A Guide to Hazard Rate Modelling

Andy Toulis

September 2015

Table of Contents

1.0 Introduction

Survival analysis is concerned with predicting the time until an event happens. It is useful for comparing the risk among alternatives, such as using a lower quality material over a better one to build a product. Models which predict events belong to the broader category of **time-to-event** modelling.

2.0 Background on Survival Analysis

2.1 Events

In order for survival to be estimated, there must be an event which can end the life of a member of a population. Multiple events can be analyzed and are said to be **competing**. This means the probabilities of each event happening, including survival, must sum to one [1]. The risk of an event happening is assumed to be independent of time passed. This assumption has been applied in many contexts, such as queuing theory in healthcare [2].

A major part of modelling is deciding which events are of interest, as there are a countless number of competing risks. Some events are nearly deterministic in nature. The more that can be explained by simple rules, the less difficult it will be for a model to be estimated on complex parts of a dataset. A modeller should create prototype models assuming events are competing and then judge if they can be combined due to similar dependency on a set of independent variables.

2.2 Snapshot and Performance Dates

Each point in time in a discrete time dataset is called a snapshot date or a **snapshot**. A snapshot contains information about a member of the population. Future points in time after a given snapshot are called **performance** dates.

2.3 Measures

An important function in survival analysis is the event density function, the probability of an event occurring at a given point in time:

$$f(T) = Probability(Event\ at\ time\ T)$$

The **survival function** is the compliment of the cumulative event density function:

$$S(T) = Probability(Survival\ up\ to\ time\ T) = 1 - F(T)$$

The **hazard rate** is the most important function for modelling purposes and is defined as the probability of an event happening at a given point in time given survival up to that point in time [3]:

$$h(T) = Probability(Event\ at\ time\ T \mid Survival\ up\ to\ time\ T) = f(T)/S(T)$$

In a modelling context, the hazard rate can be used to estimate the proportion of population that survives:

$$Expected\ Population\ at\ time\ T = P_o \prod_{i=1}^{T} [1 - h(i)]$$

To understand this, one should recall that the hazard rate at any given point in time is the proportion of an already surviving population that is predicted to experience an event.

3.0 Discrete Time

In a discrete time setting, a row of data has, by existence, survived up to a given point in time since individuals are removed from a dataset after they experience an event. Instead of modelling events directly, one must model the hazard rate of an event since all data is conditioned on survival. The response variable of interest is an event **indicator variable**:

$$\mathbf{1}_T = \begin{cases} 1\ if\ an\ event\ occurs\ at\ time\ T\ for\ a\ loan\ existing\ in\ a\ dataset \\ 0\ otherwise \end{cases}$$

To get an understanding of the difference between hazard rates and event rates, see the illustrative example in Figure 1.0 below.

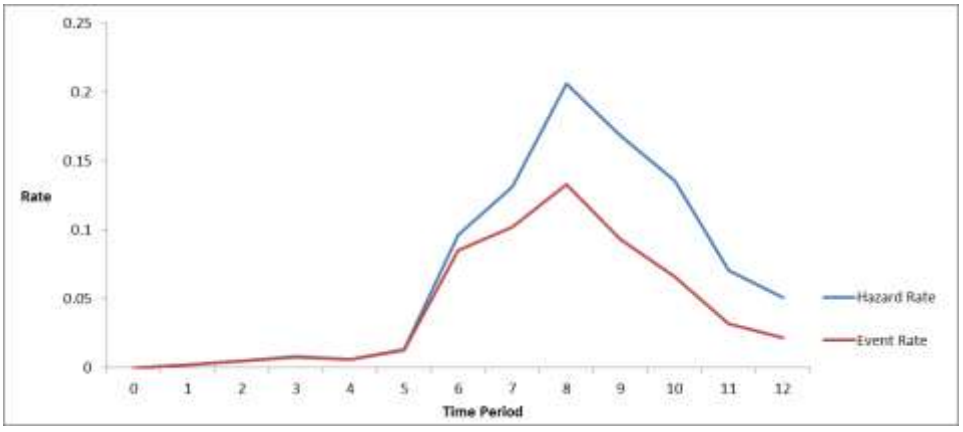

Figure 1.0:A comparison of the event and hazard rate of a population.

Table 1.0 below contains the data from which this plot was generated. The hazard rate is always greater than its corresponding event rate, since a population loses members as time passes. If the event rate spikes at a given point in time, so does the hazard rate. Visually inspecting the event rate curve can be misleading since the population which is able to experience events becomes smaller and smaller.

Table 1.0: Data illustrating the event and hazard rates corresponding to a series of events.

Time Period	Events	Alive	Event Rate	Survival Rate	Hazard Rate
0	0	1000	0.00	1.00	0.00
1	2	998	0.00	1.00	0.00
2	5	993	0.01	0.99	0.01
3	8	985	0.01	0.99	0.01
4	6	979	0.01	0.98	0.01
5	13	966	0.01	0.97	0.01
6	85	881	0.09	0.88	0.10
7	102	779	0.10	0.78	0.13
8	133	646	0.13	0.65	0.21
9	93	553	0.09	0.55	0.17
10	66	487	0.07	0.49	0.14
11	32	455	0.03	0.46	0.07
12	22	433	0.02	0.43	0.05

4.0 Selecting Baselines

Often, it is the case that the hazard rate of a population follows an expected path. This can be illustrated by analyzing the example population from Section 3.0. First, a few assumptions need to be made to focus on structure. Assume the following:

1) Independent variables affecting event rates are steady over the time periods observed;
2) The population has a max age of 12 periods;
3) The population is the same age;
4) The population has similar characteristics.

These assumptions ensure the hazard rate experienced is not influenced by differences in population characteristics or a shift in conditions. Only the age of the population has an impact on their hazard rate. One could imagine taking a slice of a full population to focus only on their intrinsic properties. The shape of the hazard rate illustrates two characteristics, summarized in Table 2.0 on the next page.

Table 2.0: General characteristics in the illustrative dataset.

Characteristic	Description
Upfront Quality	There is low inherent risk when a population member is young. As time passes, quality becomes mixed and risk increases.
Structural Death	The end of a life becomes increasingly likely for natural reasons as the maximum possible age is reached.

A set of properties like this should be constructed before modelling a population. The resulting structure is referred to as a **baseline hazard** and can be drawn from intuition and visual analysis of past hazard rates. It is important to have an idea of what factors influence hazard rates so that visualizations of past events can be understood both in terms of a baseline hazard and external factors.

The baseline is best utilized when it does not need to be forecasted. As an example, age is always known. This means that the inherent properties captured by age can be easily integrated into a forecasting model, capturing a large amount of complex information.

5.0 Alternatives for Modelling

5.1 Proportional Hazards Model

Modelling the shape captured by a baseline by using multiple independent variables is not straight forward. Instead, a famous statistician David Cox suggested using the **proportional hazards model**, which separates the baseline from independent variables [4]. In a mathematical sense, one can imagine modelling hazard rates as the product of a baseline and a function of independent variables. Cox established that a statistically appropriate function to perform regression on is the exponential function [5]:

$$h(T) = Baseline(T) * Exp(\boldsymbol{XB}),$$

where **X** is a vector of independent variables and **B** is a vector of their respective parameters. This notation will be used in the remainder of the report. The reason this model is called a proportional hazards model is since the ratio of the hazard rates of two individuals at any point in time is only proportional to their characteristics at that snapshot:

$$\frac{h^*(T)}{h(T)} = \left[\frac{Baseline(T)}{Baseline(T)}\right] * \left[\frac{Exp(X^*B)}{Exp(XB)}\right] = Exp([\boldsymbol{X}^* - \boldsymbol{X}]\boldsymbol{B})$$

Estimating the baseline is not necessary when comparing alternatives. However, a model can be made **parametric** if a function is specified for the baseline hazard. Furthermore, by taking the logarithm of each side of the above equation and incrementing a single variable, one finds a simple relation:

$$\ln h^*(T) - \ln h(T) = [x_k^* - x_k]b_k$$

Increasing any independent variable will have a linear impact on the change in the logarithm of the hazard rate. This is intuitive: the effect of increasing a single variable on the hazard rate can be significant upfront but will always taper off. Since the hazard rate is at most one, which signifies certainty, extremely slow growth is necessary after some reasonable point. This also means that multiple variables are more effective at reaching higher risk. As an example, it is intuitive that when an individual is already 300 pounds overweight, the difference in being 30 pounds more overweight is not as significant to the risk of heart disease as being a regular smoker, a separate variable.

Due to the separation of a baseline and the independent variables, parameters can be estimated using partial maximization techniques which ignore the baseline function [6]. This technique is called **Cox regression**.

5.2 Logit Model

In a discrete time setting, the hazard rate model's **maximum-likelihood estimation** is equivalent to that of another model called the logit model [7]. This means that finding the best parameters based on available data is the same problem for these models. The technique for estimating these parameters is called **logistic regression**, which is concerned with predicting the probability something is true. As an example, one may wish to model the probability that a patient will have high blood pressure given a set of attributes such as exercise and dietary habits. Many special options are implemented for logistic regression in popular statistical packages since it is such a widely applicable and well-studied technique. Cox regression is developed but does not share the same set of options in popular tools such as SAS [8, 9].

The dependent variable of a logistic regression is generated from a Bernoulli trial, taking on only two possible values instead of a continuous stream of values like the height of a person. A variable which responds this way is often called a **binary** response variable since it has two outcomes. It is intuitive that the event indicator variable, defined in Section 3.0, can be modelled as a Bernoulli trial due to the assumed independence between time periods.

When trying to model a discrete set of outcomes linearly, the residuals do not look normally distributed like one expects in a classic linear regression setting [10]. To tackle this, binary outcomes are converted into

continuous ones. This is accomplished by performing transformations. First, take the odds of an event happening:

$$Odds(Event) = \frac{Probability(Event)}{1 - Probability(Event)}$$

It is easy to see that, since probabilities range from zero to one, the odds can take on values from zero to an unbounded positive number. By taking the logarithm of the odds, this range can be extended to all real values, making the final result more appropriate for linear regression. This transformation is called the **logit** transformation, from which the logit model gets its name. Such a transformation is called the **link function** in a regression. A linear model with a link belongs to the class of **generalized linear models**.

The link function must be the same for hazard rate modelling since their estimation is mathematically the same problem [7]. Therefore, the logit of the hazard rate can be written in terms of a linear combination of independent variables:

$$\ln\left[\frac{h(T)}{1 - h(T)}\right] = \boldsymbol{XB}$$

To continue based on the notion of capturing important structural information using baselines, one can introduce a function which is separate from the independent variables.

$$\ln\left[\frac{h(T)}{1 - h(T)}\right] = f(T) + \boldsymbol{XB}$$

By incrementing any independent variable, this form yields a very similar result to the proportional hazards model, now in terms of the odds of the hazard rate:

$$\ln\left[\frac{h^*(T)}{1 - h^*(T)}\right] - \ln\left[\frac{h(T)}{1 - h(T)}\right] = [x_k^* - x_k]b_k$$

5.3 Multiple Baselines

The logit model allows for greater flexibility in defining the baseline hazard function. In particular, this discrete time approach allows for the definition of multiple baselines in a straightforward manner:

$$\ln\left[\frac{h(T)}{1 - h(T)}\right] = f(T) + g(S) + \boldsymbol{XB}$$

5.4 Interactions

When a model is developed, it may be necessary to model an **interaction** between the baseline and non-baseline variables. An interaction is used when it is hypothesized that the value of one variable affects the impact of another variable. For example, lack of change can be a good indicator for some variables. However, this is only true if enough time has passed to actually create the change.

Interactions are often introduced when a generalized linear model does not produce random errors. One method to judge this is to visually analyze the distribution of residuals across a baseline variable. If a clear trend exists in the errors, interactions may be necessary. Interactions of baselines and independent variables are not possible when performing Cox regression.

6.0 Alternative Comparison

The analysis from the previous section is summarized in Table 3.0 below, which compares the logit model to the proportional hazards model.

Table 3.0: A comparison of two primary survival analysis techniques.

	Simplicity	**Statistical Options**	**Flexibility**
Proportional Hazards	Medium-High	Few	Low
Logit	High	Many	High

Modelling hazard rates of individuals of a population is generally considered complex. The proportional hazards model is the simplest out of the two models since it strictly separates the impact of a baseline from the effects of independent variables. The baseline does not need to be specified, which makes modelling more focused on the task of finding strong independent variables. Developing a proportional hazards model first can serve as a good prototyping exercise for identifying a candidate set of variables for a logit model. Computational efficiency has been shown to be similar across the two types of models [11].

In practice, developers do not have expertise to build and test specific statistical features for a model. Therefore, the availability of statistical options for a type of model is of high importance. Logistic regression has received far more attention from researchers and modellers. In the statistical software SAS, for example, the functionality of Generalized Estimating Equations, which are discussed in Section 7.3, is only available for the logit model [12]. Finally, the structural flexibility of using a logit link is incredible: multiple baselines can be implemented and can be interacted with independent variables. This makes the logit model the better alternative out of the two.

7.0 Evaluation of Assumptions

One of the goals of developing a model is to simplify a problem by describing it in terms of a set of assumptions. In this section, the fundamental assumptions of a logit model are discussed. Many of these apply to other types of models, such as the proportional hazards model. A modeller should constantly refer to these assumptions and attempt to challenge them when designing and implementing any new feature in a hazard rate model.

7.1 Linearity

At the heart of any generalized linear model is the assumption that independent variables are linearly related to the response variable. This assumption must be challenged, as improper model design has major consequences, especially in an engineering context. One test is plotting a scatter plot of residuals across buckets of an independent variable used in a model. A linear regression expects the residuals to be uniform and random. Errors therefore should not be a function of any independent variable.

7.2 Variable Selection

Correlation does not imply causation. Any variable selected for modelling an engineering problem should be intuitive. Otherwise, one risks **overfitting** a dataset by finding a random set of variables that happen to work in the past. Additionally, independent variables selected in a final model should not be functions of one another. The easiest way to test this is plotting buckets of one variable against buckets of the other. Accidentally choosing variables that are dependent on each other can be limited by calculating the variance inflation factor of each variable in a candidate model. This is discussed in Section 10.0 which deals with measuring model performance.

One should also make efforts to minimize the number of parameters in a model. Measures, such as the **Akaike information criterion**, can be used to judge the relative quality of candidate models based on their maximum likelihood and how many variables they incorporate [14].

7.3 Independence

The independence between any two points in a population member's life is a fundamental assumption when modelling hazard rates. However, the correlation from one time period to the next for a given member is likely not zero. **Generalized Estimating Equations** (GEE) can be used to take into account correlations in a dataset to improve the accuracy of a model. By specifying subjects and a within-subject

dimension, such as a unique identifier for a population member and the age of that member, a **correlation structure** is estimated. In this case, the GEE procedure works by assuming that errors within a member are somehow correlated across age, but between members are uncorrelated [13].

A recommended exercise is to first estimate a strong candidate model using logistic regression and then to perform estimation using a GEE with a logit link on the same set of variables. The results of each model can be compared by creating confidence intervals for each variable:

$$1 - \alpha/2 \ = P\left(\frac{\theta_{i,logistic} - \theta_{i,GEE}}{standard\ error_{i,GEE}} \in \left[-z_{\alpha/2} \ \ z_{\alpha/2}\right]\right)$$

The left-hand side is a level of confidence and is typically taken to be 95% or 99%. The parameter estimates in the numerator are outputs of the regressions. The assumption is that the more accurate GEE procedure produces a true estimate. On the denominator, the standard deviation or **standard error** of the GEE model acts as a normalizing factor, since it is assumed that the GEE model produces the population mean. This is a two-sided test, but intuition can be used to hypothesize whether correlations would work in favor of a parameter estimate or not.

8.0 Assessing Data Quality

There are several data quality considerations that must be made when modelling a dataset. Recommended data quality investigations are included in this section. Often data will come from several different sources which may contradict each other. It is important to choose the best data available, which demands for careful data investigations and iterative discussions with the parties producing the data. A modeller should have an intuition of the processes affecting the population being analyzed.

8.1 Unexpected Values

It is important to understand the meaning, if any, behind an unusual or impossible data point. A variable can be inspected by plotting the distribution of values, either with histograms for continuous variables or by listing frequencies for discrete variables. These charts can be inspected for outliers that need to be understood and questioned. Additionally, one should constantly question the reason for spikes in trends. Intuition and other data sources can be used to judge whether such a change is possible to identify data quality issues.

8.2 Missing Values

It is important to ensure that missing values are kept to a minimum. In SAS and many languages, logistic regressions omit any row in a dataset that contains a missing value in any variable [15]. This means that having a small amount of missing values in random places across multiple variables can significantly reduce the size of data to model. The problem of filling in these holes is called **imputation**. It is good practice to keep track of whether a variable has been imputed by creating a corresponding binary variable, also called a **dummy variable**.

8.3 Censoring and Truncation

Left censoring happens when an event has occurred prior to the beginning observation date in the modelling set [16]. Losing this data can be random or informative. To test whether censoring had a high impact, a candidate model can be tested against a dataset with a different range of dates.

Right censoring is the by-product of cutting a dataset off, something that is unavoidable [16]. It is usually due to information about the future not existing. If a population member does not experience an event within the time period modelled, it is said to be Type 1 right censored. If, for some data quality reason, population members drop off the dataset without events, it is important to investigate whether this is truly random or not. This is since underlying equations used to estimate a hazard model do not consider data which drops off a dataset. A large amount of unexplainable censoring could result in a miscalculated model [16].

Left truncation is the idea that an older population has already been at risk before entering a study [16]. This can be accounted for by utilizing age as a variable, which is already an essential part of the hazard rate model discussed. Therefore, before removing age as a variable, one should assess whether left truncation is common or not. **Right truncation** happens when observations are included only if their time-to-event is less than a certain period [16]. In the framework discussed in this report, there is no right truncation.

A major assumption of the logit model is that censoring and truncation is independent to events [16]. Furthermore, where a dataset starts and ends will impact the final model and needs to be assessed for appropriate randomness.

9.0 Variable Treatment

9.1 Clusters

Making detailed choices among hundreds of variables can be extremely time consuming. A final model should have variables which are as independent of each other as possible, as discussed in Section 7.1. Since this is the case, it is often useful to develop **variable clusters**. The correlation within a cluster of variables should be high, but the correlation among clusters should be low. How high and low is different from case to case. There are different algorithms which attempt to find clusters of variables in a dataset. Manual clustering can be an effective starting point and is often mixed with algorithmic techniques. Intuition and modeller expertise should be used to adjust the final output of any statistical method applied.

9.2 Transformations

Variables should have an almost linear impact on the logit of the hazard rate. A population's hazard rate can be flattened across buckets, such as percentiles, of a variable to assess if this assumption is true. It is a good idea to also plot frequencies to ensure that any bucket deviating from a linear trend is comprised of a significant proportion of a population. If a trend is nonlinear, transformations must be applied to the independent variable. When transforming variables, it is important to judge whether the transformation is intuitive. Some common variable transformations are summarized in Table 4.0 below.

Table 4.0: A summary of common variable transformations.

Transform	**Purpose**
Piecewise Linear	A trend may look like a combination of linear relations with different slopes in different sections. To model this, a variable must be split into multiple variables corresponding to each section. The endpoints, or **knot points**, of these sections can be identified algorithmically or visually, based on plotting hazard rates against the variable. This is particularly useful for transforming baseline variables, which are designed to be non-linear since they capture different behaving points in time.
Cap / Floor	A trend may be mostly linear but flat on its ends. Caps can be applied to set a maximum value a variable can take on to eliminate a flat end. Floors can be applied to set a minimum to eliminate a flat start. Caps and floors can also be used to bundle outliers that are not removed.
Logarithm	A trend may have a rapidly decreasing slope which can be modelled as logarithmic growth. The set of transformations that can be applied is ln(a+x) for any value of a.

10.0 Model Selection

Selecting the best variables for a model is a complex process often faced with time limitations. Often, the set of given variables must be reduced to a reasonable size. Regressions can be run for a single variable at a time, with baselines and interactions included to help capture structural information. The outputs of these regressions can be used to shortlist variables for the final modelling stage. Some important measures for shortlisting a variable are its parameter significance [17], the sign of its parameter estimate, and the GINI statistic for its model [18].

The **significance**, or p-value, is related to the maximum confidence level a statistical procedure has with its parameter estimates. Typically speaking, the significance should be 0.05 or less. The sign of an estimate can be compared against intuition. Finally, the **GINI statistic** is unique to a given model and measures how much better a model is compared to a base model that randomly assigns events [18].

After a shortlist of variables is made, multi-factor models can be generated automatically by employing custom branch and bound methods. These algorithms only generate models if they meet certain guidelines, such as a low p-value in an entering variable [19]. Another alternative is outputting all possible models, which is typically very time costly. The outputs of this process can then be filtered based on custom rules, such as having no more than two variables from a given cluster.

Once candidate models are created, either automatically or through manual efforts, the **variance inflation factor** (VIF) of each variable in the model becomes an additional factor to check [20] along with significances, sign intuition and the overall model GINI. The VIF gets its name since correlated variables tend to shift errors in the same direction, increasing the overall variance of a model.

11.0 Conclusions

In a discrete time setting, the hazard rate must be modelled instead of an event rate. A comparison of hazard rate modelling techniques was performed. The proportional hazards model is the simplest approach but does not offer the flexibility needed in a final model. In a logit model, interactions can be made, multiple baselines can be defined, and a hidden correlation structure can be detected.

A modeller should perform extensive data investigations and challenge assumptions when forecasting in an engineering context. The following specific conclusions were made about hazard rate modelling:

1) A baseline parameter should be deterministic;

2) Censoring and truncation must be analyzed for appropriate randomness;

3) Linearity of baselines must be enforced by performing variable transformations.

Understanding these conclusions is critical to the successful modelling of hazard rates.

12.0 Recommendations

Recommendations developed throughout the report focus on optimizing the modelling process:

1) Combine events that are affected similarly by variables;

2) Plot residuals in prototype models to search for potential interactions;

3) Create variable clusters to mitigate variance inflation factors;

4) Shortlist variables using single factor regressions with baselines and interactions.

To allow readers to validate previously existing models, a major recommendation is developing a Generalized Estimating Equation model based on variables from a logit model. This will allow for the independence assumption to be relaxed and tested through confidence intervals.

13.0 References

[1] S. Hinchliffe. (2012). Competing Risks [Online].
Available: http://www2.le.ac.uk/departments/health-sciences/research/biostats/youngsurv/pdf/SHinchliffe.pdf

[2] L. Green. (2006). Queueing Analysis in Healthcare [Online].
Available: https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/4386/chapter%2011%20QueueingAnalysis.pdf

[**3**] R, MacKay. (2006). Survivor and Hazard Functions [Online].
Available: http://people.stat.sfu.ca/~raltman/stat402/402L32.pdf

[4] F. Schoonjans. (2015, August 14). Cox proportional-hazards regression [Online].
Available: https://www.medcalc.org/manual/cox_proportional_hazards.php

[5] G. Rodríguez. (2015). Survival Models [Online].
Available: http://data.princeton.edu/wws509/notes/c7.pdf

[6] D. Zhang. (2005). Modeling Survival Data with Cox Regression Models [Online].
Available: http://www4.stat.ncsu.edu/~dzhang2/st745/chap6.pdf

[7] S. Jenkins. (2015). Discrete-time survival analysis [Online].
Available: http://www.stata.com/manuals13/stdiscrete.pdf

[8] N. Fultz. (2012, November 19). Logistic (and Categorical) Regression [Online].
Available: http://www.ats.ucla.edu/stat/sas/topics/logistic.htm

[9] SAS. (2014, November 6). PHREG Procedure [Online].
Available: http://support.sas.com/rnd/app/stat/procedures/phreg.html

[10] B. Nau. (2014, July 24). Linear regression models [Online].
Available: http://people.duke.edu/~rnau/testing.htm

[11] I. Annesi. (1989, December 8). [ABSTRACT] Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies [Online].
Available: http://www.ncbi.nlm.nih.gov/pubmed/2616941

[12] SAS. (2010, May 17). Generalized Estimating Equations [Online].
Available:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_genmod_sect008.htm

[13] A. Hanley. (2000, January 7). Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation [Online].
Available: http://aje.oxfordjournals.org/content/157/4/364.full

[14] M. Mazerolle. (2004, June). APPENDIX 1: Making sense out of Akaike's Information Criterion (AIC) [Online].
Available: http://theses.ulaval.ca/archimede/fichiers/21842/apa.html

[15] SAS. (2009, September 29). Missing Values [Online].
Available:
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect021.htm

[16] M. Lunn. (2007, January 18). Definitions and Censoring [Online].
Available: http://www.stats.ox.ac.uk/~mlunn/lecturenotes1.pdf

[17] I. Ruczinski. (2015). Variable Selection [Online].
Available: http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf

[18] H. Greene. (2010, January 13). Assessing model performance: The Gini statistic and its standard error [Online].
Available: http://www.palgrave-journals.com/dbm/journal/v17/n1/full/dbm20102a.html

[19] SAS. (2009, September 29). Stepwise Logistic Regression and Predicted Values [Online].
Available:
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect052.htm

[20] Penn State University. (2015). Detecting Multicollinearity Using Variance Inflation Factors [Online].
Available: https://onlinecourses.science.psu.edu/stat501/node/347