Comparison of Survival Analysis and Logistic Regression for Correlated Data

Niloofar Ramezani[1]
[1]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO, 80639

Abstract

Time is often modeled using Survival Analysis. The ability to consider the time element of event occurrences by proportional hazards models has meant that the logistic regression has played a less important role in the analysis of survival data (Abbott, 1985). This paper, however, discusses the situations in which the censored indicator can be modeled using Correlated Logistic Regression based on the binary nature of it and compares it to the model using Survival analysis which is rarely compared.

This paper presents a comparison between Survival Analysis models and Logistic regression models for both independent and correlated observations. Applying these methods, this paper presents an example of the length of stay in the transitional housing facility in the greater Rocky Mountain Region which is an autocorrelated longitudinal data that are subject to both left truncation and right censoring. The results are explained in terms of the comparisons of models mainly based on the significance of the independent variables.

Key Words: Survival Analysis, Logistic Regression, Correlated data, Longitudinal data, Proportional Hazard Models, Right Censoring

1. Introduction

Survival analysis is a branch of applied statistics concerning the sequential occurrences of incidents to model the time to a specific event applying the probabilistic laws (Liu, 2012). Survival analysis which is a combination of statistical methods for analyzing longitudinal data on the occurrence of events derives from the historical development of the field going back to mortality tables from hundreds of years ago. This method of analysis used to be applied when studying death at the beginning but nowadays, different methods in survival analysis is being applied in modeling time not only to death but also to any event that success or failure can be defined for it. This method of analysis has the advantage of being capable of dealing with censored and truncated data which arise when partial information about the random variable of interest is available. There might be some incompletion due to factors that are random for each subject that is called censoring or due to a selection process inherent in the study design which is known as truncation (Hosmer, Lemeshow, & May, 2008).

In the other hand, in logistic regression, we are interested in studying the association of the risk factors with the occurrence or nonoccurrence of an event. However, if the interest is on the effect of the risk factor or treatment on the time to event, logistic regression would not be appropriate anymore (Hosmer & Lemeshow, 2013).

When dealing with correlated data, by violating the assumption of the observations independence, some problems such as overestimation of the statistical significance and underestimation of variance may arise (Williams, 1995). The correlated measurements

add a complexity to the statistical model which requires some adjustments. Logistic regression and survival analysis assume independence of the observations; therefore, they won't work with the correlated data anymore (Bena & Mclntyre, 2008). There are different models that can be applied when dealing with correlated data and among those, shared frailty model can be applied to autocorrelated observations to generalize the survival model. In the other hand, Generalized Linear Mixed Model (GLMM) which is a particular type of mixed models is a useful approach to be applied instead of the regular logistic regression. Shared frailty model and the GLMM are comparable and so are applied to the data used for this study to compare their results in terms of significance.

2. Survival Analysis versus Logistic Regression

Unlike linear regression, survival analysis has a dichotomous outcome but is also different from the logistic regression. Survival analysis analyzes the time to an event which is an important distinction because this is what enables us to account for censoring which takes into account the fact that each subject has its own entry time into the study.
Survival analysis is known to do a better job than logistic regression when modeling the processes with multiple time measures. While logistic regression suffers an important difficulty when dealing with multiple time measures, survival analysis and specifically the proportional hazard model is very efficient in this situation considering both the time and censoring at the same time which is a major advantage (Wang, Brown, An, Yang, & Ligmann-Zielinska, 2013).

2.1 Survival models

Survivor and hazard functions are the two functions that are of central interest in survival analysis. As mentioned by Kleinbaum and Klein (2012), the survivor function $S(t)$ gives the probability that a person survives longer than some specified time t. Survivor function is defined as $S(t) = P(T > t)$ which gives the probability of the random variable T exceeding t.
In contrast to the survivor function which focuses on not failing, the hazard function focuses on failing so gives the opposite side of the information given by the survivor function. The hazard function, denoted by $h(t)$, is given by the formula below that is difficult to explain in practical terms. The easiest way to explain this function is as stated by Liu (2012) that the hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to the specific time t or the conditional failure rate which can be from 0 to infinity.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t},$$

where

$$P(t \le T < t + \Delta t | T \ge t)$$
$$= P(individual\ fails\ in\ the\ interval\ [t, t + \Delta t] | survival\ up\ to\ time\ t).$$

Due to the skewness and censoring issues when dealing with survival data, standard techniques, such as t-tests and linear models are not usually appropriate anymore. There are distribution-free or non-parametric, parametric, and finally, semi-parametric approaches for analyzing survival data. The main focus of this paper is on the semi-parametric, Cox's Proportional Hazards (CPH), model because it has a lot similarity with logistic regression and so is comparable to it based on the purpose of this paper (Kleinbaum & Klein, 2012).

The distinguishing feature of the CPH model, sometimes referred to as the Cox model, is its demonstration that the relationship between the hazard rate and explanatory variables could be estimated without making any assumptions about the shape of the baseline hazard function (cf. the parametric models). The result derives from the use of the proportional hazard assumption with many other insights and assumptions, and a partial likelihood method of estimation rather than maximum likelihood (Jenkins, 2005).
A linear model for the log-hazard or as a multiplicative model for the hazard in a parametric model based on the exponential distribution may be written as

$$log\, h_i\,(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik},$$
or, equivalently, $$h_i(t) = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik}).$$

In contrast to the parametric models, the Cox model, leaves the baseline hazard function $\alpha(t) = log\, h_0\,(t)$ unspecified and the model may be written as

$$log\, h_i(t) = \alpha(t) + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$
or, equivalently, $$h_i(t) = h_0(t) exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}).$$

Having two observations that differ in their x-values, with the corresponding linear predictors leaves us with the hazard ratio for these two observations that is independent of time t (Kleinbaum & Klein, 2012).

2.2 Logistic models

Logistic regression, which is a multivariate analysis model, is useful for predicting the presence or absence of a characteristic or an outcome based on values of a set of predictor variables. Through the addition of an appropriate link function to the usual linear regression model, the variables may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions (Hosmer & Lemeshow, 2013).
As explained by Hosmer and Lemeshow (2013), the relationship between the occurrence of any event and its dependency on different independent variables can be expressed as

$$p = {}^{1}\!/_{1 + e^{-z}},$$

where p is the probability of the occurrence of an event. Then, logistic regression fits an equation of the following form to the data

$$\mathrm{z} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

where β_0 is the model's intercept, $\beta_j{}'$s ($\mathrm{j} = 1, 2, \ldots, \mathrm{k}$) are the slope coefficients of the logistic regression model, and $x_{ij}{}'$s ($\mathrm{i} = 1, 2, \ldots, \mathrm{n}$; $\mathrm{j} = 1, 2, \ldots, \mathrm{k}$) are the independent variables.
In logistic regression, the probability of the outcome is measured by the odds of occurrence of an event. Change in probability is not constant (linear) with constant changes in X. This means that the probability of a success given the predictor variable is a non-linear function, specifically a logistic function.
The most common form of logistic regression uses the logit link function which gives us the logistic regression equation as

$$logit\,(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

3. Correlated Data

As mentioned in Bena and Mclntyre (2008), an added complexity occurs when not all of the observations are independent. Often subjects may have multiple interventions of the same type at the same time point. When there are multiple observations per subject in a model, standard errors of the survival estimates underestimate the true amount of variability that exists.

3.1 Generalization to survival models

Using a method that employs Taylor series linear approximations makes the necessary adjustments of the standard errors after obtaining Kaplan-Meier estimates which are non-parametric estimates. Williams (1995) described a method based upon Taylor series approximations of the survival estimates for each observation followed by applying the between-cluster variance estimator that is mostly used in multi-stage surveys.

Employing adjustments by using a sandwich estimator of standard errors by Lin (1994) can be also used to fit a model which is referred to as a marginal model.

One of the other methods that can be used in this situation is based on the Jackknife variance estimator discussed by Lipsitz and Parzen (1996). Wei, Lin, and Weissfeld (1989) and Lee, Wei, Amato, and Leurgans (1992) showed that if the marginal distributions of the correlated survival times follow a proportional hazards model, then the Cox's partial likelihood gives estimates that naively treat the correlated survival times as independent and therefore give consistent estimates of the relative risk parameters. However, because there still is correlation between survival times, the inverse of the information matrix may not be a consistent estimate of the asymptotic variance. Wei et al. (1989) and Lee et al. (1992) proposed a robust variance estimate that is consistent for the asymptotic variance. They showed that a "one-step" jackknife estimator of variance is asymptotically equivalent to the other variance estimator.

In 2000, Florin and Ronghui, proposed a general proportional hazards model with random effects for handling clustered survival data that works better in terms of comparison with the generalized logistic regression. They generalized the usual frailty model by allowing a multivariate random effect with an arbitrary design matrix in the log relative risk, in a way similar to the modelling of random effects in linear, generalized linear, and non-linear mixed models. This model is sometimes called Shared Frailty Survival model. The random effects are generally assumed to have multivariate normal distribution, but other (preferably symmetrical) distributions are also possible.

Shared frailty survival model accounts for heterogeneity and random effects and captures the stochastic dependence by allowing the Gaussian random effects of the linear model to be correlated with the frailty term of the CPH model (Philipson et al., 2012) which can be expressed as

$$h_{ij}(t) = h_0(t)\exp\left(\beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + g_i\right),$$

where observations j are in cluster i, and g_i is typically normal with mean 0 and also g_i is uncorrelated with $g_{i'}$.

3.2 Generalization to logistic models

The modelling of correlated binary outcomes in a way that the marginal response probabilities are still logistic has been discussed in different articles along with the association measures for the dependence between correlated observations. For paired correlated data, the full likelihood can be evaluated and for an arbitrary number of

correlated observations, a pseudo likelihood approach for obtaining parameter estimates is proposed by Cessie and Houwelingen (1994). A discussion of the various approaches to model correlated binary observations can be found in Prentice (1988), Zeger, Liang, and Albert (1988), and also in Neuhaus, Kalbfleisch, and Hauck (1991).

In the population average model, estimation is based on Generalized Estimating Equations (GEE). Liang and Zeger (1986) and Zeger et al. (1988) first used the GEE with the binary data population average model. The set of equations used in the GEE approach look like weighted versions of the likelihood equations. Requiring an assumption about the structure of this correlation, the weights involve an approximation of the underlying covariance matrix of the correlated within-cluster observations. Under the independent model, $Cor(Y_{ij}, Y_{i1}) = 0$ for $j = 1$ and the GEE equations simplify to the likelihood equations obtained from the binomial likelihood (Hosmer & Lemeshow, 2013).

The correlation among responses depends on the lag between the observations and is assumed to be constant for equally lagged observations. Settings where there is an explicit time component are more specialized and need additional approaches to handling such data covered in texts such as Diggle et al. (2002, as cited by Hosmer & Lemeshow, 2013) or Hedeker and Gibbons (2006, as cited by Hosmer & Lemeshow, 2013).

In the unstructured correlation case, one assumes that the correlation of the possible pairs of responses is different, $Cor(Y_{ij}, Y_{i1}) = \rho_{j1}$ for $j = 1$. The disadvantage of using this method is that it requires estimating a large number of parameters that are, for the most part, of secondary importance. In most applications researchers are only interested in estimating the regression coefficients and need to account for correlation in responses to obtain the correct estimates of the standard errors of the estimated coefficients. The idea is to choose a correlation structure for estimation that seems plausible for the setting and then this structure is used in adjusting the variance's estimator (Hosmer & Lemeshow, 2013).

For data without a clear choice of structure, a reasonable and parsimonious choice is the "exchangeable correlation" structure. One of the advantages of the GEE approach is the "robustness" of the estimates to choice of correlation. In other words, even if the correlation structure chosen is not the true structure, the parameter estimates from the GEE are often still valid.

Pseudo likelihood estimation and related estimation techniques are also very helpful if the full underlying distribution of the data is unknown or if the true likelihood is difficult to evaluate. The pseudo-likelihood method is a very easy method to understand and the multivariate correlation model has the advantage that estimates of the joint probabilities can be generated relatively easily. There is some loss in efficiency by using pseudo-likelihood but because it equals the full likelihood for $p = 0$, only small losses are expected when p is small (Cessie & Houwelingen, 1994).

One of the best models to use when dealing with the correlated data is the GLMM which is a particular type of mixed model. It is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. These random effects are usually assumed to have a normal distribution. In the GLMM, the default optimization technique that is used is the Quasi-Newton method. Because a residual likelihood technique is used to compute the objective function, only the covariance parameters are participating in the optimization. This model is not complicated and more details about it can be found in Agresti (2007). The GLMM can be written as

$$\eta = X\beta + Z\gamma,$$

where link function is $g(.) = log_e(\frac{p}{1-p})$.

4. Example: Transitional Housing Facility

Data are presented from a transitional housing facility (THF) in the greater Rocky Mountain Region that works as a temporary facility that helps family units find stable housing and employment while living within the shelter. The THF aims to help these first-time and episodic (experiencing episodes of homelessness) homeless families regain stable housing and employment with the goal of mitigating long-term occurrence of the problem. Data analysis was conducted using SAS 9.3 (The SAS Institute, Cary, NC).

This research outlines the changing condition of the homeless in terms of the time it takes till finding a job and leaving the transitional house based on the job training they had at the center, the number of hours they spent with case worker each month, either they had temporary assistance for needy families or not, and some demographic variables that seem effective on the length of their stay at the transitional house like child abuse, number of children in the family, being either a single parent or not, and finally being either unemployed or employed for families living in the THF from 2006 to 2010.

To apply the survival analysis and fit logistic regression to this dataset, I defined 60 days as the end point because families were supposed to leave the transitional house within 60 days while not everyone did so. The ones who didn't leave in 60 days are considered censored because the focus of the study is 60 days of stay. The number of nights they stayed at the THF is used to define the censoring status; if this number is larger than 60, then the length of stay is defined as censored (censored = 0) for the related family, but if this number for a family is less than 60, the length of stay for that family is defined as non-censored (censored = 1). For the regular survival analysis and logistic regression the numbers of nights each family stayed at the THF in different months are aggregated that gives us one observation for each family creating independent observations in the dataset.

The non-parametric Kaplan-Meier (KM) method is applied to the aggregated data including 415 families. The results of the log-rank-test for equal survival-functions within the KM analysis shows that there is not a significant difference between the survival functions of families facing child abuse and the ones who haven't faced this issue (p-value = .148), between families who were single parent or not (p-value = .175), between unemployed or employed families (p-value = .767), among families with different number of children (p-value = .586), and between the families who received temporary assistance for needy Families or not (p-value = .0516). But there is a significant difference at the significance level of .05 among the survival functions of families with different numbers of hours spent with the case worker within the THF (p-value < .0001) and also families with different numbers of time they went through some job training at the THF (p-value < .0001). These results from the KM method are more descriptive and not used for making the final comparisons which is the main purpose of this study.

Applying the semi-parametric survival analysis, Cox regression model, to the aggregated data revealed that four of the seven variables that were chosen to be in the model because of their importance which are mentioned above were statistically significant in this model. The Cox regression model itself was significant with the P-value less than .0001. Based on the results from the CPH model which are shown on table 1, variables that are significant at the .05 significance level are child abuse (Abuse; p-value = .0277), unemployment (Unemployed; p-value = .0252), number of times families went through job training within the transitional house (Job_sum; p-value = .0002), and finally the number of hours families spent with the case worker within the transitional house (Case_sum) which was highly significant with the p-value less than .0001.

In order to check the potential outperformance of the CPH model in comparison to the logistic regression, the Cox regression's results need to be compared to the logistic regression's results in terms of the significance of variables.

Table 1: Results of CPH model analysis

Parameter	*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*	*Hazard Ratio*
TANF	1	-0.15879	0.16198	0.9611	0.3269	0.853
Abuse	1	-0.24750	0.11242	4.8472	0.0277	0.781
Num_Children	1	-0.04614	0.04232	1.1886	0.2756	0.955
Unemployed	1	0.24980	0.11157	5.0129	0.0252	1.284
Single_Parent	1	0.17972	0.12026	2.2334	0.1351	1.197
Job_sum	1	-0.26633	0.07075	14.1717	0.0002	0.766
Case_sum	1	-0.40040	0.01883	451.9504	<.0001	0.670

The logistic regression model is also applied to the aggregated data looking into the censoring variable as the binary variable showing whether the family stayed in the shelters after 60 days or as the THF wanted, they left within 60 days. The likelihood-ratio test (p-value < .0001) and Wald test (p-value < .0001) showed that the logistic regression model is significant. The model fit is also good based on Hosmer and Lemeshow goodness of fit test (p-value = .9923). To check the association of predicted probabilities and observed response, Somer's D value was calculated which showed that the proportion of variance explained by the variables in the model is 95% which is significantly high. Based on analysis of maximum likelihood estimates on table 2, only the numbers of hours families spent with the case worker within the THF is significant (Job_sum; p-value < .0001) at the significance level of .05.

Table 2: Results of logistic regression analysis

Analysis of Maximum Likelihood Estimates Logistic Regression

Parameter		*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept		1	14.3722	2.2156	42.0805	<.0001
TANF	*0*	1	0.1554	0.6818	0.0520	0.8197
Abuse	*0*	1	0.0539	0.5116	0.0111	0.9161
Num_Children		1	-0.0865	0.1925	0.2020	0.6531
Unemployed	*0*	1	-0.7524	0.5134	2.1482	0.1427
Single_Parent	*0*	1	1.0044	0.5529	3.3001	0.0693
Job_sum		1	-0.2633	0.2151	1.4976	0.2210
Case_sum		1	-0.6833	0.1029	44.0752	<.0001

As expected the results from the logistic regression are different from the ones from the CPH in terms of the number of significant variables in the model. Fewer variables are significant applying the logistic regression model in comparison to the CPH model.

Methods applied above assume independence of the observations. As mentioned in the previous section, correlated measurements require adjustment to avoid underestimation of the variance and overestimation of the statistical significance (Bena & Mclntyre, 2008). An added complexity occurs when not all the observations are independent. This should be addressed by applying appropriate models to the correlated data from the THF which does not satisfy the independence assumption anymore because of not aggregating different observations of the same family over time for the second analysis to keep the repeated measure nature of the data that gives more observations including 926 families.
To generalize the survival model in order to work with the correlated data, shared frailty model is applied to the correlated observations. 75.5% of the observations were censored. The shared frailty model is significant based on the likelihood ratio test (Chi-square = 95.5112, p-value < .0001) and Wald test (Chi-square = 85.2975, p-value < .0001) results. Finally based on table 3, the results from the shared frailty model analysis of maximum likelihood estimate, number of hours families spent with case worker each month is highly significant (Case_Hours; p-value < .0001); also employment is significant (Unemployed; p-value = .0320) at the significance level of .05.

Table 3: Results of the shared frailty analysis

Analysis of Maximum Likelihood Estimates						
Parameter	*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*	*Hazard Ratio*
Child_Abuse	1	-0.02981	0.17429	0.0292	0.8642	0.971
Case_Hours	1	-0.24826	0.03202	60.1114	<.0001	0.780
Job_Training	1	0.11611	0.15936	0.5308	0.4663	1.123
Single_Parent	1	0.07826	0.14659	0.2850	0.5935	1.081
Unemployed	1	-0.34605	0.16133	4.6008	0.0320	0.707
Num_Children	1	0.07333	0.04907	2.2334	0.1351	1.076
TANF	1	0.17635	0.18793	0.8806	0.3480	1.193

As mentioned before to generalize the logistic regression for the autocorrelated data, one of the best models is the GLMM which is a particular type of mixed model. A lower boundary constraint is placed on the variance component for the random center effect. The solution for this variance cannot be less than zero. After the initial optimization, the GLIMMIX procedure performed 16 updates before the convergence criterion was met. At convergence, the largest absolute value of the gradient was near zero which indicates that the process stopped at an extremum of the objective function. This model is statistically significant as twice the negative of the residual log likelihood in the final pseudo-model equaled 4764.52. The ratio of the generalized chi-square statistic and its degrees of freedom is close to 0.3. This is a measure of the residual variability in the marginal distribution of the data. From the covariance parameter estimation procedure, the variance of the random center intercepts on the logit scale is estimated as $\hat{\sigma}_c^2$=5.2757. Finally looking into table 4 to check the significance of different variables, it is obvious that the number of hours families spent with case worker each month is highly significant (Case_Hours; p-value = .0032) and employment is also significant at the .05 significance level (Unemployed; p-value = .0211). After doing the comparison, the results in terms of significance from both the frailty models and the GLMM are the same.

Table 4: Results of the GLMM analysis

Solutions for Fixed Effects					
Effect	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*
Intercept	3.2436	0.5461	413	5.94	<.0001
Child_Abuse	-0.2864	0.4396	505	-0.65	0.5150
Case_Hours	-0.1541	0.05205	505	-2.96	0.0032
Job_Training	-0.6988	0.4233	505	-1.65	0.0994
Single_Parent	-0.03898	0.3794	505	-0.10	0.9182
Unemployed	0.7847	0.3391	505	2.31	0.0211
Num_Children	-0.1358	0.1308	505	-1.04	0.2998
TANF	-0.3568	0.4594	505	-0.78	0.4378

5. Conclusion

In this paper a comparison between survival analysis models and logistic regression models is presented along with the comparison of the generalized methods when dealing with the correlated data applying the shared frailty model and the GLMM.

Due to the real situations dealt with when analyzing the THF data and based on the lack of studies comparing the logistic regression and survival analysis, this study focused on this comparison. The biggest gap was the comparison of these two methods when there is clustering and so some correlation in the dataset among the subjects. Although there are many different methods of analyses for both models such as the methods based on the Jackknife estimation, Tayler series, and frailty models to handle survival models and the GEE or mixed models for Logistic models, not all of them are comparable and there should be methods applied to both analyses that make the results equivalent. Based on the review of literature for this study and the data analysis, the best way to do this comparison for correlated data is to add a random term to both models so the analyses would be comparable. The appropriate model as an extension to logistic regression is generalized linear mixed model and the extension to the Cox survival model is made by adding a random effect to the model which is called shared frailty model and they are comparable as discussed above.

Analyzing the real data approved getting different results using logistic regression in comparison to survival analysis for the aggregated data as expected since survival model performed better by not losing information about length of stay in the THF. For the correlated data, getting the same results based on the different nature of the two models applied to the dataset is promising to be able to finally make the distributional comparisons.

Ongoing research includes the distributional comparisons through simulation study and some data analysis

References

Abbott R. D. 1985. Logistic regression in survival analysis. *American Journal of Epidemiol*.121(3):465-71.

Agresti, A. 2007. An Introduction to Categorical Data Analysis. *Willey Series in Probability and Statistics*, second edition.

Bena, J., McIntyre, Sh. 2008. Survival Methods for Correlated Time-to-Event Data.

Florin, V., Ronghui, X. 2000. Proportional hazards model with random effects, *Statistics in Medicine*, Volume 19, Issue 24, pages 3309–3324.

Hosmer, D., Lemeshow, S. 2013. Applied Logistic Regression. *Willey Series in Probability and Statistics*, third edition.

Hosmer, D., Lemeshow, S., and May, S. 2008. Applied Survival Analysis. Regression Modeling of Time-to-Event Data . *Willey Series in Probability and Statistics*, second edition.

Jenkins, S. P. July 2005. Survival Analysis, *Institute for Social and Economic Research, University of Essex.*

Kleinbaum, D., Klein, M. 2012. Survival Analysis. A Self-Learning Text, *Springer series*, third edition.

Le Cessie, S., and Van Houwelingen, J. C. 1994. Logistic regression for correlated binary data. *Applied Statistics*, 95-108.

Lee, E. W., Wei, L. J., Amato, D. A., and Leurgans, S. 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival analysis: state of the art* (pp. 237-247). Springer Netherlands.

Liang, K. Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Lin, D. Y. 1994. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine*, *13*(21), 2233-2247.

Lipsitz, S. R., and Parzen, M. (1996). A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics*, 291-298.

Liu, X. 2012. *Survival analysis: models and applications*. John Wiley & Sons.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, 25-35.

Philipson, P., Diggle, P., Sousa, I., Kolamunnage-Dona, R., Williamson, P., and Henderson, R. 2012. joineR: Joint modelling of repeated measurements and time-to-event data. http://cran.r-project.org/web/packages/joineR/vignettes/joineR.pdf. [Last checked: 07.07.2014].

Prentice, R. L. 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

Wang, N., Brown, D. G., An, L., Yang, S., & Ligmann-Zielinska, A. 2013. Comparative performance of logistic regression and survival analysis for detecting spatial predictors of land-use change. *International Journal of Geographical Information Science*, *27*(10), 1960-1982.

Wei, L. J., Lin, D. Y., & Weissfeld, L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, *84*(408), 1065-1073.

Williams, R. L. 1995. Product-limit survival functions with correlated survival times. *Lifetime data analysis*, *1*(2), 171-186.

Zeger, S. L., Liang, K. Y., & Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.