

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS**

**ΣΧΟΛΗ
ΕΠΙΣΤΗΜΩΝ &
ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ**
SCHOOL OF
INFORMATION
SCIENCES &
TECHNOLOGY

**ΜΕΤΑΠΤΥΧΙΑΚΟ
ΣΤΑΤΙΣΤΙΚΗ**
MSc IN
STATISTICS

Mixture Cure Models for credit scoring

By

Athina Papageorgiou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece

February 2025



**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS**

**ΣΧΟΛΗ
ΕΠΙΣΤΗΜΩΝ &
ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ**
SCHOOL OF
INFORMATION
SCIENCES &
TECHNOLOGY

**ΜΕΤΑΠΤΥΧΙΑΚΟ
ΣΤΑΤΙΣΤΙΚΗ**
MSc IN
STATISTICS

Μικτά Μοντέλα Θεραπείας για την Αξιολόγηση Πιστοληπτικής Ικανότητας

Αθηνά Παπαγεωργίου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Φεβρουάριος 2025



DEDICATION

*To my beloved grandparents,
who are no longer here but always watching over me*



ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks to my supervisor, Professor Dimitris Karlis, for his unwavering support and guidance throughout my academic years. His remarks have always been to the point during my thesis, and I am truly grateful for the time he has dedicated to helping me throughout this journey. It is an honor to have had his support and guidance through different stages over the years. When things became unclear, his kind words and encouragement were invaluable. I consider him an inspiration and a mentor who has shaped me as a statistician.

To my parents, who have supported all my decisions and always believed in me. A big part of who I am today is because of the way they raised me, and I am forever grateful for their unconditional love. To my brother and my whole family, who have been a true source of strength when things got difficult, for lighting my way and reminding me of my worth and capabilities - I am deeply thankful.

I would also like to express my gratitude to my friends who have stood by my side throughout my academic journey, giving me the strength to keep going no matter where I am or what my next step may be.

Finally, I would like to express my love and admiration for Thanasis, who has played a significant role in my growth and has always encouraged me to push my limits further. He himself is proof of the saying: “He who has a why to live can bear almost any how.”





ABSTRACT

Athina Papageorgiou

Mixture Cure Models for credit scoring purposes

February 2025

Credit scoring has been a crucial concern for banks aiming to identify creditworthy borrowers and minimize potential financial losses caused by default risk. However, the focus has now shifted from merely determining whether a borrower is eligible for a loan to estimating the probabilities of default at different time points.

To address this challenge, this thesis proposes the use of survival analysis, a powerful statistical tool primarily used in biostatistics to model the time until an event occurs. In this context, survival analysis is applied to model the time until a borrower defaults. Notably, it is reasonable to assume the existence of a subpopulation of borrowers who will never default, often referred to as “cured”. To account for this, mixture cure models are utilized which combine a component to estimate the cure rate and a survival component to model the time-to-event for susceptible individuals.

In this study, we propose two models: a classical Weibull model and a mixture cure model. The mixture cure model consists of a logistic component for the cure rate and a Weibull component associated for the survival probabilities of susceptible individuals. Both models are implemented on a dataset of mortgage loans with diverse characteristics. The models are evaluated based on their goodness of fit using the Kaplan-Meier estimator and residuals graphs. Furthermore, their predictive performance is assessed using a train-test split approach. The results indicate that the mixture cure model not only provides a better fit to the data but also delivers more accurate predictions compared to the classical Weibull model.





ΠΕΡΙΛΗΨΗ

Αθηνά Παπαγεωργίου

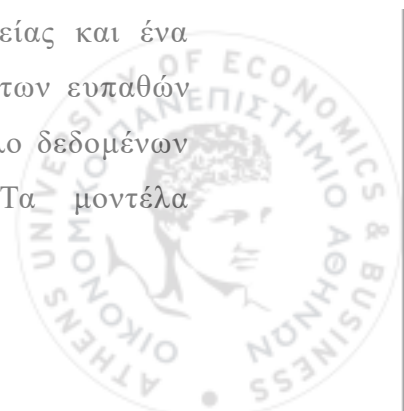
Μικτά Μοντέλα Θεραπείας για την Αξιολόγηση Πιστοληπτικής Ικανότητας

Φεβρουάριος 2025

Η αξιολόγηση πιστοληπτικής ικανότητας αποτελεί ένα κρίσιμο ζήτημα για τις τράπεζες, οι οποίες επιδιώκουν να αναγνωρίζουν δανειολήπτες με καλή πιστοληπτική ικανότητα και να ελαχιστοποιούν τις πιθανές οικονομικές απώλειες που προκαλούνται από τον κίνδυνο αθέτησης. Ωστόσο, η προσοχή έχει μετατοπιστεί από τον απλό καθορισμό της καταλληλότητας ενός δανειολήπτη για δάνειο, στην εκτίμηση των πιθανοτήτων αθέτησης σε διαφορετικά χρονικά σημεία.

Για την αντιμετώπιση αυτής της πρόκλησης, αυτή η διπλωματική εργασία προτείνει τη χρήση της ανάλυσης επιβίωσης, ένα ισχυρό στατιστικό εργαλείο που χρησιμοποιείται πρωτίστως στη βιοστατιστική για τη μοντελοποίηση του χρόνου μέχρι την εμφάνιση ενός γεγονότος. Σε αυτό το πλαίσιο, η ανάλυση επιβίωσης εφαρμόζεται για τη μοντελοποίηση του χρόνου μέχρι την αθέτηση ενός δανειολήπτη. Είναι σημαντικό να σημειωθεί ότι υπάρχει ένας υποπληθυσμός δανειοληπτών που δεν θα αθετήσουν ποτέ, γνωστός και ως «θεραπευμένος» (cured). Για να ληφθεί υπόψη αυτή η περίπτωση, χρησιμοποιούνται μοντέλα μικτής θεραπείας (mixture cure models), τα οποία συνδυάζουν ένα μέρος για την εκτίμηση του ποσοστού θεραπείας (cure rate) και ένα μέρος επιβίωσης για τη μοντελοποίηση του χρόνου μέχρι το γεγονός για τους ευπαθείς δανειολήπτες.

Σε αυτή τη μελέτη, προτείνουμε δύο μοντέλα: ένα κλασικό μοντέλο Weibull και ένα μικτό μοντέλο θεραπείας. Το μικτό μοντέλο θεραπείας αποτελείται από ένα μέρος λογιστικής παλινδρόμησης για το ποσοστό θεραπείας και ένα στοιχείο με Weibull κατανομή για τις πιθανότητες επιβίωσης των ευπαθών δανειοληπτών. Και τα δύο μοντέλα εφαρμόζονται σε ένα σύνολο δεδομένων στεγαστικών δανείων με διαφορετικά χαρακτηριστικά. Τα μοντέλα



αξιολογούνται ως προς της προσαρμοστικότητας στα δεδομένα χρησιμοποιώντας τον εκτιμητή Kaplan-Meier και γραφήματα καταλοίπων. Επιπλέον, η προβλεπτική τους ικανότητα αξιολογείται με τη χρήση διχοτόμησης των δεδομένων σε training και test datasets. Τα αποτελέσματα δείχνουν ότι το μικτό μοντέλο θεραπείας όχι μόνο έχει καλύτερη προσαρμογή στα δεδομένα, αλλά παρέχει και πιο ακριβείς προβλέψεις σε σύγκριση με το κλασικό μοντέλο Weibull.





Contents

Introduction	1
Methodology	5
2.1 Survival Analysis – definitions and notations	5
2.2 Weibull survival model	10
2.3 Mixture Cure Model.....	19
Real life problem and Dataset description.....	31
3.1 Dataset Overview and Preparation.....	31
3.2 Numerical Variable Analysis	32
3.3 Categorical Variable Analysis and Cured Population Assessment.....	36
Modeling and Results	41
4.1 Results of the Classical Weibull Survival Model.....	41
4.2 Results of the Mixture Cure Model.....	43
4.3 Model Evaluation and Comparison.....	46
4.4 Predictive Performance on Test Dataset.....	50
4.5 Cure Probability Analysis.....	52
Conclusions, Discussions and Future Work	57
5.1 Summary of Findings	57
5.2 Discussion and Future Work	59





List of Tables

3.1 Dataset variables overview	32
3.2 Descriptive statistics of numerical variables	33
4.1 Results of the Weibull Survival Model.....	43
4.2 Results of the Mixture Cure Model.....	45





List of Figures

Figure 3.1 Histogram of FICO Scores	34
Figure 3.2 Histogram of time until default or censoring (in months).....	35
Figure 3.3 Barplot of Occupancy status (I: Investment property, P: Primary Residence, S: Second Home).....	36
Figure 3.4 Barplot of Property Type (CO: Condominium, PU: Planned Unit Developments, SF: Single-family Homes, Other: Cooperative shares and manufactured housing.....	37
Figure 3.5 Barplot of the Loan origination source (R: Lender or its affiliates and no involvement of a third party, T: Broker or Correspondent, or a third-party involvement not specified.....	38
Figure 3.6 Barplot of Default status of borrowers (0: censored, 1: uncensored)	39
Figure 4.1 Kaplan-Meier curves with Estimated Survival curves from Weibull model (left) and Mixture Cure Model (right) for the whole dataset	47
Figure 4.2 Cox-Snell Residuals for Weibull Model (left) and Mixture Cure Model (right) vs unit exponential line (red)	48
Figure 4.3 Martingale Residuals with explanatory variables from the Weibull Model	49
Figure 4.4 Martingale Residuals with explanatory variables from the Mixture Cure Model.....	50
Figure 4.5 Kaplan-Meier curves with Predicted Survival curves from Weibull model (left) and Mixture Cure Model (right) for the test dataset.....	51
Figure 4.6 Probability of being cured for 6 random borrowers from the test dataset at different time points (in months)	53
Figure 4.7 Average probability of cure for the test data at different time points (months).....	54
Figure 4.8 Cure rate for the test data	55





Chapter 1

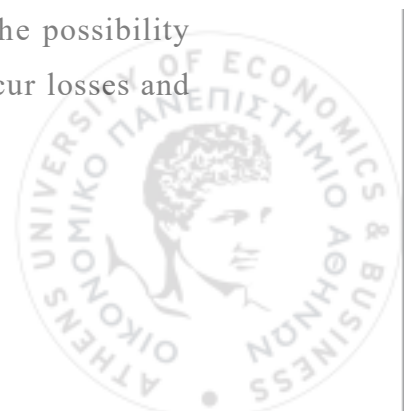
Introduction

Borrowing and lending are fundamental components of modern financial systems, providing businesses and individuals access to credit for various purposes. Loans are a crucial tool for firms to launch start-ups, implement new ideas and invest in growth. Individuals seek financial support from banks to obtain capital for purchasing cars, homes, or funding specific needs such as education. Among the different types of loans, mortgage loans are the most common, mainly used for home purchases.

On the other hand, banks issue loans not only to support and facilitate economic activities but also because they benefit financially. When banks provide businesses and individuals with credit, they practically agree to receive back the original money they lent along with an additional amount, mainly generated through interest rates. Interest rates are usually expressed as a percentage of the initial loan amount and serve as a cost for borrowing money. The cumulative interest payments received from borrowers lead to long-term revenue growth for banks.

In addition to interest income, banks also make profits through other fees such as origination and servicing costs. Another significant source of revenue comes from loan securitization where banks sell loans to government-backed entities or private investors. These transactions allow them to generate immediate profits rather than waiting for future repayments from borrowers.

Despite the benefits of issuing loans, there are many lurking dangers for banking institutions in this activity. Among the risks deriving from interest rate changes and liquidity limitations, the most significant challenge for banks is credit risk, also known as default risk. This is associated with the possibility that borrowers may fail to repay their loans, causing banks to incur losses and restrict their lending capacity.

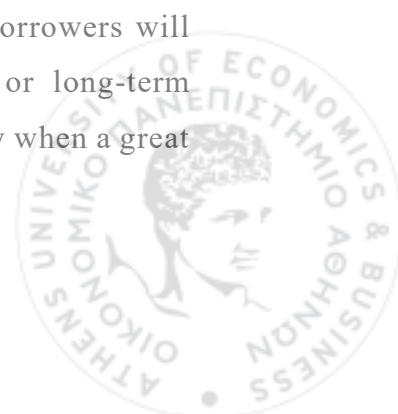


Today, concerns over credit risk have intensified due to financial instability caused by increasing political and geopolitical uncertainty, stricter regulatory compliance rules and credit standards for housing loans. Rising default risks, along with these challenges, have made it essential for banks to develop structured procedures to quantify the default risk. Accurate risk models assist banks to decide whether to approve a loan and determine suitable interest rates to limit potential losses.

Traditionally, to evaluate a borrower's creditworthiness, banks based their assessment on credit scores. They used real-time and historical data, such as income, employment status, monthly expenditures and overall credit history to identify reliable individuals. However, after the Basel II Accord (Basel Committee on Banking Supervision, 2004), banks are no longer interested only in distinguishing between trusted and non-trusted borrowers. Instead, they now focus on predicting when a borrower is more likely to default on a loan. As a result, it is important to quantify risk by estimating the probability of default at different time points.

A powerful method for this is survival analysis, a statistical tool mainly used in medical studies to model the time until an event occurs. The main concept of survival analysis is the survival function, defined as $S(t) = P(T > t)$, which denotes the probability of an individual surviving beyond a specific time point t . In the case of credit risk, if we set loan default as the event of interest, survival analysis proves to be a valuable tool for modeling the time until a borrower defaults on a loan. Through the survival function, banks can predict the probability that individuals will meet their financial obligations of the loan at any time point.

However, classical survival analysis assumes that after a very long time, all individuals participating in a study will eventually experience the event. In contrast, Farewell (1982) proposed that a subpopulation may exist what will never experience the event. In our case, this means that some borrowers will never default on their loans and can be considered "cured" or long-term survivors. Mixture cure models account for these cases, especially when a great



number of individuals are right-censored because they continue repaying their loans on a long-term horizon.

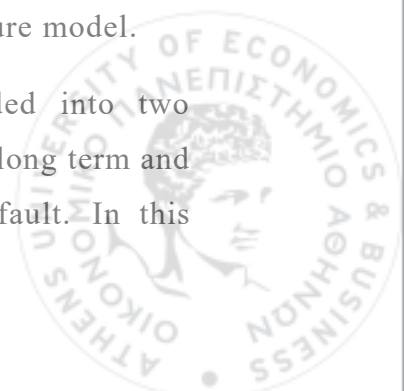
Mixture cure models incorporate this idea in survival analysis by assuming that there are two subpopulations: one that will eventually experience the event and another that will never default, making them non-susceptible. The model consists of two parts. The first, known as the incidence component, is responsible for predicting whether an individual will default or not. The most common model for the first part is logistic regression, which assigns individuals to one of the two groups, usually based on specific variables. The second component, called the latency model, is a survival model that estimates the probability of survival, given that the individual belongs to the susceptible group.

Farewell (1982) implemented a mixture cure model with logistic regression for the incidence part and a Weibull distribution for the latency. Since a specific distribution is defined for the latency part, this model falls within the parametric family. However, semi- and non-parametric models are also used, with the Cox proportional hazards model being one of the most popular alternatives (Peng and Dear, 2000; Sy and Taylor, 2000).

Another key advantage of mixture cure models is their flexibility in including different explanatory variables in each component. This means that one set of characteristics can be used for the classification of individuals into the cured or uncured group, while another set can influence the survival probability and time to default.

In the present thesis, we implement and evaluate the performance of survival analysis models for default risk, using a large dataset of mortgage loans. Estimating the probability of loan default is crucial for banks, as it allows them to take appropriate measures and prevent potential losses. Given this need for default risk, we assess the goodness of fit of a parametric survival model with Weibull distribution and compare its performance to a mixture cure model.

The mixture cure model assumes that borrowers are divided into two subpopulations: those who will never default on their loan in the long term and those who are susceptible and will eventually experience default. In this



context, the incidence part is modeled through a logistic regression, while the latency part, employs a Weibull survival model to estimate the probability of default.

The goal of this thesis is to apply survival analysis, which is mainly used in medical research, to financial data of mortgage loans. The mixture cure model is compared to a simple survival model, with the expectation that it will provide more accurate estimations when two distinct groups of borrowers exist. The results and the performance of the two models are evaluated through the Kaplan-Meier estimator, survival curves, and residuals plots. Additionally, their predictive accuracy is evaluated using a train-test split approach.

The remainder of the thesis is organized as follows. In Chapter 2, we present the basic concepts of survival analysis as a statistical method. Furthermore, we describe the two survival models, simple and mixture cure, along with their estimation and evaluation methods. In Chapter 3, we present the dataset with mortgage loans and its characteristics. After that, the results of the models under investigation are discussed in Chapter 4. Finally, Chapter 5 summarizes conclusions and initiate discussions about further research.



Chapter 2

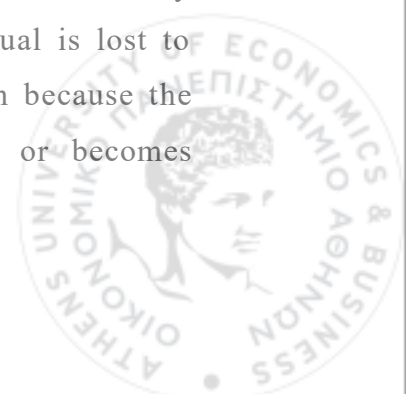
Methodology

2.1 Survival Analysis – definitions and notations

Survival analysis is a powerful statistical tool, formally introduced by Kaplan and Meier (1958). It consists of a pool of statistical methods whose outcome of interest is the time until an event occurs. By event, also expressed as failure, we could mean death, disease, recovery or any other specific event aligned with our objectives. In this study, for example, the event is the default of the loan by a borrower. Note that, despite the term ‘failure’, an event can also be positive, such as in the case of recovery. Time until the occurrence of an event can be measured in days, weeks, months or years from the beginning of observation of a subject. It is also referred to as survival time.

One of the key characteristics of survival analysis which makes it stick out among other statistical analyses is censoring (Cox, 1959; Kaplan and Meier, 1958). Censoring is the problem that occurs when we do not know exactly the time when the event occurs, but we do have some information about it. Consider the following simple example in the medical field; we have patients in a clinic, and we observe them until they recover. If the study ends and the event has not yet occurred, i.e., recovery, this patient is considered censored. While we know that their survival study is at least as long as the period of their follow-up, we will not know the complete survival time if the patient recovers after the study period. On the other hand, if a patient recovers during the study period, we know the exact survival time and they are considered uncensored.

Censoring can mainly occur for three reasons. The first one, described above, is when the event does not occur during the study period. Another reason why we will not obtain the exact survival time is when an individual is lost to follow-up while the study is still in progress. This can happen because the subject may not want to contribute anymore to the study or becomes



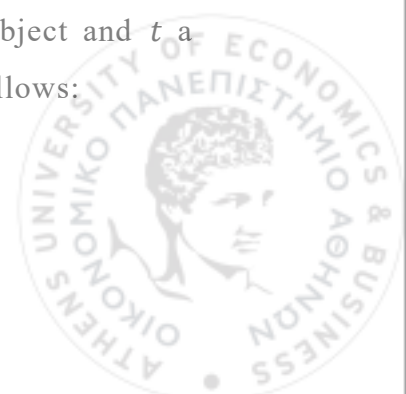
unreachable. It is also possible that a participant withdraws from the survey due to death or other causes.

Censorship of an individual can also be divided into three distinct categories. The most common category is right-censored and as its name indicates, the complete survival time becomes unavailable at the right side of the study period. More specifically, this is the case when an individual is lost, withdraws or the event does not occur during the follow-up. As in the medical example that we described above, in the right-censoring we know that the complete survival time is equal to or greater than the observed time. On the contrary, in the left-censored case the complete survival time is less than or equal to the time that the subject is observed. The third scenario of censorship is the interval-censored data. In this case, the true survival time is within a specific and known time interval. Note that this study is focused on and explores right-censored survival data.

The problem of censoring is unique in survival data. If we knew exactly the survival time of everyone, then the classical statistical methods should work efficiently for analyzing this data. However, since this is not the case, a new approach is needed to capture this special characteristic. Statistical methods in survival analysis employ incomplete information of censorship to model the time-to-event adequately.

A common trait between survival and classical models, though, is the inclusion of explanatory variables. Like in a multiple regression model, if there is evidence that there are covariates who influence the outcome, they can be incorporated in the survival model too. For example, if we study the time until death occurs, we might include variables such as age, treatment, and other characteristics of the patients which potentially influence their survival times.

After presenting key characteristics of survival analysis, we will now present the core functions which describe the distribution of survival times. Let's denote as T the random variable for the survival time of a subject and t a specific value of it. Also, let δ be an indicator of censoring as follows:



$$\delta = \begin{cases} 1, & \text{if the event is observed (uncensored)} \\ 0, & \text{if the observation is censored} \end{cases}$$

The survival function, $S(t)$:

$$S(t) = P(T > t)$$

describes the probability that a person survives longer than a specific time point t . Through the survival function and by changing the values of t , we can obtain the survival probabilities for different moments. It is clear that, the higher the value of $S(t)$, the higher the prospect of survival. Some characteristics of the survival function are the following:

- I. Since it denotes the probability of survival, it is bounded between 0 and 1
- II. For $t = 0$: $S(0) = 1$. At the start of the study, no event has occurred so the probability of surviving beyond that is equal to one.
- III. For $t = \infty$: $S(\infty) = 0$. Theoretically, if the time horizon is extended to infinity, we expect that everyone will eventually experience the event of interest. As a result, the probability of surviving after a very long time is equal to zero.

Taking into consideration the above properties, when we plot the probability of surviving for multiple time points t , we obtain a nonincreasing curve, called the survival curve. Theoretically, the survival curve is smooth and while the time increases, the probability of survival decreases until it reaches zero. However, in practice, the graph that we obtain is a step function because the estimated survival probability remains stable from one discrete time of an event to another.

The next key function in survival analysis is the hazard function which has the following formula:

$$H(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

where Δt is a small interval of time.



The interpretation of the hazard function is not as straight-forward as the survival function due to the limit component. In a simplified approach, hazard function denotes the momentary possibility per unit of time for the event to occur, given that the subject of the study has survived up to time point t . Note that, in contrast to the survival function, the hazard function is related to the failure of a subject.

Focusing on the nominator of the formula above, we consider a conditional probability. This represents the probability of an individual failing within a time interval $[t, t + \Delta t]$, given that they have survived up to time t . By dividing this conditional probability with a small-time interval Δt , we obtain a probability per unit of time, which is interpreted as a rate. Due to this reasoning, the hazard function is also called a conditional failure rate. Finally, by taking the limit of this fraction, we obtain the momentary potential for failing at time t per unit of time, given surviving until time t .

The hazard function has the following characteristics:

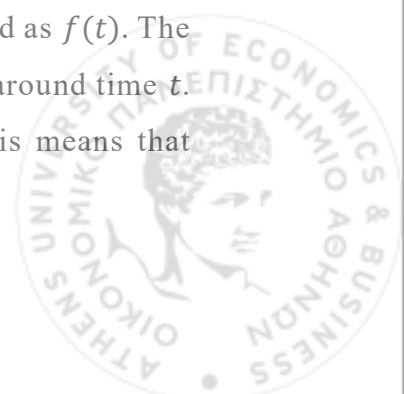
- I. As a rate, it can take values up to infinity
- II. It is always nonnegative: $h(t) \geq 0$
- III. It can change direction over time with no upper bound

Although survival and hazard functions provide different types of information in survival analysis, there is a structured relationship between them. This means that if we know one of the two functions, we can easily determine the other. More explicitly, they are connected through the following formulae:

$$S(t) = \exp \left[- \int_0^t h(u) du \right]$$

$$h(t) = - \left[\frac{dS(t)/dt}{S(t)} \right]$$

Survival and hazard functions are also related through a relationship which also involves the probability density function of survival times, denoted as $f(t)$. The latter represents the likelihood of the event of interest occurring around time t . For example, if $f(t)$ reaches its highest point at time $t = 5$, this means that



survival times are most likely to take place around 5 units of time. The relationship between all the above-mentioned functions is the following:

$$h(t) = \frac{f(t)}{S(t)}$$

Before moving to the last important feature of survival analysis, we also refer to the cumulative hazard function, $H(t)$, which allows us to compute the total accumulated risk of occurrence of the event by time t :

$$H(t) = \int_0^t h(u)du$$

Another key concept in survival analysis is the Kaplan-Meier (KM) method (Kaplan and Meier, 1958). It is a non-parametric technique used to estimate the survival probabilities for different times t , without assuming a specific distribution for the survival data. The formula for the KM estimator for the survival function is the following:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (2.1)$$

where t_i are the time points when events occur, d_i is the number of events at time t_i and n_i is the risk set in time t_i . The risk set represents the number of subjects who have survived at least to time t_i and are still at risk of experiencing the event. This set includes not only individuals for which the event has not occurred then but also those who are under observation and uncensored before t_i . It becomes clear that the KM estimator makes use of all the available information, including censorship, by means of the risk set.

If we plot the survival probabilities derived from the KM estimator against the ordered failure time, we get the Kaplan Meier survival curve. As mentioned earlier, the estimated survival curve, in practice, is a step function rather than a smooth curve. The KM estimator is a useful tool for evaluating the fit of a survival model, by comparing the KM curve and the one derived from the fitted model. This application will be explored in the following sections.



2.2 Weibull survival model

In this study, we will focus solely on the parametric family for modeling survival data. The most common and widely used distribution in parametric survival analysis is the Weibull distribution. Introduced by Swedish engineer and scientist Waloddi Weibull in 1937, it has gained popularity over the years, and it has been recognized as a fundamental tool for modeling time-to-event data. More details about the advantages and disadvantages of the use of a parametric model, and especially Weibull, are mentioned in the next section.

Let $T > 0$, be a random variable which expresses the survival time until an individual experiences the event of interest. Let t be a specific value of T . The survival function of the Weibull model with two parameters is given by:

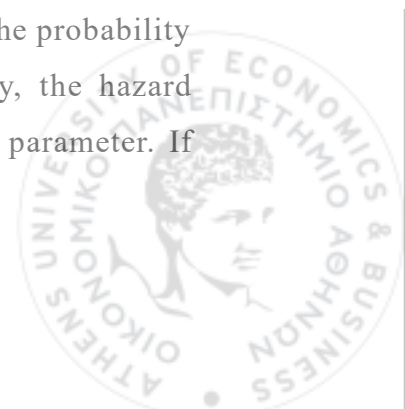
$$S(t) = \exp(-\lambda t^p)$$

while the hazard function is:

$$H(t) = \lambda p t^{p-1}$$

where $\lambda > 0$ and $p > 0$ are the scale and shape parameters respectively. In our working framework, we are interested in including covariates and exploring whether they have effect on the time until the event occurs. For this reason, λ will be later reparameterized with a regression form.

The shape parameter p of the Weibull distribution determines the shape of the hazard function. If $p > 1$, the hazard increases as time passes, while if $p < 1$, the hazard decreases. When $p = 1$, the Weibull distribution reduces to a special case, the exponential model, where the hazard is constant over time. Like the scale parameter λ , the shape parameter p can also be incorporated with regression coefficients. However, in this study, p will be considered constant across different values of variables. One key property of the Weibull model related to the shape parameter is that the hazard and survival functions do not change direction across time. This means that as time increases, the probability of surviving beyond time t decreases monotonically. Similarly, the hazard function form is also monotonic based on the value of the shape parameter. If



this is not the case for the behavior of the hazard function, a log-logistic model should be used instead, as it allows for nonmonotonic hazard behavior.

Another key characteristic of the Weibull model is the Probability Density Function (PDF) which can be easily derived from survival and hazard functions:

$$f(t) = h(t)S(t) = \lambda p t^{p-1} \times \exp(-\lambda t^p)$$

Similarly, the Cumulative Distribution Function (CDF) which represents the probability that the event of interest has occurred by a specific time t , is given by:

$$F(t) = 1 - S(t) = 1 - \exp(-\lambda t^p)$$

It is evident that both PDF and CDF are influenced by the values of the scale and shape parameters of the distribution.

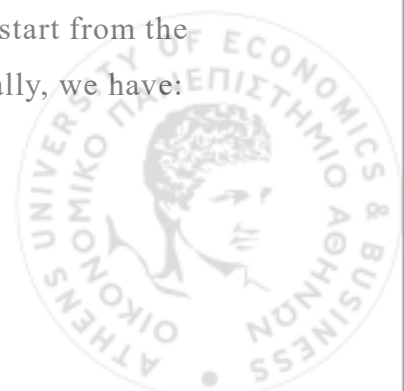
Popularity of the Weibull model in survival analysis lies to the fact that it can be written as a proportional hazards (PH) model and as an Accelerated Failure Time (ATF) model. These two assumptions describe the same model using different parameterizations. Proportional hazards models are related to the hazard and the multiplicative effect of covariates with respect to it. On the other hand, the Accelerated Failure Time assumption is that the impact of explanatory variables is again multiplicative but with respect to survival. More specifically, the latter can be expressed in terms of survival time as following:

$$S_2(t) = S_1(\gamma t)$$

where $S_2(t)$ and $S_1(t)$ are the survival functions of two different groups. The constant γ , called the acceleration factor, describes the expansion or contraction of survival between two groups. A unique property that the Weibull distribution holds is that if the AFT assumption holds then the PH assumption is also valid, and vice versa, given that the shape parameter is fixed.

To formulate the AFT parameterization in the Weibull model, we start from the survival function of the PH model and solve for t . More specifically, we have:

$$S(t) = \exp(-\lambda t^p)$$



$$\ln S(t) = -\lambda t^p$$

$$t^p = -\frac{\ln S(t)}{\lambda}$$

$$t = [-\ln S(t)]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

Returning to a previously discussed topic, we wish to include explanatory variables in the Weibull distribution and explore to what degree they may affect the time until a borrower defaults on their loan. Covariates can be integrated into both the PH and AFT parameterizations of the model, but in a different way. In greater detail, starting with the AFT parameterization previously written, we got the following relationship:

$$t = [-\ln S(t)]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

By letting $\frac{1}{\lambda^{1/p}} = \exp(-x'\alpha)$, we have:

$$t = [-\ln S(t)]^{1/p} \times \exp(-x'\alpha)$$

where x is a set of predictors and α is a vector of their corresponding coefficients, including the intercept. In the above relationship, x 's are integrated directly in the survival time within the AFT framework. Conversely, in the context of PH parameterization, the dependent variables are considered through the hazard function, in the following way:

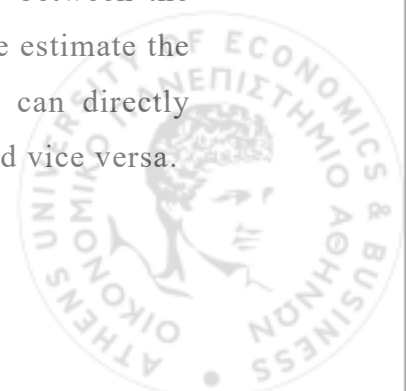
$$h(t) = \lambda p t^{p-1}$$

$$\lambda = \exp(x'\beta) .$$

The scale parameters in the two models are connected as follows:

$$\lambda_{AFT} = \lambda_{PH}^{-1/p}$$

Considering the above equation, we can derive the relationship between the coefficients in the PH model and the AFT model. This way, if we estimate the coefficients of the Weibull model in the AFT framework, we can directly compute the matching parameters of the PH parameterization, and vice versa.



In the AFT:

$$\begin{aligned}\lambda^{1/p} &= \exp(-x'\alpha) \\ 1/p \ln\lambda &= -\exp(x'\alpha) \\ \ln\lambda &= -p \times x'\alpha\end{aligned}$$

And in the PH:

$$\begin{aligned}\lambda &= \exp(x'\beta) \\ \ln\lambda &= x'\beta\end{aligned}\tag{2.2}$$

From the above, we obtain the following relationship:

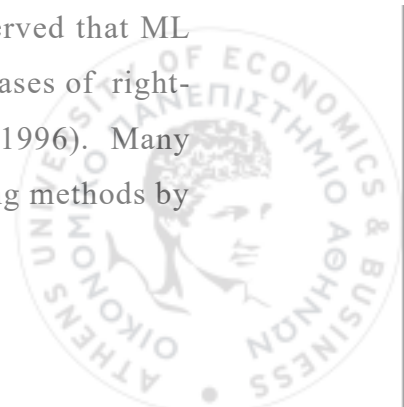
$$\beta_j = -\alpha_j p .$$

Up to this point, we have defined the two-parameter Weibull model in survival analysis and explored its two reparameterizations and the relationship between them. The next step is to estimate the necessary parameters. More precisely, the hazard function of the Weibull distribution needs the estimation of β 's and the shape parameter p , to be fully specified.

The most common estimation approach for the parameters of the Weibull distribution is the classic Maximum Likelihood (ML) method. According to this method, the optimal parameter values are those that maximize the log-likelihood of the data.

However, a significant difference between survival analysis and other statistical analysis is the presence of censoring, which cannot be overlooked. As mentioned in the previous section, there are three types of censoring: left, interval and right. Due to this characteristic, the computations required for the direct ML method, such as second derivatives, can become very complicated or remain in an open form.

In addition to this problem of complexity, researchers have observed that ML estimators for Weibull parameters can be biased, particularly in cases of right-censored data or small data sets (Mackisack and Stillman, 1996). Many adjustments have been proposed to address this problem, including methods by



Fei, Kong, and Tang (1995) and Makalic and Schmidt (2022). Other approaches also include Bayes estimators (Lu, 1992) and the Partial Imputation Expectation-Maximization (PIEM) algorithm (Choi et al., 2020).

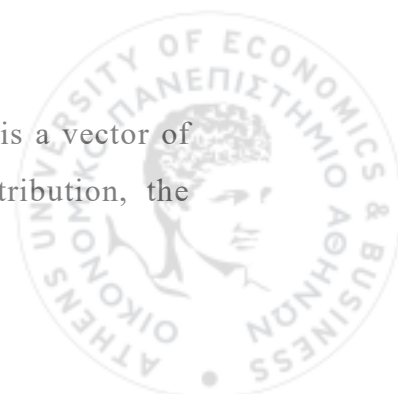
In this study, the Expectation-Maximization algorithm is proposed as an alternative method to overcome these limitations. The EM algorithm is an iterative approach for maximum likelihood estimators of parameters. It is very effective and widely used in cases of latent variables or incomplete data, such as in censoring survival data. As implied by the name, the algorithm consists of two steps in each iteration, the Expectation (E-step) and the Maximization step (M-step) (Haugh, 2015). In the first step, the formula computes the expected value of the complete log-likelihood given the data and some initial values of the parameters that we wish to estimate. In the M-step, we maximize over the parameters the expectation computed in the E-step. This two-step process is repeated until the algorithm converges.

The focus of this study is to implement the EM algorithm to obtain the maximum likelihood estimators (MLEs) of the β 's and the shape parameter p in the Weibull distribution. Note that in our case, we focus on right-censored data. This partial information about the survival time of a censored individual should not be ignored but rather incorporated in the estimation of the parameters (Ferreira and Silva, 2017).

Assuming that we have n observations, let t_i be the observed survival time or censoring time for the i -th individual. The hidden information is the exact failure time for borrowers who have been censored since we only know for sure that the time is greater than its observed time. Let δ_i be a censorship indicator where:

$$\delta_i = \begin{cases} 1, & \text{if } t_i \text{ is uncensored (the event has occurred)} \\ 0, & \text{otherwise (censored)} \end{cases}$$

The observed data for the i -th individual is (x_i, t_i, δ_i) where x_i is a vector of covariates included in the scale parameter λ . In Weibull distribution, the



unknown parameters are the shape parameter p and the regression coefficients, β , that give the value for the scale parameter λ implicitly, from relationship (2.2).

In Weibull distribution, we have that the observed full likelihood is:

$$L(\beta, p) = \prod_{i=1}^n f(t_i)^{\delta_i} \times S(t_i)^{1-\delta_i}$$

$$= \prod_{i=1}^n [\lambda_i p t_i^{p-1} \times \exp(-\lambda_i t_i^p)]^{\delta_i} \times [(\exp(-\lambda_i t_i^p))]^{1-\delta_i}$$

And the observed log-likelihood:

$$l(\beta, p) = \sum_{i=1}^n \{ \delta_i [(p-1) \ln t_i + \ln \lambda_i + \ln p - \lambda_i t_i^p] + (1 - \delta_i) \lambda_i t_i^p \}$$

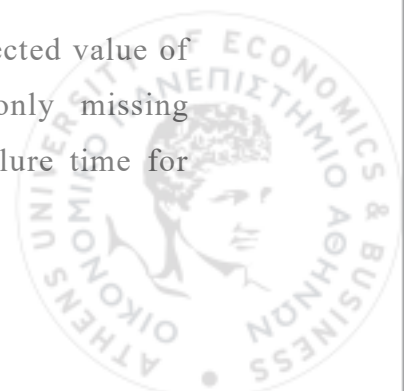
$$= \sum_{i=1}^n \delta_i [(p-1) \ln t_i + X_i \beta + \ln p] - \sum_{i=1}^n \lambda_i t_i^p$$

where $\lambda_i = \exp(x_i' \beta)$.

It is clear that $f(t_i)$ is the contribution of the uncensored individuals to the likelihood, while $S(t_i)$ is the corresponding part for the censored subjects. The next step is to incorporate the hidden information, which is the real failure time for censored individuals, denoted as T_i . As we stated above, the only thing we know about T_i is that it is equal or greater than the corresponding observed failure time, t_i . As a result, we obtain the complete data log-likelihood:

$$l(\beta, p) = \sum_{i=1}^n \delta_i [(p-1) \ln t_i + x_i \beta + \ln p] - \sum_{i=1}^n \lambda_i T_i^p$$

In the E-step of the EM algorithm, we wish to compute the expected value of the complete log-likelihood. From this computation, the only missing information is the expected value of the unobserved exact failure time for



censored individuals. From the memoryless property of Weibull distribution, we have the following:

$$E[T_i^p | T_i > t_i] = t_i^p + \frac{1}{\lambda_i p}$$

We substitute the above result into the full log-likelihood for the censored default times. In the M-step of the algorithm, we maximize the log-likelihood obtained in the E-step, with respect to the parameters of our interest, β and p . To solve for β , we replace the scale parameter λ with $\exp(x'\beta)$ (see relationship 2.2).

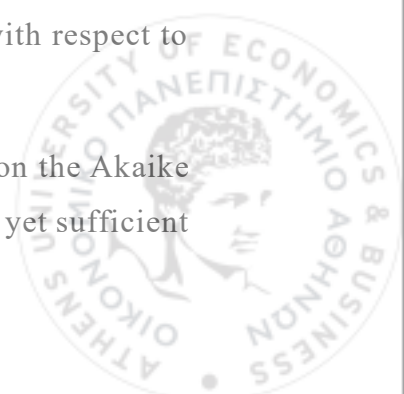
To update our parameters, we derive the system of partial derivatives with respect to β and p . Since the system does not have a closed-form solution, it is solved numerically using techniques such as Newton-Raphson. At each iteration, we update the maximum likelihood estimates for β and p and check if convergence is reached by:

$$|\beta(k+1) - \beta(k)| < \epsilon \quad |p(k+1) - p(k)| < \epsilon$$

where k is the number of the current iteration and ϵ a small number. If convergence is not reached, we move to the next iteration of the algorithm starting again from the E-step. Otherwise, the EM algorithm terminates, and we have obtained the maximum likelihood estimates for β and p . An important note about the EM algorithm is that starting values must be set for its implementation. This step is crucial because reasonable values help the algorithm converge more quickly and avoid local optima.

Now that we have demonstrated the method to acquire the estimates for the parameters of the Weibull distribution, another issue should be investigated. Regarding the scale parameter λ , we should account for the choice of variables included in its modeling. For this reason, a variable selection method will be employed to keep in the model the most relative characteristics with respect to the survival time of the individuals.

In this study, we focus on the backward selection method based on the Akaike Information Criterion (AIC). The goal is to obtain a parsimonious yet sufficient



model that fits the data good. According to this approach, a lower value of AIC recommends a better model.

The process is the same as in a classic regression model. First, we start by fitting the full Weibull model. This means that all available covariates X_1, \dots, X_k are modeling the scale parameter λ (relationship 2.2). For this model, we compute the AIC value:

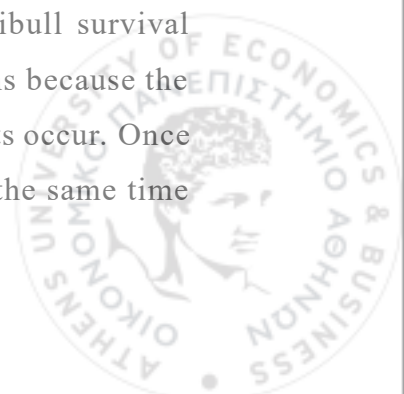
$$AIC = 2k - 2l$$

where k is the number of parameters in the model and l is the log-likelihood of the applicable model. We start by removing a covariate of the full model and computing the AIC for the new model. If the new AIC value is lower than before, we exclude this variable conclusively. This procedure is repeated until there is no improvement in the AIC value when removing any other feature.

Once we have selected the variables for the optimal Weibull model for our problem, we move on to the evaluation of its goodness-of-fit. One way to check whether the final model fits the survival data well is through graphical representation. More precisely, we compare the Kaplan-Meier curve with the survival curve obtained from the Weibull model. The closer the two curves align, the better the fitted model describes the data.

To generate these graphical results, we begin by creating the Weibull curve. Using the Weibull survival function $S(t) = \exp(-\lambda_i t_i^p)$ as denoted earlier, we compute the survival probabilities at different time points t . Note that the scale parameter λ (through the β 's) and the shape parameter p are the ML estimators derived from the EM algorithm. By plotting these survival probabilities against time, we obtain the Weibull survival curve.

Next, we generate the Kaplan-Meier (KM) curve. As discussed in the previous section, this curve is a non-parametric estimate of the survival function which is based purely on the available survival data. Using the KM estimator, we compute empirical survival probabilities. In contrast to the Weibull survival curve, which is smooth, KM curve is a step function. This happens because the survival probability remains constant between the times that events occur. Once both survival curves are computed, we plot them together along the same time



axis. The closer the curves appear to be, the better the fit of the underlying Weibull model is to this specific survival data.

In addition to the graphical approach mentioned earlier, it is advisable to take into consideration the residuals of the model. Residuals are a widely used tool in statistical modelling, helping to assess the goodness of fit of a model. They provide valuable graphs which can help determine the sufficiency of the model or to examine if underlying assumptions hold. In survival analysis, residuals have also been proposed to evaluate the adequacy of the fitted model. Notably, Cox-Snell residuals and Martingale residuals are commonly applied for this purpose (Cox and Snell, 1968; Lundborg, 2015; Collett, 2014, ch. 4).

Cox-Snell residuals can be calculated as follows:

$$r_{ci} = \hat{H}(t_i) = -\ln \hat{S}(t_i), \quad i = 1, \dots, n$$

where $\hat{S}(t_i)$ is the survival probability for individual i at their corresponding observed time t_i , estimated by the model under study. In our case, these probabilities are obtained from the survival function of the Weibull model. The main property of Cox-Snell residuals is that, under the assumption of an adequately fitted model, they follow a unit exponential distribution, i.e.,

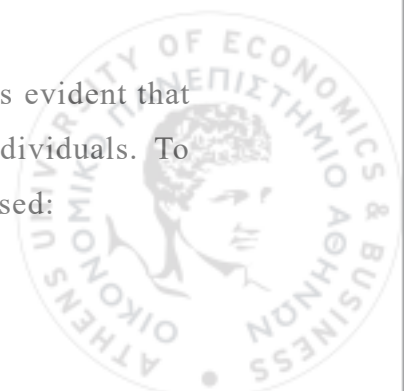
$$r_{ci} \sim \exp(1)$$

This means that:

$$H(r_c) = -\ln(S(r_c)) = r_c$$

As a result, if we plot the Cox-Snell residuals, r_{ci} , against their estimated cumulative hazard rate, $\hat{H}(r_{ci})$, we expect to obtain the straight line $x = y$. This has been proposed because residuals corresponding to censored individuals also form a censored sample from the unit exponential distribution. The mathematical details and proofs can be found in Collett (2014, ch.4). Note that to estimate $H(r_c)$, the Kaplan-Meier (KM) and Nelson-Aalen (NA) estimators are recommended.

Regarding the Cox-Snell residuals for censored observations, it is evident that they are not directly comparable with those of uncensored individuals. To address this issue, a modification of the residuals has been proposed:



$$r_{ci} = \begin{cases} r_{ci}, & \text{for uncensored} \\ r_{ci} + 1, & \text{for censored} \end{cases}$$

Unlike residuals in linear regression, Cox-Snell residuals in survival analysis have different characteristics. Specifically, they are non-negative and not symmetrically distributed around zero. Additionally, when the model is specified correctly, they follow a unit exponential distribution, meaning the expected value and variance of each r_{ci} is expected to be equal to one.

From the Cox-Snell residuals, we can derive the Martingale residuals as follows:

$$r_{mi} = d_i - r_{ci}$$

Martingale residuals have an expected value and summation of zero. They range from negative infinity to one, with negative values corresponding to censored observations. However, since they are not symmetrically distributed around their expected value, their interpretation and the conclusions drawn from them are not straightforward.

2.3 Mixture Cure Model

In many applications of survival analysis, particularly in medical studies, the concept of a “cured” population was introduced by Joseph Berkson and Robert P. Gage in 1952 (Berkson and Gage, 1952). The main idea is that a substantial proportion of individuals participating in a study may never experience the event of interest. In classical survival analysis, it is assumed that if individuals are observed for a sufficiently long period, all will eventually experience the event. On the contrary, mixture cure models suggest that a fraction of the population will never experience the event of interest in their lifetime or over an extended time horizon. These individuals are considered “cured” or non-susceptible, meaning their probability of survival is equal to one. The remaining individuals are viewed as susceptible since they are still at risk of experiencing the event.



Mixture cure models account for these two subpopulations by combining separate models for cured and susceptible individuals. Their objectives can be divided into two parts. The first one is estimating the proportion of long-term survivors, known as the cure rate. The second one is investigating the effects of individual characteristics on both the cure rate and the failure time of short-term survivors.

Let T , where $T > 0$, be a random variable representing the survival time until an individual experiences the event of interest. Let t be a specific value of T . Additionally, let $f(t|x, z)$ and $S(t|x, z)$ be the overall probability density function and the overall survival function of T for the entire population, respectively. Here, x and z represent vectors of explanatory variable that may influence T . Specifically, z is associated with the cure rate, while x affects the failure time distribution of the susceptible group.

The overall survival function in the mixture cure model is given by:

$$S(t|x, z) = \pi(z) + [1 - \pi(z)] S_u(t|x)$$

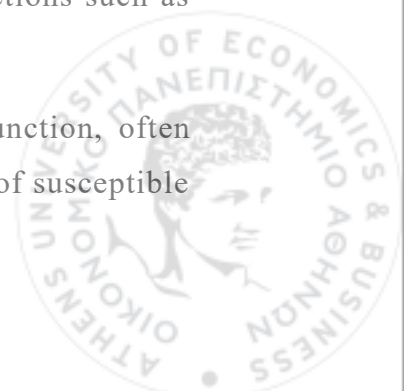
where $\pi(z)$ represents the cure rate, i.e., the probability that an individual is “cured”, and $S_u(t|x)$ is the survival function for the uncured (susceptible) group.

The survival function of the entire population consists of two components. The first, related to the probability of being cured, is called the incidence model and is often modeled using a logistic regression approach. Let Y be a binary variable, where $Y = 0$ indicates an individual who remains susceptible to the event of interest, while $Y = 1$ denotes a long-term survivor. Using a logit model, the probability of being cured is given by:

$$\pi(z) = P(Y = 1) = \exp(z'b) / (1 + \exp(z'b)) \quad (2.3)$$

where b is a vector of parameters associated with the covariate vector z . In addition to the logit link function, shown above, other link functions such as probit and log-log link can also be applied.

The second component of the mixture cure model’s survival function, often referred to as the latency model, describes the survival function of susceptible



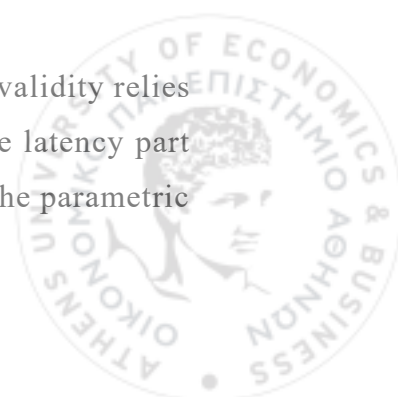
individuals, i.e. $S_u(t|x)$. The modeling of this component can follow one of three different approaches, resulting in nonparametric, semiparametric or fully parametric mixture cure models. This study primarily focuses on the fully parametric approach.

In semi-parametric mixture cure models, Kuk and Chen (1992) proposed a proportional hazards (PH) model to describe the failure time distribution of susceptible individuals, combined with logistic regression for modeling the cure probability. A similar approach was introduced by Peng and Dear (2000) within the framework of non-parametric mixture cure models. While both studies investigated PH mixture models, they used different estimation methods.

In parametric mixture cure models, the time until susceptible individuals experience the event is assumed to follow a specific probability distribution. Commonly used distributions in this setting include the Weibull, exponential (a special case of Weibull), Generalized Gamma, generalized F, log-logistic and log-normal distributions. A significant characteristic of parametric models, including parametric mixture cure models, is that once one function out of the probability density function, survival function or hazard functions, is specified, then the other two are fully determined. The relationships between these functions were previously discussed in Chapter 1.

Parametric mixture cure models hold many advantages, provided that the failure time distribution is correct. Since these models rely on a well-defined distributional assumption, they enable precise estimation of survival and hazard functions. Their structured form allows for straightforward statistical testing and interpretation of results. Assuming a known distribution for failures of short-term survivors, they facilitate the inclusion of explanatory variables without creating an overcomplicated model. This makes it easier to identify and interpret the impact of a pool of characteristics on the survival probabilities of uncured individuals.

However, parametric models also yield certain limitations. Their validity relies explicitly on the assumption that the distribution assumed for the latency part is correct. If the assumed function is misspecified, the results of the parametric



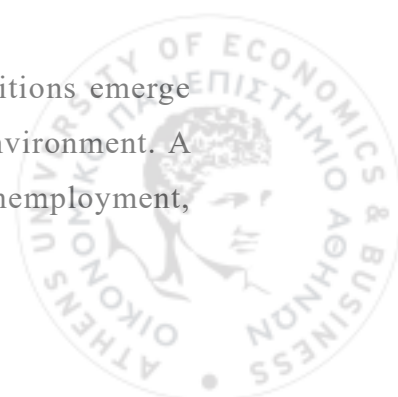
mixture cure model may be inaccurate and misleading. Therefore, model validation constitutes an essential step of the process before trusting the estimations of the fitted model and the conclusions derived from it. Furthermore, the fact that a specific distribution is assumed for the failure time can limit their flexibility, preventing them from capturing more complex information from the data. To address this limitation might, non-parametric models, which require fewer parametric constraints and assumptions, may provide a more flexible alternative.

Taking into consideration all the above, we focus on investigating the parametric family in the frame of mixture cure models. More specifically, the main objective of this study is to define and examine a parametric mixture cure model using the Weibull distribution for the latency part. The choice of Weibull distribution to model the time until susceptible individuals default on their loan, is motivated by its flexibility, introduced through the additional shape parameter p . As previously discussed, the shape of the hazard function changes depending on p . If $p > 1$, the hazard increases over time while if $p < 1$, it decreases. If $p = 1$, it simplifies to the exponential model.

This study aims to model the time until an individual defaults on their loan. In the financial framework, it is reasonable to assume that default risk does not remain constant over time. Instead, hazard rates may decrease or increase due to various endogenous and exogenous factors.

In the first scenario, default risk declines over time. Borrowers who manage to meet their early loan obligations, such as up-front payments, bank fees, fixed costs, interest rates and insurance, are more likely to continue repaying over time. Conversely, financially unstable borrowers who struggle with these initial costs are at higher risk of early default. On the other hand, individuals who are financially stable are expected not only to navigate this early period successfully but also to maintain their repayments for the whole duration of the loan.

However, default risk may also increase over time as new conditions emerge both at an individual level and within the broader economic environment. A borrower's ability to repay their loan may weaken due to unemployment,



excessive debt accumulation or appearing overoptimistic about their ability to meet loan repayments. Even borrowers who initially appeared creditworthy may face financial distress if their income declines. Additionally, other factors in an economic framework, such as rising interest rates or regulatory changes, can also impede loan repayments, increasing default risk over time.

These two scenarios suggest that default risk changes across time and the hazard rate could either decrease or increase. This variability is incorporated into the mixture cure model by the Weibull distribution's shape parameter p .

The overall survival function in the mixture cure model using the Weibull distribution takes the following form:

$$S(t|x, z) = \pi(z) + [1 - \pi(z)] \exp(-\lambda t^p)$$

By incorporating the logistic model for $\pi(z)$ (relation 2.3), we obtain:

$$S(t|x, z) = \exp(z'b) / (1 + \exp(z'b)) + [1 - [\exp(z'b) / (1 + \exp(z'b))]] \exp(-\lambda t^p)$$

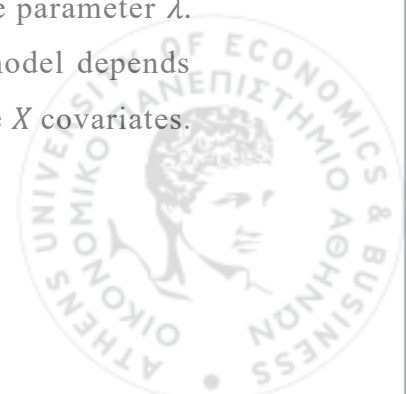
where p and λ are the shape and scale parameters, respectively.

One of the goals of this study is to investigate the effect of explanatory variables on the survival of uncured individuals. To include the set of available variables in the parameterization of the Weibull distribution, we set a regression framework for the scale parameter λ :

$$\lambda = g(x'\beta)$$

where x represents the set of covariates associated with the latency part of the model, and β is the vector of regression coefficients related to x . A commonly used function for g is the exponential, $g(u) = \exp(u)$ which facilitates the interpretation of β 's.

While one could also add a regression framework for the shape parameter p , in this study, we focus on a regression framework only for the scale parameter λ . Additionally, note that the incidence part of the mixture cure model depends only on Z covariates, while the latency model depends only on the X covariates.



These two sets of exploratory variables, X and Z , may be entirely or partly different, or identical.

After specifying the form of the mixture cure model with the Weibull distribution in the latency part, the next step is to estimate the coefficients of the model. More specifically, we aim to obtain the estimates \hat{b} related to the incidence and $\hat{\beta}$ for the latency part that maximize the observed full likelihood. The Expectation-Maximization (EM) algorithm will also be implemented in this case.

As previously denoted, Y is a binary variable, where $Y = 0$ indicates that the individual remains susceptible and will eventually experience the event of interest, i.e. default on the loan, while $Y = 1$ represents a long-term survivor. Additionally, let δ be an indicator variable that denotes whether the borrower is uncensored (i.e. has defaulted on their loan, $\delta_i = 1$), or censored ($\delta_i = 0$). Given the values of these two indicators, it is clear that if $\delta_i = 1$, then $y_i = 0$. However, if $\delta_i = 0$, then y_i is unobserved for this individual. It can either be $y_i = 0$ if the borrower eventually defaults or $y_i = 1$ if the borrower remains a long-term survivor, meaning that they never stop repaying their loan.

Let the observed data be denoted as $(t_i, \delta_i, x_i, z_i)$ for $i = 1, \dots, n$. The observed full likelihood for the mixture cure model is:

$$L(b, \beta, p) = \prod_{i=1}^n \{(1 - \pi(z_i))f(t_i|Y_i = 0; x_i)\}^{\delta_i} \\ \times \{\pi(z_i) + (1 - \pi(z_i))S_u(t_i|Y_i = 0; x_i)\}^{1-\delta_i}$$

where $f(t_i|Y_i = 0; x_i)$ and $S_u(t_i|Y_i = 0; x_i)$ are the probability density and survival functions derived from the Weibull distribution, respectively.

The complete-data log-likelihood is given by:



$$l(b; y) = \sum_{i=1}^n y_i \ln [\pi(z_i)] + (1 - y_i) \ln(1 - \pi(z_i))$$

$$+$$

$$l(\beta, p; y) = \sum_{i=1}^n (1 - y_i) [\delta_i \ln f(t_i | Y_i = 0; x_i) + (1 - \delta_i) S_u(t_i | Y_i = 0; x_i)]$$

As shown above, the likelihood function is the sum of two log-likelihood functions: one from the logistic regression model and one from the Weibull model. However, as previously noted, the variable y_i is partially observed, only for uncensored borrowers.

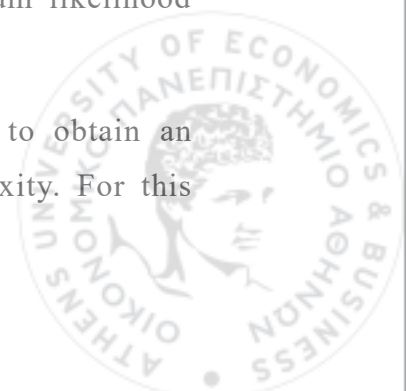
The E-step replaces the unobserved y_i 's in the complete data full-likelihood with their expected values given the observed data. Thus, we update as follows:

$$E(Y_i) = \begin{cases} 0, & \text{if } \delta_i = 1 \\ \frac{\pi(z_i)}{\pi(z_i) + (1 - \pi(z_i)) S_u(t_i | Y_i = 0; x_i)}, & \text{if } \delta_i = 0 \end{cases}$$

This function can be considered as a weight for each individual in the likelihood function. In the E-step, the unobserved information y_i for censored borrowers is replaced by the best estimate given the observed data.

In the M-step of the algorithm, we maximize the likelihood function obtained from the E-step, with $E(Y_i)$ replacing the missing information, with respect to the parameters b, β and p . This maximization process is typically solved using numerical methods, such as the Newton-Raphson algorithm. As with the simple Weibull survival model, the steps of the EM algorithm are repeated until convergence is reached, at which point we obtain the maximum likelihood estimates for b, β and p .

In the mixture cure model, the objective remains the same: to obtain an effective model for survival data without unnecessary complexity. For this



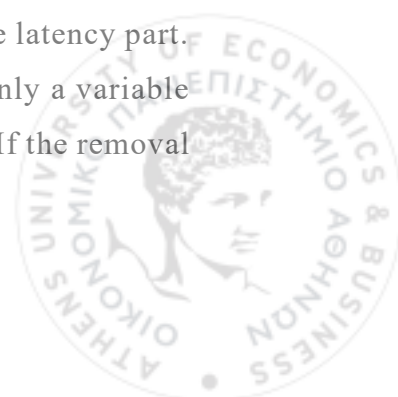
reason, it is advisable to proceed with a selection method to identify the most relevant variables for the mixture cure model. However, a slight increase in complexity arises because we need to account for two parts, each containing explanatory variables.

First, we aim to develop a parsimonious model for the logistic part by selecting the most relevant variables in terms of the cure rate. Additionally, we are interested in keeping the most significant variables in the Weibull model for the latency part of the mixture. As discussed earlier, explanatory variables assumed to influence the survival probability of uncured individuals are incorporated into the Weibull model through the scale parameter λ . It is important to remember that the variables in the incidence part, Z , and those in the latency part, X , can be either entirely or partly different, or identical. Thus, there are three distinct scenarios to consider when selecting variables.

We have chosen to implement a backward feature selection process based on AIC, similar to the method used in the simple Weibull model. However, to explore the three scenarios mentioned above, we will apply the variable selection procedure within three different frameworks. The goal, though, remains the same: to identify the model with the lowest AIC value, regardless of the scenario.

In the first scenario, we assume that the same variables should be included in both the incidence and latency parts of the mixture cure model. We begin by fitting the full model, which includes all available features and compute the AIC value. We then iteratively remove the same variable from both parts of the model and recalculate the AIC. If the removal of a particular variable leads to a lower AIC, it is excluded from both components. We repeat this procedure until there is no further reduction in AIC, at which point we obtain the best model under this assumption.

The second scenario considers the case where all available variables remain in the incidence part while we start removing variables only from the latency part. The process begins again with the full model, but in each step, only a variable from the Weibull model is removed before recalculating the AIC. If the removal



of a variable results in a lower AIC, it is permanently excluded and the procedure continues until there is no further improvement.

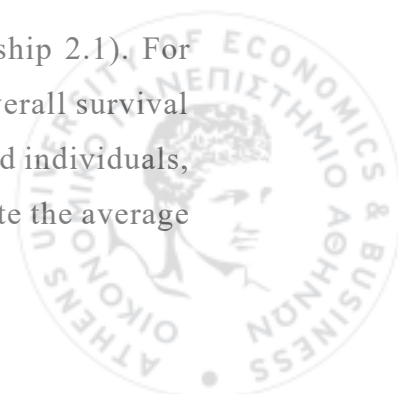
Similarly, the procedure is repeated by removing variables only from the incidence part of the mixture cure model while keeping all variables in the latency part. Starting with the full model, we compute the AIC, then iteratively remove one feature at a time from the logistic component and recalculate the AIC. If the removal of a variable results in a lower AIC, it is removed, and the method continues until no further improvement is possible, yielding the final model.

Once the best models from each of these three approaches have been identified, the selection process does not end there. We repeat it using the final model obtained from the previous step as the new full model. The three selection procedures are then reapplied iteratively. After several applications of this formula, we reach a point where no further reduction in AIC is possible, yielding the most efficient mixture cure model for our problem proposed by this backward selection process.

Up to this point, we have determined the structure of the mixture cure model, incorporating a logistic regression for the incidence part and a Weibull distribution for the latency part. We then described the EM algorithm for estimating the model parameters through maximum likelihood and proposed a backward selection method based on AIC to identify the most relevant explanatory variables for both the cure rate and survival time.

The next step in our analysis is to assess how well the final selected mixture cure model fits the available survival data. Similar to the standard Weibull case, a straightforward way to evaluate the goodness-of-fit of the mixture cure model is through a comparison of two survival curves: the Kaplan-Meier (KM) and the one generated from the mixture model. If the two curves closely align, it indicates a good fit.

We first obtain the KM curve using the KM estimator (relationship 2.1). For the parametric survival curve based on our model, we compute overall survival probabilities, i.e., survival probabilities for both cured and uncured individuals, at various time points. To summarize this information, we calculate the average



overall survival probability across the entire population at each time point. The mixture cure model curve is then plotted using these values. By visually comparing the two curves, we can assess the model's performance based on their proximity.

Similar to the simple Weibull case discussed in the previous section, martingale and Cox-Snell residuals can also be used to assess the fit of the mixture cure model. We begin by computing the negative logarithm of the overall survival probabilities for each individual based on the fitted mixture cure model:

$$-\ln[P(T > t_i | x_i, z_i)] = -\ln\hat{S}(t_i | x_i, z_i) = \ln[\pi(z_i) + (1 - \pi(z_i))\exp(-\lambda_i t_i p)]$$

From this, we obtain the estimated Cox-Snell residuals:

$$r_i = -\ln\hat{S}(t_i | x_i, z_i) \quad i = 1, \dots, n$$

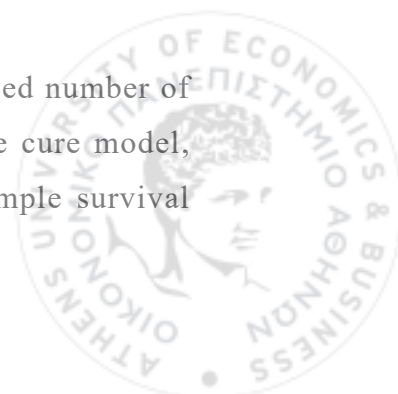
In the mixture cure model, the interpretation of Cox-Snell residuals is slightly different from the simple case. When the fitted model is adequate, the residuals follow a mixed type distribution consisting of a unit exponential component and a probability mass. However, to assess the overall goodness-of fit, we can plot the Cox-Snell residuals against their estimated cumulative hazard function and compare it to the unit exponential distribution. The closer the two curves align, the better the model fit.

Peng and Taylor (2017) provided key insights into mixture cure model diagnostics. They noted that even when the model is incorrect, Cox-Snell residuals will still follow a mixed-type distribution but with different components. They also suggested using residuals to assess the adequacy of the incidence and latency parts separately. However, in this study, we focus on the overall performance of the mixture cure model.

From the Cox-Snell residuals, we can derive the Martingale residuals using the relationship:

$$m_i = d_i - r_i$$

Martingale residuals represent the difference between the observed number of events and the expected number of events based on the mixture cure model, given the follow-up time and included variables. Similar to simple survival



models, martingale residuals in mixture cure models have an expected value of zero, and their estimates sum to zero. Additionally, they satisfy

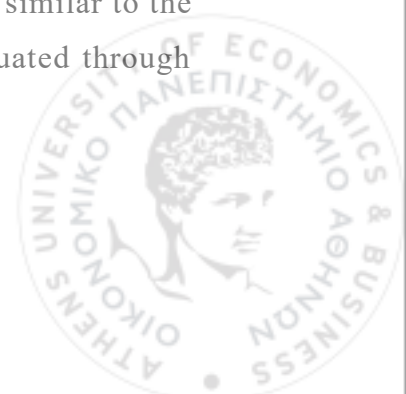
$$\text{Cov}(m_i, m_j) = 0, \text{Cov}(\hat{m}_i, \hat{m}_j) < 0.$$

However, a key difference lies in the range. In contrast to standard survival models, where martingale residuals span from negative infinity to one, in mixture cure models, they have a lower bound of

$$\ln(1 - \pi(z_i)).$$

In this Chapter, we provided an overview of the survival analysis methodology adopted in this study. We began by presenting the main definitions and concepts about survival analysis and explained its core functions. After that, we introduced the Weibull distribution and its key properties, focusing on its flexibility through the shape parameter, which allows the hazard function to either increase or decrease over time. We then presented the simple Weibull model to estimate the survival probabilities by incorporating explanatory variables through a regression framework. Additionally, we introduced the EM algorithm approach for estimating the model's parameters and described the use of backward variable selection based on AIC to identify the most relevant covariates. For model evaluation, we presented a graphical comparison between the model's survival curve and the Kaplan-Meier curve, along with residual analysis of Cox-Snell and Martingale residuals.

Moving forward, we introduced the mixture cure model, which extends classical survival models by allowing for the presence of a cured, or non-susceptible, subpopulation. We described how the model combines a logistic regression for estimating the cure probability and a Weibull distribution to model the survival of uncured individuals, or long-term survivors. The estimation of model parameters using the Expectation-Maximization algorithm was outlined, along with a tailored backward variable selection process based on AIC that accounts for both components of the model. Finally, similar to the simple Weibull model, we described how model fit can be evaluated through survival curve comparisons and residual diagnostics.



With the methodological framework described, the next chapter introduces the dataset used in the empirical analysis. A general description of the data will be provided, followed by descriptive statistics for the explanatory variables that will be used to estimate and interpret the survival models of interest.



Chapter 3

Real life problem and Dataset description

3.1 Dataset Overview and Preparation

The objective of this study is to analyze the behavior and effectiveness of the previously discussed methods in modeling the time until a borrower defaults on their loan. To address this, we use a publicly available dataset of monthly loans from Freddie Mac (<https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset>). Specifically, we examine a subset of the full dataset which consists of a simple random sample of 50000 loans and in a pre-processed form from prior studies. This sample contains monthly data on single-family mortgage loans, with both quantitative and qualitative characteristics. To enhance the dataset's accuracy, certain variables required reformatting due to incorrect labeling. After these adjustments, we obtained the final set of variables, presented in Table 3.1. However, it is important to note that not all listed variables will be included in the subsequent analysis and modeling process.

The next crucial step is to thoroughly examine the dataset for potential issues or errors that could compromise the analysis. After ensuring the absence of duplicate observations and invalid or extreme values, we proceed to assess missing data. One possible approach to handling missing values involves imputation using simple statistical measures such as the mean or median, or more advanced methods such as Random Forest. However, upon careful examination, we notice that only 2.79% of the dataset contains missing values in at least one variable. Given this low percentage, we decide to remove these rows rather than impute the missing values. The final version of the data set consists of 48607 observations, i.e., monthly mortgage loans.

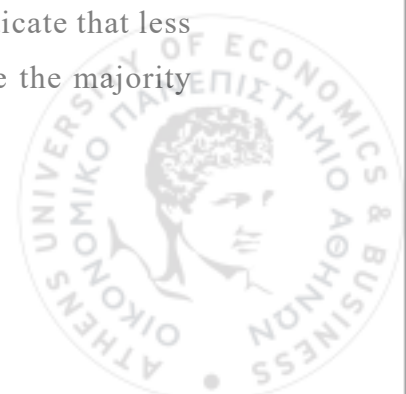


Name	Description	Type
ID loan	Unique loan identifier number	Character
Credit Score	Credit score	Numeric
MI %	Mortgage insurance percentage (%)	Numeric
Occupancy	Occupancy status	Categorical
CLTV %	Combined loan-to-value ratio (%)	Numeric
DTI %	Debt-to-income ratio (%)	Numeric
Original Balance	Original UPB	Numeric
LTV %	Loan-to-value ratio (%)	Numeric
Channel	Loan origination source	Categorical
Property Type	Property type	Categorical
Loan Term (months)	Scheduled monthly payments	Numeric
Origination Month	Loan issuance month	Categorical
Default status	Default status	Categorical
Months	Time until default or censorship in months	Numeric

Table 3.1 Dataset variables overview

3.2 Numerical Variable Analysis

Turning to the numerical variables, we analyze their key characteristics. According to the descriptive statistics in Table 3.2, the dataset consists of both short-term and long-term mortgage loans. Specifically, loan durations range from 48 months (four years) to 360 months (30 years) with an average term of 25 years. Another characteristic is the mortgage insurance percentage, which represents the proportion of loss coverage on a loan. The data indicate that less than 25% of loans are partially covered in case of default, while the majority remain uninsured.

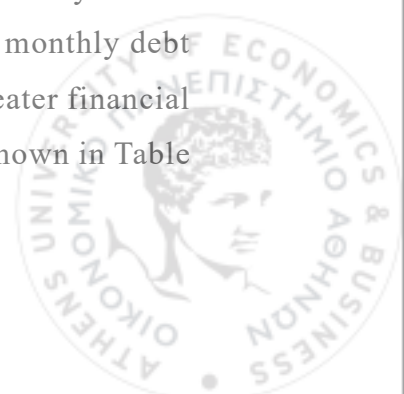


	min	1 st q.	median	mean	3 rd q.	max	sd
Credit score	300.00	683.00	729.00	719.90	765.00	830.00	56.98
MI %	0.00	0.00	0.00	3.85	0.00	50.00	9.16
LTV %	6.00	62.00	75.00	70.57	80.00	100.00	16.46
CLTV %	6.00	62.00	75.00	71.12	80.00	160.00	16.89
DTI %	1.00	24.00	32.00	31.93	40.00	65.00	11.33
Original balance	9.31	11.35	11.74	11.71	12.10	13.27	0.52
Loan Term (months)	48.00	180.00	360.00	295.80	360.00	360.00	84.90
Months	0.00	24.00	30.00	47.05	47.00	291.00	47.05

Table 3.2 Descriptive statistics of numerical variables

The dataset also includes the original combined loan-to-value (CLTV) and loan-to-value (LTV) ratios for mortgage loans. The CLTV is calculated by dividing the original loan amount, including any secondary loans, by the property value. This metric assesses the risk associated with multiple loans on the same property. A lower CLTV indicates reduced risk for lenders, who generally prefer it to be no more than 80-85%. In contrast, the simple LTV ratio does not take into consideration any additional loans. Since we observed that for most borrowers the two ratios are identical, we decided to proceed with only the CLTV as it provides a more complete measure of loan risk. In our dataset, at least one-third of the observations fall within the desirable CLTV range (below 80%), though some values reach as high as 160%.

Beyond these ratios, the debt-to-income (DTI) ratio is another key factor available in our dataset. It is calculated by dividing a borrower's monthly debt payments by their total monthly income. A lower DTI suggests greater financial stability and a higher ability to manage monthly obligations. As shown in Table



3.2, the average DTI in our dataset is 32%, which is considered a relatively low value, indicating a moderate level of default risk.

Another key metric for assessing a borrower's creditworthiness is the credit score, commonly expressed as the FICO score in the U.S. This score is based on a combination of factors related to an individual's payment and credit history. Mortgage lenders heavily use credit scores as an indicator to evaluate a borrower's likelihood of repaying a loan. The score ranges from 300 to 850, with higher values reflecting greater reliability. Figure 3.1 shows that the majority of borrowers in our dataset have high creditworthiness with an average FICO score of 719.9. Based on this information, we would expect a relatively low default rate in our sample.

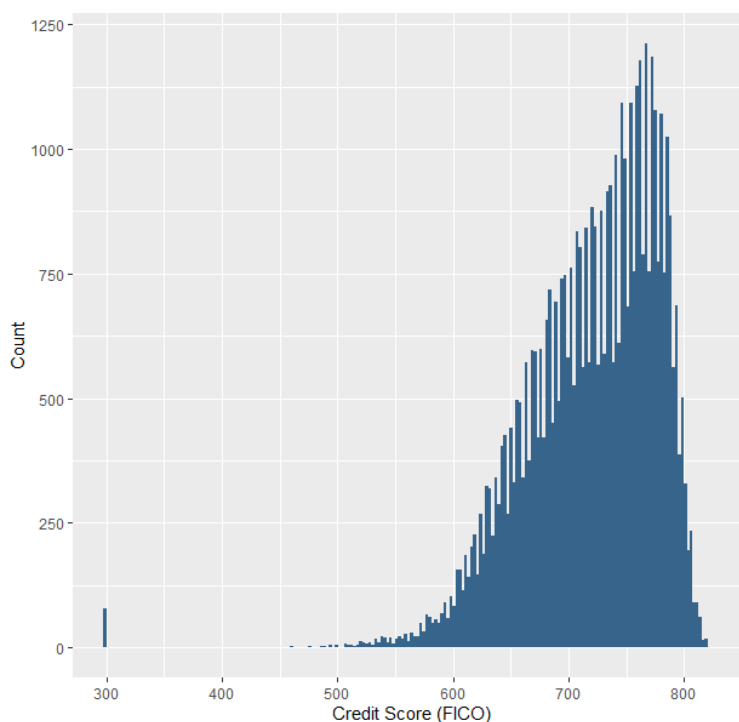
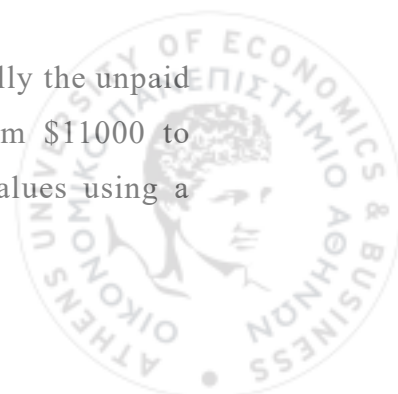


Figure 3.1 Histogram of FICO Scores

The dataset also includes information on loan amounts, specifically the unpaid principal balance of issued mortgage loans, which ranges from \$11000 to \$578000. Given this wide range, we opt to transform these values using a



logarithmic scale for better comparability. Therefore, any reference to the original unpaid principal balance from this point forward will correspond to its logarithmic value.

Finally, we analyze the time until default or censorship, which is the variable of interest in this survival study. Figure 3.2 shows its distribution, showing that the minimum value is zero, while the maximum extends to 291 months. This implies that some borrowers either defaulted or were censored at the very start of their follow-up period, while others were observed (or defaulted) after nearly 24 years. The median time until default or censorship is 30 months.

Since the Weibull distribution requires that $T > 0$, we replaced loans with a time until default or censorship of zero with a very small positive value (0.0001) to ensure proper implementation of the survival models.

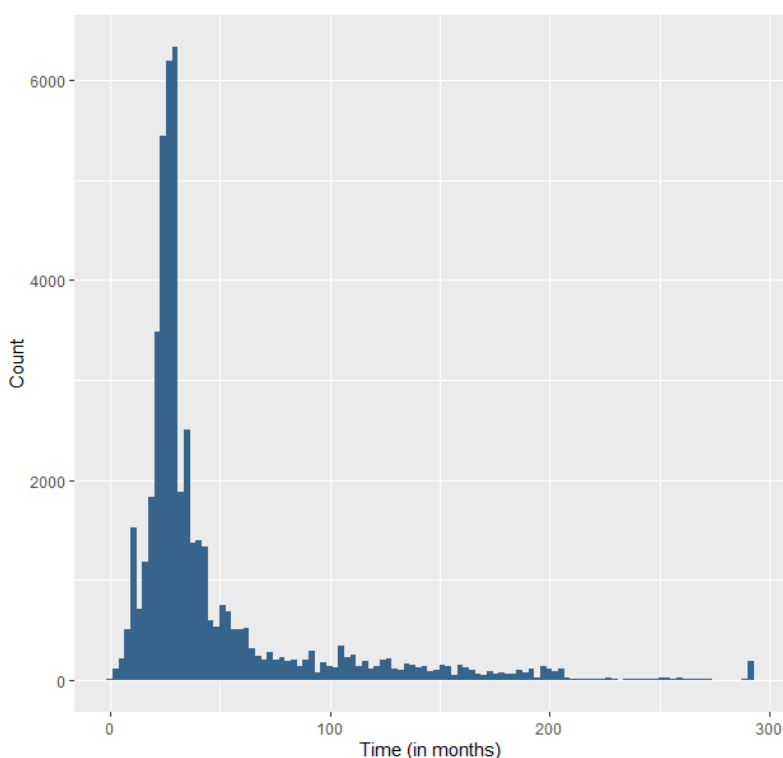


Figure 3.2 Histogram of time until default or censoring (in months)



3.3 Categorical Variable Analysis and Cured Population Assessment

We next examine the categorical characteristics of the dataset, beginning with the occupation status of the loans, which indicates the mortgage type. As shown in Figure 3.3, the largest proportion, 93%, of the loans are issued for the borrower's primary residence (P). Investment properties (I) account for 4.1% of loans, while the remaining share corresponds to mortgages for second homes (S).

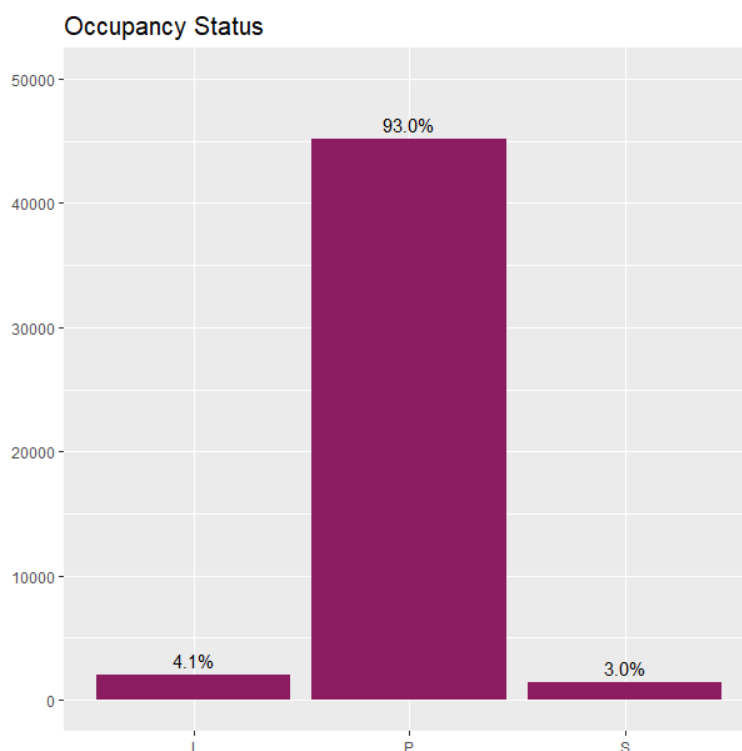
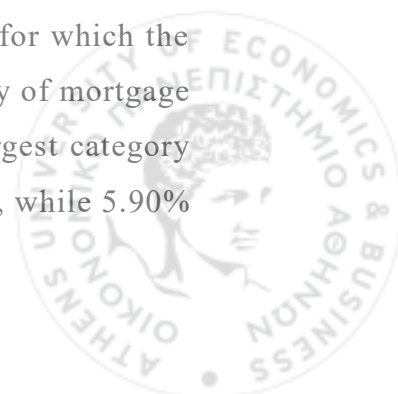


Figure 3.3 Barplot of Occupancy status (I: Investment property, P: Primary Residence, S: Second Home)

Another key characteristic in the dataset is the type of property for which the loan is issued. According to the descriptive statistics, the majority of mortgage loans, 83.96%, are for single-family homes (SF). The second largest category includes properties classified as planned unit developments (PU), while 5.90%



of the loans are for condominiums (CO) (Figure 3.4). The remaining category, labeled as “Other”, includes cooperative shares and manufactured housing.

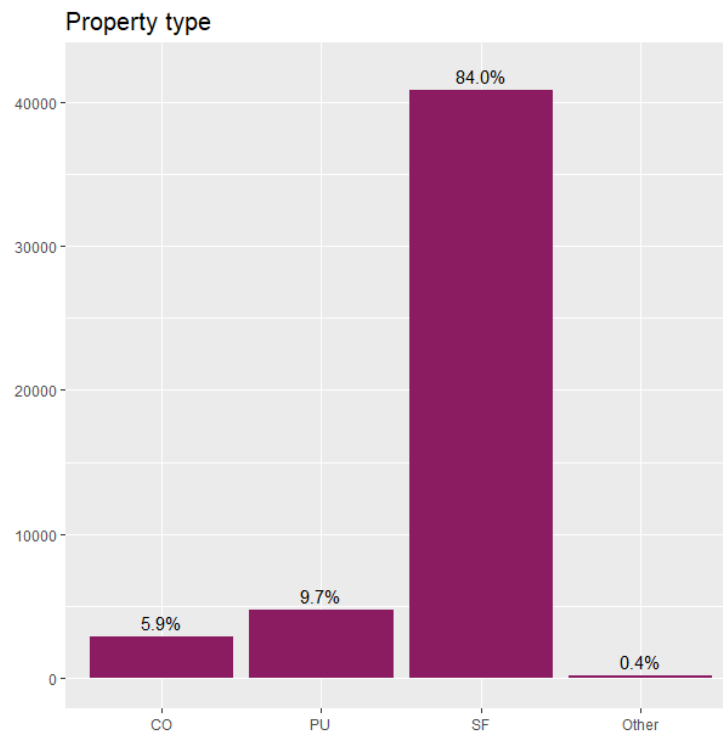
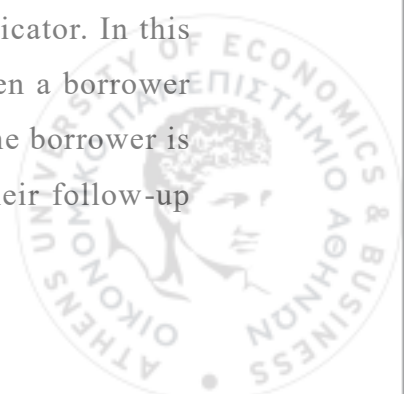


Figure 3.4 Barplot of Property Type (CO: Condominiums, PU: Planned Unit Developments, SF: Single-family Homes, Other: Cooperative shares and manufactured housing)

Additionally, the dataset includes information on the origination channel of the mortgage loan (Figure 3.5.). More than half of the borrowers (52%) obtained their loans through a Broker or Correspondent, or did not specify whether a third party was involved in the process (Category T). The remaining loans were originated directly by the lender or its affiliates, without the involvement of a third party (Category R).

Finally, a crucial aspect of survival analysis is the censoring indicator. In this study, as previously discussed, this indicator is equal to one when a borrower defaults on their loan. Conversely, it takes the value zero when the borrower is censored, meaning that they have not defaulted by the end of their follow-up



period. As shown in Figure 3.6., only 3% of the borrowers in our dataset have defaulted, corresponding to 1482 observations.

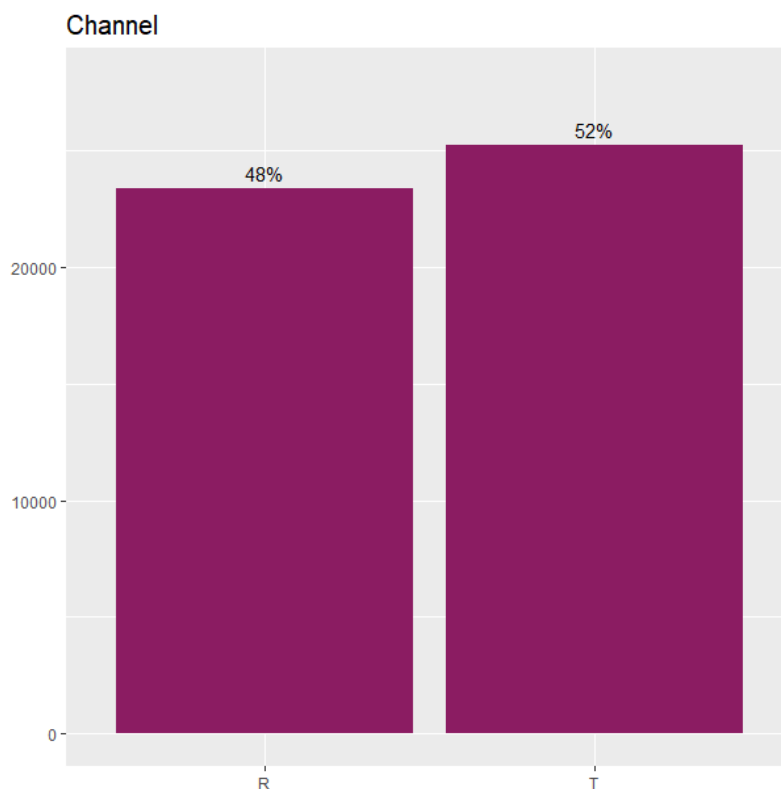
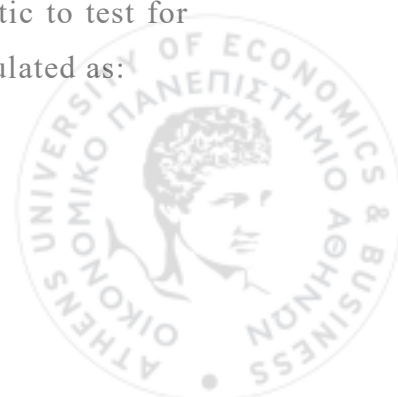


Figure 3.5 Barplot of the Loan origination source (R: Lender or its affiliates and no involvement of a third party, T: Broker or Correspondent, or a third-party involvement not specified)

In addition to exploring the dataset’s characteristics to understand its nature, it is also important to examine whether the assumption of a cured population holds in our case. Since our goal is to fit a mixture cure model for default time, we need to determine whether a proportion of borrowers will continue meeting their financial obligations in the long term.

Maller and Zhou (1992) proposed a simple nonparametric statistic to test for the presence of long-term survivors. The null hypothesis is formulated as:

$$H_0: \tau_{Fu} > \tau_G$$



where $F_u = 1 - S_u$ represents the distribution of survival times for uncured individuals, G denotes the distribution of censoring times, and τ_{F_u} and τ_G are their respective supports.

This hypothesis implies that the maximum survival time of uncured borrowers is longer than the maximum censoring time, meaning that there are no distinct

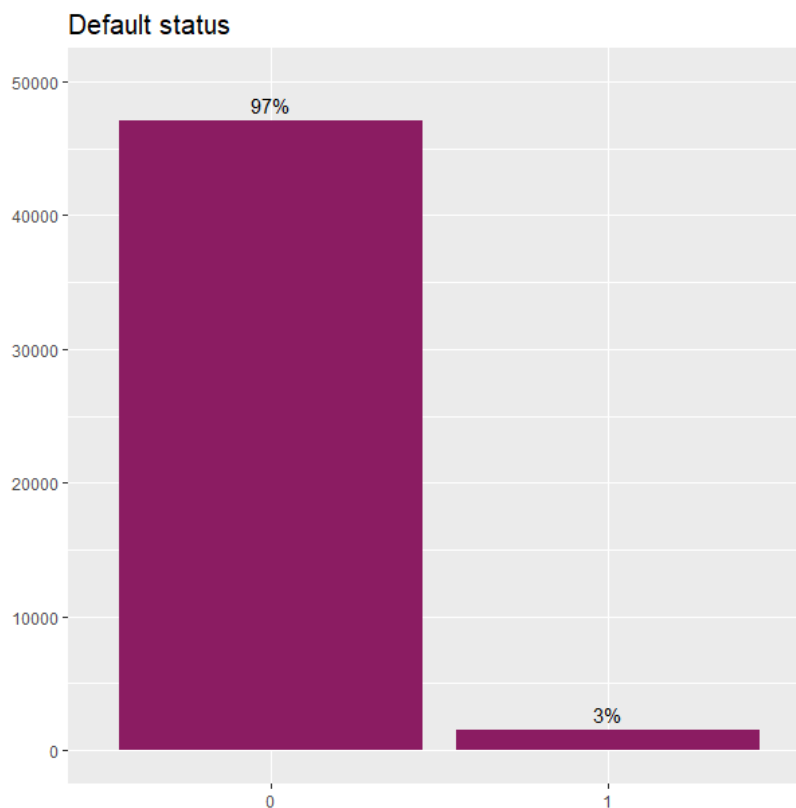
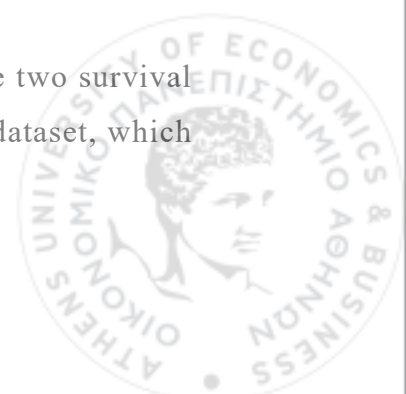


Figure 3.6 Barplot of Default status of borrowers (0: censored, 1: uncensored)

groups of long- and short-term survivors in the dataset. However, in our case, we reject the null hypothesis ($p < 0.001$), indicating the presence of a cured subpopulation in our dataset. Hence, it is meaningful to proceed to a mixture cure model which accounts for the cured individuals.

In the following chapter, we will proceed with implementing the two survival models, the simple Weibull and the mixture cure model, on our dataset, which has been described in detail above.





Chapter 4

Modeling and Results

In the previous chapters, we thoroughly presented the methodology of the two survival models, as well as the dataset used in our analysis. Now, we present the results of applying these methods to model the time until a borrower defaults on their loan.

4.1 Results of the Classical Weibull Survival Model

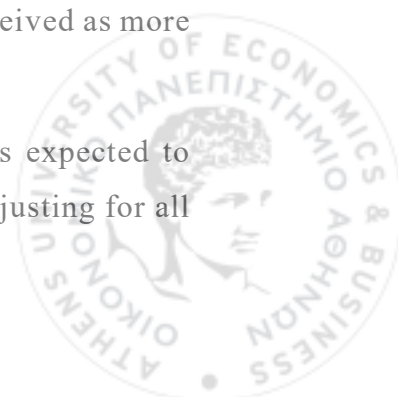
The first survival model we implement is the Weibull survival model, which, as a classic parametric model, assumes that eventually, every borrower will default given a sufficiently long (infinite) period. After implementing backward selection based on AIC, we arrive at the final model, which includes all the available explanatory variables from the dataset.

The results of the Weibull survival model, presented in the Accelerated Failure Time (AFT) parameterization, are presented in Table 4.1. Since these results are obtained in the AFT form, the interpretation is related to the survival time of borrowers. Specifically, the effect (or factor) of each characteristic on survival time can be determined by exponentiating its coefficient, which stretches or shrinks the time until default. The change in survival time is also calculated as follows:

$$T = [e^{\beta} - 1] \times 100$$

As shown in Table 4.1, both primary and second-home owners are expected to have longer survival times, by 23.4% and 74.08%, respectively, compared to those who have taken out loans for investment properties. This result aligns with expectations, as borrowers typically prioritize repaying loans associated with their primary homes, and second-home owners are often perceived as more financially stable.

Regarding loan duration, a 1-month increase in the loan term is expected to accelerate default, reducing survival time by a factor of 0.997, adjusting for all



other characteristics in the model. Furthermore, for each 1% increase in the debt-to-income ratio and combined loan-to-value ratio, survival time decreases by 1.2% and 1.6%, respectively. These results show that borrowers with higher debts or loan-to-value ratios are at an increased risk of default. Another expected finding comes from the FICO score: a 10-point increase in the credit score results in a 6.1% increase in survival time. This result confirms that higher creditworthiness is associated with lower default risk.

Beyond the estimations of coefficients (β) that relate to the scale parameter (λ) of the Weibull survival model, we also estimate the shape parameter (p). Unlike the scale parameter, the shape parameter is not modeled through a regression framework but is instead considered constant across all borrowers. In our model, the estimated shape parameter value is 1.49, which is greater than 1. This indicates that the hazard (or risk of default) increases over time. In other words, while borrowers may initially meet their financial obligations, the risk of default rises as time progresses. This result is crucial for financial institutions, as it highlights the importance of early intervention, such as loan modifications, to reduce the risk of defaults occurring in later periods.

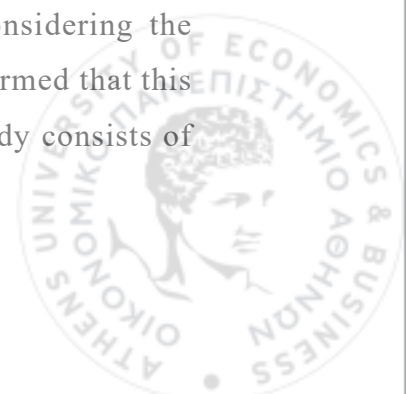


Weibull model					
	Coefficient	% Change in Survival Time	Std error	zvalue	pvalue
Intercept	1.263	253.670	0.454	2.78	0.005
Credit Score	0.006	0.613	0.001	28.55	<0.001
MI %	-0.006	-0.587	0.002	-3.07	0.002
Occupancy (P)	0.212	23.646	0.069	3.08	0.002
Occupancy (S)	0.555	74.135	0.161	3.45	0.001
CLTV %	-0.016	-1.604	0.002	-9.00	<0.001
DTI %	-0.012	-1.148	0.002	-7.24	<0.001
Original Balance	0.328	38.798	0.040	8.29	<0.001
Channel (T)	-0.377	-31.432	0.037	-10.10	<0.001
Property Type (PU)	-0.140	-13.070	0.121	-1.16	0.247
Property Type (SF)	-0.376	-31.326	0.098	-3.82	0.001
Property Type (Other)	-0.381	-31.702	0.160	-2.39	0.017
Loan Term	-0.003	-0.301	0.001	-11.35	<0.001
Shape parameter p	1.495				
Log-Likelihood	-11186.1				
AIC	22400.22				

Table 4.1 Results of the Weibull Survival Model

4.2 Results of the Mixture Cure Model

After implementing the classical Weibull survival model, we proceed to the mixture cure model to analyze the time until loan default, considering the existence of a cured population. In the previous section, we confirmed that this assumption is valid, and the fitted mixture cure model in our study consists of



a logistic component for the incidence part and a Weibull component for the latency part.

To obtain the final model, we apply backward selection based on AIC across the three scenarios outlined in Chapter 3. This method reveals that, unlike the classical Weibull model, some variables should be eliminated from the parts of the mixture cure model. Specifically, the final mixture cure model with the lowest AIC, includes all variables in both parts, except for the Mortgage insurance percentage which was excluded from both components. The results of the mixture cure model are presented in Table 4.2.

The coefficients for each characteristic with the prefix π - correspond to the logistic component and are associated with the cure rate of the model. Meanwhile, the coefficients with the prefix θ 1- are connected to the Weibull part of the model, describing the survival times of the susceptible subpopulation. By exponentiating the θ 2 intercept, we obtain the estimation for the shape parameter p , which is 1.627. Similar to the result in the classical Weibull model, since $p > 1$, this indicates that the hazard increases as time passes.

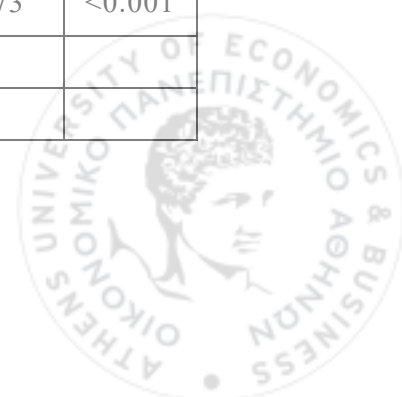
Regarding the explanatory variables, it is interesting to note that those included in the logistic regression (incidence part) are not statistically significant according to their p -values. This suggests that the logistic part of the model does not identify significant factors for the cure rate.

On the other hand, in the Weibull component (latency part), we observe more meaningful results. There are several characteristics that show significant effect on the survival time of uncured borrowers. The latency component is modeled using the Weibull distribution and follows the Proportional Hazards (PH) parameterization. Exponentiating the coefficients β gives us the hazard ratios, which represent a multiplicative effect on the hazard for borrowers in the susceptible subgroup, who will eventually default.



Mixture Cure Model					
	Coefficient	% Change in Hazard	Std error	zvalue	pvalue
p.Intercept	0.112	11.803	9.142	0.012	0.990
p.Credit Score	-0.011	-1.045	0.014	-0.748	0.454
p.Occupancy (P)	0.769	115.683	1.358	0.566	0.571
p.Occupancy (S)	-0.294	-25.453	4.240	-0.069	0.945
p.CLTV %	0.033	3.344	0.051	0.643	0.520
p.DTI %	-0.010	-0.991	0.027	-0.374	0.708
p.Original Balance	0.178	19.482	0.769	0.231	0.817
p.Channel (T)	0.106	11.138	1.282	0.082	0.934
p.Property Type (PU)	-0.146	-13.618	0.884	-0.166	0.868
p.Property Type (SF)	-0.902	-59.436	0.961	-0.939	0.348
p.Property Type (Other)	0.319	37.544	1.416	0.225	0.822
p.Loan Term	0.002	0.241	0.008	0.286	0.775
th1.Intercept	-0.302	-26.096	1.578	-0.192	0.848
th1.Credit Score	-0.015	-1.532	0.001	-18.182	<0.001
th1.Occupancy (P)	0.005	0.479	0.323	0.015	0.988
th1.Occupancy (S)	-0.219	-19.687	0.776	-0.282	0.778
th1.CLTV %	0.035	3.529	0.004	8.382	<0.001
th1.DTI %	0.013	1.289	0.007	1.899	0.058
th1.Original Balance	-0.321	-27.428	0.163	-1.971	0.049
th1.Channel (T)	0.467	59.458	0.205	2.275	0.023
th1.Property Type (PU)	-0.227	-20.340	0.306	-0.744	0.457
th1.Property Type (SF)	-0.085	-8.119	0.234	-0.362	0.717
th1.Property Type (Other)	-0.325	-27.774	0.492	-0.661	0.509
th1.Loan Term	0.005	0.508	0.001	7.979	<0.001
th2.Intercept	0.487	62.776	0.053	9.273	<0.001
Log-Likelihood	-11071.14				
AIC	22192.27				

Table 4.2 Results of the Mixture Cure Model



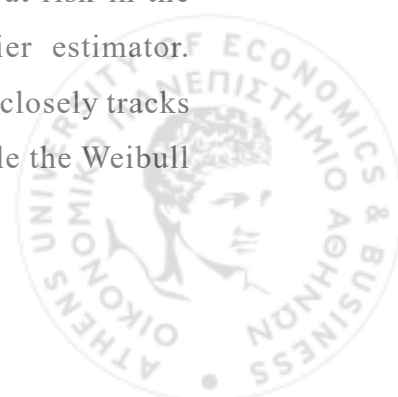
For instance, for a unit increase in credit score, the hazard decreases by 1.53%, indicating a longer survival time and a lower risk of default. Conversely, when the combined-loan-to-value ratio increases by 1%, the hazard increases by 3.53%, meaning that higher loan-to-value ratios are associated with a greater risk of default. The interpretation of the other coefficients follows a similar logic.

In summary, although there were no statistically significant characteristics in the logistic component for the cure rate, the Weibull component provided valuable insights into the factors that influence the survival time of susceptible borrowers. These results underscore the importance of considering both cured and uncured subpopulations in the modeling process and emphasize the ability of the mixture cure model to offer a better understanding of loan default.

4.3 Model Evaluation and Comparison

Now that we have presented the results from both models, the next step is to evaluate their goodness of fit using the Kaplan-Meier estimator. In Figure 4.1, the survival curves from each model are compared against the Kaplan-Meier curve, which is based on the entire dataset. The closer a model's survival curve is to the Kaplan-Meier curve, the better it fits the data. From the comparison, we can see that the survival curve derived from the mixture cure model aligns more closely with the Kaplan-Meier curve, indicating a better fit when modeling the time until default. While the Weibull model also provides a reasonable fit, the mixture cure model outperforms it, especially in the early to middle time periods, where it more accurately captures borrowers' survival probabilities.

However, both models struggle to correctly capture survival at later time points. This might be due to the limited follow-up period, which prevents us from clearly observing the portion of the cured subpopulation, even though it likely exists. Furthermore, the small number of individuals remaining at risk in the later time periods could cause instability in the Kaplan-Meier estimator. Despite these challenges, the mixture cure model's survival curve closely tracks Kaplan-Meier curve throughout most of the observed period, while the Weibull model deviates more in the later stages.



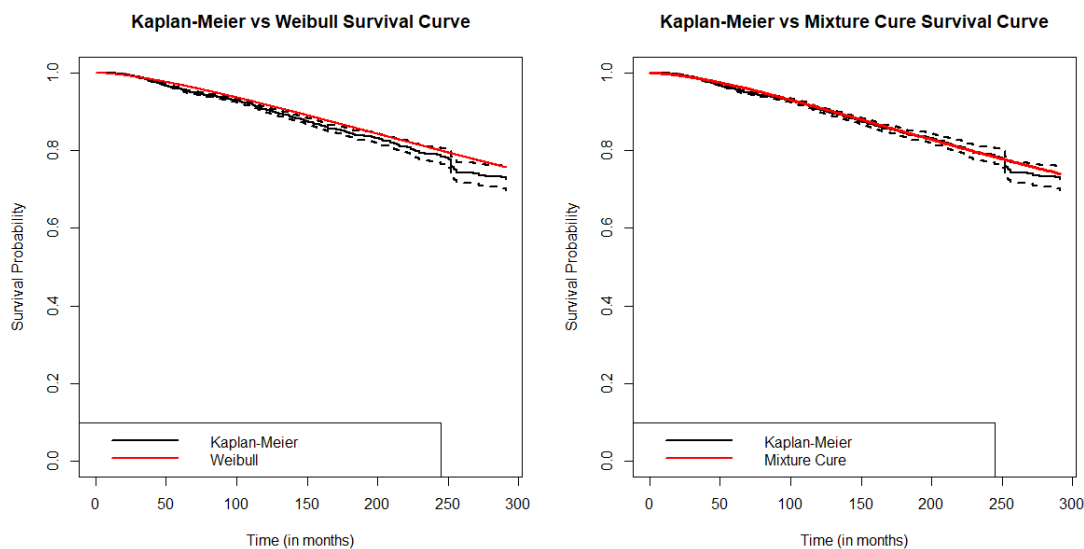
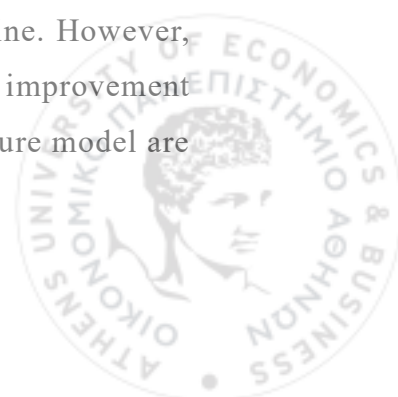


Figure 4.1 Kaplan-Meier curves with Estimated Survival curves from Weibull model (left) and Mixture Cure Model (right) for the whole dataset

Another useful metric for evaluating the performance of both models is the Cox-Snell residuals. In Figure 4.2, we present the Cox-Snell residuals derived from the Weibull model (on the left) and the mixture cure model (on the right). As discussed earlier, if the models fit the data well, we expect the cumulative hazard rate of the corresponding Cox-Snell residuals, estimated using the KM estimator, to follow a unit exponential distribution. This is reflected in the graph as the straight line $y = x$. It is important to note that this assumption holds for the mixture cure model when the overall fit is good. If we are interested in assessing the fit of the incidence and the latency components separately, the Cox-Snell residuals should be compared to other distributions (Peng and Taylor, 2017).

Looking at both graphs, we observe a similar behavior: the two curves align closely with the unit exponential line, but at later time points, they start to deviate. Given that there are fewer loans remaining at risk in the later stages of the dataset, it is expected that the curves differ from the $y = x$ line. However, the graph derived from the mixture cure model shows a slight improvement over the Weibull model. The Cox-Snell residuals in the mixture cure model are



closer to the unit exponential line, indicating a better overall fit. This observation is consistent with the conclusions drawn from Figures 4.1.

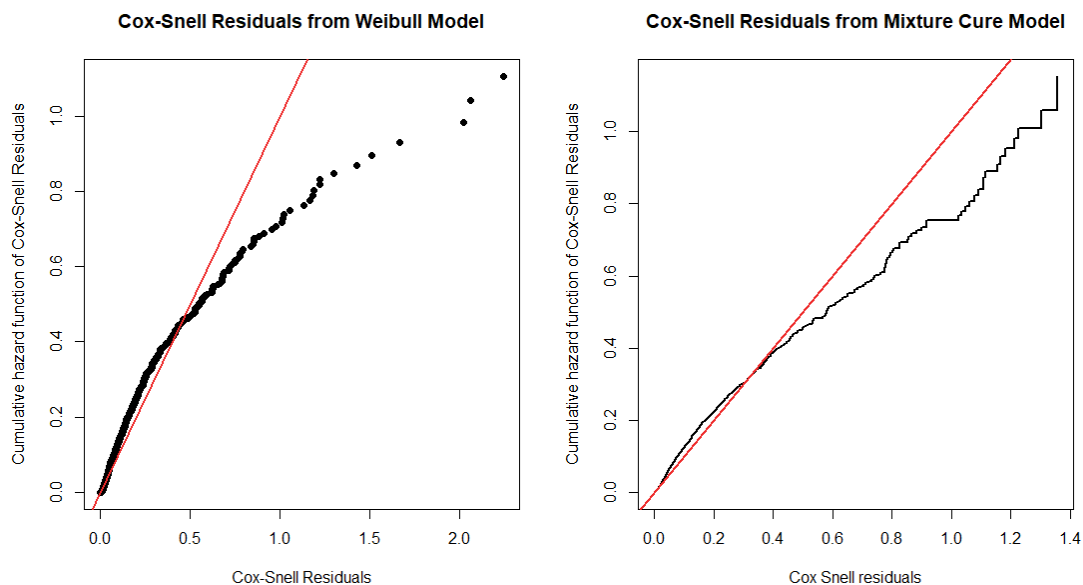
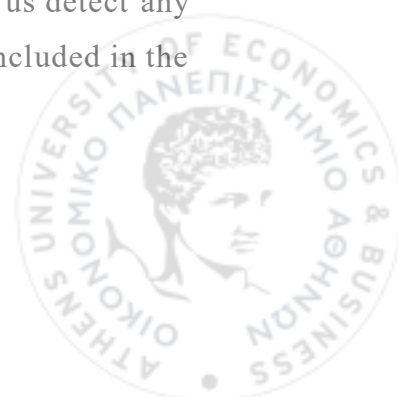


Figure 4.2 Cox-Snell Residuals for Weibull Model (left) and Mixture Cure Model (right) vs unit exponential line (red)

Several factors could explain these results. First, as mentioned earlier, the limited follow-up period could be a contributing factor. Additionally, if the Weibull distribution is appropriate for both the classical model and in the latency component of the mixture cure model, there may still be underlying heterogeneity not captured by these models. More complex models or modifications, such as including time-varying covariates, frailty terms or splines, could potentially enhance the performance of the models.

Using the Cox-Snell residuals, we also calculate the Martingale residuals for both models. Figures 4.3. and 4.4., show the estimated values of these residuals along with numeric variables from the dataset. These graphs help us detect any nonlinearity or assess whether additional interactions should be included in the model.



In the Weibull model, when a covariate is modeled appropriately, we expect the Martingale residuals for censored cases to be scattered around zero or negative values, while uncensored observations should be closer to one or negative. The graphs in the Weibull Model seem to be generally adequate, considering the significant amount of censoring in our dataset. However, they also highlight potential outliers with higher negative values.

In contrast, the mixture cure model in Figure 4.4 shows some improvement. Here, the spread of the Martingale residuals across the explanatory values is more consistent and the values have a tighter lower bound, which aligns with expectations since this model accounts for the cured fraction of borrowers.

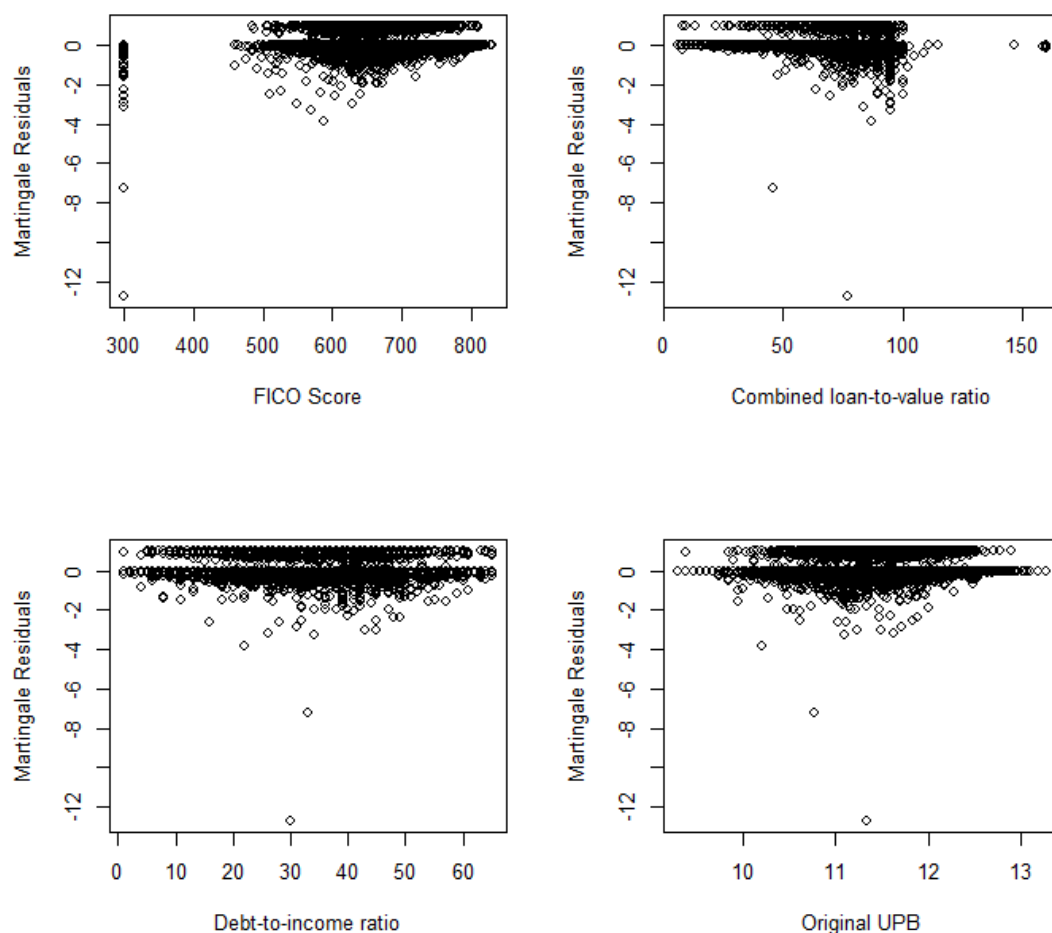
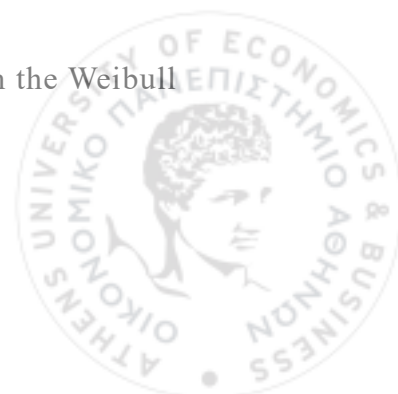


Figure 4.3 Martingale Residuals with explanatory variables from the Weibull Model



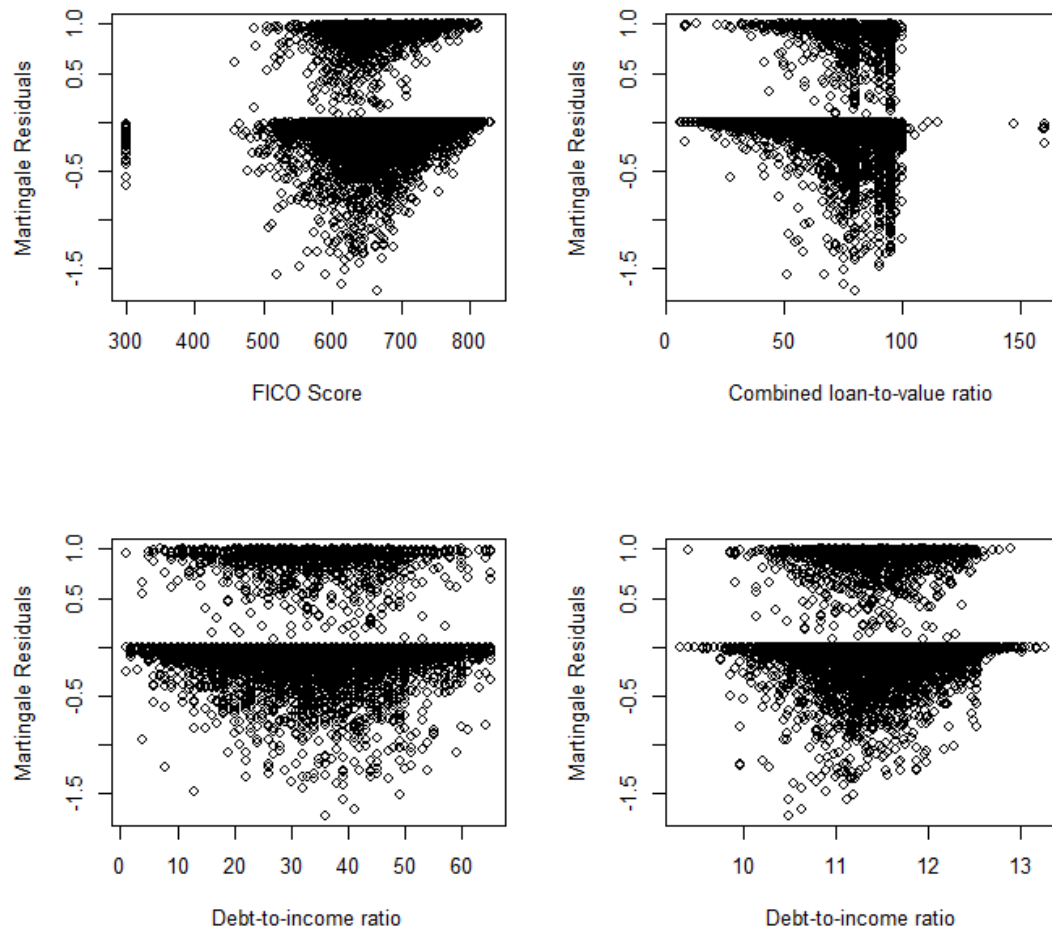


Figure 4.4 Martingale Residuals with explanatory variables from the Mixture Cure Model

4.4 Predictive Performance on Test Dataset

In addition to evaluating the goodness-of-fit of both the Weibull and the mixture cure model, it is also important to assess their predictive ability. Since banks are looking for tools that can accurately predict the default probabilities of borrowers, we investigate the models' performance in predicting these probabilities.

To do so, we perform a training and test analysis. Specifically, we split the dataset into two parts: the training dataset, which consists of 70% of the observations of the initial dataset, and the test dataset, which make up the



remaining 30%. The training dataset is used to fit both the Weibull model and the mixture cure model using the same set of variables as before.

Once the models are fitted, we use them to estimate the survival probabilities of the borrowers in the test dataset. To evaluate the predictive performance of each model, we derive the survival curve for the test observations using the Kaplan-Meier (KM) estimator and compare it with the survival curves generated by the models. The closer the two curves are, the better the models' predictions for the survival probabilities in the test set.

The results of the predictive analysis, showing the survival curve derived from the Weibull model (left) and the mixture cure model (right), are presented in Figure 4.5.

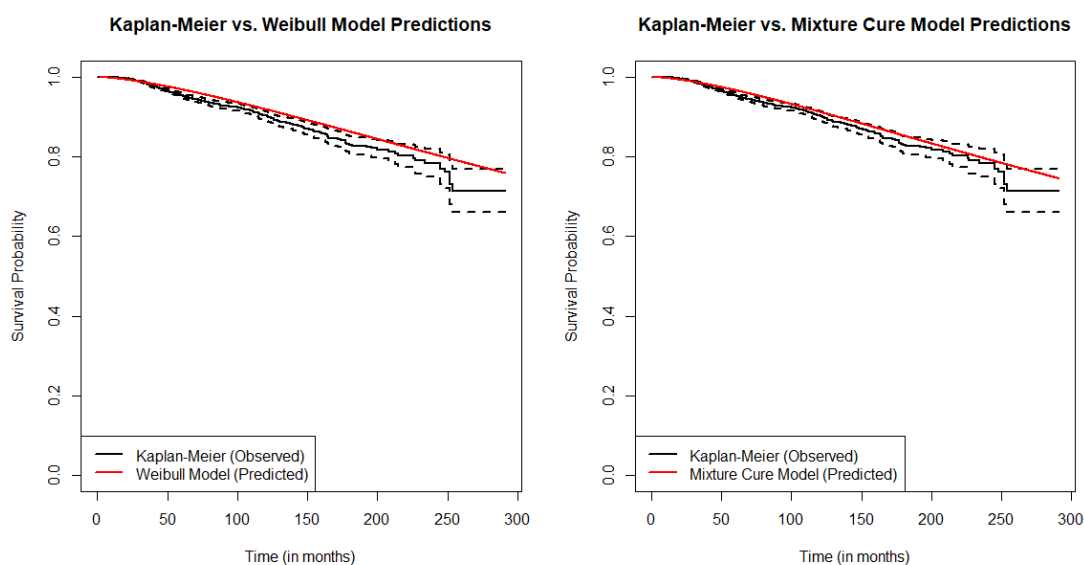
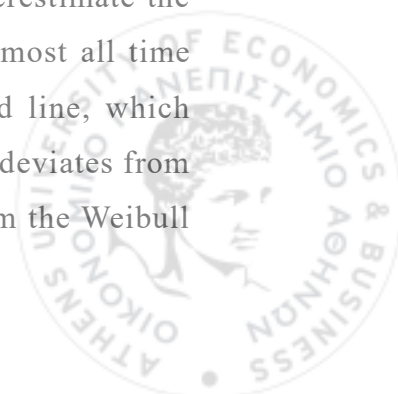


Figure 4.5 Kaplan-Meier curves with Predicted Survival curves from Weibull model (left) and Mixture Cure Model (right) for the test dataset

As shown in the graph, the classical Weibull model tends to overestimate the survival probabilities for borrowers in the test dataset across almost all time points, except in the early period. This is evident from the red line, which represents the survival estimates of the model, as it significantly deviates from the KM curve, indicating a poor fit. Moreover, the estimates from the Weibull



model generally fall outside the 95% confidence interval of the KM estimator for most of the time points, further highlighting the model's misfit.

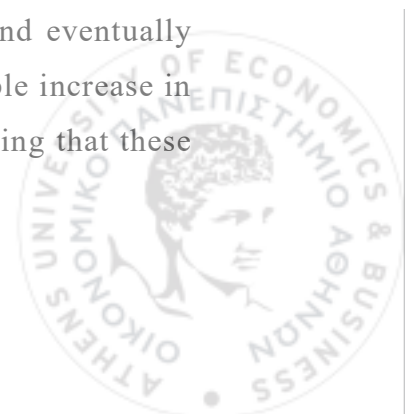
In contrast, the results generated by the mixture cure model appear to be more optimistic. The model's estimated survival curve of the test aligns much more closely with the KM curve over time, suggesting that it offers better predictive performance than the classical Weibull model, particularly in accounting for the cured subpopulation. Additionally, the 95% confidence interval of the KM estimator includes the survival probabilities estimated by the mixture cure model for most time points. Unlike the classical Weibull model, the mixture cure model provides more accurate predictions not only in the early stages of the loan, but also in later time periods, for the probability of not defaulting.

4.5 Cure Probability Analysis

In addition to providing survival probabilities for the susceptible group, the mixture cure model offers another key advantage over classical survival models. More specifically, it estimates the cure rate at different time points. This is particularly useful because, except for the probabilities of default, it also provides the probability that a borrower belongs to the cured group, meaning they will never default.

In our analysis, this cure rate is estimated through the logistic component of the mixture cure model. To demonstrate this, we use the same training and test datasets as before. After fitting the mixture cure model to the training set, we estimate the cure probabilities for the test dataset at different time points. Figure 4.6 presents the results of the probability of being cured for six randomly selected borrowers from the test dataset.

As shown in the figure, the probability of never defaulting increases over time for all borrowers. However, there are differences in the rate of cure among borrowers. For borrowers 1, 4 and 5 the probabilities of cure remain close to zero, suggesting that they are more likely to stay susceptible and eventually default. On the other hand, borrowers 2, 3 and 6 show a noticeable increase in these probabilities, with values rising to around 25-30%, indicating that these individuals are more likely to belong to the cured group.



We can also examine the average probability of being in the cured fraction across time for all borrowers in the test dataset (Figure 4.7). This provides an overall sense of how the probability of never defaulting, or being cured, evolves as time passes. For instance, at 70 months, the average probability of being cured is around 1%, while at 240 months, it increases to approximately 2%. This suggests that over time, a larger proportion of borrowers are likely to be “cured”, indicating a lower probability of default among borrowers as the loan progresses.

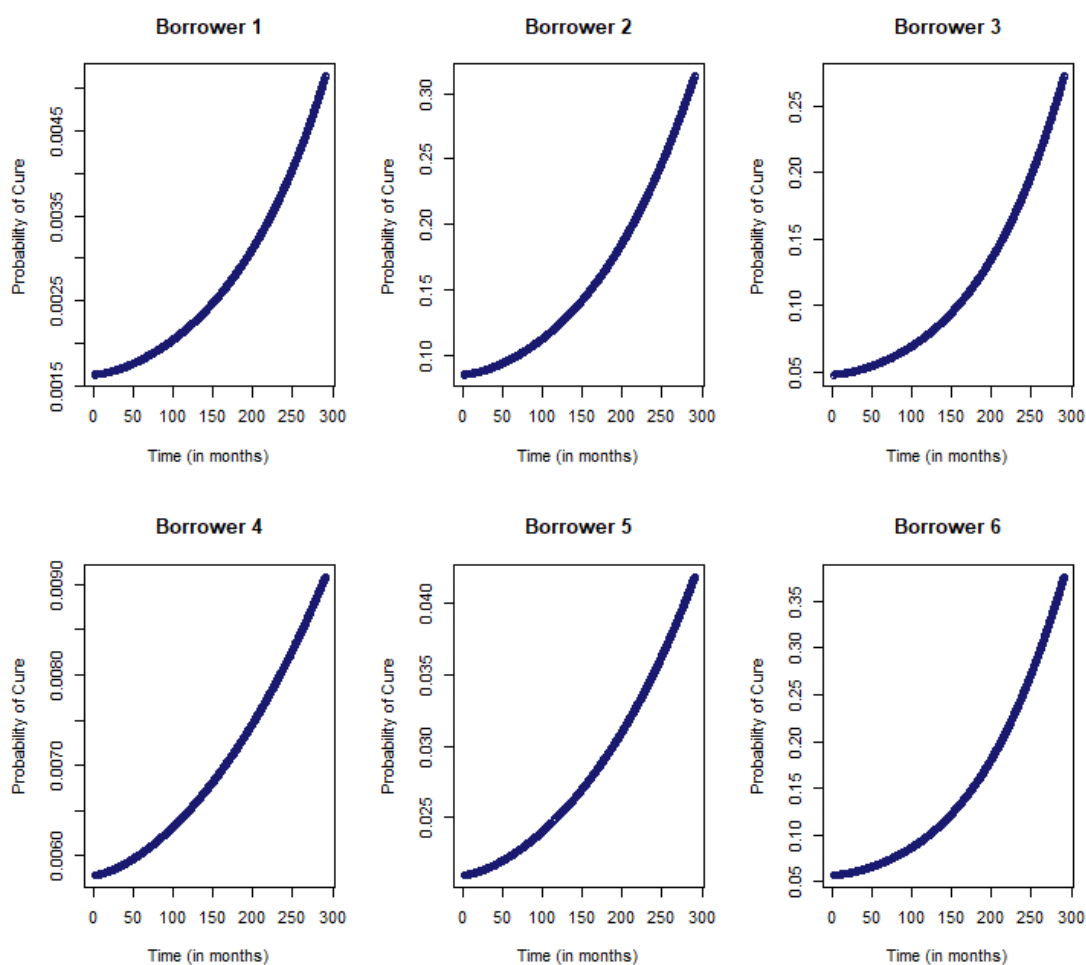


Figure 4.6 Probability of being cured for 6 random borrowers from the test dataset at different time points (in months)



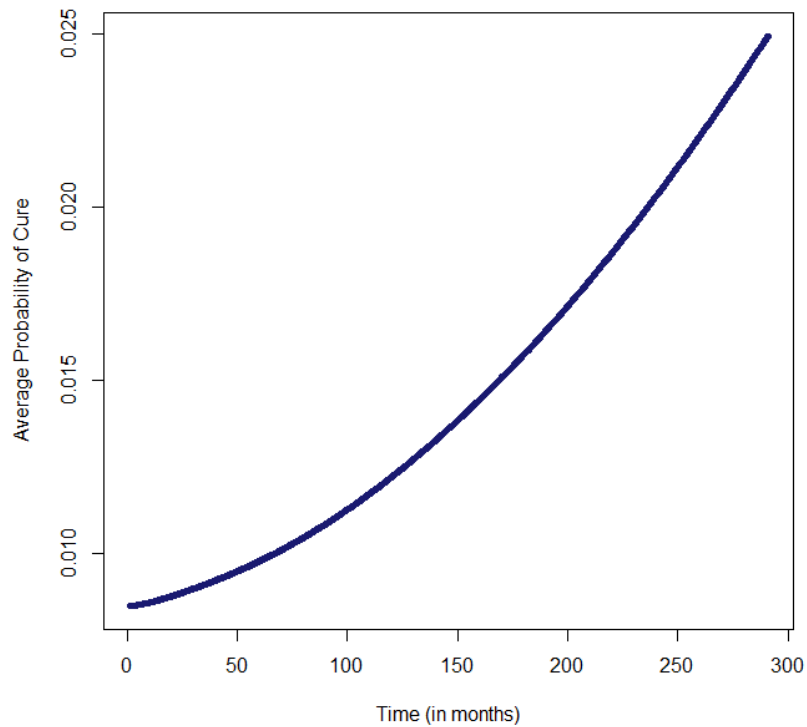


Figure 4.7 Average probability of cure for the test data at different time points (months)

Finally, in Figure 4.8, we present the estimated cure rate for borrowers in the test dataset. This rate represents the overall probability of an individual belonging to the “cured” subpopulation and is derived from the mixture cure model as a whole. As observed, some individuals exhibit a high cure rate, indicating a strong likelihood of being unsusceptible to default based on the estimated model and its explanatory variables.



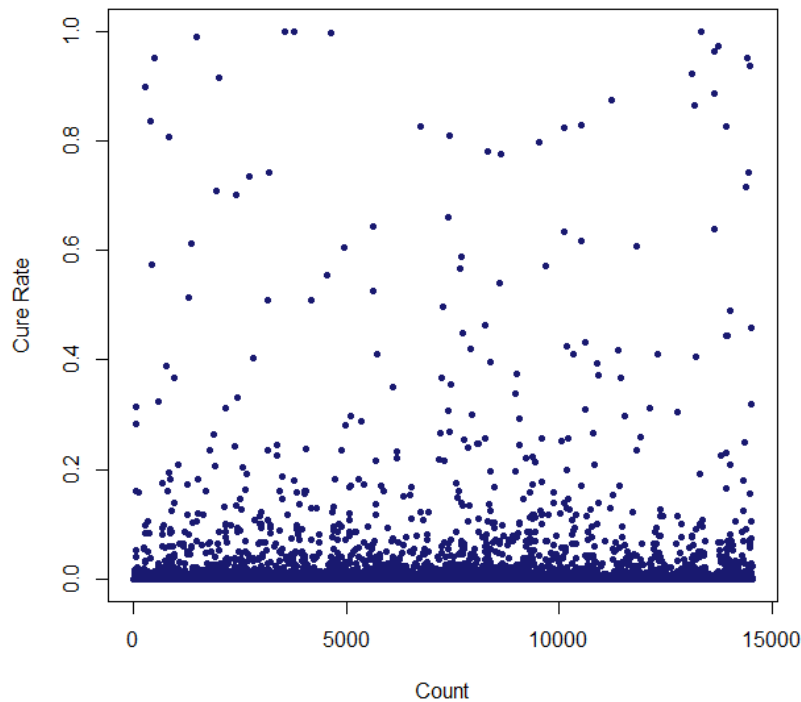
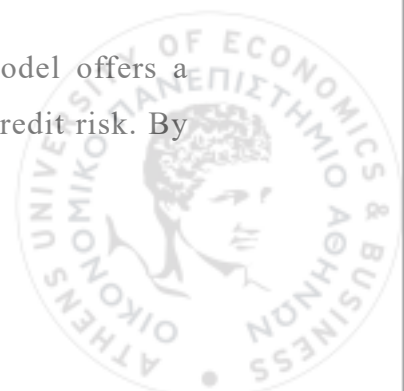


Figure 4.8 Cure rate for the test data

In this chapter, we have presented and analyzed the results from the classical Weibull survival model and the mixture cure model, which are used to estimate the time until loan default. While both models provide meaningful insights, the mixture cure model demonstrates a superior fit to the data, particularly by taking into account borrowers who may never default.

Evaluation through goodness-of-fit diagnostics, including Kaplan-Meier comparisons and Cox-Snell and Martingale residuals, confirmed that the mixture cure model more accurately captures the survival behavior of borrowers over time. Moreover, its ability to distinguish between cured and susceptible subpopulations offers valuable predictive advantages, particularly for long-term forecasting.

In conclusion, this chapter highlights how the mixture cure model offers a meaningful improvement over traditional methods in modeling credit risk. By



accounting for the cured fraction, the model provides deeper insights into borrower and enhances predictive accuracy.



Chapter 5

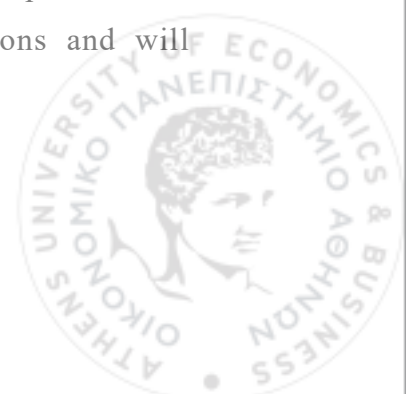
Conclusions, Discussions and Future Work

5.1 Summary of Findings

Default risk is a crucial topic which concerns banks, as it has a major impact on their financial stability and plays a key role in their liquidity adequacy. Banks are interested in distinguishing creditworthy individuals, as it helps them decide whether to issue a loan and determine details such as the interest rate and the loan amount. However, today's conditions have shifted their focus not only to whether a borrower might default but also to when they are more likely to do so. Therefore, estimating the probability of default at different time points is essential. To address this problem, survival analysis, a powerful tool for modeling the time until an event occurs, is proposed.

In this thesis, we present and implement two different approaches in survival analysis to estimate the time until a borrower defaults. More specifically, we propose a classical Weibull survival model and a mixture cure model with logistic and Weibull components. Since both models assume a specific distribution for the time until default, they fall into the category of parametric models. Our goal is to investigate and compare their performance in fitting the data well and predicting the time until a borrower defaults on their loan.

The first method assumes that all borrowers will eventually default on their loans after a sufficiently long (theoretically infinite) period of time. In our case, the time in months until the event follows a Weibull distribution. On the other hand, the mixture cure model suggests that there is a subpopulation of borrowers who will continue meeting their financial obligations and will remain “cured”, meaning they will never default.



The Weibull distribution is chosen for both models because it accounts for changes in the probability of survival over time, due to external and individual factors that influence the borrower's ability to repay. The shape parameter p captures changes in the hazard rate, where $p > 1$ indicates an increasing hazard and $p < 1$ a decreasing one.

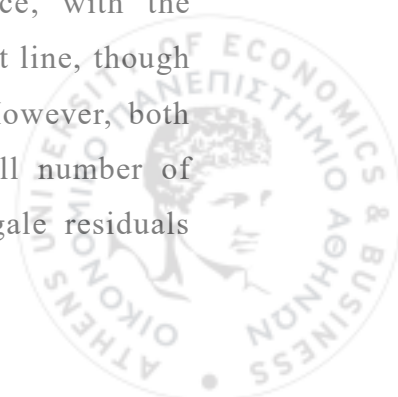
The scale parameter λ allows us to include variables that possibly influence the survival probabilities of the whole population and the susceptible subpopulation in the second case of the mixture cure model.

In the Weibull survival model, the backward variable selection method indicated that the best model includes all available explanatory variables. However, in the mixture cure model, the model with the lowest AIC eliminates mortgage insurance percentage from both the logistic and Weibull components.

Using the Expectation-Maximization (EM) algorithm, we estimate the coefficients β and the shape parameter p for both models. While all variables appear statistically significant in the classical Weibull model, this is not the case in the final mixture cure model, where many characteristics in both components do not have significant impact. In both models, the shape parameter p is estimated to be greater than 1, confirming that the hazard rate increases over time, i.e., borrowers become more likely to default as time progresses.

To evaluate model fit, we compare the survival curves derived from each model to the Kaplan-Meier estimator. The Weibull model aligns well with the Kaplan-Meier curve only in the early periods, but the distance between them becomes wider, indicating a poorer fit. In contrast, the survival curve from the mixture cure model remains closely aligned with the Kaplan-Meier curve across all time points, suggesting a superior fit.

In the same context, we examine Cox-Snell and Martingale residuals for both models. The Cox-Snell residuals suggest similar performance, with the cumulative hazard rate aligning closely with the exponential unit line, though the mixture cure model demonstrates a slight improvement. However, both models struggle in later time periods, likely due to the small number of remaining observations and increased variability. The Martingale residuals



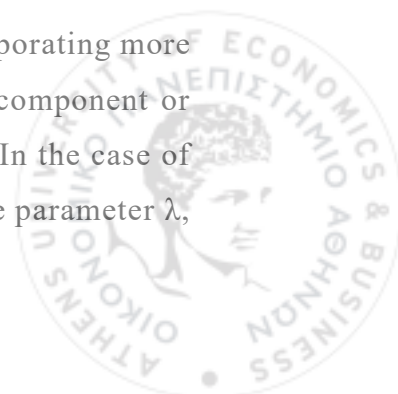
show no strong trends, indicating that the covariate effects are well-specified in both models, with the mixture cure showing slight improvements. In the Weibull model, there may be indications of potential outliers.

To evaluate predictive performance, we implement a train-test split, using 70% of observations for model fitting and the remaining 30% for testing. We then estimate survival probabilities for the test group and compare the predicted survival curves to the Kaplan-Meier estimator. The results further confirm the superiority of the mixture cure model. Its survival curve aligns closely with the KM curve across time and remains within the 95% confidence interval. On the other hand, the Weibull model only performs well in the early periods before diverging significantly. This suggests that the mixture cure model provides more accurate survival probability estimates over time.

Beyond its stronger predictive ability, the mixture cure model offers the additional advantage that we can estimate the cure rate and the probability of borrowers belonging to the cured subpopulation at different time points. In our case, we observe that some borrowers in the test group have a above 50%. Additionally, the probability of never defaulting, as estimated from the logistic component, increases over time. These insights are valuable for banks since they can refine their policy decisions and strengthen risk management strategies.

5.2 Discussion and Future Work

Based on this analysis, several challenges and potential extensions arise for further investigation. A natural next step would be to explore models beyond the parametric framework, which is more restricted. One possible alternative is the Cox-Proportional Hazards Model, which, while more flexible in fitting the data, introduces additional complexity. Additionally, comparing the mixture cure model with other parametric distributions apart from the Weibull, such as Generalized Gamma, log-logistic and log-normal, could provide further insights. The mixture cure model could also be enhanced by incorporating more advanced techniques, such as random forests for the incidence component or alternative models for the latency part, as previously suggested. In the case of the Weibull model, we included explanatory variables in the scale parameter λ ,

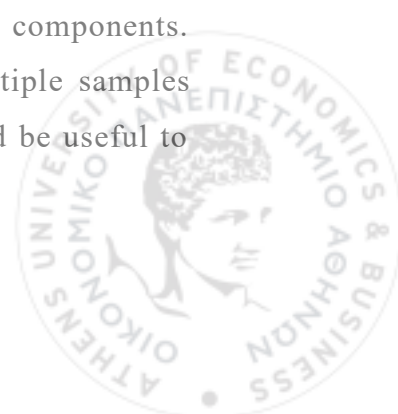


but an alternative could be to allow the shape parameter p to vary based on covariates, instead of assuming it remains constant.

Another important consideration is the censoring assumption. In our analysis, we treated observations as right-censored, assuming that the time until default or censoring is measured continuously. However, in reality, defaults are recorded at discrete intervals, suggesting that interval-censoring might be more appropriate. Instead of knowing the exact time of default, we only observe that it occurred within a specific time interval. Due to this discreteness, the standard survival analysis methods assuming continuous time might not be fully appropriate. Instead, models specifically designed for interval-censored data should be explored to better capture the nature of the data.

Regarding the dataset, while the analysis was conducted using approximately 50000 mortgage loans, it would be useful to expand it further both in terms of sample size and explanatory variables. Specifically, including loans with longer follow-up periods would improve the accuracy of estimates for later time points. Additionally, incorporating time-varying covariates could provide more accurate predictions, though it would also add complexity in interpretation. As for variable selection, while we applied a backward selection approach, a bidirectional method that includes both forward and backward steps may provide a clearer understanding of each variable's impact by examining more scenarios. Regularization methods, such as LASSO and Ridge Regression, could also be alternative selection methods.

For evaluating model performance, supplementing the goodness-of-fit measures used in this study with the Kolmogorov-Smirnov test could help compare the observed and predicted survival distributions more effectively. Further residual analysis, such as the Schoenfeld and Deviance residuals, are useful tools to detect any patterns or potential model misspecifications. In the case of mixture cure models, it would be insightful to investigate further the Cox-Snell residuals separately for the incidence and latency components. Within the prediction framework, validating results across multiple samples and assessing the accuracy of predictions in extrapolation would be useful to ensure the generalizability and reliability of the findings.



References

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5, 311-342. <https://doi.org/10.1146/annurev-statistics-031017-100101>

Ansin, E. (2015). An evaluation of the Cox-Snell residuals. *Master Thesis, Department of Statistics, Uppsala University*. <https://www.diva-portal.org/smash/get/diva2:826234/FULLTEXT01.pdf>.

Basel Committee on Banking Supervision (2004). International convergence of capital measurement and capital standards. *Bank for International Settlements*. <https://www.bis.org/publ/bcbs118.pdf>

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(12), 1699–1707. <https://doi.org/10.1057/jors.2008.130>

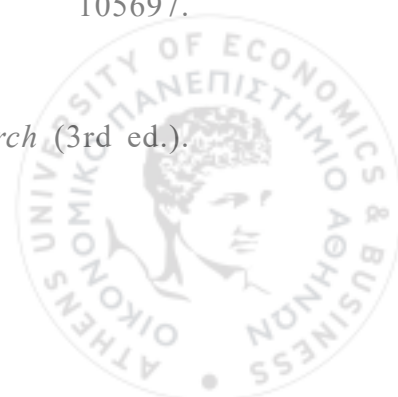
Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501-515. <https://doi.org/10.1080/01621459.1952.10501187>

Cao, R., Vilar, J. M., & Devía, A. (2009). Modelling consumer credit risk via survival analysis. *SORT: Statistics and Operations Research Transactions*, 33(1), 3–30. <https://ddd.uab.cat/record/97695>

Chen, S. & Yang, F. (2023). Expectation-Maximization Algorithm for the Weibull Proportional Hazard Model under Current Status Data. *Mathematics*, 11(23), 4826. <https://doi.org/10.3390/math11234826>

Choi, K., Park, S.M., Han, S., and Yim, D-S. (2020). A partial imputation EM-algorithm to adjust the overestimated shape parameter of the Weibull distribution fitted to the clinical time-to-event data. *Computer Methods and Programs in Biomedicine*, 196, 105697. <https://doi.org/10.1016/j.cmpb.2020.105697>

Collett, D. (2014). *Modelling Survival Data in Medical Research* (3rd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/b18041>



Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21, 411-421. <https://doi.org/10.1111/j.2517-6161.1959.tb00349.x>

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30, 248-275. <https://www.jstor.org/stable/2984505>

defi SOLUTIONS. (n.d.). Top lending risks for banks & how to prevent them. *defi SOLUTIONS*. Available at: <https://defisolutions.com/answers/top-lending-risks-for-banks-and-how-to-prevent-them/>

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: A benchmark study. *Journal of the Operational Research Society*, 68(6), 652–665. <https://doi.org/10.1057/s41274-016-0128-9>

Dvořák, L. (2021). How using data helps banks minimize the risks of loan default. *Big Data for Banking*. Available at: <https://bigdataforbanking.com/blog/how-using-data-helps-banks-minimize-the-risks-of-loan-default/>

Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041-1046. <https://doi.org/10.2307/2529885>

Fei, H., Kong, F., & Tang, Y. (1995). Estimation for two-parameter Weibull distribution and extreme-value distribution under multiply type-II censoring. *Communications in Statistics - Theory and Methods*, 24(9), 2087-2104. <https://doi.org/10.1080/03610929508831604>

Ferreira, L. and Silva, J. (2017). Parameter estimation for Weibull distribution with right censored data using EM algorithm. *Eksploatacja i Niezawodnosc - Maintenance and Reliability*, 19, 310-315. <https://doi.org/10.17531/ein.2017.2.20>

Freddie Mac. (n.d.). *Single-Family Loan-Level Dataset*. Retrieved from <https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset>



Gandy, A. (2012). Performance monitoring of credit portfolios using survival analysis. *International Journal of Forecasting*, 28(1), 139–144. <https://doi.org/10.1016/j.ijforecast.2010.08.006>

Haugh, M. (2015). The EM algorithm, *IEOR E4570: Machine Learning for OR&FE*. Spring 2015. Columbia University. https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf

J.P. and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56, 227-236. <https://doi.org/10.1111/j.0006-341X.2000.00227.x>

Jensen, R. K., Clements, M., Gjørde, L. K., & Jakobsen, L. H. (2022). Fitting parametric cure models in R using the packages cuRe and rstpm2. *Computer Methods and Programs in Biomedicine*, 226, Article 107125. <https://doi.org/10.1016/j.cmpb.2022.107125>

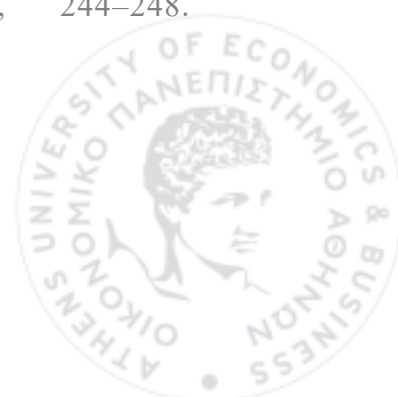
Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481. <https://doi.org/10.1080/01621459.1958.10501452>

Kuk, A.Y.C. and Chen, C.H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79, 531-541. <https://doi.org/10.1093/biomet/79.3.531>

Lu, J.C. (1992). Bayes parameter estimation for the bivariate Weibull model of Marshall–Olkin for censored data. *IEEE Transactions on Reliability*, 41(4), 608-615. <https://doi.org/10.1109/24.249597>

Lundborg, A. (2015). Survival analysis: A self-study introduction. *DIVA Portal*. <https://www.diva-portal.org/smash/get/diva2:826234/FULLTEXT01.pdf>

Mackisack, M.S., & Stillman, R.H. (1996). A cautionary tale about Weibull analysis. *IEEE Transactions on Reliability*, 45(2), 244–248. <https://doi.org/10.1109/24.510809>



Makalic, E. and Schmidt, D.F. (2022). Maximum likelihood estimation of the Weibull distribution with reduced bias. *arXiv preprint*, arXiv:2209.14567. Available at: <https://arxiv.org/abs/2209.14567>

Maller, R.A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79(4), 731–739. <https://doi.org/10.1093/biomet/79.4.731>

Peng, Y. and Dear, K.B.G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56, 237-243. <https://doi.org/10.1111/j.0006-341x.2000.00237.x>

Peng, Y. and Taylor, J.M.G. (2017). Residual-based model diagnosis methods for mixture cure models. *Biometrics*, 73, 495-505. <https://doi.org/10.1111/biom.12582Sy>

Tong, E., Mues, C. and Thomas, L. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218, 132-139. <https://doi.org/10.1016/j.ejor.2011.10.007>

Zhang, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. *Big-data Clinical Trial Column*. Available at: <http://dx.doi.org/10.21037/atm.2016.08.45>

