

Application of the Generalized Linear Model to the Prediction of Lung Cancer Survival

Andrew Karem, University of Louisville, Louisville, KY

ABSTRACT

The study of cancer prognosis in patients is a crucial element of the medical research field. The focus of this report is one such study: the first documented lung cancer trial in history. The data involved in this study were generated in a two-year period by the US Veterans Administration with the hopes of revealing patterns among patient survival statistics.

The specific goal of this project was to determine whether survival could be adequately predicted based on other patient characteristics. In order to solve this predictive task, the general and generalized linear models of regression were applied to the data. Potential effects were added and removed from models to enhance performance; following estimation of the optimal models, comparisons were made between the general and generalized approach with respect to the accuracy in prediction.

A number of conclusions may be drawn from this study. First of all, manipulation of the generalized model indicates that the optimal generalized solution implements the Poisson distribution and log link function. As well, the assessment statistics imply both the inadequacy of the general model and adequacy of the generalized model in prediction. Finally, as is discussed at the end of the document, a strong dilemma exists with the potential of over-fitting of the data.

INTRODUCTION

Statistical modeling is an integral tool in the analysis of medical studies. In many cases, particularly with older medical data, the standard general linear model fails to provide adequate support for potential irregularities in distribution; in such cases, the generalized linear model is a much more effective tool. This report deals with the analysis of one such scenario: the assessment of survival rates in an early lung cancer study. The generalized linear model was found to provide a much more precise analysis than a less flexible method.

The data of interest in this project come from an early Veteran's Administration lung cancer trial featuring 137 patients and several variables, including survival in days [1]. These data were first analyzed using a traditional, general, linear regression approach, with an interest in determining which combination of effects allows for the optimal prediction of survival for the patients. Following this evaluation, a similar routine was followed using the generalized linear model approach.

It should be noted that the generalized linear model methodology was not conducted with a preset notion as to distribution or necessary link function. Instead, analysis of the data revealed a number of potential link functions and candidate distributions to address the challenge. In order to synthesize an optimal choice from among these possibilities, several models were evaluated; assessment statistics, most notably Pearson's residuals, were used to choose the most effective construct in fitting the data. Of particular interest in this project was the comparison of this optimal generalized model with the prime general linear model.

The results of this study are interpreted with respect to a number of details. First of all, it is made clear that the best general linear model available is a poor predictor of survival. It is additionally shown that a full-effects generalized linear model easily outperforms this inadequate general model. More specifically, it is illustrated that a selection of the Poisson distribution and log link function leads to a model with excellent assessment. Lastly, issues related to over-fitting of data are examined for both models.

METHOD

This section of the report focuses on the mechanics involved in both data analysis and data acquisition, the latter of which is the initial focus. One of the first and most commonly studied cancer-related datasets is based on the 1971 US Veteran's Administration Lung Cancer Survey. The data are comprised of vital statistics for patients diagnosed and treated for lung cancer at several prominent medical centers between the years 1968 and 1971. The data were first made available to the public via [1].

The majority of data in this survey were acquired at the time of patient visitation, with the exception of the outcome variable -- survival in days. This particular statistic was recorded at trial's end, with some values being censored due to a patient's still being alive. All other variables in the dataset are complete, with no missing values. The succeeding data summary section should be consulted for detailed information regarding the variables.

Prior to the process of data analysis, some simple data pre-processing was performed. Because the integrity of the survival variable was of prime interest, a portion of samples that appeared to violate its stability were removed. The first decision made was to remove samples for which the individuals had not yet died, since these survival values are inaccurate. Additionally, a simple and lenient outlier test was performed to remove those records with survival beyond three standard deviations of the mean. These steps were taken in attempt to make the resulting models less erratic. Furthermore, this procedure only resulted in removal of ten samples, which should have a negligible effect on the quantity of data available.

Following preprocessing, the actual process of analysis could begin. The first action taken in this regard was to perform basic general linear regression on the data. Each discrete and continuous variable was used to predict survival separately. Those effects passing a simple p-value threshold of 90% confidence were then included in a combination model, along with all two-level interactions among the variables. Based on these results, interactions and poorly fitting variables were removed from the model, and an optimal general model solution was found.

The second phase of the analysis process involved application of the generalized linear model. In most respects, this process mirrored that of the general model phase, with the exception that a number of potential link functions were applied to the regression. Furthermore, based on visualization of the data, a number of possible distributions were found to fit the data; these, too, were considered in the potential generalized models. Pearson residual and deviance statistics were assessed for the prospective models in an attempt to isolate the optimal structure and set of effects.

With an optimal general linear model and generalized linear model computed, the last step of this project was to compare the results of the two. Pearson residuals and deviance statistics were computed for the general model of choice, and an evaluation of the disparity between the two models was then conducted. In particular, a judgment was made as to whether the benefit in accuracy garnered in employing the generalized model was worth the additional complication in structure.

DATA SUMMARY

One of the fundamental tasks of this project was to examine the structure of the dataset in an effort to extract some useful information that directs the process of constructing an optimal generalized linear model. This section covers visualization and basic relationships within the dataset, which are crucial in determining potential model parameters. Prior to accomplishing this, however, a brief and practical description of the variables in the dataset is provided.

The most important feature of the data is the survival characteristic. This continuous variable details the number of days a particular patient survived beyond the day in which the participant began the trial. Survival is treated as the outcome variable in generating the models, and hence, much attention is paid to its distribution and analysis. Three other continuous variables within the dataset were of crucial import to this report. The first of these is the Karnofsky Score – an assessment as to a patient's general health and well-being at the time of visitation. It should be noted that this is a subjective value, gauged by the patient's doctor on a scale from 0 – 100. A score of 0 indicates death and a score of 100 indicates perfect health. Specific details regarding the Karnofsky score assessment may be found online [2]. The third continuous variable in the study corresponds quite simply to a patient's age at the time of visitation. The last continuous variable corresponds to the number of months between the trial's end and the patient's initial diagnosis.

The attributes in this study are not limited to a continuous nature; several discrete variables are of key interest. The first of these corresponds to the type of carcinoma a particular patient was diagnosed as possessing. This may be one of squamous-cell, large-cell, small-cell, and adeno carcinoma. While modern procedures and terminology have changed with regards to carcinoma diagnosis in the present, the variable was treated precisely as recorded for the purposes of this project. The second discrete variable is a simple binary attribute corresponding to whether a patient received a conventional form of treatment, or was involved in a test for a newer drug/therapy. Again, this is a subjective detail for which the values may not be applicable today. The last binary records whether or not a patient received prior cancer treatment before the trial.

With the variables of interest defined, an important step is to visualize basic statistics and distribution. To that end, figures 1a-1d show kernel density estimates of the continuous variables in the study. As can be seen, all four variables have relatively skewed distributions; of particular note is that the survival variable does not appear to follow a normal distribution. It is apparent that the enhanced flexibility of the generalized model would be appealing. Either a Poisson distribution or an exponential distribution would be a more appropriate fit to the data. Also evident is that the majority of attributes are unimodal in nature, with the exception of the Karnofsky score, which peaks around values of 40 and 60. Finally, the variance in distribution does not appear uniform among the continuous variables. While the possibility still exists for strong predictors in the dataset, one should not expect simplistic regression to reveal an overwhelming correlation among any two attributes.

Figure 1a. Distribution of Survival in Days

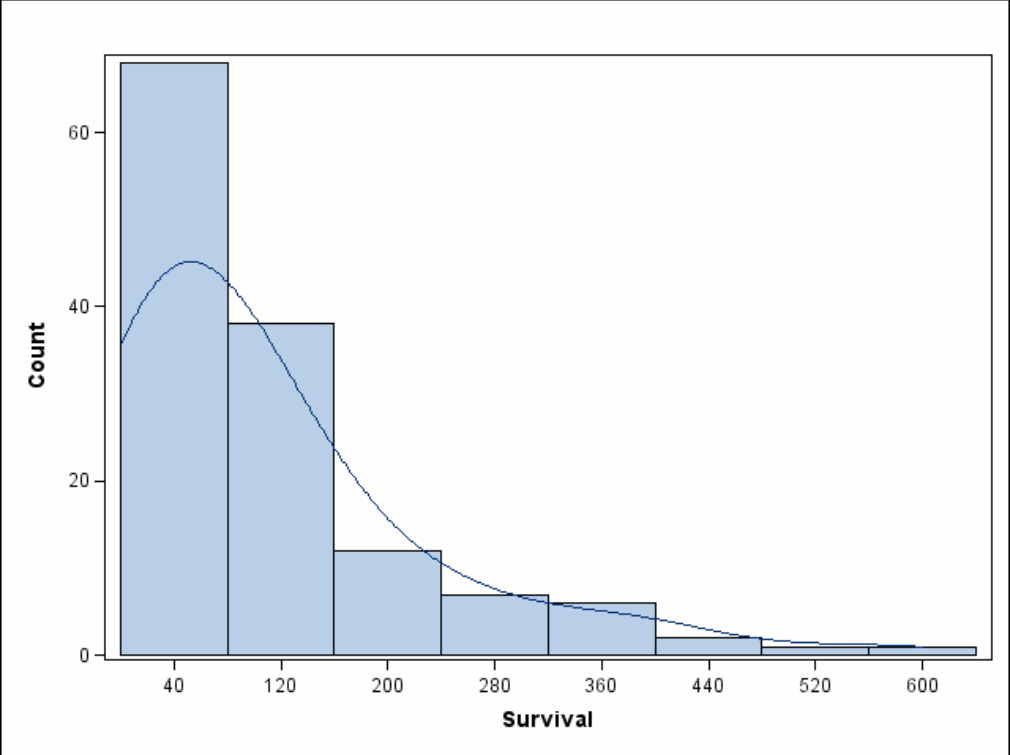


Figure 1b. Distribution of Karnofsky Score Statistic

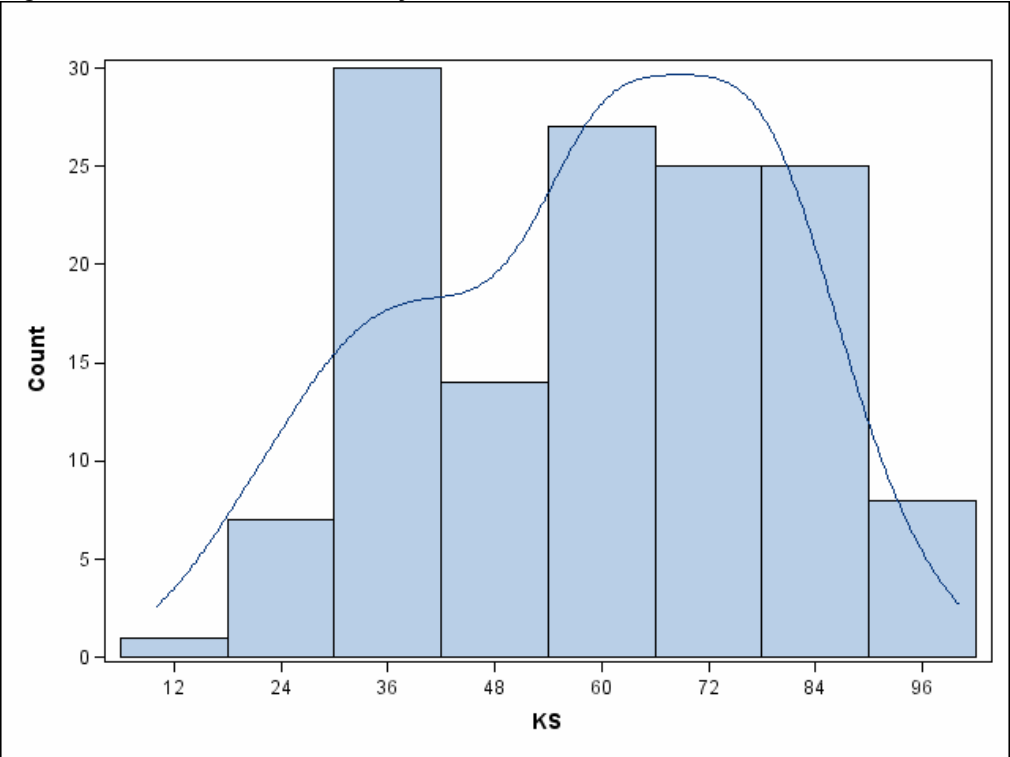


Figure 1c. Distribution of Age in Years

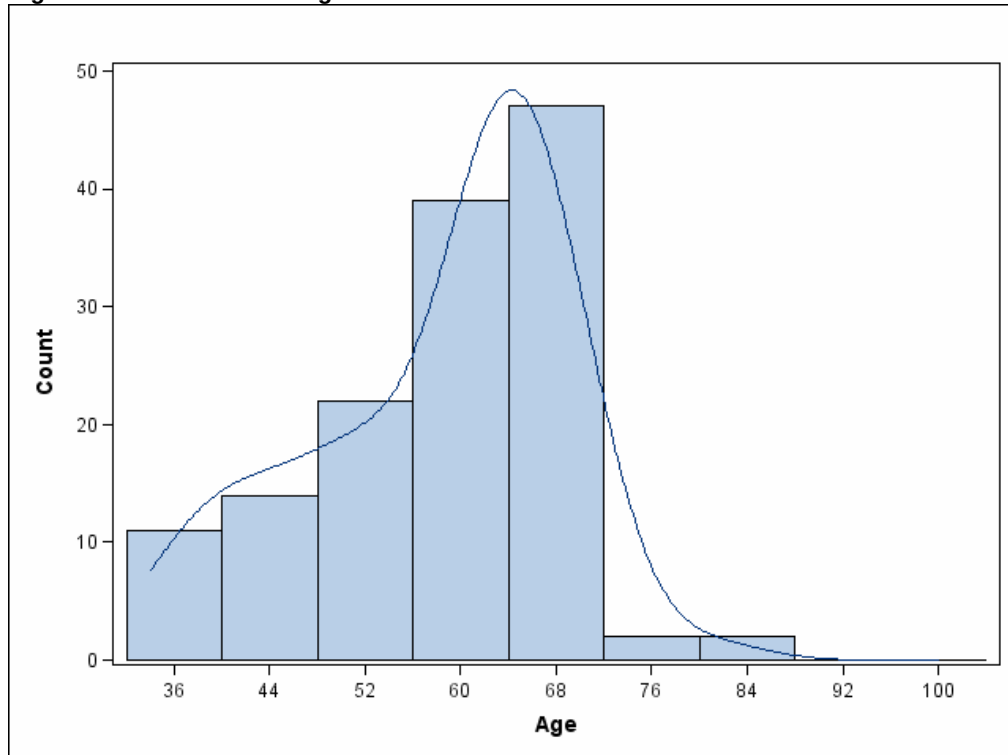
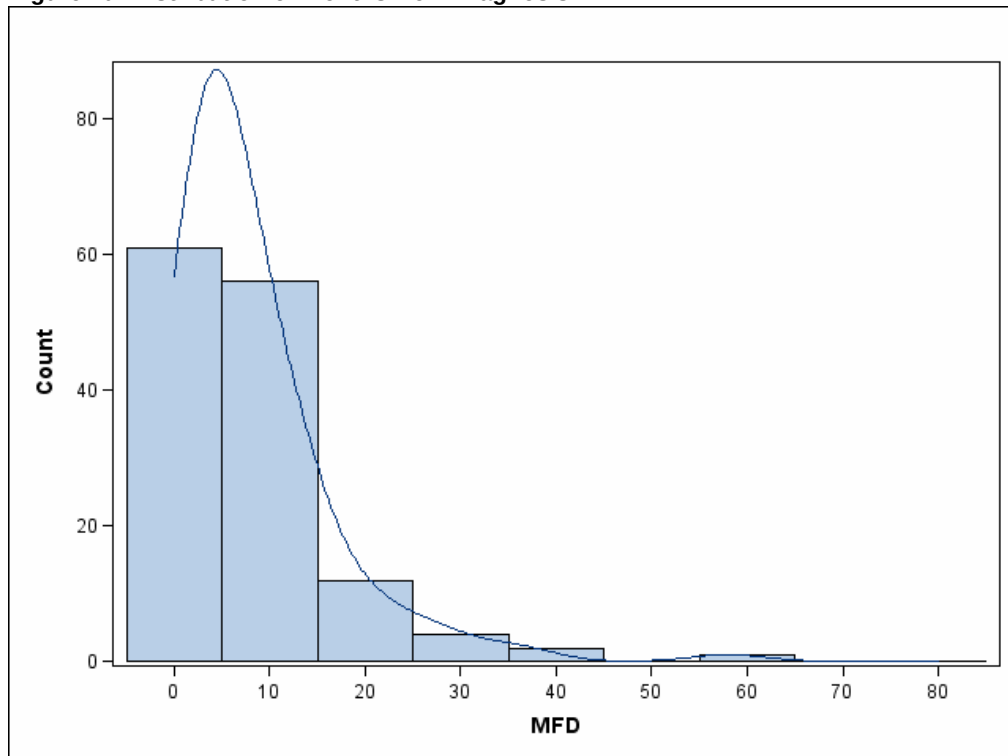


Figure 1d. Distribution of Months from Diagnosis



The other three variables to analyze are discrete in nature. Table 1 summarizes these features. One notable feature regarding the binary attributes is that the proportion of cases with standard and test treatments is even, with approximately half the samples representing each. On the other hand, the prior therapy variable is not so even, with about one-fourth of the samples having the “yes” response. Finally, the Type of Carcinoma (ToC) attribute entertains a fairly well-balanced distribution, with none of the four values representing less than one-fifth the total population.

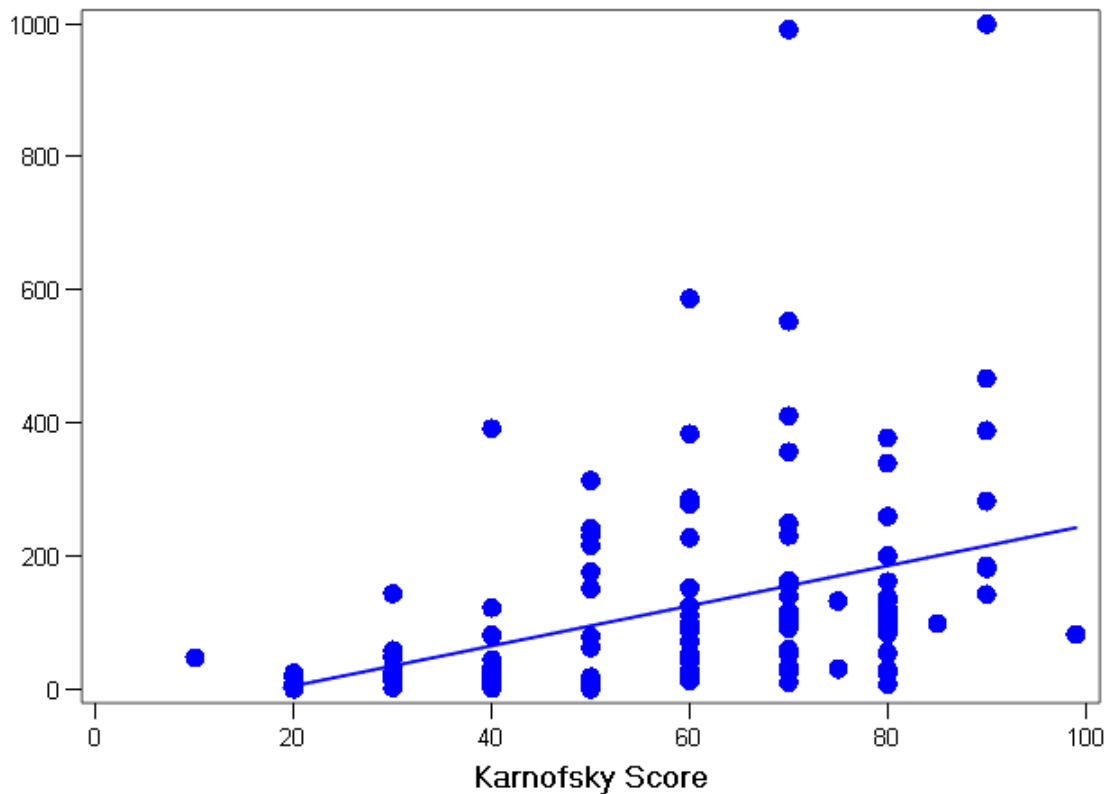
Table 1. Distribution of Discrete Attributes

		Frequency	Total Percentage
Treatment	Standard	69	51.11
	Test	66	48.89
Prior Therapy	No	97	71.85
	Yes	38	28.15
Type of Carcinoma	Adeno	27	20.00
	Large	27	20.00
	Small	48	35.56
	Squamous	33	24.44

Having presented the data distributions, it may be useful to evaluate a few basic relationships among the variables in the dataset. Figure 2 presents an integral component of this study – the correlation between the Survival and Karnofsky score variables. Even when ignoring the potential for link functions/abnormal distributions, a tangible relationship exists between the two variables. Specifically, there is a gradual increase in expected survival as the Karnofsky score increases. This is most notable when examining scores with a high discrepancy. For instance, the highest survival statistic with a corresponding Karnofsky score of 30 is on par with the lowest survival statistic at a score of 90. Visualizations concerning other relationships among continuous variables do not reveal strong or obvious connections.

Figure 2. Survival vs. Karnofsky Score

Survival in Days



Another category of relationships to evaluate corresponds to the discrete attributes. Table 2 represents contingencies among the discrete attributes in the form of percentages satisfying given values. Several quick observations can be garnered from this visualization. First of all, in both cases of standard and test treatments, prior therapy has a large relationship to the ToC present. To be specific, prior therapy is proportionally more likely to occur in those cases dealing with large-cell and squamous carcinomas. Comparisons similarly demonstrate that the treatment used appears related to the ToC involved. Proportionally, more Adeno-carcinoma cases are associated with test treatment, and small-cell cases are associated with standard treatment. Finally, we note that there exists the strong potential for a three-way interaction involving all three attributes, because the ratios across the table are not consistent.

Table 2: Type of Carcinoma vs. Treatment vs. Prior Therapy

		Treatment			
		Standard		Test	
		Prior Therapy?		Prior Therapy?	
		No	Yes	No	Yes
Type of Carcinoma	Adeno	14.58	9.52	30.61	17.65
	Large	20.83	23.81	14.29	29.41
	Small	47.92	33.33	28.57	23.53
	Squamous	16.67	33.33	26.53	29.41

Having examined relationships among continuous variables and among discrete variables separately, it would be useful to consider interactions among the two. The last form of data visualization addresses this issue. Figure 3a addresses the crucial relationship of ToC to expected survival. There is a compelling increase in life expectancy for someone who has squamous carcinoma as compared to someone who has small-cell or adeno carcinoma. In particular, roughly 50% of individuals with squamous carcinoma survive the 200 day mark, whereas only 10% or so of the other two types do so. Large-cell carcinoma is somewhere between these two distinctions.

Figure 3b compares ToC to the Karnofsky score attribute. As with ToC vs. survival, there is an apparent disparity among the carcinomas. However, the relationship does not appear to mirror the survival completely vs. ToC results. Whereas squamous-cell carcinoma indicates the strongest value for life-expectancy, large-cell carcinoma has the highest Karnofsky score on average. This is crucial because it indicates the Karnofsky score and ToC effects in our model may not have a redundant effect in predicting survival. Other discrete vs. continuous attribute comparisons show potential interaction, but none as pronounced as in the two cases examined above.

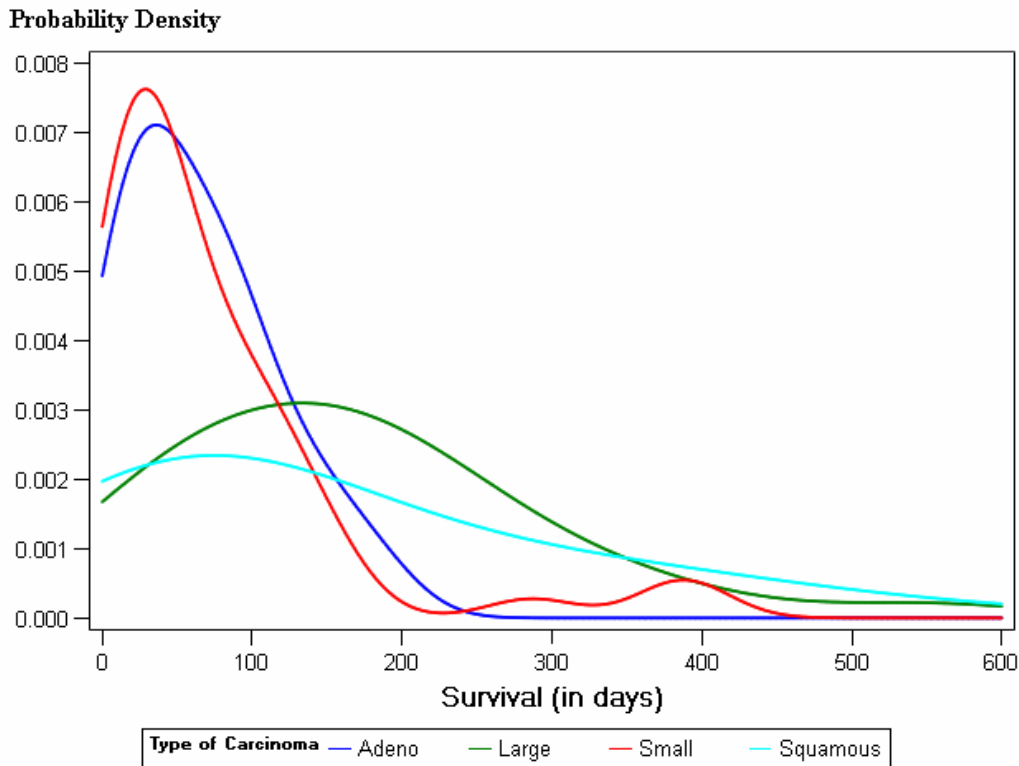
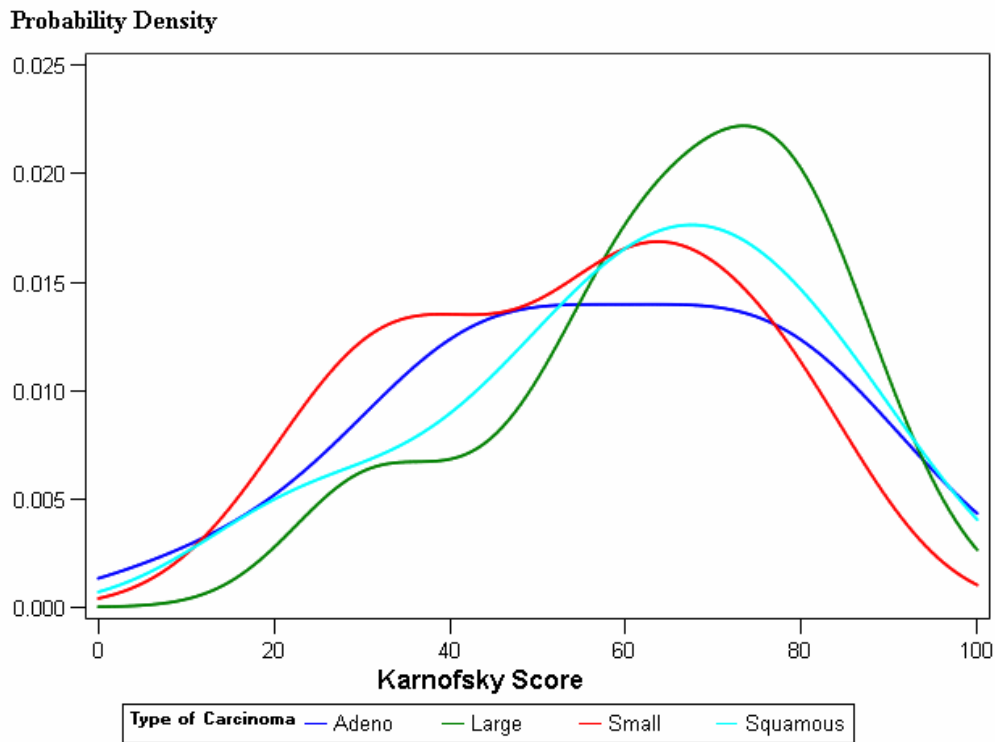
Figure 3a. Survival vs. Type of Carcinoma

Figure 3b. Karnofsky Score vs. Type of Carcinoma



DATA ANALYSIS AND RESULTS

With visualization and evaluation of the data complete, it becomes necessary to document the results of the modeling process. The first step taken in regards to this goal is to perform a simple general linear analysis. To that end, the first model constructed was created using the general linear model with all potential single-attribute effects and all potential two-way interactions among variables. Tables 3a and 3b present these preliminary results. The significant p-value indicates the null-hypothesis is strongly rejected. Furthermore, the R-square value indicates that, based on the demographic nature of the dataset, this model does appear to possess some adequacy in addressing the task. Table 3b indicates the significant effects in the model. Both Karnofsky score and type of carcinoma have very strong individual effects (>99% confidence) in the model; the interaction of the age and prior therapy variable is also significant (>95% confidence). All other effects and interactions are insignificant at this stage.

Table 3a. Preliminary General Model p-value and R-Square

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	33	747334.404	22646.497	2.10	0.0026
Error	101	1091706.529	10808.976		
Corrected Total	134	1839040.933			
R-Square	Coeff Var	Root MSE	Survival Mean		
0.406372	95.65488	103.9662	108.6889		

Table 3b. Preliminary General Model Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	1228.4906	1228.4906	0.11	0.7367
KS	1	296074.2265	296074.2265	27.39	<.0001
MFD	1	57.1225	57.1225	0.01	0.9422
PT	1	465.8665	465.8665	0.04	0.8360
ToC	3	199232.0876	66410.6959	6.14	0.0007
Treatment	1	3083.9405	3083.9405	0.29	0.5944
Age*KS	1	4965.8209	4965.8209	0.46	0.4994
Age*MFD	1	4022.4901	4022.4901	0.37	0.5432
Age*PT	1	42914.4229	42914.4229	3.97	0.0490
Age*ToC	3	9407.4604	3135.8201	0.29	0.8324
Age*Treatment	1	19916.1618	19916.1618	1.84	0.1777
KS*MFD	1	77.3646	77.3646	0.01	0.9327
KS*PT	1	429.0159	429.0159	0.04	0.8425
KS*ToC	3	41702.5556	13900.8519	1.29	0.2833
KS*Treatment	1	18890.0810	18890.0810	1.75	0.1892
MFD*PT	1	19.2462	19.2462	0.00	0.9664
MFD*ToC	3	32671.3243	10890.4414	1.01	0.3927
MFD*Treatment	1	1847.3385	1847.3385	0.17	0.6802
PT*ToC	3	18526.5430	6175.5143	0.57	0.6352
PT*Treatment	1	20434.9036	20434.9036	1.89	0.1722
Treatment*ToC	3	31367.9410	10455.9803	0.97	0.4113

The next task faced involves refining our set of effects. This implies removing all two-way interactions for those effects which lack a significant p-value (< 0.10) for both these interactions as well as when taken as a single variable. This results in the set of effects is given in table 4. Our new set of effects is still far from optimal, and one of the few potential interactions (Age*PT) is no longer valid. We now remove all interaction effects, since none are significant. This gives us the set of effects in table 5.

Table 4. Intermediate General Model Effects I

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	1228.4906	1228.4906	0.11	0.7355

Source	DF	Type III SS	Mean Square	F Value	Pr > F
KS	1	296074.2265	296074.2265	27.64	<.0001
PT	1	272.7074	272.7074	0.03	0.8735
ToC	3	199336.9591	66445.6530	6.20	0.0006
Age*KS	1	5271.1636	5271.1636	0.49	0.4844
Age*PT	1	21304.9971	21304.9971	1.99	0.1612
Age*ToC	3	15229.0114	5076.3371	0.47	0.7011
KS*PT	1	4422.3676	4422.3676	0.41	0.5218
KS*ToC	3	46082.0476	15360.6825	1.43	0.2366
PT*ToC	3	26891.3836	8963.7945	0.84	0.4764
MFD	1	339.9978	339.9978	0.03	0.8589
Treatment	1	1313.1312	1313.1312	0.12	0.7269

Table 5. Intermediate General Model Effects, Step II

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	1228.4906	1228.4906	0.12	0.7344
KS	1	296074.2265	296074.2265	27.86	<.0001
PT	1	272.7074	272.7074	0.03	0.8730
ToC	3	199336.9591	66445.6530	6.25	0.0005
MFD	1	145.4101	145.4101	0.01	0.9071
Treatment	1	3083.9405	3083.9405	0.29	0.5910

We now remove those variables which do not have a significant effect in the model; this results in our final, optimal general linear solution. Tables 6a and 6b display the statistical characteristics of this model. Note that the R-square value has been reduced, but the F-value and p-value are substantially improved for this model; this indicates that the initial model substantially over-fits the data. Furthermore, the much lesser value for R-square implies that the general linear model may not be adequate in explaining this prediction problem.

Table 6a. Final General Model p-value and R-Square

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	486106.722	121526.681	11.68	<.0001
Error	130	1352934.211	10407.186		
Corrected Total	134	1839040.933			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
R-Square	Coeff Var	Root MSE	Survival Mean		
0.264326	93.86021	102.0156	108.6889		

Table 6b. Final General Model Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
KS	1	215644.7587	215644.7587	20.72	<.0001
ToC	3	188972.2374	62990.7458	6.05	0.0007

Having explored potential general linear models, we now apply generalized linear methodology. Because there exists no trivial way to compare a generalized model to a general model, our first step is to use the effects of the optimal general model in a generalized linear model for assessment comparison. This is accomplished with the assumption of the normal distribution among the data and the identity link function. The results of this computation are given in tables 7a and 7b. The scaled deviance and scaled Pearson chi-square values indicate that this model is marginal in terms of assessment. As this corresponds to the optimal general linear model, it will be the standard by which generalized models will be compared.

Table 7a. Assessment of General to Generalized Model Conversion

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	130	1352934.2110	10407.1862
Scaled Deviance	130	135.0000	1.0385
Pearson Chi-Square	130	1352934.2110	10407.1862
Scaled Pearson X2	130	135.0000	1.0385
Log Likelihood		-813.4012	

Table 7b. General to Generalized Model Conversion: Effects

Analysis Of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept		1	29.419	31.659	-32.632	91.471	0.8	0.352
KS		1	2.053	0.442	1.186	2.921	21.5	<.000
ToC	Adeno	1	-84.662	25.987	-135.59	-33.727	10.6	0.001
ToC	Large	1	3.189	26.084	-47.934	54.312	0.0	0.902
ToC	Small	1	-67.721	22.801	-112.41	-23.031	8.8	0.003
ToC	Squamous	0	0.000	0.000	0.000	0.000		

Analysis Of Parameter Estimates						
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square Pr > ChiSq
Scale	1	100.108	6.092	88.852	112.790	

The first step in the process of defining a generalized model is to consider potential distributions and link functions. Based on examination of the data, it appears that either the Poisson or Gamma distribution would be an appropriate choice. It appears that either an identity or log function would be suitable as a link function. As well, it may be of analytic use to further examine generalized models corresponding to the normal distribution using both the identity and log link function. This yields six potential generalized model structures to examine. The nature of these structures is provided in table 8.

Table 8. Generalized Model Reference

Structural Label	Distribution	Link Function
Model 1A	Normal	Identity
Model 1B	Normal	Log
Model 2A	Poisson	Identity
Model 2B	Poisson	Log
Model 3A	Gamma	Identity
Model 3B	Gamma	Log

The process of evaluating effects for these potential models follows the same procedure as in the general case. Prior to this process, however, examination of the different structures based on a single effect provides some insight into how accurate the potential models are. The appropriate choice for an individual effect is the Karnofsky score, as it appears to have the greatest correlation with survival. To this end, tables 9a-11a and 9b-11b present assessment of the six base models using this effect. Note that in each case for which the generalized model converged to solution, the p-value of the Karnofsky score is less than 0.0001.

We can draw a few inferences from the single-effect generalized linear models. It can be noted that model 1B, which is identical to 1A, but uses the log link function, offers no better assessment. On the other hand, this change in model 3B does induce a slightly better estimate of the data compared to model 3A. On a related note, models 3A and 3B are superior to models 1A and 1B with respect to residuals. As was surmised in our initial evaluation of the data, using the Poisson distribution in combination with the log link function (model 2B) appears to offer the best model for the data. In fact, the sheer discrepancy in magnitude among the chi-square and deviance tests implies that using the Poisson distribution is probably the only reasonable choice in addressing this dataset.

Table 9a. Single effect GLIM analysis: Model1A

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	133	1541906.4484	11593.2816
Scaled Deviance	133	135.0000	1.0150
Pearson Chi-Square	133	1541906.4484	11593.2816
Scaled Pearson X2	133	135.0000	1.0150
Log Likelihood		-822.2264	

Table 9a. Single effect GLIM analysis: Model1B

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	133	1558874.3044	11720.8594
Scaled Deviance	133	135.0000	1.0150
Pearson Chi-Square	133	1558874.3044	11720.8594
Scaled Pearson X2	133	135.0000	1.0150
Log Likelihood		-822.9652	

Table 10a. Single effect GLIM analysis: Model2A

Algorithm Failed to Converge

Table 10b. Single effect GLIM analysis: Model2B

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/D F
Deviance	133	5613.7382	42.20855
Scaled Deviance	133	5613.7382	42.20855
Pearson Chi-Square	133	7239.0526	54.42895
Scaled Pearson X2	133	7239.0526	54.42895
Log Likelihood		27790.6798	

Table 11a. Single effect GLIM analysis: Model3A

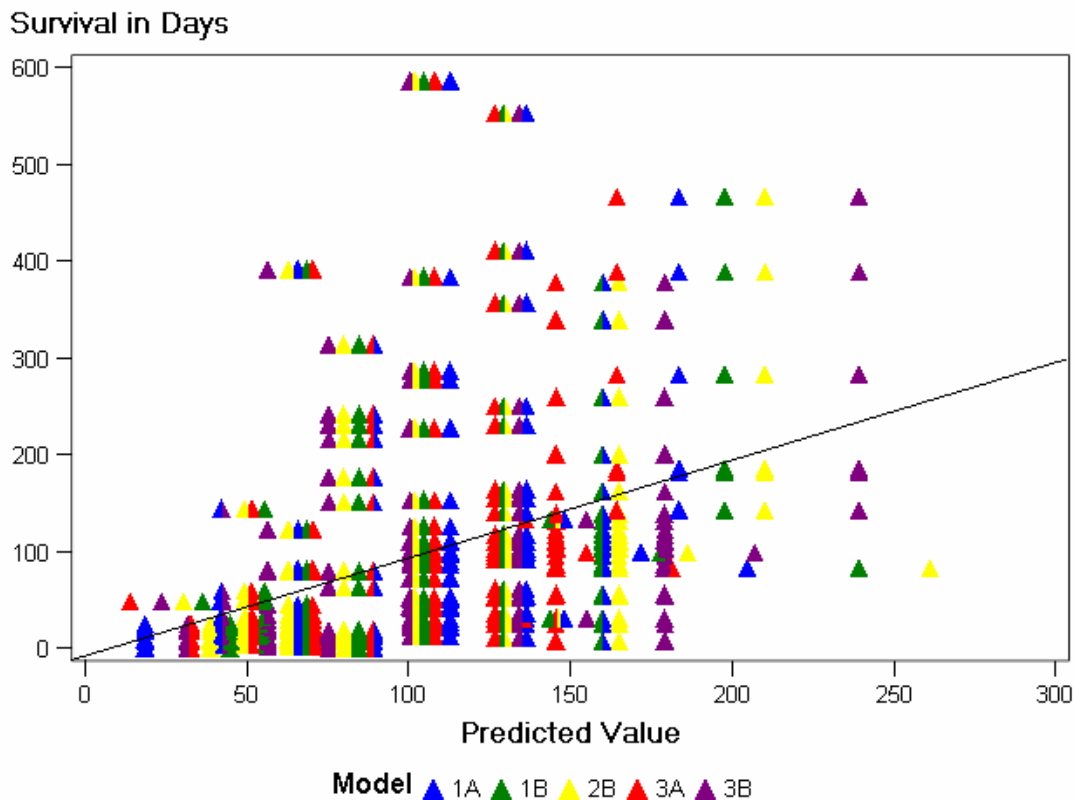
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/D F
Deviance	133	138.4687	1.0411
Scaled Deviance	133	153.9775	1.1577
Pearson Chi-Square	133	145.4591	1.0937
Scaled Pearson X2	133	161.7509	1.2162
Log Likelihood		-750.7505	

Table 11b. Single effect GLIM analysis: Model3B

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/D F
Deviance	133	137.0803	1.0307
Scaled Deviance	133	153.8240	1.1566
Pearson Chi-Square	133	165.0918	1.2413
Scaled Pearson X2	133	185.2569	1.3929
Log Likelihood		-749.9751	

While deviance and Pearson chi-square figures provide valid statistical assessment, visualization of the five models as applied to the data may also be of interest. Figure 4 provides a graph depicting survival prediction vs. actual figures for the five convergent models. Note that frequently used Karnofsky scores form linear data clusters for predicted survival – assessment of these clusters is crucial in visualizing the model. We can immediately see a disparity in model performance. For instance, model 1A appears to function quite well in predicting the smaller survival figures. However, as the survival figure increases, the other models appear to straddle the score clusters better. This is particularly true of model 2B, which corresponds to the Poisson distribution. Finally, the sheer variance in prediction for the score clusters in each model demonstrates the relative inadequacy of using the Karnofsky score as a sole predictor of survival.

Figure 4. Single Effect Models: Predicted Survival vs. Actual Survival



Based on statistics and visualization, it is clear that model 2B is best when the Karnofsky score effect is used alone. We cannot assume that this would be the case with the inclusion of other attributes; nonetheless, we may use single-effect assessments and full effects assessment to draw reasonable conclusions about potential partial

effect models. First of all, the log link function appears most suitable for the task. Full effects comparisons, with the inclusion of all 2-way interactions, are provided in table 12; these results appear to confirm the suitability of the log link function. Furthermore, it appears that the assessment overwhelmingly favors the Poisson distribution. As a result, structure 2B will be the basis of our optimal generalized linear model.

Table 12. Full effect GLIM assessment

Structural Label	Scaled Deviance			Scaled Pearson Chi-Square		
	DF	Value	Value/D F	DF	Value	Value/D F
Model 1A	10	135.0000	1.3366	10	135.0000	1.3366
Model 1B	10	135.0000	1.3366	10	135.0000	1.3366
Model 2A	*	*	*	*	*	*
Model 2B	10	7295.150	72.2292	10	8258.496	81.7673
Model 3A	12	151.5471	1.2028	12	145.1155	1.1517
Model 3B	10	148.5998	1.4713	10	137.0886	1.3573
*Model does not converge						

With the distribution and link function chosen, it becomes necessary to consider potential effects within the model. Tables 13a and 13b provide complete assessment and effects details for the full model (with 2-way effects) using the Poisson distribution and log link function. The assessment of this model is superior to that of the single effect Poisson model, and there are no single parameters or interactions that are insignificant for all potential p-values. As a result, the task of selecting effects is quite simple: all merit inclusion, and our full effects model is the optimal selection in prediction.

Table 13a. Optimal Generalized Model Assessment

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/D F
Deviance	101	7295.1506	72.2292
Scaled Deviance	101	7295.1506	72.2292
Pearson Chi-Square	101	8258.4961	81.7673
Scaled Pearson X2	101	8258.4961	81.7673
Log Likelihood		57547.5225	

Table 13b. Optimal Generalized Model Effects

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.9673	0.3459	32.35	<.0001
Age	1	0.0795	0.0051	241.19	<.0001
KS	1	0.0602	0.0041	214.99	<.0001
MFD	1	0.0407	0.0126	10.52	0.0012

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
PT	no	1	3.7253	0.2255	273.00	<.0001
PT	yes	0	0.0000	0.0000	.	.
ToC	Adeno	1	-4.1782	0.2697	239.92	<.0001
ToC	Large	1	-0.0810	0.1927	0.18	0.6741
ToC	Small	1	-1.8538	0.1981	87.53	<.0001
ToC	Squamous	0	0.0000	0.0000	.	.
Treatment	Standard	1	2.1288	0.1635	169.61	<.0001
Treatment	Test	0	0.0000	0.0000	.	.
Age*KS		1	-0.0003	0.0001	25.55	<.0001
Age*MFD		1	-0.0015	0.0002	57.98	<.0001
Age*PT	no	1	-0.0396	0.0028	198.57	<.0001
Age*PT	yes	0	0.0000	0.0000	.	.
Age*ToC	Adeno	1	0.0202	0.0031	42.50	<.0001
Age*ToC	Large	1	0.0031	0.0023	1.80	0.1799
Age*ToC	Small	1	0.0102	0.0025	17.10	<.0001
Age*ToC	Squamous	0	0.0000	0.0000	.	.
Age*Treatment	Standard	1	-0.0295	0.0021	200.26	<.0001
Age*Treatment	Test	0	0.0000	0.0000	.	.
KS*MFD		1	0.0002	0.0001	10.21	0.0014
KS*PT	no	1	-0.0231	0.0017	195.55	<.0001
KS*PT	yes	0	0.0000	0.0000	.	.
KS*ToC	Adeno	1	0.0126	0.0016	60.90	<.0001
KS*ToC	Large	1	-0.0044	0.0016	7.07	0.0078
KS*ToC	Small	1	-0.0013	0.0014	0.89	0.3453
KS*ToC	Squamous	0	0.0000	0.0000	.	.
KS*Treatment	Standard	1	-0.0169	0.0013	181.94	<.0001
KS*Treatment	Test	0	0.0000	0.0000	.	.
MFD*PT	no	1	-0.0073	0.0034	4.49	0.0341
MFD*PT	yes	0	0.0000	0.0000	.	.

Parameter			DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
MFD*ToC	Adeno		1	0.0894	0.0077	133.69	<.0001
MFD*ToC	Large		1	-0.0064	0.0050	1.63	0.2012
MFD*ToC	Small		1	0.0518	0.0029	309.16	<.0001
MFD*ToC	Squamous		0	0.0000	0.0000	.	.
MFD*Treatment	Standard		1	0.0183	0.0032	33.35	<.0001
MFD*Treatment	Test		0	0.0000	0.0000	.	.
PT*ToC	no	Adeno	1	0.6290	0.0976	41.51	<.0001
PT*ToC	no	Large	1	-0.2097	0.0608	11.91	0.0006
PT*ToC	no	Small	1	-0.2555	0.0638	16.06	<.0001
PT*ToC	no	Squamous	0	0.0000	0.0000	.	.
PT*ToC	yes	Adeno	0	0.0000	0.0000	.	.
PT*ToC	yes	Large	0	0.0000	0.0000	.	.
PT*ToC	yes	Small	0	0.0000	0.0000	.	.
PT*ToC	yes	Squamous	0	0.0000	0.0000	.	.
PT*Treatment	no	Standard	1	0.4247	0.0604	49.41	<.0001
PT*Treatment	no	Test	0	0.0000	0.0000	.	.
PT*Treatment	yes	Standard	0	0.0000	0.0000	.	.
PT*Treatment	yes	Test	0	0.0000	0.0000	.	.
ToC*Treatment	Adeno	Standard	1	0.4963	0.0674	54.15	<.0001
ToC*Treatment	Adeno	Test	0	0.0000	0.0000	.	.
ToC*Treatment	Large	Standard	1	0.7499	0.0535	196.49	<.0001
ToC*Treatment	Large	Test	0	0.0000	0.0000	.	.
ToC*Treatment	Small	Standard	1	0.6776	0.0552	150.52	<.0001
ToC*Treatment	Small	Test	0	0.0000	0.0000	.	.
ToC*Treatment	Squamous	Standard	0	0.0000	0.0000	.	.
ToC*Treatment	Squamous	Test	0	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	.	.

CONCLUSION

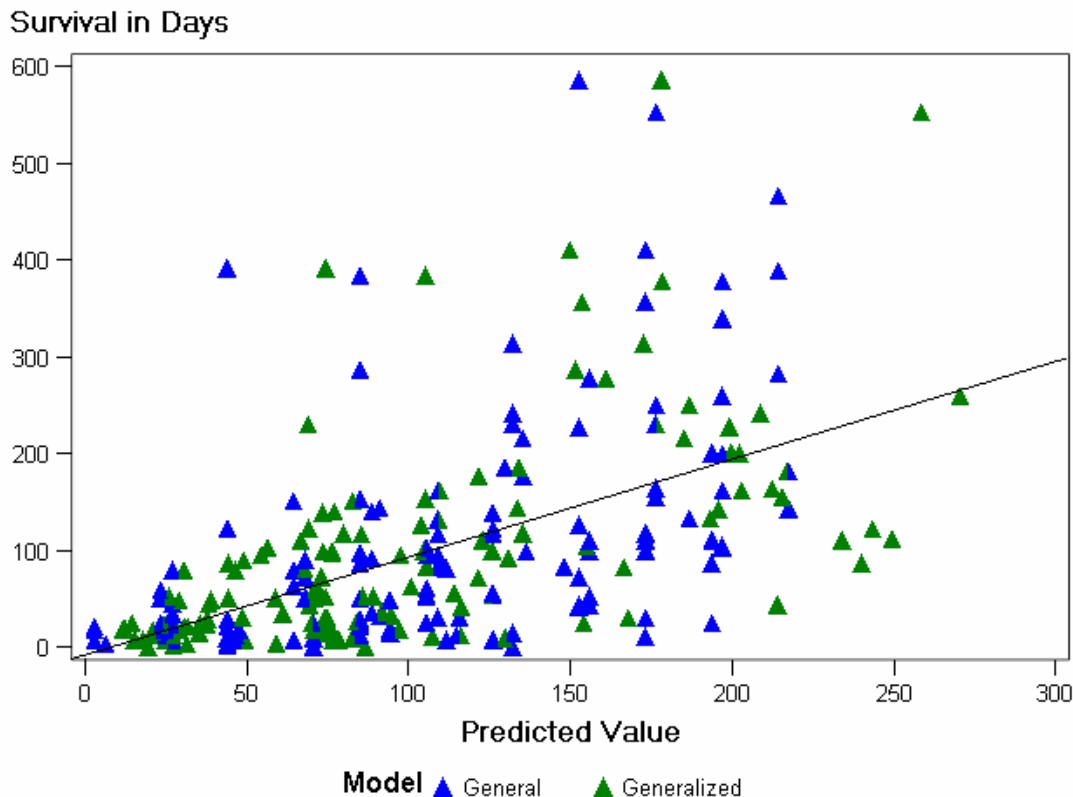
Results for both the general and generalized model have been produced, and these results have been assessed for performance strength. It has furthermore been demonstrated that the generalized model is statistically superior in predicting the survival data. However, a few more points need to be made regarding the two models and potential comparison. First of all, it needs to be pointed out that either model suffers from the potential of over-fitting the data. This is clear not only from the prior statistical analysis performed, but also from the limited scope of the data used. Were more data available, and were practical application of these results sought, it would be quite useful to separate the data into training and validation datasets. The consequential assessments would then suffer far less from over-fitting.

Another weakness of this experiment rests in the contemporary relevance of the data itself, as is common to most medical data analyses. Definitions of what was “cutting-edge” treatment for cancer in 1970 would hardly apply today. Again, any practical application of the data would require more recent data, or perhaps a time series analysis of a number of lung cancer trial datasets taken over the years. Unfortunately, due to issues of privacy, these data are still not as plentiful or available as would be desired.

An additional complication arises from the fact that many definitions of attributes in the dataset are subjective at some level. For instance, the Karnofsky score is based on physician judgment, which may vary widely from doctor to doctor. As such, an ideal dataset would also include information regarding the physician, which could in turn potentially be used to “normalize” the Karnofsky scores. Similar examples include the fact that contemporary doctors use a slightly different methodology in labeling types of carcinoma.

These caveats aside, this experiment does demonstrate that the generalized linear model is statistically superior in analysis of the underlying data. Figure 5 should help to visualize the difference in performance. It can be seen that, with the exception of a few outlying points, the predicted survival in the generalized model shows much less spread from the actual result. This is especially true for larger survival values, for which the general solution tends to produce very inaccurate predictions.

Figure 5



Having compared the models produced for this report, all that remains is to consider potential improvements to our methodology. As noted above, for these results to apply to a practical endeavor, it would be necessary to gather more recent trial data and perhaps use a larger number of test subjects. Another potential change to make would be to evaluate patients on multiple visits throughout the course of their treatment, potentially allowing for a time-series

analysis of the data. With regards to the attributes in the dataset, it would be appropriate to specifically define the type of treatment (rather than “standard” or “test”) used on a particular patient. If some or all of these extensions were applied, it is quite feasible that the results could enhance the ability for physicians to predict survival.

REFERENCES

- [1] J.D. Kalbfleisch and R.L. Prentice. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York: 1980.
- [2] <http://www.hospicepatients.org/karnofsky.html>. Accessed May 10, 2006.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Andrew Karem
2439 Ransdell Ave.
Louisville, KY 40204
(502)454-4174
Email: adkare01@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.