Survival Analysis Applied to Agricultural Sciences

Jung Ae Lee
Agricultural Statistics Laboratory
University of Arkansas, Fayetteville, AR 72701

September 11, 2018

Abstract

Time-to-event outcomes are common in agricultural sciences. For example, how long it takes until flowering is one of the critical research questions for plant scientists. Despite the popularity of survival analysis in medical studies for past decades, application in agricultural sciences has been less discussed. The main strength of this method is their ability to handle missing data over time, namely, right-censored data. Even well-designed experimental data may encounter drop-out, which can be ignored in methods such as analysis of variance (t-test). Survival models also tend to have greater statistical power to detect a significant treatment effect than methods for binary response such as logistic regression. The goal of this study is to review basic concepts of survival analysis, importantly to discuss the benefit of this method when it comes to agricultural research applications. Cox regression and alternative regression models are compared to demonstrate the advantage (or disadvantage) of each method through both simulation study and real data examples.

Key words: survival analysis, cox regression, logistic regression, agriculture

1 Introduction

The survival analyses, such as Kaplan-Meier estimation, Cox regression, and Accelerated failure time model, have achieved great popularity in epidemiological and medical studies for past decades. However, agricultural applications are rarely found in literature (e.g., plant or animal studies) simply because the nature of experimental data fits other method such as analysis of variance (t-test). Agricultural sciences are evolving and diverse experimental designs bring the need of diverse statistical analysis as appropriate. In particular, when "repeated measures" over time is involved in an experimental setting, the choice of statistical methods is neither trivial nor unique. This study explores survival analysis to investigate time-to-event response where regular long-term follow-up is allowed. Examples are various such as time till damage of seed in storage and time till clean of water. This paper is organized as follows: Section 2 illustrates two examples of agricultural studies that survival analysis can be useful, Section 3 reviews the survival analysis, Section 4 compares the cox and the logistic regressions, and in Section 5, we discuss the practical reasons for cox or logistic regression models, respectively.

2 Illustrative Examples

2.1 Soybean storage

Soybean farmers are interested in knowing, how long the stored seed can maintain fine condition for planting, whether some treatment group perseveres longer than others, and what environment, such as temperature, is needed to preserve the quality of the seeds. To analyze this, define a critical event, *damage* if accelerated aging is less than 60 out of 100, due to any cause or a specific cause. If necessary, other seed vigor measure(s) can be used, e.g., damage = 1 if germination is under 70, otherwise 0. Therefore,

> The *survival time* of soybean seed is the time in storage until seed vigor falls below a "damage" threshold or censoring,

which is depicted in Figure 1. A bag (1kg) is the experimental unit, and total 116 bags are investigated under different year, cultivar, and warehouse as a factorial design. Researchers regularly checked up on seed condition of each bag during 18 weeks, and a binary response of damage=1 or 0, can be recorded accordingly. Therefore, 116 bag's survival time can be expressed as two weeks interval up to 18 weeks (or 126 d). In Figure 1 (a), a bag's survival time is 28 days, whereas, in (b), some other bag's survival time is 84 days, showing better maintenance. Would this gap come from different cultivar, environment, or simply initial quality?

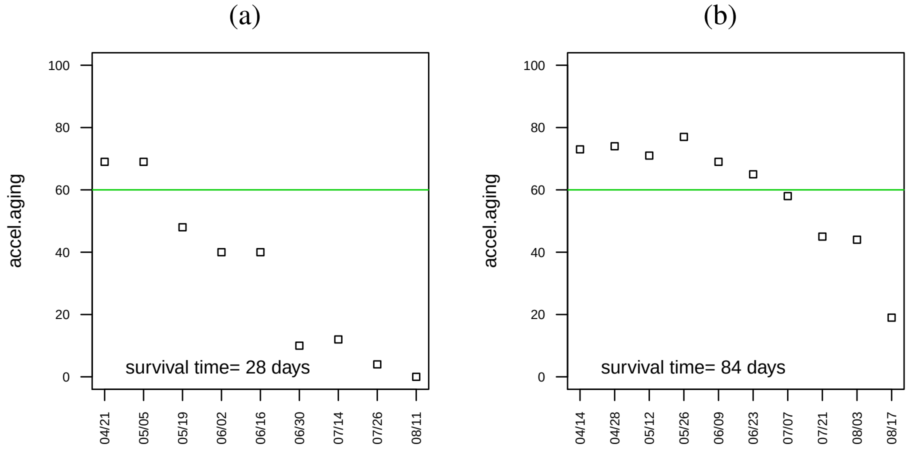


Figure 1: Illustration of survival time of two different seed bags

2.2 Microbe survival in mesocosm

Water quality has been gaining attention recently in Arkansa area in the United States concerning food safety and public health. Researchers in the Department of Food Science at the University of Arkansas conducted a mesocosm experiment where the concentration of 18 kinds of microbes are measured in different conditions. Presence (or amount) of microbe such as E.coli can be an indication of whether (or how much) water is contaminated.

Survival analysis can be applied to answer the following questions: how long it takes till die off microbes, whether seasonal or environmental factors affect the survival (or death) of microbes, and whether there are specific microbes that are predictive of each other in their presence. Let us define a critical event as follows: *death* = 1 if microbes are not detected, zero LOG10 CFU or PFU remaining in a sample, otherwise 0. Accordingly,

> The *survival time* of microbes is the time in mesocosm until the first detection of zero LOG10 CFU or PFU or censoring,

A mesocosm sample is an observation unit. Total 464 samples are collected from 18 different microbes, two mesocosm types (river and lake), two sample types (water and sediment), and four seasons. Our inspection occurred daily for seven days, then in seven days interval up to 28 days. In some cases, we stopped the inspection at 14 or 21 days because it was "dead" or close to be dead. Thus, the survival time points that we observe are:

1, 2, 3, 4, 5, 6, 7, 14, 14+, 21, 21+, 28, 28+

Here 14+ indicates censoring, that is, the inspection stopped at 14 days while microbes are still alive (we don't know the future but what we can say is that it survived more than 14 days).

3 Why Survival Analysis?

3.1 Types of response

Survival analysis is used for time-to-event outcomes. In addition to previous examples, the examples vary, for instance how long it takes until flowering among different genotypes of plants, time-to-disease of calves after exposure to certain viruses. Experimental data tend to be collected by regular follow-up, rather than just record the end point, survival time. Therefore, the repeated measures of binary response and possibly explanatory variable are available as well. In such situation, both logistic regression and cox regression are appropriate. These two methods are compared in this work.

3.2 Censoring

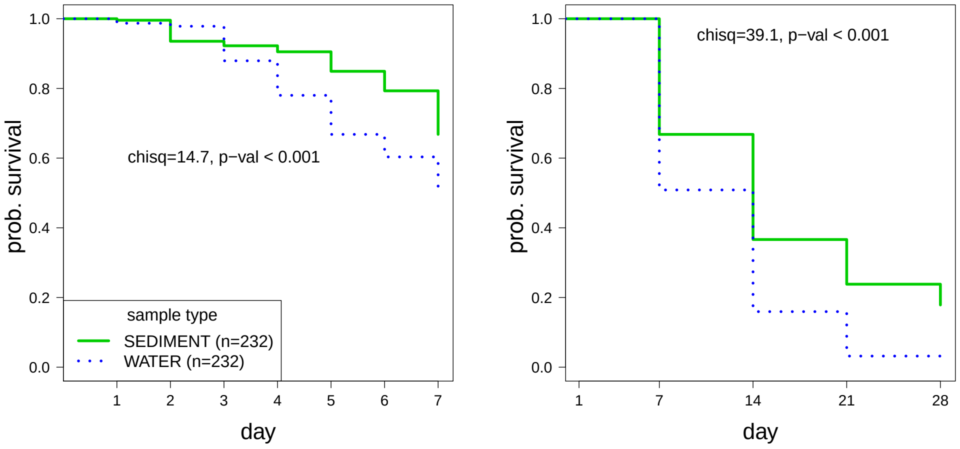
One of the difficulties in longitudinal study is the drop-outs, only some individuals have experienced the event (death) and, subsequently, survival times will be unknown for a subset of the study group. This phenomenon is called *right-censoring*. Censoring can also occur if we observe the presence of a subject but do not know where it began, namely, *left-censoring*. *Interval censoring* means that individual data only appear during partial follow-ups without knowing where it began and how it ended. The main strength of survival model is their ability to handle such missing values in a study design with long-term follow-up. Even well-designed experimental data may encounter drop-out, which can be handled by unbalanced design or simply be ignored in other methods. In general, however, censoring is less problematic in experimental data due to the fact that the experimental subject is a plant or an animal that has no intentional drop-out.

3.3 Kaplan-Meier (KM) curve and Log-rank test

The basic idea of Kaplan-Meier (KM) curve is that we count the number of terminations at some point of interest and divide by the number still unterminated. The graph looks like stairs in Figure 2. The KM curve is most useful when comparing two groups. i.e., treatment vs. control.

In Figure 2(a), the proportion of survival significantly drops for the water compared to the sediment as time goes. We separate the short term (daily up to 7 days) vs. the long term (weekly up to 28 days) analysis due to the way that our experiment has been conducted, and the fact that equally-spaced time interval is appropriate for Cox model that assumes a constant proportional hazard over time. Like the short term, the long term survival curve in (b) shows the difference between two sample types at weekly time points. The logrank test is a formal comparison of two KM curves, with a null hypothesis of no difference between two or more groups. All logrank tests in (a) and (b) reveal that there is a significant difference between the sediment and the water in the proportion of survival.

An important quantity in survival analysis is the survival function, denoted by $S(t)$, which provides the probability of survival at a given time $S(t) := P(T \geq t) = 1 - F(t)$ where T is a random variable of survival time. The estimate of this quantity is obtained by calculating cumulative proportion of survival as in Table 1, and the estimates by two groups form the KM survival curves in Figure 2. Note that there is no special parametric form in calculating the estimate of $S(t)$ and the subsequent hazard function, $h(t) = -d\log(S(t))/dt$. This simplicity with nonparametric test is one of the reasons for the popularity of KM estimation.



(a) Short term survival

(b) Long term survival

Figure 2: Kaplan-Meier curve by sample type

Table 1: Probability of survival at weekly time points

time	N	no. censor	no. death	proportion of survival at this interval	cumulative prop. survival
7	464	0	191	1 - (191/464) =0.59	0.59
14	273	12	151	1 - (151/273) =0.45	0.26
21	110	44	52	1 - (52/110) =0.53	0.14
28	14	10	4	1 - (4/14) =0.71	0.10

3.4 Cox models

Let $S(t)$ be the probability of being event-free up to time t and let $h(t)$ denote the instantaneous hazard rate at time t by the definition, $h(t) = -d\log S(t)/dt$. Cox regression [1] expresses the $h(t)$ with a function of covariate $\mathbf{X}$,

$$h(t|\mathbf{X}) = h_0(t)\exp(\mathbf{X}\beta), \tag{1}$$

where $h_0(t)$ is the baseline hazard function, and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the regression coefficient corresponding to the predictors of $\mathbf{X} = (X_1, \ldots, X_p)$. The covariates in (1) do not change over time while the baseline hazard function $h_0(t)$ holds the effect of time. For example, two sample types (sediment vs water), microbes types (E.coli vs Enterococcus), or cultivars (Osage vs Delta in soybean) are the time-invariant covariates. If researchers are interested in time-variant factors, which is not uncommon in many experiment such as temperature, humidity, or UV intensity, then time-dependent cox model is appropriate as given by:

$$h(t|\mathbf{X}(t)) = h_0(t)\exp(\mathbf{X}(t)\beta). \tag{2}$$

Inclusion of the time-variant covariate $\mathbf{X}(t)$ in (2) can prevent the violation of proportional hazard assumption in some situations. If departures from proportional hazards are substantial, one may consider the cox model that allows the time-variant coefficients as follows:

$$h(t|\mathbf{X}(t)) = h_0(t)\exp(\mathbf{X}(t)\beta(t)). \tag{3}$$

The model (3) contains interactions with parametric functions of time, and the time block provides interpretation, for instance, the effect of temperature on microbe survival may differ before or after 7 days. There are many discussion on how to estimate those interaction term for the time blocks, such as fractional polynomials and penalized likelihood approaches [3]. Alternatively, we may consider partition the time axis and do piecewise estimation of parameter, such as cross-sectional logistic regression at different time interval. Natural question arises if cox models are superior to logistic regression, or vice versa in certain situation.

4 Cox vs Logistic regression

Cox model and its variations differ from the linear logistic regression both in terms of the distributional assumptions and also in terms of the functional relationship with the predictors. Importantly, a cox model analyzes the time to an event outcomes, whereas a logistic regression is a direct method for binary outcome at a fixed time point. However, these two methods are asymptotically equivalent in the sense that the time-to-event response and associated predictors can be easily transformed to repeated measures of both binary responses and the predictors during the follow-up period.

4.1 Relative risk, odds ratio, and hazard ratio

The relative risk rate is an important estimate as it provides clinical interpretation between two treatment groups. Let P_0 and P_1 denote the probability of the event occurring under the control and the treatment group, respectively. The relative risk (RR), the odds ratio (OR) and the hazard ratio (HR) are given by:

$$RR = \frac{P_1}{P_0}, \quad OR = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}, \quad HR = e^{\beta}.$$

In our soybean storage example, we compare these relative risks and Wald statistics using a binary predictor of relative humidity ($\leq 61, > 61$). The estimates of RR and OR are obtained by the logistic regression and HR is estimated by cox model. The result is shown in Table 2, and clinical conclusion for humidity effect based on Wald statistics is similar in each follow-up. HR is always in the middle of RR and OR and in theory these become close each other under rare event and short term follow-up.

Table 2: Comparison of three relative risk measures and the Wald test

Follow-up	70 d		98 d	
	Risk	Wald	Risk	Wald
Relative Risk (RR)	1.49	-	1.47	-
Hazard Ratio (HR)	2.05	1.48	2.69	3.41*
Odds Ratio (OR)	2.75	1.77	9.39	3.72*

*statistical significance at 0.05 level

The asymptotic relationship has been shown in Symons and Moore [4]. Briefly, let $H(T) = \int_0^T h(t)dt$, and let the probability of event during the follow-up period [0,T] denote $P_0 = 1 - \exp\{-H(T)\}$ and $P_1 = 1 - \exp\{-H(T)e^{\beta}\}$. Then, the inequality holds to be $1 <$ RR $\leq$ HR $\leq$ OR for $\beta > 0$, which is explained by the asymptotic relations as follows:

$$RR \doteqdot \frac{1-[1-H(T)e^{\beta}]}{1-[1-H(T)]} = e^{\beta} = HR,$$

$$OR \doteqdot \frac{[1+H(T)e^{\beta}]-1}{[1+H(T)]-1} = e^{\beta} = HR.$$

4.2 Pooled logistic regression

When repeated measurements on predictors are obtained and one measurement of response is recorded in time, time-dependent covariate cox model (TDCX) in (2) is appropriate. Equivalently, logistic regression can be analyzed by combining the repeatedly measured dichotomous outcomes at each interval instead of one end point, survival time. This pooled logistic regression (PLR) is extensively discussed in D'Agostino et al. [2], in which binary outcomes are obtained in sequential time blocks after dropping both terminated and censored subjects. Mathematically, TDCX and PLR is equivalent in terms of regression coefficient, score test statistics for the coefficient, and likelihood function. To see this relation, let $p_i(\mathbf{X}(t_{i-1}))$ denote the conditional probability of observing an event by time t_i, given that the individual is event free at time t_{i-1} at any interval I_i. Then, an equivalent form of the logistic regression model at I_i is given by:

$$\begin{aligned} p_i(X(t_{i-1})) &= \frac{1}{1+\exp[\alpha_i + \mathbf{X}(t_{i-1})\beta]} \\ &\approx 1 - G_i(\mathbf{X}(t_{i-1})). \qquad (4) \end{aligned}$$

The approximation (4) is valid for small values of $G_i(\mathbf{X}(t_{i-1})) = \exp[\alpha_i + \mathbf{X}(t_{i-1})\beta]$ by the Taylor expansion. A similar approach applies to TDCX model. Let us denote $H_i(X(t_{i-1})) =$

$\int_{t_{i-1}}^{t_{i-1}} h_0(u)\exp[\mathbf{X}(t_{i-1})\beta]du$ for any interval I_i. For the small value of $H_i(X(t_{i-1}))$, the conditional probability $p_i(X(t_{i-1}))$ can be approximated by its Taylor expansion,

$$\begin{aligned} p_i(X(t_{i-1})) &= \exp(-H_i(X(t_{i-1}))) \\ &\approx 1 - H_i(X(t_{i-1})). \end{aligned} \tag{5}$$

Therefore, the equation (4) and (5) will hold for small values of both $G_i(\mathbf{X}(t_{i-1}))$ and $H_i(X(t_{i-1}))$. The requirement of small values of both quantities can be justified by assuming rare event and short term follow-up.

4.3 Simulation Study

A simulation study is performed to see the empirical difference between cox and logistic regression. First we generate two Weibull random samples ($n_i = 200$, $i = 1,2$) representing survival time. The shape parameter is one and the scale parameters are set to 5 and 3, respectively. The simulation repeats 1000 times. Three different follow-up time points (short, medium, long) and two censoring rates (0, 0.5) are investigated. The results are summarized in Table 3.

Table 3: Comparison of statistical power and risk ratios

		Power		Risk ratio		
Setting	Follow-up	cox	logit	RR	HR	OR
censoring rate 0.0	short	0.88	0.88	1.49	1.68	1.96
	middle	0.96	0.95	1.40	1.67	2.14
	long	1.00	0.98	1.20	1.67	3.17
censoring rate 0.5	short	0.67	0.77	1.56	1.63	1.98
	middle	0.83	0.94	1.41	1.62	2.76
	long	0.88	0.54	1.10	1.62	$> 20 \times 10^6$

In cox regression, the statistical power increases for later follow-up, and this pattern is the same under 0% and 50% censoring rate. At higher censoring rate, the statistical power is low in general. In logistic regression, however, long-term follow-up is not necessarily beneficial especially when censoring rate is high. This is because under many drop-out situation, longer follow up also implies more drop-outs, and there is a trade-off in statistical power between censoring and follow-up. On the other hand, three risk ratios are compared, HR is always in the middle of RR and OR, but this divergence becomes substantial as follow-up period increases and censoring rate increases. Note that the values of HR are consistent regardless, providing reliable clinical interpretation.

5 Conclusion

Despite the similarity between cox and regression in theory, there are reasons in practice to prefer cox model to logistic regression, or vice versa. In general, time to an event response contains more useful clinical information than whether or not the event occurred. The hazard ratio is a reliable measure of risk ratio regardless of follow-up time and censoring. Meanwhile, the estimation of odds ratio is inconsistent under different settings and is hard to defend for different clinical meaning. Survival analysis tends to have more statistical power than a logistic regression, especially with a long-term follow up. In the presence of drop-outs, which is almost always true in longitudinal studies, the cox regression that handles censoring well is advantageous for reliable estimation.

However, when a high rate of drop-out is problematic, a longer follow-up is rather disadvantageous because of the reduced sample size at the end. There exist a trade-off between censoring rate and follow-up period, and therefore a logistic regression within short follow-up period may be better in statistical power. When some covariate, like age or current temperature, can be confounded with follow-up time, the logistic regression at a fixed time can effectively detect the covariate effect than the cox regression. In addition, data for logistic regression are easier to achieve for some experiment, for example, to investigate mortality of bees, counting the number of death is much more convenient than collecting individual records of bees at all time points.

Acknowledgement

The author is grateful to John Rupe, Department of of Plant Pathology, U of Arkansas, and to Kristen Gibbson, Department of Food Science, U of Arkansas for providing the datasets of this article.

References

1. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34:187–220.

2. D'Agostino, R. B., Lee, M., Belanger, A. J., Cupples, L. A., Anderson, K., and Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. *Statistics in Medicine*, 9:1501–1515.

3. Lehr, S. and Schemper, M. (2007). Parsimonious analysis of time-dependent effects in the cox model. *Statistics in Medicine*, 26:2686–2698.

4. Symons, M. J. and Moore, D. T. (2002). Hazard ratio and prospective epidemiological studies. *Journal of Clinical Epidemiology*, 55:893–899.