SESUG Paper 160-2019

Generalized Linear Mixed Model Approach to Time-to-Event Data with Censored Observations

Kathleen Yeater, USDA-ARS; George Yocum, USDA-ARS; Kendra Greenlee, North Dakota State University; Julia Bowsher, North Dakota State University; Arun Rajamohan, USDA-ARS; and Joseph Rinehart, USDA-ARS

ABSTRACT

The time-to-event response is commonly thought of as survival analysis, and typically concerns statistical modeling of expected life span. In the example presented here, alfalfa leafcutting bees, *Megachile rotundata*, were randomly exposed to one of eight experimental thermoprofiles or two control thermoprofiles, for one to eight weeks. The incorporation of these fluctuating thermoprofiles in the management of the bees increases survival and blocks the development of sub-lethal effects, such as delayed emergence. The data collected here investigates the question of whether any experimental thermoprofile provides better overall survival, with a reduction and delay of sub-lethal effects. The study design incorporates typical aspects of agricultural research; random blocking effects. All *M. rotundata* prepupae brood cells were randomly placed in individual wells of 24-well culture plates. Plates were randomly assigned to thermoprofile and exposure duration, with three plate replicates per thermoprofile x exposure time. Bees were observed for emergence for 40 days. All bees that were not yet emerged prior to fixed end of study were considered to be censored observations. We fit a generalized linear mixed model (GLMM), using the SAS® GLIMMIX Procedure to the censored data and obtained time-to-emergence function estimates. As opposed to a typical survival analysis approach, such as Kaplan-Meier curve, in the GLMM we were able to include the random model effects from the study design. This is an important inclusion in the model, such that correct standard error and test statistics are generated for mixed models with non-Gaussian data.

INTRODUCTION

Survival analysis is a class of methods for which the outcome variable of interest is time until an event occurs. Time is measured from beginning (time=0) until the event occurs or the observation time ends. All subjects are observed, even if the subject does not experience the event, the length of time in the study is also recorded. A common goal in a survival analysis study is not only whether an event occurred, but also when it occurred. For example, a subject that lives 5 years after an experimental treatment is different from a subject that lives only 1 month after treatment. An analysis that only counted death events would ignore the equally valuable information about survival time.

Additionally, survival analysis methods allow for some incomplete time to event information in the study. These observations are referred to as *censored* observations, and they occur when a subject does not experience the event before the end of the study, the event occurs before the indicated start of the study, or if the observations are assessed at infrequent intervals such that the exact timing of the event is unknown. Censoring is uninformative if it occurs when the reasons for removal are unrelated to the event and it does not bias the parameter estimates and statistical inference. Informative censoring occurs when the reasons for removal are related to the event.

Conventional statistical methods are not appropriate for analysis of the time-to-event and censoring response variables. Logistic regression ignores the timing of events, and cannot handle time-dependent variables. Linear regression cannot handle censored observations or time-dependent variables, and is also inappropriate because time-to-event data often has a non-Gaussian distribution. Additionally, we introduce random model effects, which if not accounted for appropriately, we will introduce conditional vs. marginal model issues, standard error issues, and test statistic issues that is observed for other mixed models with non-Gaussian data (Gbur et al., 2012).

A successful survival analysis might provide the researcher with the ability to estimate and interpret survival probability; compare survival among different groups; assess the relationship between the survival time distribution and the time-independent and time-dependent explanatory variables; and possibly predict the time until the event. The methods described in this paper analyze the time-to-event response with left censored observations, within the generalized linear mixed model (GLMM), to estimate survival probability. Additionally, the relationship of temperature regime treatments to the time-to-event response are explored.

RESEARCH PROBLEM AND STUDY DESIGN

The researchers of this study hypothesized that exposing the solitary alfalfa leafcutting bee (*Megachile rotundata*) to an optimal temperature thermoprofile improves survival and decreases the development of sub-lethal effects (such as wing deformity). The researchers are also interested in any possible delay in the time to emergence as a first screening for sublethal effects. Thorough details of this research and the data referenced in this proceedings are provided in Yocum et al. (2019).

The study design: In the prepupae stage, each individual brood cell was inspected for developmental stage. Cells were placed individually in wells of 24-well culture plates. Plates were randomly assigned to thermoprofile treatment. There were 10 thermoprofiles (eight experimental and two controls), eight exposure duration to the thermoprofiles (one to eight weeks), and initiation of the exposure at two developmental stages (eye-pigmented pupae and emergence-ready adults). Once a week, three plates from each temperature treatment were transferred to constant 29°C to resume development. The individual bees within plates were observed weekly, every other day, and adult emergence date and sex were recorded for approximately 40 days. Our statistical model includes the treatment x week factors as fixed effects, and the individual plates observed are random blocking effects. The treatment and design features are necessary to include in the model to adequately interpret the relationship of the thermoprofile exposure to emergence.

The continuous day of emergence is the time point at which the event occurred. It is the "survival" time that is of interest. During the observation period, if a wasp emerged, these are recorded as a death (or *censored)* event at day = *t*. If emergence did not occur prior to the end of the study time period, the death is presumed to have occurred during the treatment phase of the study, prior to day 0 of the observation stage for emergence. These observations are left censored. A binary censored variable observation is recorded for all subjects, either the subject is censored or not censored. In our data, we coded *C=0* if the observation is censored, and *C=1* if the observation is not censored. For the following data exploration and analysis components, results from the emergence-ready adults is presented.

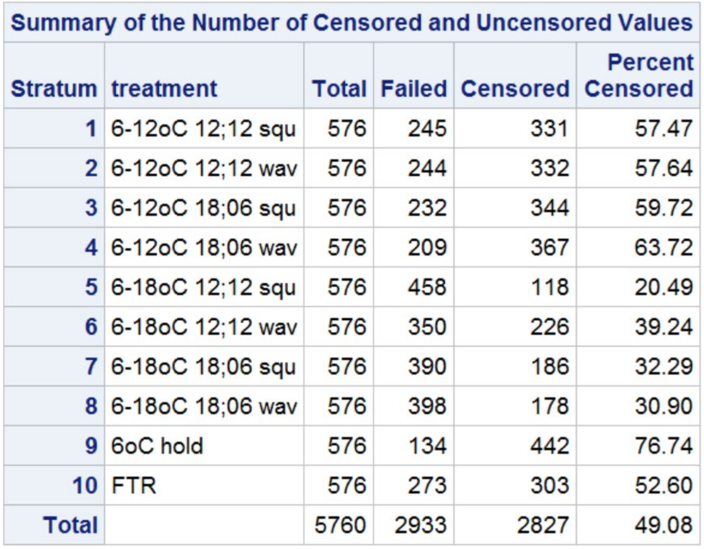
DATA EXPLORATION

Usually the first step in the analysis of time-to-event ("survival") data is to estimate and plot the survival function. The Kaplan-Meier method is able to incorporate the continuous days to emergence and binary censored observations to compute the probability of emergence at a given time *t.* The LIFETEST procedure computes and plots survival functions, and also tests for differences between survival functions. The TEST functionality within PROC LIFETEST is not appropriate for this example, as the random block effect is not taken into account with this method. You are able to identify the temperature treatment, week in storage, or a treatment x week combined variable to define the strata for the analysis. This approach is recommended here to explore a simplistic survival curve and tabulate summary statistics prior to building the GLMM with post-hoc comparisons. Details of this procedure and approach are thoroughly described in Allison (2010).

Here is an example of PROC LIFETEST code where the temperature treatment is the variable defining the strata. The TIME statement indicates the time-to-event variable and the censoring variable. The parenthetical 0 value corresponds to the censoring value used in the data file to identify censored observations:

```
proc lifetest data=data plots=survival;
   time TimetoEmerge * C_Emerge(0);
   strata treatment;
run;
```

Table 1 is a Censored Summary output table generated from the preceding code. For each of the 10 temperature treatments, you can explore the total emergence and censored observations. Recall that in the system being studied here, the FAILED column is actually the bees that successfully emerged, and the CENSORED column is indicative of those observations that did not emerge during the observation period.



Summary of the Number of Censored and Uncensored Values					
Stratum	**treatment**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	6-12oC 12;12 squ	576	245	331	57.47
2	6-12oC 12;12 wav	576	244	332	57.64
3	6-12oC 18;06 squ	576	232	344	59.72
4	6-12oC 18;06 wav	576	209	367	63.72
5	6-18oC 12;12 squ	576	458	118	20.49
6	6-18oC 12;12 wav	576	350	226	39.24
7	6-18oC 18;06 squ	576	390	186	32.29
8	6-18oC 18;06 wav	576	398	178	30.90
9	6oC hold	576	134	442	76.74
10	FTR	576	273	303	52.60
Total		5760	2933	2827	49.08

Table 1. Summary of the Number of Censored and Uncensored Values from PROC LIFETEST

The summary table does not account for random effects, and only the information based upon the temperature treatment is reported; however, you can still utilize this information to validate the input data. We expected stratum 5 treatment (6-18°C 12;12 squ) to have a high level of emergence, and it does in comparison to the other treatments. We need to know more about possible interactions with the storage time duration, which is explored with the GLMM.

The corresponding Kaplan-Meier survival curve (Figure 1) is the graphical representation of the same input data. You can identify when initial emergence occurs for the 10 temperature treatments, and follow each emergence pattern. The default Y-axis label 'Survival Probability' is the Probability of Emergence Event. With 10 possible stratum in the plot, the extreme probabilities are perhaps the easiest to interpret. Emergence appears to begin on or about day 15 in treatment '6-18° 12:12 squ' with 50% emergence on or about day 16. Emergence for treatment '6°C hold' appears to begin on or about day 16 with 50% emergence on or about day 21:

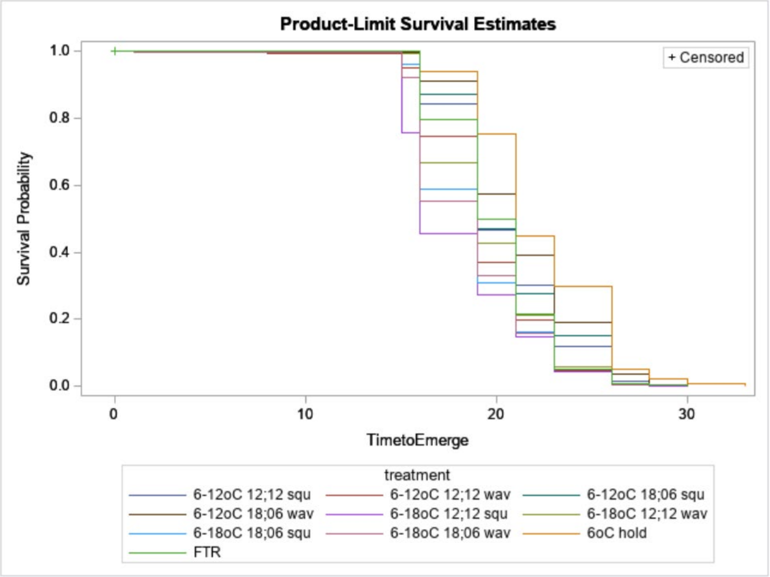


Figure 1. Kaplan-Meier Survival Curve from PROC LIFETEST

EXPONENTIAL SURVIVAL GLMM FOR CENSORED DATA

For the treatment and experiment designs that are described previously, the GLMM becomes a useful tool for obtaining estimates of the survivor and hazard functions. The hazard function is the instantaneous risk or potential that an event will occur at time *t*, given that a subject has survived up to time *t*. It takes the form of the number of events per interval of time. It is a constant rate *(λ)* regardless of the time, not a probability, which ranges from zero to infinity. The purpose of the data inquiry that we present does not focus on interpretation of hazard rate, so little discussion on the topic occurs here, except to mention where calculations are accomplished. For examples on hazard function and fitting proportional hazards regression models, see Hosmer et al. (2008).

The elements of a GLMM remain unchanged for time-to-event data. The approach and description provided here closely follows examples provided in Stroup (2013). The components of the model for these data are:

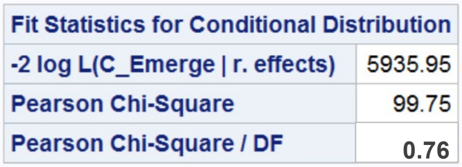
1. Linear predictor: this is the structure of the experimental design, randomization of how the treatment is applied; in model form, $\eta_{ijk} = \eta_{ij} + r(ab)_{ijk} + ab_{ij}$, where r is the k^{th} random block effect ‘plate’, a is the i^{th} temperature thermoprofile ‘treatment’, and b is the j^{th} duration of exposure ‘week’. We use the “cell means” form of the model since our interest lies on the interaction of temperature treatments and duration of exposure, rather than on the partitioned main effects.
2. Distribution(s): encompasses all random effects in the linear predictor, here $r(ab)_{ijk}$ and ab_{ij} are i.i.d. $N(0, \sigma^2)$; the observations are conditional on the random model effects, and the response variable, survival time (y_{ijk}) ~ independent $Exponential(\mu_{ijk})$. Note on distributions of survival analysis: The Poisson provides a theoretical starting point for developing time-to-event distributions. This process leads directly to the exponential distribution, which can be easily generalized to the gamma distribution.
3. Link Function: $\eta_{ijk} = \log(\mu_{ijk})$, this is the natural or canonical parameter – always a function of the mean, but a better candidate for regression and ANOVA-like models than the mean.

Initially, you should verify that the exponential (or gamma) distribution provides an adequate fit for the time-to-event response variable. This is achieved by determining if Φ = 1 is a plausible value for the scale parameter, i.e. fit the exponential distribution and determine if there is evidence of over- or under-dispersion.

The GLIMMIX statements for the exponential model are:

```
 proc glimmix method=laplace data=data;
    class treatment week plate;
    model TimetoEmerge = treatment*week / dist=exponential;
    random intercept / subject = plate(treatment*week);
    covtest/cl(type=plr);
run;
```

The only output of interest for this question is the scale parameter ($\widehat{\Phi}$), which is the Pearson X^2/df, located in the 'Fit statistics for conditional distribution' output table (Table 2):



Fit Statistics for Conditional Distribution	
-2 log L(C_Emerge | r. effects)	5935.95
Pearson Chi-Square	99.75
Pearson Chi-Square / DF	0.76

Table 2. Fit Statistics for Conditional Distribution.

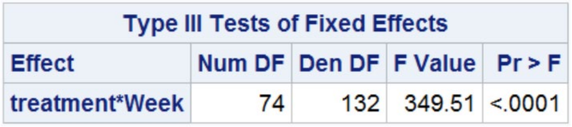
Values close to 1, with a confidence interval containing 1 are optimal. Our data appears to err on the under-dispersion boundary, with the Pearson X^2/df = 0.76, 95% confidence interval: [0.51, 0.97]. Perhaps the potential under-dispersion is influenced by the large proportion of censored observations, we have less variation in the data than the model predicted. However, the proportion of censored observations in this study was not unremarkable in the context of the biology of the system. More importantly, there is no evidence of over-dispersion, which is often encountered when fitting simple parametric models, such as those based on the Poisson distribution. There is no evidence to suggest that the gamma distribution is a better fit, and no evidence of over-dispersion, so the analysis moves forward utilizing the exponential model approach.

As mentioned previously, the exponential model is a generalization of the Poisson; therefore, the resulting log-likelihood for the exponential is the Poisson. The censoring random variable *C* has a Poisson distribution with rate parameter λ. At given time *t*, $E(C|t) = \mu_c = \lambda t$. In the time-to-event GLMM with censored observations, we use *C* as the primary response variable. The Poisson distribution is the conditional distribution of *C* given the random model effects. The link is $Log(\mu_c) = \log(\lambda) + \log(t)$. Because the form of the log-likelihood works within the GLMM estimating equations, we use $\log(t)$ as an *offset.* This is all accomplished in the GLIMMIX procedure:

```
  proc glimmix method=laplace data=data;
    class treatment week plate;
    logt = log(TimetoEmerge);
    model c_emerge=treatment*week / noint d=poisson offset=logt ;
    random intercept / subject=plate(treatment*week);
  run;
```

Notice we use the METHOD=LAPLACE option in our time-to-event data analysis. The class of models for which a Laplace approximation can be applied in PROC GLIMMIX is few compared to models to which Pseudo-Likelihood (PL) can be applied. Laplace works here because we have a conditional log-likelihood (d=poisson) and G-side random effects that are assumed to be normal (0, σ^2) (RANDOM statement).

The Type III tests (Table 3) tell us that the interaction of treatment x week is statistically significant at α=0.05.



Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
treatment*Week	74	132	349.51	<.0001

Table 3. Type III Tests of Fixed Effects

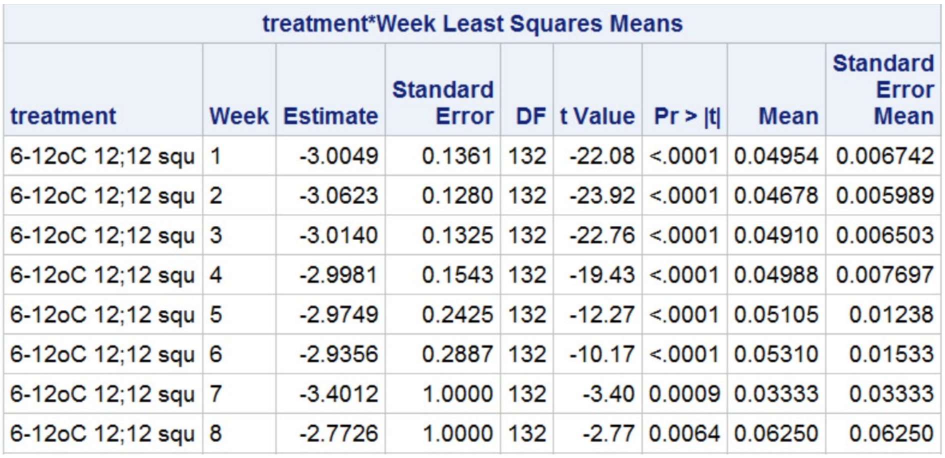
ESTIMATING MEANS

Our main steps with the analysis is to estimate μ, the temperature x week treatment combination means, and use the $\hat{\mu}$ to determine the hazard and survivor functions for each treatment. Once we generate parameter estimates, we apply the inverse link to give us the $\hat{\mu}$ for the functions of interest. Hence, taking $(\hat{\mu})^{-1}$ yields $\hat{\lambda}$ for the estimable functions of interests, and in turn we can calculate the survivor and hazard functions. This is all accomplished using LSMEANS and LSMESTIMATES statements.

The LSMEANS statement provides estimates of $\log(\lambda)$ – the rate parameter; the ILINK option gives actual values of the estimated hazard function for each treatment x week. The PLOTS option provides a graphic representation of the hazard function over the duration weeks for each temperature treatment:

```
proc glimmix method=laplace data=data;
    class treatment week plate;
    logt = log(TimetoEmerge);
    model c_emerge=treatment*week / noint d=poisson offset=logt ;
    random intercept / subject=plate(treatment*week);
    lsmeans treatment*week / ilink plots=meanplot(sliceby=treatment ilink);
run;
```

The output from the LSMEANS statement (first 8 observations) appears as Table 4:



treatment*Week Least Squares Means								
treatment	**Week**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Mean**	**Standard Error Mean**
6-12oC 12;12 squ	1	-3.0049	0.1361	132	-22.08	<.0001	0.04954	0.006742
6-12oC 12;12 squ	2	-3.0623	0.1280	132	-23.92	<.0001	0.04678	0.005989
6-12oC 12;12 squ	3	-3.0140	0.1325	132	-22.76	<.0001	0.04910	0.006503
6-12oC 12;12 squ	4	-2.9981	0.1543	132	-19.43	<.0001	0.04988	0.007697
6-12oC 12;12 squ	5	-2.9749	0.2425	132	-12.27	<.0001	0.05105	0.01238
6-12oC 12;12 squ	6	-2.9356	0.2887	132	-10.17	<.0001	0.05310	0.01533
6-12oC 12;12 squ	7	-3.4012	1.0000	132	-3.40	0.0009	0.03333	0.03333
6-12oC 12;12 squ	8	-2.7726	1.0000	132	-2.77	0.0064	0.06250	0.06250

Table 4. Treatment*Week Least Squares Means

The MEAN column shows the estimates of the hazard functions for each temperature x week treatment combination. The ESTIMATE column shows the link predictor function – which is a value of very little interest for the current interpretation. The following plot (Figure 2) show that across many of the treatments (except for 6°C hold) the hazard function increases as the number of weeks exposed to temperature treatment also increases. But what does the model tell us about emergence?

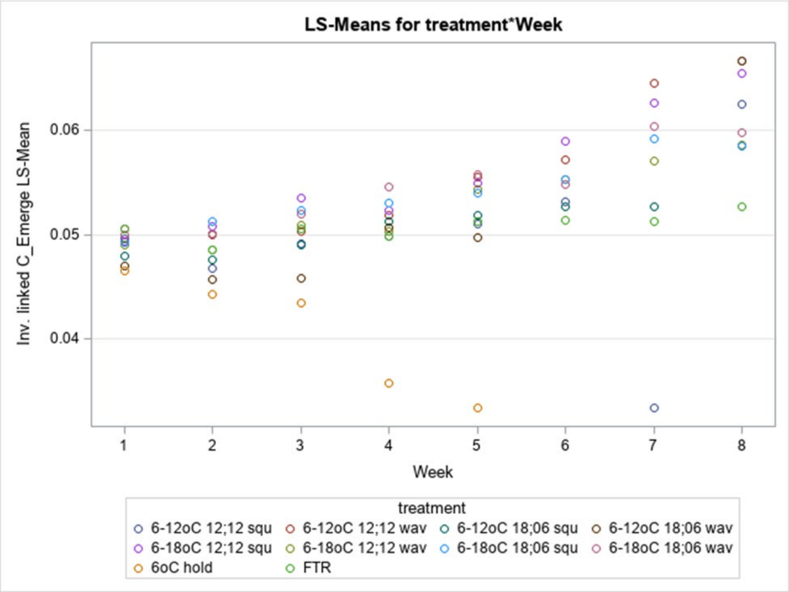


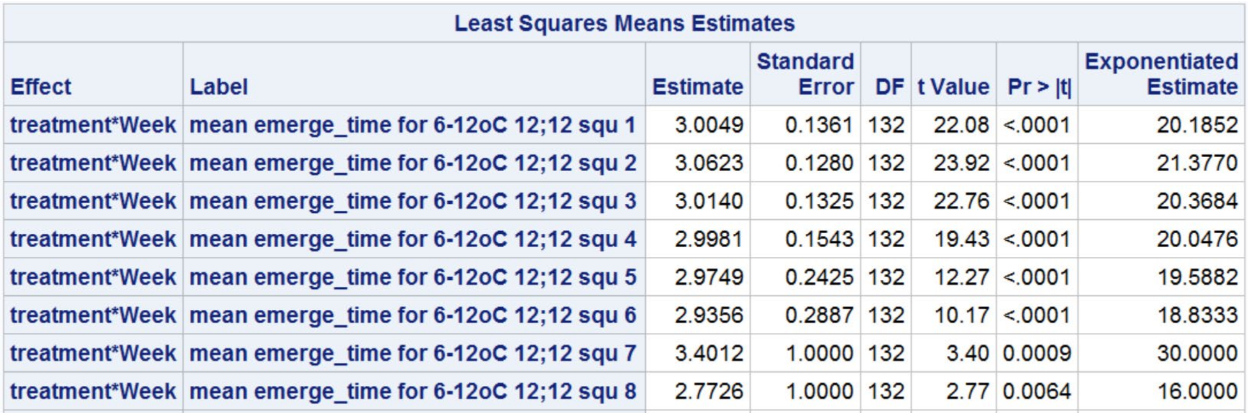
Figure 2. Plot of inverse link (hazard function) by duration of exposure 'Week' sliced by temperature thermoprofile 'Treatment'

The statement/option LSMESTIMATES / EXP provides us with exponentiated estimates of mean survival time – which we interpret as the mean emergence. The first set of 8 weeks for temperature regime treatment '6-12°C 12:12 squ' are as follows:

```
proc glimmix method=laplace data=data;
    class treatment week plate;
    logt = log(TimetoEmerge);
    model c_emerge=treatment*week / noint d=poisson offset=logt ;
    random intercept / subject=plate(treatment*week);
    lsmestimate treatment*Week
          'mean emerge_time for 6-12oC 12;12 squ wk1' -1,
          'mean emerge_time for 6-12oC 12;12 squ wk2' 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk3' 0 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk4' 0 0 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk5' 0 0 0 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk6' 0 0 0 0 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk7' 0 0 0 0 0 0 -1,
          'mean emerge_time for 6-12oC 12;12 squ wk8' 0 0 0 0 0 0 0 -1/exp;

run;
```

Table 5 is the LSMESTIMATE output table for the subset of observations provided in the preceding PROC GLIMMIX statements:



Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Exponentiated Estimate
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 1	3.0049	0.1361	132	22.08	<.0001	20.1852
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 2	3.0623	0.1280	132	23.92	<.0001	21.3770
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 3	3.0140	0.1325	132	22.76	<.0001	20.3684
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 4	2.9981	0.1543	132	19.43	<.0001	20.0476
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 5	2.9749	0.2425	132	12.27	<.0001	19.5882
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 6	2.9356	0.2887	132	10.17	<.0001	18.8333
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 7	3.4012	1.0000	132	3.40	0.0009	30.0000
treatment*Week	mean emerge_time for 6-12oC 12;12 squ 8	2.7726	1.0000	132	2.77	0.0064	16.0000

Table 5. Least Squares Means Estimates of Treatment ‘6-12°C 12:12 squ’ for Weeks 1 through 8

The EXPONENTIATED ESTIMATE column shows the estimated mean survival times, which we interpret as the estimated mean emergence day. We use these estimates to plot the survivor (emergence) functions for each temperature X week treatment combination. Figure 3 is modified from Yocum et al. (2019).

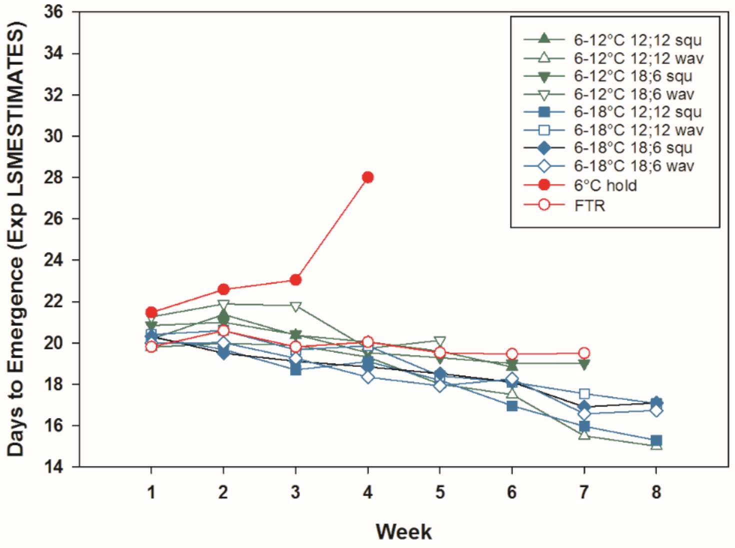


Figure 3. Plot of ‘Days to Emergence’ by the duration of exposure ‘Week’ sliced by the temperature thermoprofile treatment

SUMMARY OF GLMM RESULTS

Some emergence-ready bees emerged during the exposure period to treatment ‘6°C hold’; these observations are left censored, and this is exhibited in Figure 3 with no apparent emergence after week 4 of exposure. Bees emerged early from the following thermoprofiles: ‘6-12°C 12;12 squ’, ‘6-18°C 12;12 squ’ and ‘6-18°C 18;6 wav’ and are further investigated for biological effect in Yocum et al. (2019). Early emergence started between exposure weeks 3 and 4 in the thermoprofiles ‘6-18°C 12;12 squ’ and ‘6-18°C 18;6 wav’, with the rate of emergence increasing in the following weeks. Early emergence began between exposure weeks 6 and 7 in the ‘6-12°C 12;12 squ’ thermoprofile. By week 8 of exposure, emergence during low-temperature exposure reached 62% ± 8.5% for ‘6-18°C 12;12 squ’, 12.7% ±6.5% for ‘6-12°C 12;12 squ’ and 9.8% ± 3.6% for ‘6-18°C 18;6 wav’, respectively (modified from Yocum et al., 2019).

CONCLUSION

Building a survival analysis, or time-to-event, model with censored observations is achievable via the generalized linear mixed model approach described here. Time-to-event response with censored observations are easily analyzed using methods commonly addressed in introductory survival analysis texts. The major difference shown here is the addition of the random model effects. This example shows how they exist because of the study design, and it is important to account for them in the model.

REFERENCES

Allison, P. 2010. *Survival Analysis Using SAS®: A Practical Guide, Second Edition.* Cary, NC: SAS Institute Inc.

Gbur, E.E., W.W. Stroup, K.S. McCarter, S. Durham, L.J. Young, M. Christman, M. West, and M. Kramer. 2012. *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences.* Madison, WI: ASA, CSSA, SSSA.

Hosmer, D.W., S. Lemeshow, and S. May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition.* Hoboken, NJ: John Wiley & Sons, Inc.

Stroup, W.W. 2013. *Generalized Linear Mixed Models: Modern Concepts, Method and Applications,* p. 375-395*.* Boca Raton, FL: CRC Press.

Yocum, G.D., J.P. Rinehart, A. Rajamohan, J.H. Bowsher, K.M. Yeater, and K.J. Greenlee. 2019. “Thermoprofile parameters affect survival of *Megachile rotundata* during exposure to low-temperatures.” *Integrative and Comparative Biology*, https://doi.org/10.1093/icb/icz126.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kathleen M. Yeater
USDA Agricultural Research Service
kathleen.yeater@usda.gov

USDA-ARS does not promote or otherwise recommend any particular commercial product.