Assignment 9: Final – DA Report – Duration Model: DA

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

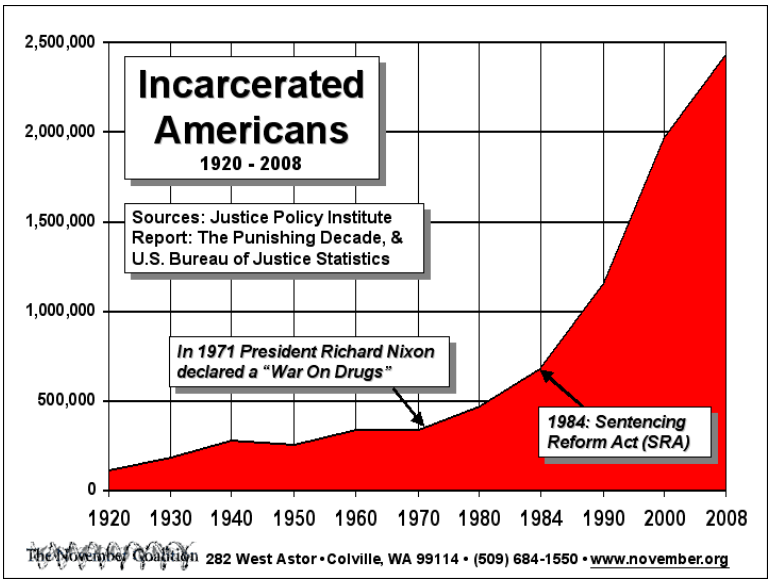
Eden Prairie, MN 55344

Executive Summary

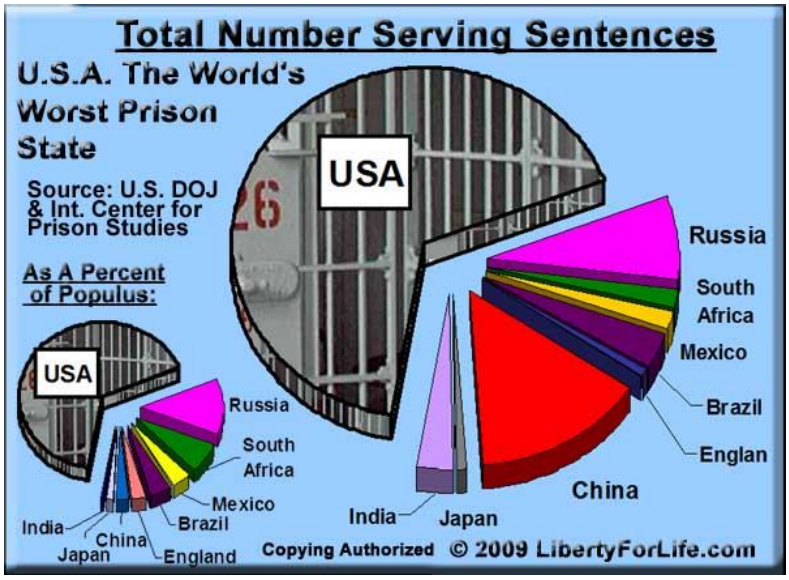
The United States has more of its citizens behind bars than any other country. It is estimated that over 63 billion dollars of US taxpayer money went to the prison system in 2010. Fighting the recidivism rate is a great way to curb the tax payer expense. This exploratory data analysis (EDA) analyzes recidivism through duration analysis. In total, 12 variables were analyzed, of which 11 were predictor variables. Censoring occurred at the 81 month interval and 60% of prisoners had not gone back to prison. The normal log distribution proved to be the most accurate distribution for the response variable, continuous variables, as well as the whole mode given that it yielded the lowest AIC values. It was found that individuals that were black and had drug and alcohol problems had the highest rate of recidivism. While, individuals that were married, and had previous felons were least likely to go back to prison within the time period of the study. Overall, working to curb recidivism is a plan that will save taxpayers billions of dollars.

Introduction

According to the Oxford dictionary, the word recidivist means, "a convicted criminal who reoffends, especially repeatedly." The goal of a corrections institution is to rehabilitate inmates such that restoration and reformation lead to changed behaviors that result in crime free living. Sadly, the inverse of this goal is often what happens to inmates that spend time behind bars. Since the war on drugs in 1971, the incarceration rate has grown faster than any other time in the United States. Over 2.5 million Americans are behind bars. Over the last two decades, legislation has been put into law that results in harsher sentences. According to the Vera institute of Justice, in 2010 the average incarceration cost in the US was over $31,000. This has resulted in over 63 billion dollars on the US taxpayer (cbsnews.com).

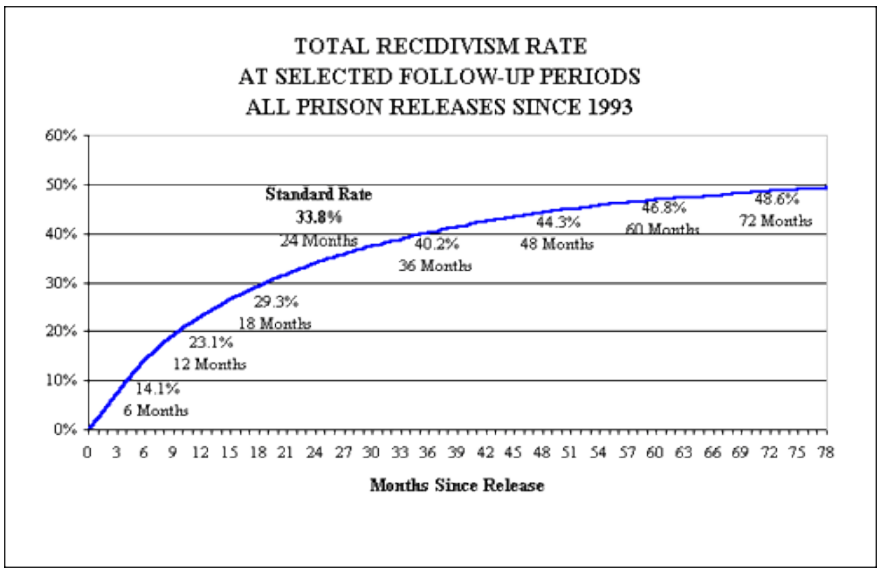


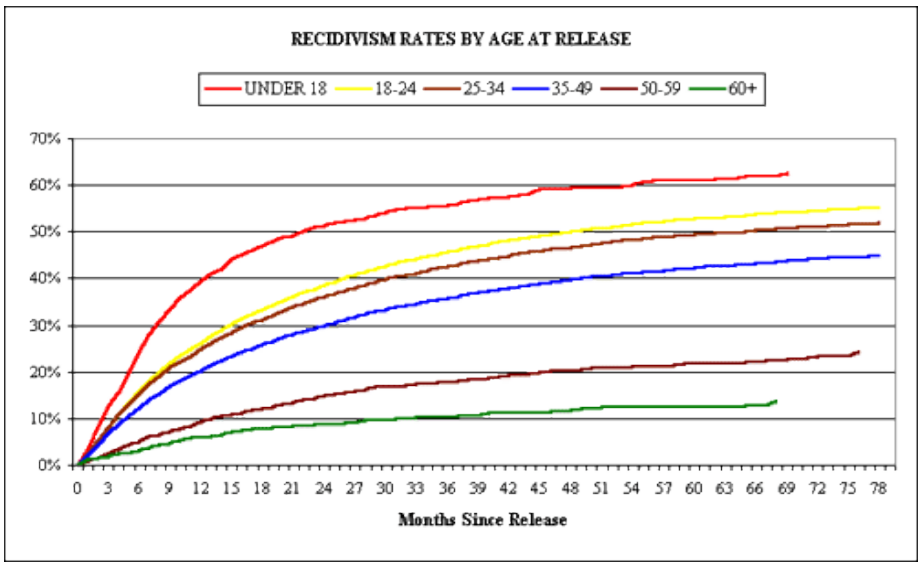
Compared to the world, the US has five percent of the population, but 25 percent of the world's prisoners. From the pie graph to the right, it can be seen that the US also has the largest number of



people serving sentences as well.

One of the greatest contributing factors to the growing prison population is recidivism. In my opinion, recidivism highlights the major flaw in our correctional system and has created a codependent relationship between prisoners and prisons such that "once a prisoner always a prisoner." The graph to the left highlights the recidivism rate for Florida since 1993. It can be seen that within 6 years of release, 48% of released prisoners had been convicted of another crime and were back in prison. To take this study one step further, one would need to break down the ages of the recidivism rates to gauge the probability of how long the average person in prison will be in prison throughout their lifetime. The graph below shows the recidivism rate by age for the state of Florida. One can see that age is a direct correlation for increased recidivism. Statistically, one could draw the conclusion that based





on a prisoners age, he or she is more likely to spend the majority of their life in prison if they were in prison from a younger age. Studies that further diagnose recidivism and suggest ways to cut recidivism are very helpful at the individual level but also at the macro level for increasing economic output.

The dataset considered in this EDA is analyzed in Wooldridge (2002) and credited to Chung, Schmidt, and Witte (1991). The data pertain to a random sample of convicts released from prison between July 1, 1977 and June 30, 1978. The objective of this EDA is to analyze different variables in correlation with the recidivism rate, and this will be done utilizing duration analysis models. Listed below are my first impressions of how the variables will interact with the recidivism rate:

Variable	First Impression of Interaction with Recidivism
Durat - the duration in months until return to prison.	The shorter the duration the more likely to recidivate more frequently.
Cens - the censoring indicator variable.	If censored, greater chance of recidivism.
Workprg - an indicator of participation in a work program.	If in a work program the longer one will go from recidivating.
Priors - the number of previous convictions.	The greater the priors the sooner the recidivism.
Tserved - the time served rounded to months.	The more time served the longer the recidivism.
Felon - an indicator of felony sentences.	If a felon, greater chance of recidivism.
Alcohol - an indicator of alcohol problems.	If alcohol is a problem, recidivism will be more frequent.
Drugs - an indicator of drug use history.	If drugs are a problem, recidivism will be more frequent.
Black - an indicator for African Americans.	If Black, recidivism will be more frequent.
Married - an indicator if married when incarcerated.	If Married, recidivism will be less frequent.
Educ - the number of years of schooling.	The less education, the more frequently recidivism will occur.

Age - in months.	The younger an individual the more frequently recidivism will occur.

Analysis

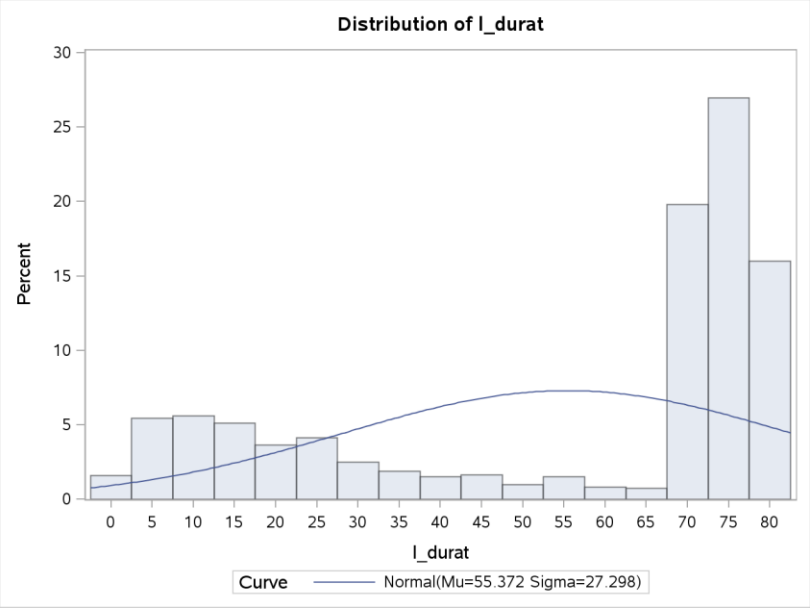
In order to meet the objective of exploring the relationship between the recidivism rate and the 12 variables, an exploratory data analysis (EDA) must be conducted. This EDA will start with a basic Proc Lifereg in SAS to begin the modeling process. From that model, analysis will be made based on the distribution that best fits the data. As a data scientist in training, I am inculcating a paradigm of which to study data. While this paradigm is redundant report to report, it is training me to have the correct mindset. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

> Problem/Objective: Explore the relationship between the recidivism rate and 12 variables. Their interpretations were explained above.
>
> Data: The data has been aggregated and has been supplied from management. There are no missing values.
>
> Analysis: I will explore the data via simple statistics. After the initial analysis, I will start with fitting the data with a Proc Lifereg in SAS to pick a distribution.
>
> Model: Proc Phreg in SAS will be used to fit a proportional hazard model to the dataset. In addition, I will analyze additional variables and weigh the efficacy of adding them to the model.
>
> Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the model fits that data and the statistical backing of the model.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best model, and the analyst's personal bias is mitigated.

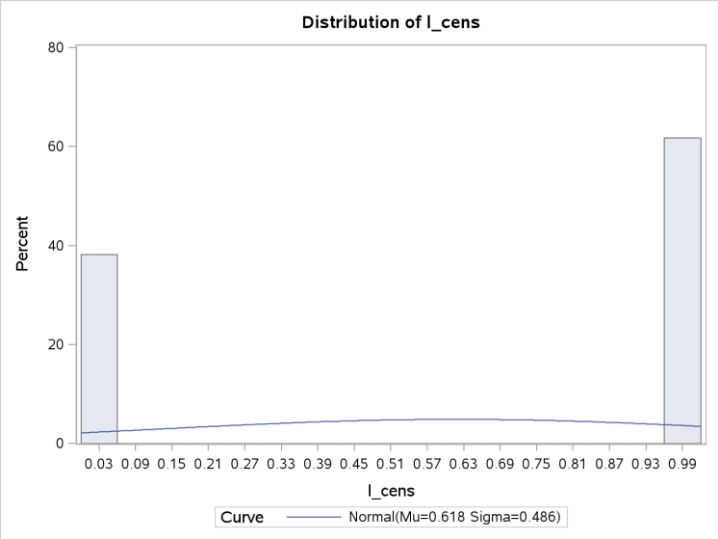
Data

Of interest is the time until inmates return to prison. The information was collected retrospectively by looking at records in April 1984, so the maximum possible length of observation is 81 months. There are a total of 1445 observations with 0 completely missing values per row. Management has requested the focus to be on the following variables: durat - the duration in months until return to prison, cens - the censoring indicator variable, workprg - an indicator of participation in a work program, priors - the number of previous convictions, tserved - the time served rounded to months, felon - an indicator of felony sentences, alcohol - an indicator of alcohol problems, drugs - an indicator of drug use history, black - an indicator for African Americans, married - an indicator if married when incarcerated, educ - the number of years of schooling, and age - in months. The data is comprised of 18 total variables over 81 months. Thus, each observation, which is an individual, has 18 data values and there are 1,445 observations in this dataset which equals a total of 26,010 data values being studied in this EDA. Below you will find general descriptive statistics of the variables and their correlation with the response variable.



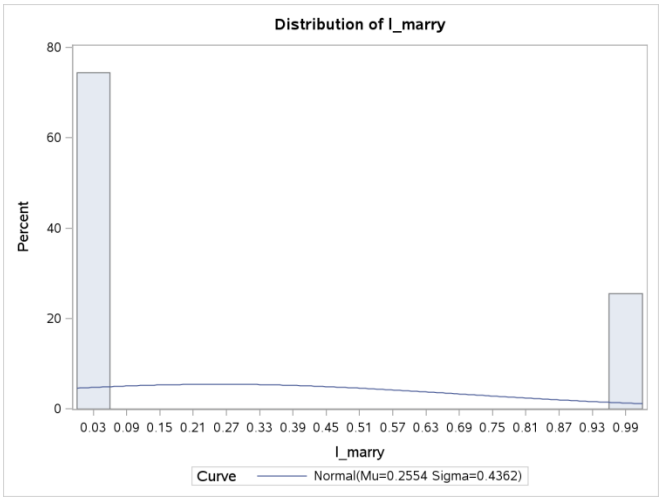
Durat - the duration in months until return to prison. The average duration of time for recidivism was 55.372 months, and a median of 71. There is a negative skew of .819, which is quite large, and a standard deviation of 27.298 months. As seen by the histogram,

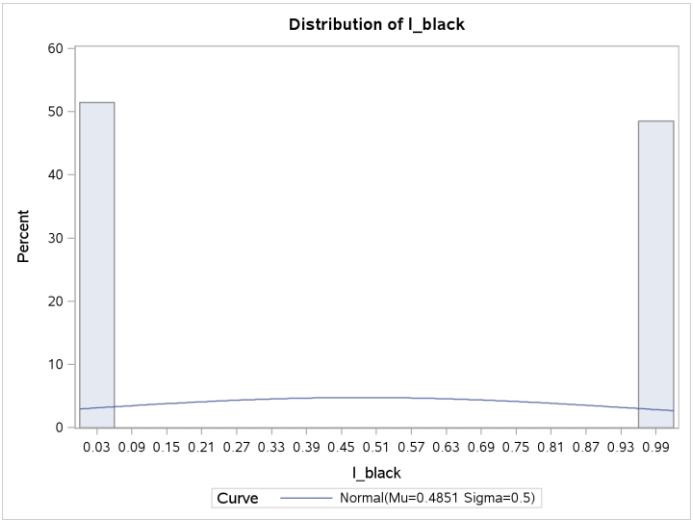
the last three months account for more than 60% of the values. It can be seen that the majority of released prisoners lasted more than 5 years.



Cens, is the censoring indicator variable and indicates how many complete observations are available at the 81 month interval. From the histogram, it can be seen there are about 60% censored values, which mean that 60% did not go back to prison.

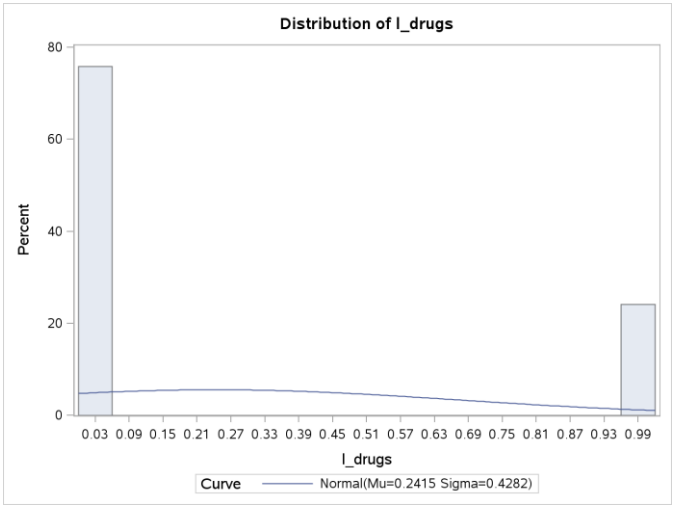
Married, an indicator variable if married when incarcerated. This variable is binary, which represents 0 if not married and 1 if married. The mean is .255 which shows that .255 of the prisoners were married when incarcerated, and the median is 0 because of the binary response. There is a positive skew of 1.123, which demonstrates that majority of the inmates were not married.



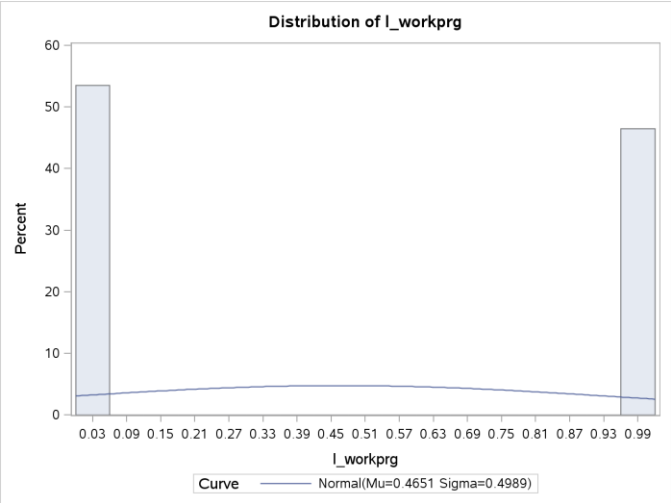


From the histogram, it can be seen that about half of the incarcerated population is black. This variable is binary, which represents 0 if not black and 1 if black. The mean is .485 which shows that 48% of

the prisoners were not black, and the median is 0 because of the binary response. There is a slight positive skew of .06, which demonstrates that the data is well distributed for this variable, but there are a few more black inmates than non-black.

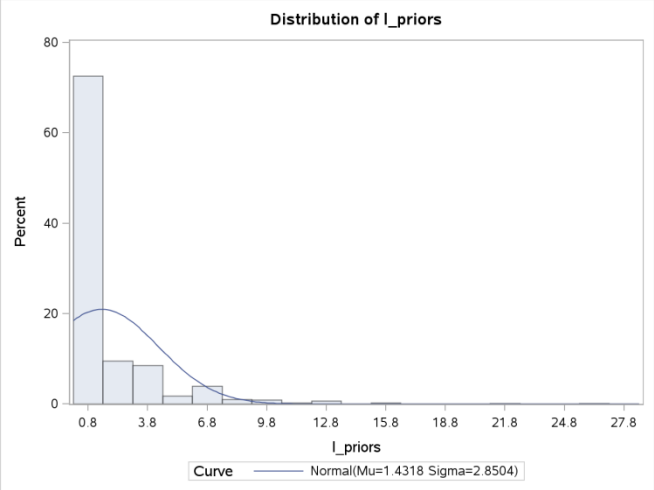


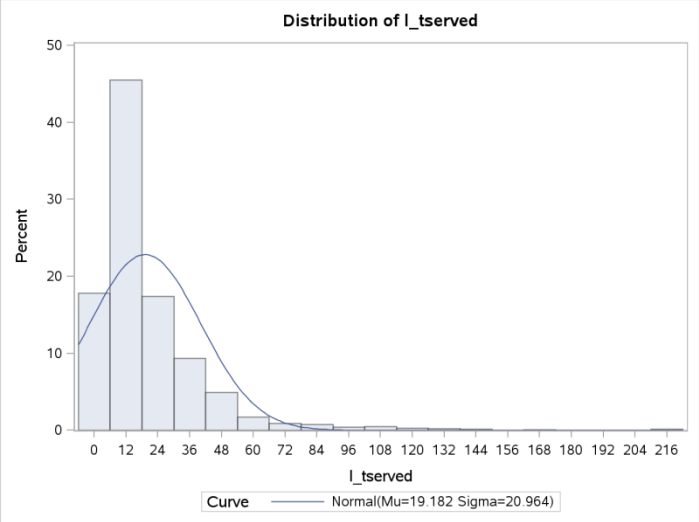
Drugs, an indicator variable of drug use history. From the histogram, it can be seen that about 75% of the incarcerated population does not have a drug problem. This variable is binary, which represents 0 if no drug problem and 1 if drug problem exists. The mean is .241 which shows that 24% of the prisoners have a drug problem, and the median is 0 because of the binary response. There is a positive skew of 1.209, which demonstrates that majority of the observations are 0.



Workprg, is an indicator variable of participation in a work program. From the histogram, it can be seen that about 53% of the incarcerated population did not partake in this program. This variable is binary, which represents 0 if one did not participate and 1 if there was participation. The mean is .465 which shows that 47% of the prisoners partook in the program, and the median is 0 because of the binary response. There is a slight positive skew of .140, which demonstrates that the data is well distributed for this variable, but less participation in the program.

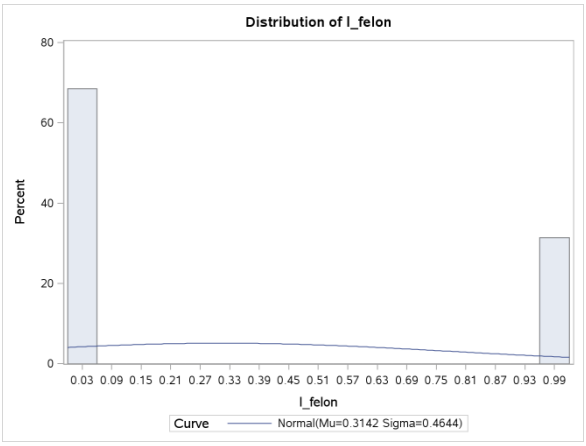
Priors, is the indicator variable that shows the number of previous convictions. The average number of previous convictions per inmate was 1.431, and a median of 0. The median shows that the vast majority of personnel did not have a prior conviction; in fact over 70% did not have a prior conviction. There is a positive skew of 3.995, which is quite large, and a standard deviation of 2.850. As seen by the histogram, the first value of 0 represents over 75% of the values.





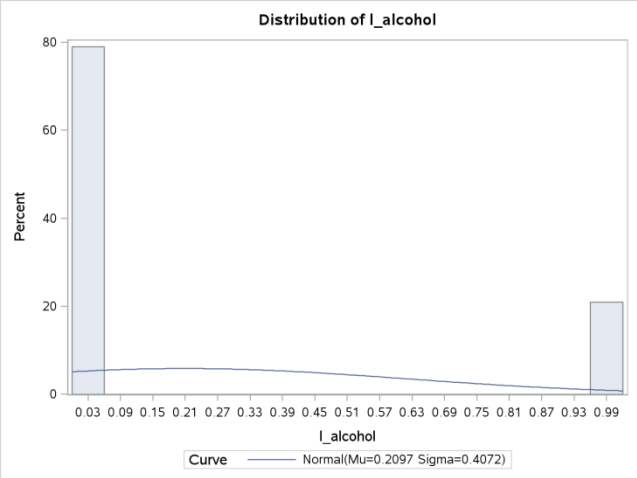
Tserved, is the indicator variable that is the time served rounded to months. The average number of months served is 19.182 months, and a median of 12. The median shows that the vast majority of personnel served 12 months; in fact around 65% served 12 months or less. There is a positive skew of 3.412, which is quite large, and a standard deviation of 20.964. As seen by the histogram, the first 12 months represents over 65% of the values.

Felon is an indicator variable that delineates whether the personnel has a felony. From the histogram, it can be seen that about 65% of the incarcerated population does not have a felony. This variable is binary, which



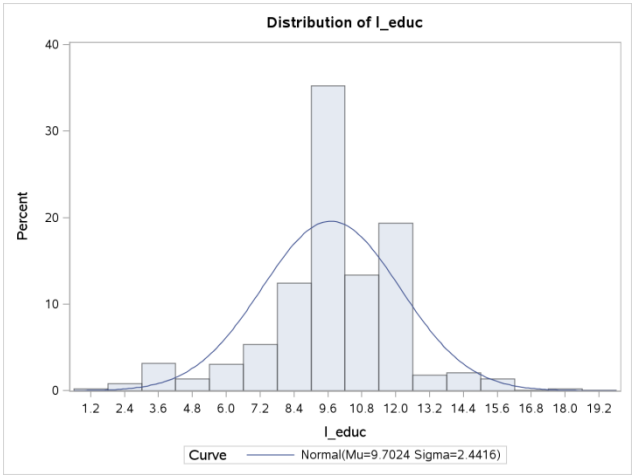
represents 0 for no felony and 1 if convicted of a felony. The mean is .314 which shows that 31% of the prisoners have a felony, and the median is 0 because of the binary response. There is a positive skew of .801, which demonstrates that majority of the observations are 0.

Alcohol is an indicator variable of an alcohol problem for the personnel. From the histogram, it



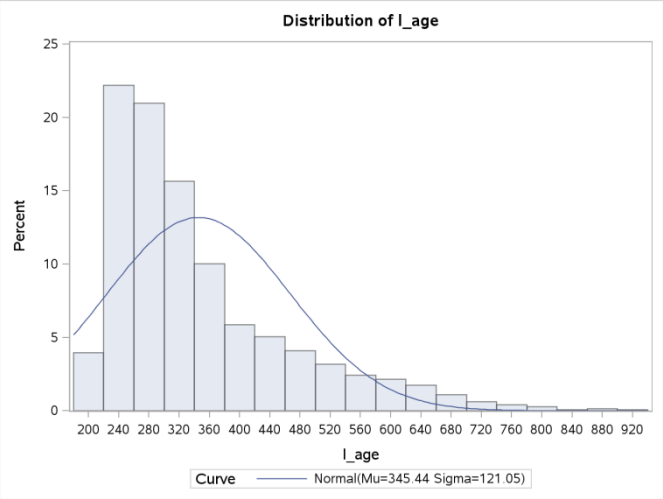
can be seen that about 80% of the incarcerated population does not have an alcohol problem. This variable is binary, which represents 0 for no alcohol problem and 1 for alcoholism. The mean is .210 which shows that 21% of the prisoners have an alcohol problem, and the median is 0 because of the binary response. There is a positive skew of 1.428, which demonstrates that majority of the observations are 0.

Educ is the indicator variable that shows the number of years of education. The average number of years of education per inmate was 9.702, and a median of 10. The close value between the mean and median demonstrates that the distribution is fairly normal. There is



a negative skew of .533, which is not too bad, and a standard deviation of 2.442. As seen by the histogram, 35% percent of the overall values were 9.6 years of education.

Age is the indicator for age expressed in months. The average age is 345.436 months, and a median of 307. The median shows that the vast majority of personnel are younger than the mean; in fact around 60% are less than 345 months old. There is a positive skew of 1.489, which is okay, and a standard deviation of 121.050. As seen by the histogram, the first 320 months represents over 60% of the values.



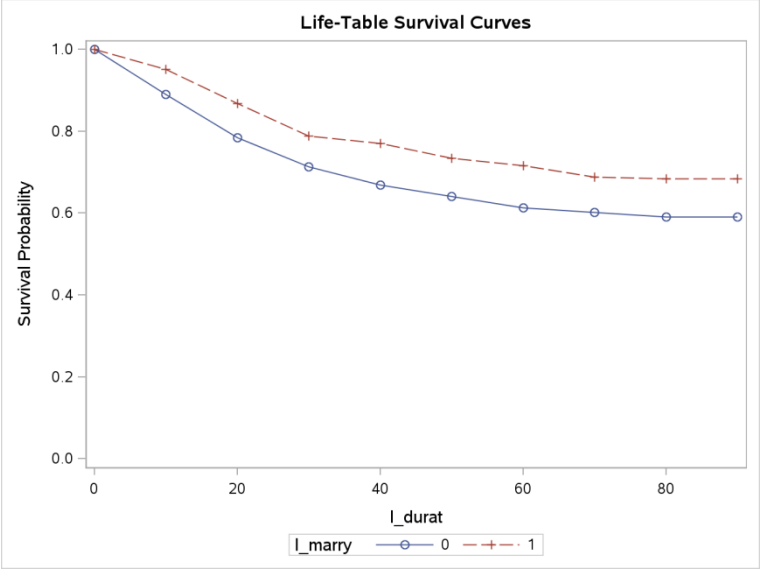
Results

The first step in building a statistically sound model is to understand the response variable and its distribution. General descriptive statistics for durat are listed above. There are three distributions that a response variable has in duration analysis, and based on the goodness of fit statistics, one is chosen. Utilizing Proc Lifereg in SAS allows one to see all three distributions for durat. Appendix 1 shows the entire output for the three distributions. To recap, there are a total of 1445 individuals that this EDA is studying, of which their duration until recidivism is of interest. There are a total of 552 uncensored values, meaning that the full duration is followed including their relapse back in prison. The rest of the observations, 893, did not relapse and are considered right censored after the 81 month time period. In essence, the majority of personnel did not end of back in prison. When analyzing different distributions, one desires a tight fit which boils down to lower fit statistics. Overall, analyzing the AIC is helpful for assessing the fit; the normal distribution had a value of 7172, exponential had a value of 6599, Weibull had a value of 6553, and taking the log of the normal distribution had a value of 6483. The distribution that will be best to use for the response variable is the log distribution.

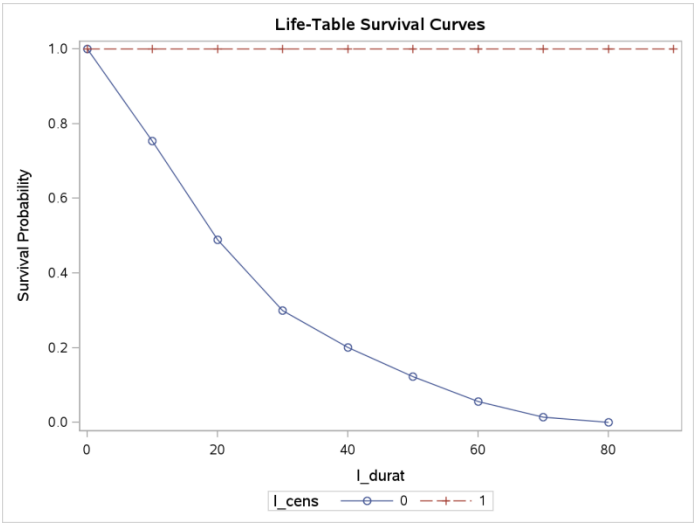
Understanding how each indicator variable interacts with the response variable is the next step in the EDA. Utilizing the Proc LifeTest with the Strata statement in SAS is a great tool to assess the relationship between the qualitative variables and the response variable. For example, here is consolidated output for marry:

Life Table Survival Estimates							
Non-Married					Married		
[Lower,	Upper)	Number Failed	Number Censored	Conditional Probability of Failure	Number Failed	Number Censored	Conditional Probability of Failure
0	10	118	0	0.1097	18	0	0.0488
10	20	115	0	0.1200	31	0	0.0883
20	30	76	0	0.0902	29	0	0.0906
30	40	47	0	0.0613	7	0	0.0241
40	50	30	0	0.0417	13	0	0.0458
50	60	30	0	0.0435	7	0	0.0258
60	70	13	0	0.0197	10	0	0.0379
70	80	7	552	0.0189	1	224	0.00704
80	90	0	88	0	0	29	0
90	.	0	0	0	0	0	0

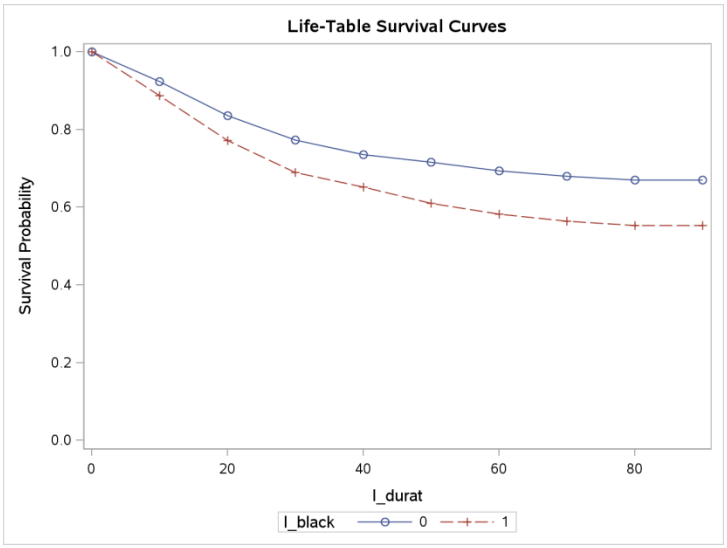
From the output above, it can be seen that married inmates have less of a probability of recidivism than the general inmate population except for the months between the highlighted sections. The life-table survival curve visually shows the recidivism rate over time comparing married and non-married inmates. In conclusion, inmates that are married will have less of a recidivism rate than non-married inmates.



Appendix 2 shows all the output for the following qualitative variables: cens, black, drugs, workprg, felon, and alcohol. I will highlight the influence each variable has on recidivism over time.

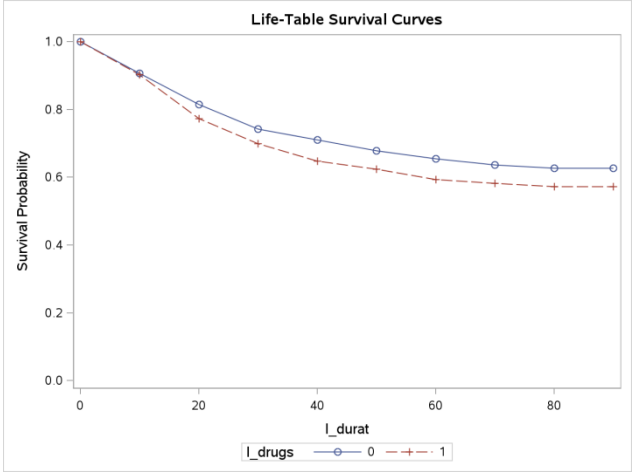


For inmates that were censored, they did not go back to prison, while those that were censored went back to prison. This can be seen in the plot to the right. The red line represents the inmates that were censored, and the blue line represents those that went back to prison.

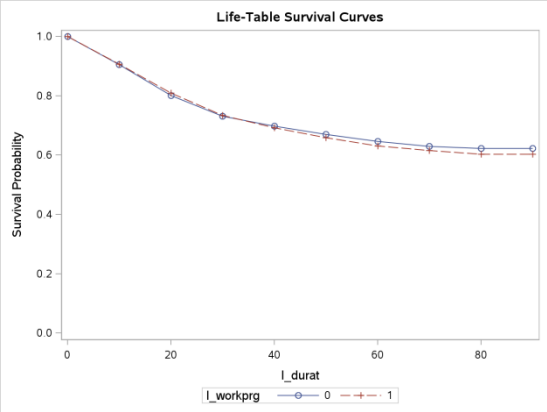


For inmates that were black, they have a higher probability of going back to prison during the time of this study. This can be seen in the plot to the left. The red line represents the inmates that were black, and the blue line represents those that were not black.

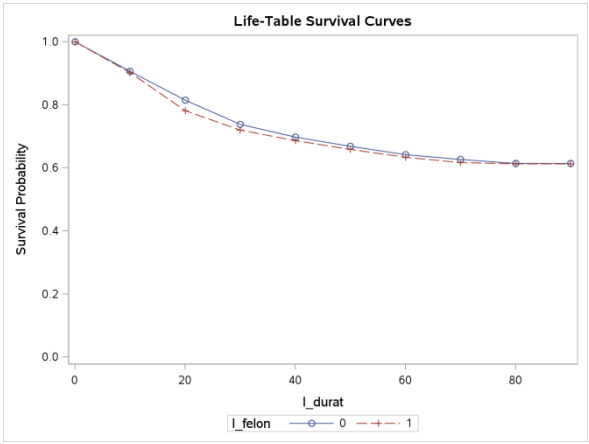
For inmates that had drug problems, they have a higher probability of going back to prison during the time of this study. This can be seen in the plot to the left. The red line represents the inmates that had drug



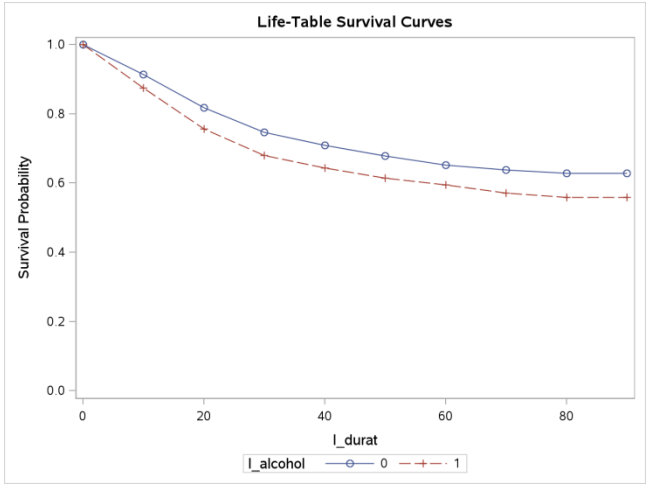
problems, and the blue line represents those that did not.



For inmates that participated in the work program, they have a higher probability of going back to prison during the time of this study. This can be seen in the plot to the left. The red line represents the inmates that were black, and the blue line represents those that were not black, which shows very little change or delineation. In my opinion, this variable is helpful in determining recidivism.



For inmates that had a felon, they have a slightly higher probability of going back to prison during the time of this study. This can be seen in the plot. The red line represents the felons, and the blue line represents those that were not black, which shows very little change or delineation. In fact, as time increases the two lines intersect.



For inmates that had an alcohol problem, they have a higher probability of going back to prison during the time of this study. This can be seen in the plot. The red line represents the alcoholics, and the blue line represents those that were not alcoholics.

Alcoholism, drug issues, being black, and being married are all strong indicators for recidivism, except for marriage which is inversely related.

The continuous variables utilize the same approach as the analysis for the response variable. Analyzing the correct distribution is the first step in fitting the variables. Appendix 3 highlights the output form running Proc Lifereg in SAS, and I will highlight the findings.

Distribution	AIC
Log Normal	6379.725
Exponential	6477.473
Weibull	6438.956

From my table chart to the left, it can be seen that using a log normal distribution will fit the data the best for the continuous variables. The next step is to analyze the variables. Maximum likelihood was used to fit the variables.

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.8512	0.3488	3.1676	4.5348	121.92	<.0001
l_priors	1	-0.1562	0.0219	-0.1991	-0.1134	51.06	<.0001
l_tserved	1	-0.0147	0.0026	-0.0198	-0.0096	32.08	<.0001
l_educ	1	0.0167	0.0257	-0.0336	0.0670	0.42	0.5158
l_age	1	0.0038	0.0006	0.0027	0.0050	42.66	<.0001
Scale	1	1.8590	0.0642	1.7374	1.9892		

Education is the only variable that is not statistically significant. Interpreting coefficients for Logistic Regression (LR) is not as straight forward as in Ordinary Least Squares Regression. In LR a logit coefficient of .5 can be interpreted as .5 log odds increase for every 1-unit increase in the explanatory variable, assuming all the other coefficients are held constant (Allison). It is really hard to conceptualize a .5 log-odds increase, which can be explained by the fact that LR captures a nonlinear relationship. From the Maximum Likelihood Estimates output, what can be gleaned is the statistical significance as well as the sign of the estimate. A positive or negative sign indicates the direction of the relationship (uoregon.edu). In addition to assessing the sign of a coefficient, analyzing the

statistical significance is important. P-values assess the probability that your sample results are chance or extreme given that the null hypothesis is true. As the p-value increases, the probability increases that the sample estimate is based on pure chance. Lower p-values are an indicator of a statistically solid coefficient.

The log-odds ratio is a far better output for understanding the coefficients. The odds ratio is simply computed by taking the natural log of e^estimate. In addition to the odds, this calculation is also adjusted since it controls for other variables. Once the odds ratio has been calculated, the output is easier to understand. For example, the odds ratio for married can be calculated as 2.7182 (e) ^ -.156 = .856. This can be interpreted as an individual that has prior convictions has .856 times the odds of recidivism than non-prior conviction. I personally prefer probability to odds. This can be translated from the following calculation, probability = -.156/(1 + -.156) = -.18. Therefore, a person with prior convictions has a -18% chance of recidivism than those without prior convictions. To the left is a break down for each coefficient and the corresponding odds ratio, in addition I calculated the probability. See appendix 4 for the proof done in excel. Please note, the statistical validation was interpreted in the preceding

Odds Ratio Estimates					
Variable	**Log Odds**	**Probability**	**95% Wald Confidence Limits**		**Description**
l_priors	0.855392	0.46103037	-0.1991	-0.1134	This variable has strong odds, and is statistically significant.
l_tserved	0.985407	0.49632518	-0.0198	-0.0096	There is very little predictive significance with this variable.
l_educ	1.016839	0.50417478	-0.0336	0.0670	This variable does not have strong odds, and is not statistically significant.
l_age	1.003807	0.50094997	0.0027	0.0050	There is very little predictive significance with this variable.

section, and this section is simply interpreting the odds ratio.

Now that all the variables have been defined, I will use Proc LifeReg to fit all the variables. As with the response variable and continuous variables, a proper distribution will have to be picked first, and then the interpretation of the variables will follow. Appendix 5 shows the complete output for assessing which distribution should fit the data. From the chart to the left, it can be seen that the Log Normal distribution fits the data the best. This follows in line with the entire EDA thus far.

Distribution	AIC
Log Normal	3218.118
Exponential	3321.506
Weibull	3290.065

	Analysis of Maximum Likelihood Parameter Estimates					
Parameter	**DF**	**Estimate**	**Log Odds**	**Probability**	**Pr > ChiSq**	Description
Intercept	1	4.0994	60.2966528	0.983686	<.0001	NA
l_black	1	-0.5427	0.58118645	0.367564	<.0001	Statistically sound, and negative relationship.
l_drugs	1	-0.2982	0.74215956	0.426	0.0247	Statistically okay, and negative relationship.
l_workprg	1	-0.0626	0.9393209	0.484356	0.6022	Statistically Invalid
l_priors	1	-0.1373	0.87171228	0.46573	<.0001	Statistically sound, and negative relationship.
l_tserved	1	-0.0193	0.98088562	0.495175	<.0001	Statistically sound, and negative relationship.
l_felon	1	0.4440	1.55890965	0.609209	0.0022	Statistically sound, and positive relationship.
l_alcohol	1	-0.6349	0.52999861	0.346405	<.0001	Statistically sound, and negative relationship.
l_marry	1	0.3407	1.40591698	0.584358	0.0148	Statistically sound, and positive relationship.
l_educ	1	0.0229	1.02316351	0.505725	0.3668	Statistically Invalid
l_age	1	0.0039	1.0039075	0.500975	<.0001	Statistically sound, and positive relationship.
Scale	1	1.8105	60.2966528			

As seen above, two variables are not statistically significant at the 95% confidence level. The variables that are significant demonstrate the predictive power of the model. Another approach to analyzing the data is to utilize the Proc Phreg function in SAS, which models the data from a hazard perspective rather than the survival. Appendix 6 lists the full output for this procedure, and the models AIC was 7788.360, which is more than double the model fit above.

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Log Odds**	**Probablity**	**Pr > ChiSq**	**Hazard Ratio**
l_black	1	0.43257	1.54119329	0.60648409	<.0001	1.541
l_drugs	1	0.27558	1.31728356	0.56846024	0.0049	1.317
l_workprg	1	0.08403	1.08765877	0.52099452	0.3548	1.088
l_priors	1	0.08759	1.09153762	0.52188285	<.0001	1.092
l_tserved	1	0.01296	1.01304395	0.50323986	<.0001	1.013
l_felon	1	-0.28284	0.75364678	0.42975974	0.0077	0.754
l_alcohol	1	0.43063	1.53820636	0.606021	<.0001	1.538
l_marry	1	-0.1549	0.85650482	0.4613534	0.1561	0.857
l_educ	1	-0.02133	0.9788965	0.49466786	0.2728	0.979
l_age	1	-0.00358	0.99642651	0.49910503	<.0001	0.996

It can be seen from the output above that workprg, marry, and educ are not statistically significant. Alcohol, drugs, and being black are the highest indicators for recidivism. While this model is another perspective on the data, I would be hesitant to utilize it for a predictive model given that it does not fit as well as the lifereg model.

Through this EDA, the relationship between recidivism and specific variables became apparent in the initial descriptive breakdown. The models further validated these findings.

Future Work

Further recommendations on how this study can be improved upon are the following:

- Around 60% of the inmates did not go back to prison over this time period, and perhaps having a greater time interval that demonstrated a higher rate of recidivism would increase the model building learning experience.
- The Cox model proposed an interesting perspective of the data, but it would be helpful to know how to refine the AIC such that the model is a better fit.
- Conducting this study on a more recent data set would allow for a more cutting edge analysis in regard to assessing prison rehabilitation programs in conjunction with recidivism.

Through this initial EDA coupled with the future work recommendations, public corrections officials would gather pertinent information in regard to recidivism rates and program efficacy.

References

Aaron, . "Logistic Regression." *University of Oregon*. N.p., n.d. Web. 26 Jan. 2013. <http://pages.uoregon.edu/aarong/teaching/G4075_Outline/node16.html>.

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.

Allison, Paul David. *Logistic regression using SAS theory and application, second edition*. 2nd ed. Cary, N.C.: SAS Institute, 2012. Print.

Pearsall, Judy. "Oxford Dictionaries." *Recidivist*. Oxford University Press, n.d. Web. 18 Feb. 2013. <a convicted criminal who reoffends, especially repeatedly>.

Ratner, Bruce. *Statistical and machine-learning data mining techniques for better predictive modeling and analysis of big data*. 2nd ed. Boca Raton, FL: CRC Press, 2012. Print.

"Recidivist". Oxford Dictionaries. April 2010. Oxford Dictionaries. April 2010. Oxford University Press. 18 February 2013 http://oxforddictionaries.com/definition/english/recidivist