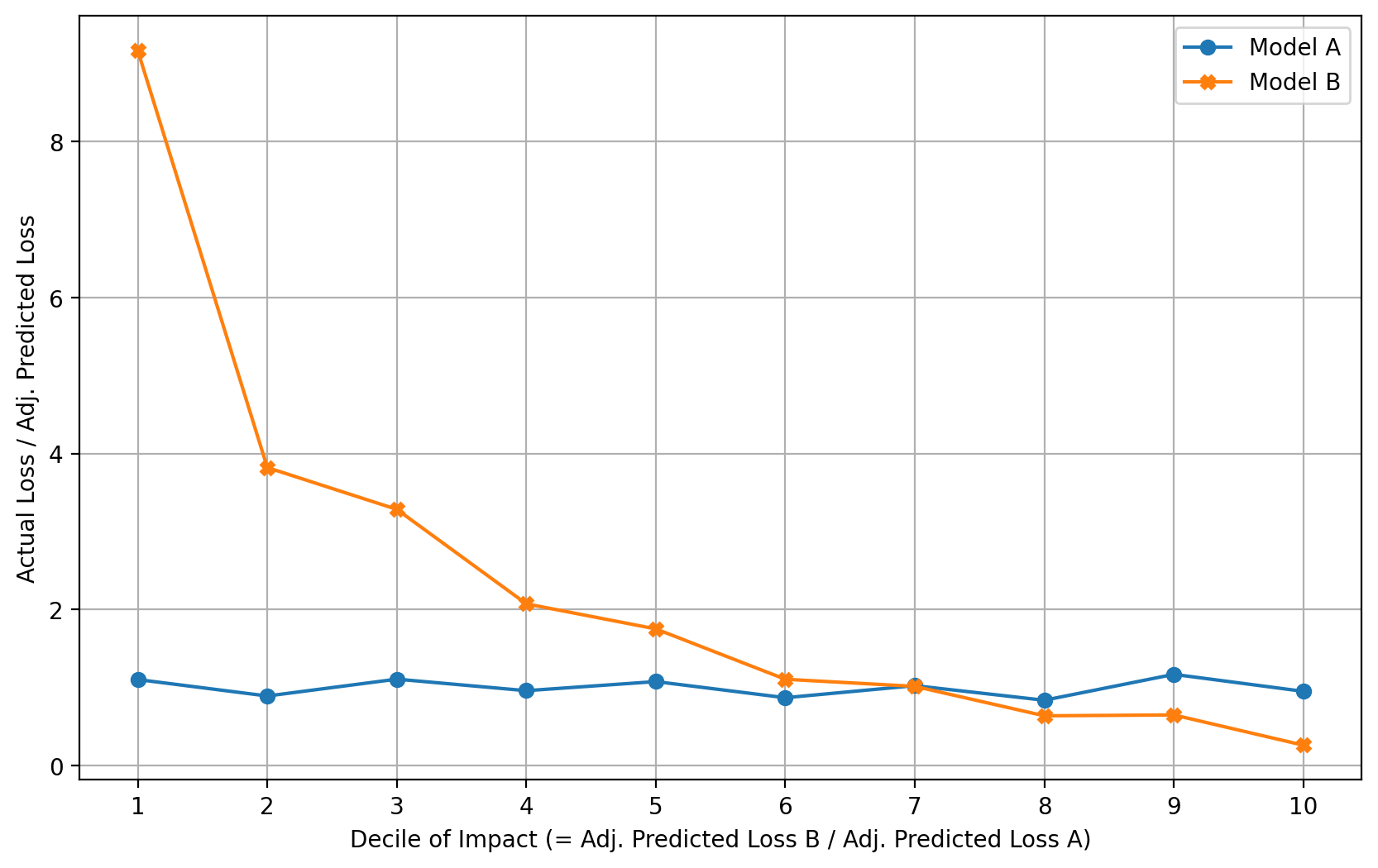
The Pearson correlation is positive and low which means there isn't much correlation between the variables.

The Distance correlation is also positive and low which means there isn't much correlation between the variables.

1. **(10 points). Please show a table of the complete set of parameters of your final model (including the aliased parameters). Besides the parameter estimates, please also include the standard errors, the 95% asymptotic confidence intervals, and the exponentiated parameter estimates. Conventionally, aliased parameters have zero standard errors and confidence intervals. Please also provide us with the final estimate of the Tweedie distribution’s scale parameter 𝜙.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Standard Error** | **Lower 95% CI** | **Upper 95% CI** | **Exponentiated** |
| Intercept | 8.0045945 | 0.0070866 | 7.9907050 | 8.0184839 | 2994.6854164 |
| URBANICITY\_Highly Rural/ Rural | -1.6681625 | 0.0027758 | -1.6736030 | -1.6627221 | 0.1885933 |
| URBANICITY\_Highly Urban/ Urban | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| EDUCATION\_Bachelors | -0.1407007 | 0.0035220 | -0.1476037 | -0.1337977 | 0.8687493 |
| EDUCATION\_Below High Sc | 0.2015884 | 0.0043330 | 0.1930958 | 0.2100810 | 1.2233443 |
| EDUCATION\_High School | 0.0966325 | 0.0039945 | 0.0888033 | 0.1044616 | 1.1014555 |
| EDUCATION\_Masters | -0.1387404 | 0.0034407 | -0.1454841 | -0.1319967 | 0.8704540 |
| EDUCATION\_PhD | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| CAR\_TYPE\_Minivan | -0.7502918 | 0.0030645 | -0.7562980 | -0.7442855 | 0.4722287 |
| CAR\_TYPE\_Panel Truck | 0.0175044 | 0.0032363 | 0.0111614 | 0.0238474 | 1.0176585 |
| CAR\_TYPE\_Pickup | -0.2670473 | 0.0029189 | -0.2727682 | -0.2613264 | 0.7656369 |
| CAR\_TYPE\_SUV | 0.0528518 | 0.0033656 | 0.0462553 | 0.0594483 | 1.0542734 |
| CAR\_TYPE\_Sports Car | 0.1142253 | 0.0037577 | 0.1068604 | 0.1215903 | 1.1210047 |
| CAR\_TYPE\_Van | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| PARENT1\_No | -0.4879962 | 0.0031300 | -0.4941308 | -0.4818615 | 0.6138552 |
| PARENT1\_Yes | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| MVR\_PTS | 0.0962476 | 0.0003094 | 0.0956412 | 0.0968539 | 1.1010316 |
| TRAVTIME | 0.0114279 | 0.0000462 | 0.0113373 | 0.0115185 | 1.0114934 |
| CAR\_USE\_Commercial | 0.5044786 | 0.0018748 | 0.5008040 | 0.5081533 | 1.6561219 |
| CAR\_USE\_Private | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| REVOKED\_No | -0.4380320 | 0.0019645 | -0.4418823 | -0.4341817 | 0.6453051 |
| REVOKED\_Yes | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| KIDSDRIV | 0.2687011 | 0.0013539 | 0.2660475 | 0.2713547 | 1.3082641 |
| TIF | -0.0417956 | 0.0001885 | -0.0421651 | -0.0414261 | 0.9590658 |
| INCOME | -0.0056110 | 0.0000238 | -0.0056575 | -0.0055644 | 0.9944047 |
| MSTATUS\_No | 0.4518538 | 0.0021474 | 0.4476450 | 0.4560626 | 1.5712222 |
| MSTATUS\_Yes | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| CAR\_AGE | -0.0230979 | 0.0001860 | -0.0234626 | -0.0227333 | 0.9771668 |
| YOJ | 0.0233943 | 0.0002144 | 0.0229741 | 0.0238145 | 1.0236701 |
| HOMEKIDS | 0.0572518 | 0.0008010 | 0.0556817 | 0.0588218 | 1.0589224 |
| GENDER\_F | -0.1941537 | 0.0025385 | -0.1991291 | -0.1891782 | 0.8235313 |
| GENDER\_M | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |
| RED\_CAR\_no | 0.1281310 | 0.0021378 | 0.1239410 | 0.1323210 | 1.1367019 |
| RED\_CAR\_yes | 0.0000000 | 0.0000000 | 0.0000000 | 0.0000000 | 1.0000000 |

1. **(10 points). Please generate a Two-way Lift chart for comparing your final model with the Intercept-only model. Based on the chart, what will you conclude about your final model?**



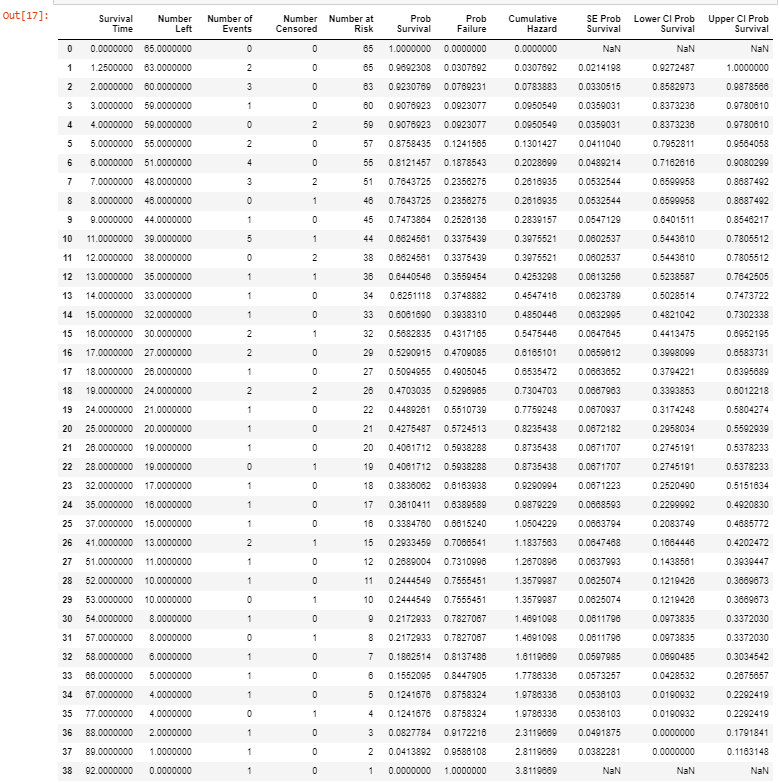
At Decile #1, Model A (The Intercept-only model) is preferred as its Loss Ratio is much lower and Model A is preferred as its Loss Ratio is closer to the ideal loss ratio of 1.

**QUESTION 2**

1. **(10 points). How many risk sets are there?**

The number of risk sets = 38

1. **(10 points). We will use the Kaplan-Meier Product Limit Estimator to create the life table. Please provide us with the life table.**



1. **(10 points). According to the life table, what is the Probability of Survival and the Cumulative Hazard at a survival time of 18 months? What do these two values mean to a layperson?**

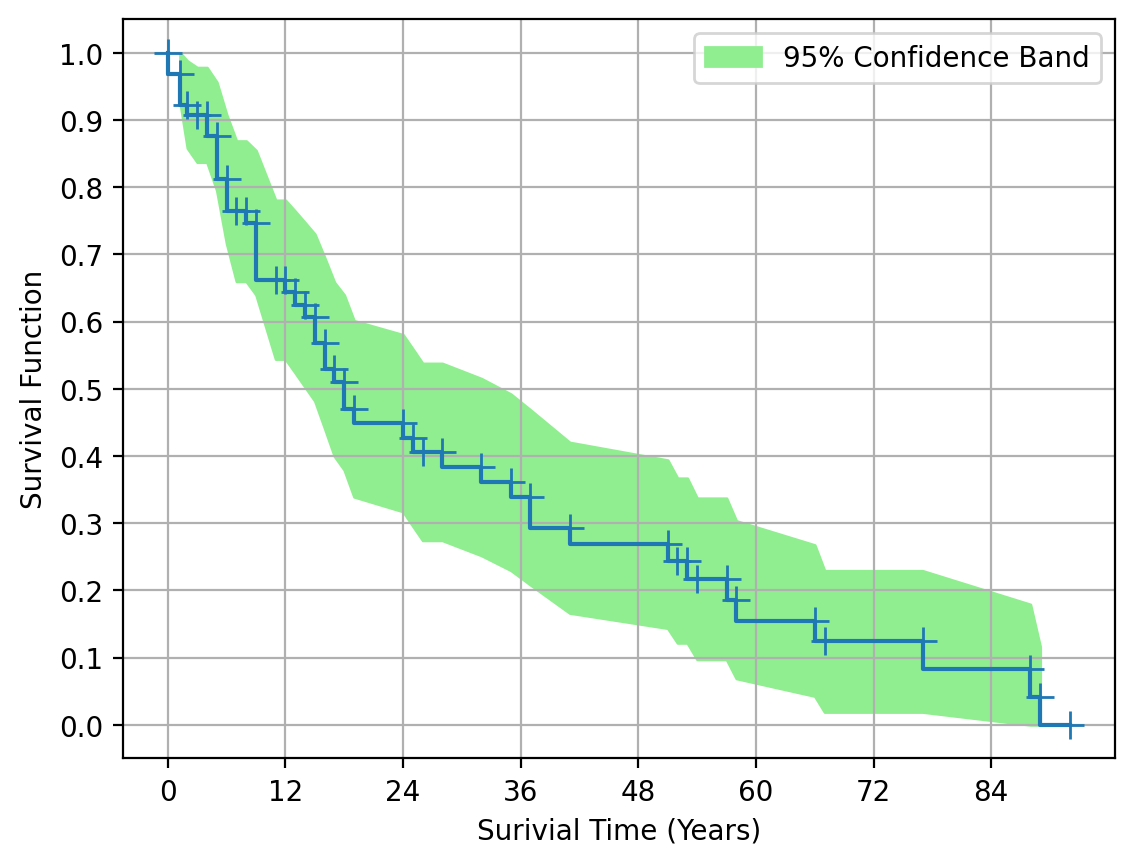
At a survival time of 18 months:

The Probability of Survival = 0.5094955102428145

The Cumulative Hazard = 0.6535471794159207

Patients with multiple myeloma and being treated with the alkylating agents have a probability of survival of 0.5094955102428145, but the average number of deaths at the time of 18 months is 0.6535471794159207.

1. **(10 points). Please generate the Survival Function graph using the Kaplan-Meier Product Limit Estimator life table. Since we measure the Time variable in the number of months, we will specify the x-axis ticks from 0 with an increment of 12. Besides plotting the Survival Function versus Time, you must also add the 95% Confidence Band. You might use the matplotlib fill\_between() function to generate the Confidence Band as a band around the Survival Function. To receive the full credits, you must label the chart elements properly.**



1. **(10 points). Use Linear Interpolation to determine the Median Survival Time (in number of months) from the Kaplan-Meier Product Limit Estimator life table. Please round your answer up to the tenths place.**

The Median Survival Time = 18.2 months