DONNER PARTY SURVIVAL: A REASSESSMENT WITH NONLINEAR REGRESSION WITH R

Eugene D. Gallagher
School for the Environment
UMass Boston
Boston MA 02125
8/20/23

Abstract

The survivorship data for the 1846-1847 Donner Party is used to demonstrate the use of restricted cubic splines (rcs), Generalized Additive Models (GAMs), and survivorship curves for the analyses of survival, a binary response variable. The Donner Party data are challenging because there were only 87 members of the traveling party. Of those 87 travelers, 8 died of causes other than weather and starvation, leaving only 79 travelers for the key analyses. Despite the small sample size, the rcs and GAM analyses reveal striking curvilinear patterns in survivorship, patterns not previously noted. Survivorship was nearly perfect for females aged 5 to 40, whereas males had lower survival decreasing linearly with age. Of the 79 snow-trapped travelers, no females and only one male past age 40 survived. Only 1 female (Age 25) in the range of 4 to 40 died. Survival was perfect for two families of size 6 and 9, but much lower for larger and smaller groups. Kaplan-Meier survivorship curves revealed the poor survival of males and especially employees, like teamsters. Males and employees died earlier and at much higher rates than females and family members.

Key words

Cox survivorship, GAM, Generalized Additive Models, *k*-fold cross-validation, Kaplan-Meir, rcs, restricted cubic splines

1 Introduction

The Donner Party expedition is an American tragedy. The Donner-Reed wagon train with 87 travelers followed the recommendation of Lansford Hastings to take an untried shortcut to California across the Sierra Nevada Mountains and were trapped by a huge 28 October 1846 snowstorm. Only 47 members of the party survived until the final rescue party on 22 April 1847. Most Donner Party members, survivors and non-survivors, resorted to cannibalism.

Grayson (1990, 1994, 1997, 2018) analyzed the demographics of Donner Party survival, concluding in 1990 (p 223) "...survivorship within the party was mediated almost entirely by three factors: age, sex and the size of the kin group with which each member traveled." Grayson's later papers and 2018 provide more support for the importance of those variables. Rarick (2008) also added a fourth important pattern, the low rates of survival of the teamsters and servants who worked for the families. Note, that I follow Grayson (2018) in using Sex rather than Gender to analyze the differential mortality among Donner Party travelers, acknowledging that gender-based differences in behavior could very well have played a role in survival.

Ramsey & Schafer (2013) analyzed a subset of the data for individuals aged greater than or equal to 15 and used the Donner Party data in their textbook Statistical Sleuth to introduce binary logistic regression. Their final model of Donner Party survival was an additive binary generalized linear logistic regression model using Age and Sex as explanatory variables. They briefly discuss an interaction model, but they did not analyze the survivorship of the 42 individuals deleted from their analysis nor did they consider Family Group Size or the survivorship of teamsters and servants. The use of those deleted travelers will allow me to present curvilinear regression and survivorship analyses, the keys to understanding Donner Party survival.

I'll analyze the effects of Sex, Age, and Family Group Size using restricted cubic spline regressions within a binomial Generalized Linear Model (Harrell 2015) and Generalized Additive Models (GAMs, Wood 2017, 2019). I'll then analyze survivorship with Cox Proportional Hazards models and Kaplan-Meier survivorship curves.

For the statistics instructor, these data provide an interesting case study for the introduction of binary logistic regression, nonlinear regression using restricted cubic splines, generalized additive models (GAMs) and survival analysis. The following analyses are in the spirit of Harrell (2015) and Andrews (2021) whose texts introduce binary logistic regression with both restricted cubic splines and GAMs in their textbooks.

2 Methods

All analyses were performed with R (R Core Team 2023) with the following R packages: caret for GAM *k*-fold cross-validation, mgcv (Wood 2017, 2019) for GAMs, rms (Harrell 2015) for Generalized linear models and restricted cubic splines, survival & survminer for survivorship analyses, and the tidyverse (Wickham & Grolemund 2017) which includes dplyr and ggplot2.

The code and data are available from Gallagher's github site (links at end). The demographic data were taken from Grayson (2018) Table 2.1. Age, occupation, cause of death, and date of death were obtained are described in Stewart (1960), Rarick (2008), Brown (2009), and Grayson (2018). Rarick's (2008, p. x-xi) ages and occupations largely match Grayson (2018), but when they conflicted, Grayson's (2018) dates based on more recent primary sources were used.

I'll test combinations of four covariates—Age, Sex, Family Group Size and Teamster & Servant status— on pared data (n=79) from which the 8 travelers who died or crossed the Sierra Nevada mountains before the first major snowfall on 28 October 1846 have been deleted. The primary statistical method is binomial logistic regression using Harrell's (2015) Generalized Linear Model function (rms::Glm). Family Group Size differs from Family Size based on surnames. For example, Elizabeth Donner traveled with two children from a previous marriage: Solomon Hook (14) and William Hook (12); the Fosdicks traveled with the Graves family, and the Fosters traveled with the Murphy family, and so on. Age and Family Group Size will be tested as restricted cubic splines using Harrell's (2015) rms::rcs function and as GAMs using Wood's (2019) mgcv package. As recommended by Harrell (2021), the Akaike Information Content (AIC) was used to find the appropriate number of knots for the restricted cubic spline regression. That *k* was usually the lowest AIC, hence the highest likelihood after penalization for the number of parameters. Preserving degrees of freedom and reducing overfitting was a high priority, so 3 knots, the minimum allowed in rcs regression, was used if its AIC was within 4 units of the minimum AIC found with a higher number of knots. The AIC < 4 threshold was chosen because Burnham and Anderson (2004 p. 271) and Bolker (2015 p. 210) regard AICs less than 2 apart as roughly equivalent, AICs 4-7 apart as clearly distinguishable, and models with AICs more than 10 apart as definitely different. This 4-unit AIC threshold serves as an additional penalty for the number of knots in the rcs model above that incorporated in the AIC calculation.

There are 3 different parameters in the analyses reported here designated *k*. First, the number of knots in rcs regression is *k*. Second, *k* is the basis function parameter controlling the smoothing of the GAMs, which Wood (2017) sometimes refers to as knots. Finally, there is *k*-fold cross validation using the caret package, which is used in this paper to find the appropriate number of knots or the GAM smoothing parameter that produced the lowest log loss statistic. For *k*-fold cross validation, ten folds were used, so that a tenth of the data were removed for each of 10 analyses and the average log loss was used to judge the fit of the model given the number of knots or the GAMs *k* parameter. Then the full data (n = 79 usually) is fit with the selected optimal number of knots or with the appropriate GAM basis function.

Tests of null hypotheses use Wald Chi-square tests and Wilks' Chi-square drop-in-deviance test, described in Ramsey & Schafer (2013).

Cox proportional hazards models and Kaplan-Meier survivorship curves were generated using the survival and survminer packages. Gallagher's R coding was assisted by Open AI's GPT-4.

3 Results

3.1 Binary logistic regression of 79 Travelers

3.1.1 Effects of Age and Sex analyzed with restricted cubic splines

The full data are the 87 original members of the Party minus five who died before the major snowfall (Halloran, Hardkoop, Pike, Snyder, Wolfinger) and the three (William McCutcheon, James Reed & Walter Herron) who crossed the Sierras before the 28 October 1846 storm that stopped the Donner Party's advance. The effects of Age and Sex and their interaction were analyzed with a binomial logistic regression with a restricted cubic spline regression for Age with three knots. The age variable was fit with a 3-knot restricted cubic spline with AIC=95.408, which was within 4 AIC units of the 4-knot AIC=92.358. The results are shown in Table 1 and Figure 1.

The most striking pattern in Figure 1 is the strong curvilinear pattern in female survivorship. From ages 4 to 40, only one female died, Eleanor Eddy aged 25. There was a tremendous difference between female and male survival, with only one male in his 20's surviving, William Eddy aged 28, the husband of Eleanor Eddy.

Table 1. Wald & Wilks statistics and effect sizes for the rcs(Age, 3 knot) * Sex model.

1.1 Formulae: Null Model: Status ~ 1
Model 1: Status ~ rcs(Age, 3) * Sex
Model 2: Status ~ rcs(Age, 3) + Sex

1.2 Wald Statistics from glm

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8355	0.8105	-1.031	0.30260
rcs(Age, 3)Age	0.3912	0.1435	2.727	0.00640
rcs(Age, 3)Age'	-0.9380	0.3359	-2.792	0.00523
SexMale	0.8534	1.0756	0.793	0.42754
rcs(Age, 3)Age:SexMale	-0.4029	0.1577	-2.554	0.01064
rcs(Age, 3)Age':SexMale	0.8973	0.3588	2.501	0.01239

1.3 Wilks statistics (Drop-in-Deviance Tests) The interaction model had considerably more explanatory value than the null model, and the interaction model had more than the additive model.

Deviance	Chi-square	df	P
Null Model:	109.201	78	
Residual Model 1:	83.408	73	
Model 1 vs. Null:	**25.793**	**5**	**<0.0001**
Residual Model 2:	92.651	75	
Models 1 vs. 2:	**9.244**	**2**	**<0.01**

1.4 Effect sizes from Harrel's rms::Glm

From Harrell's (2015) Glm. A 14-y old female's odds of survival were 45 times higher (exp(3.814)) than a 14-y old male (95% CI: 3.5 to 590 times). A 24-y old male has odds of survivorship only 27% of a 14 year old (exp(10*-.59288)*100), but the 95% CI's are huge: <1e6 to 88,000). Examine the male curve at ages 14 and 24 to see why the intervals are so broad.

Factor	Effect	S.E.	Lower 0.95	Upper 0.95
Age	-0.59288	0.63738	-1.8632	0.67743
Sex - Female:Male	3.81400	1.28540	1.2523	6.37570

Adjusted to: Age=14 Sex=Male

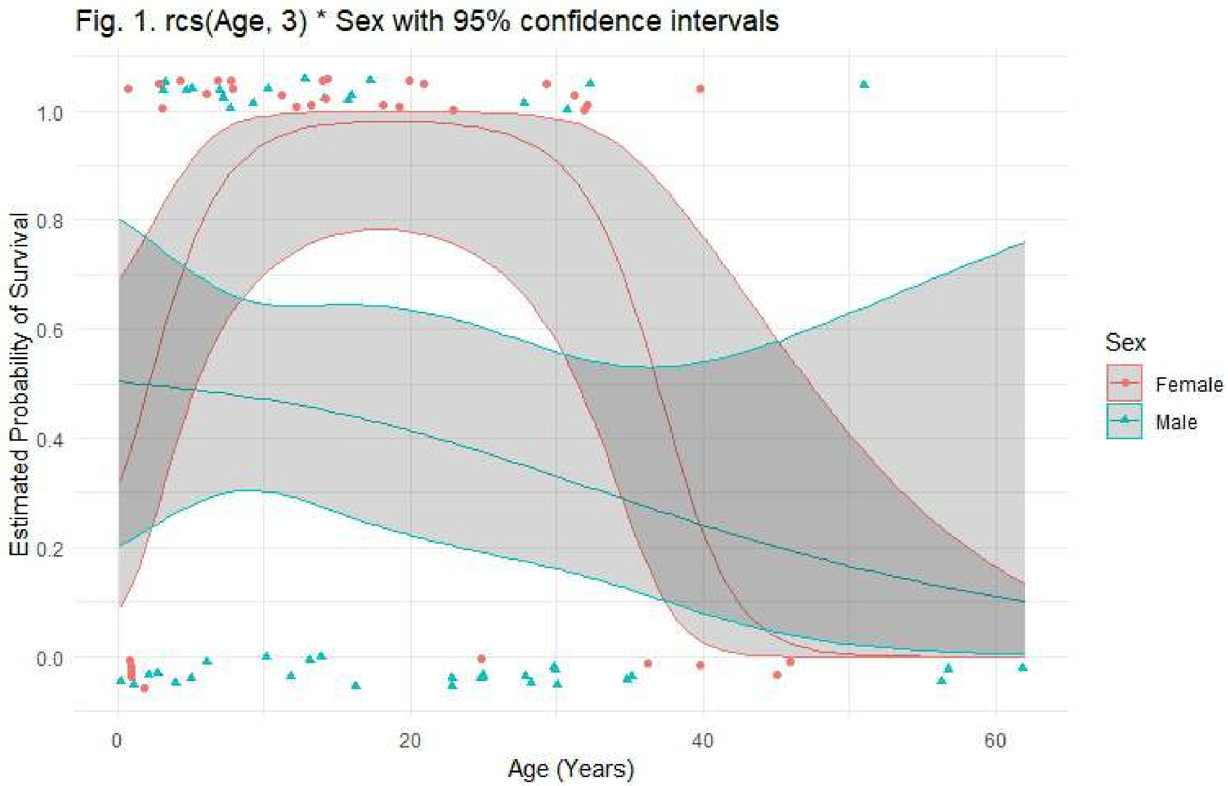

Figure 1. Display of the restricted 3-knot cubic spline model of the interaction of Age and Sex. Between the ages of 4 and 42 only one Female died (Eleanor Eddy, Age 25). Only two travelers survived past age 40: Margaret Breen (40) and Patrick Breen (51).

3.1.2 Effects of Age and Sex analyzed with a Generalized Additive Model

The effects of Age and Sex on survivorship were analyzed using a GAM. A *k*-fold cross-validation determined the best basis function or smoothing parameter *k* for the GAM. With $k = 2$, the Root Mean Square Error (RMES) was minimized. The results of the GAM analysis of Age, Sex and their interaction are shown in Table 2 and Figure 2.

Table 2. Statistics for the binomial GAM model of Age and Sex with a logit link function with *k* = 2 for the GAM smoother, determined by a *k*-fold cross validation. The effective degrees of freedom or edf indicates a quadratic pattern (Wood 2017, p 83). UBRE is the unbiased risk estimator (Wood, 2017, p 255).

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    9.193      3.127    2.94  0.00328

Approximate significance of smooth terms:
                                   edf Ref.df Chi.sq p-value
s(Age):as.numeric(Sex == "Male")   2.0   2.00 11.264 0.00358
s(Age):as.numeric(Sex == "Female") 1.9   1.99  7.495 0.02286

R-sq.(adj) =  0.229   Deviance explained = 21.9%
UBRE = 0.20352  Scale est. = 1         n = 79
```

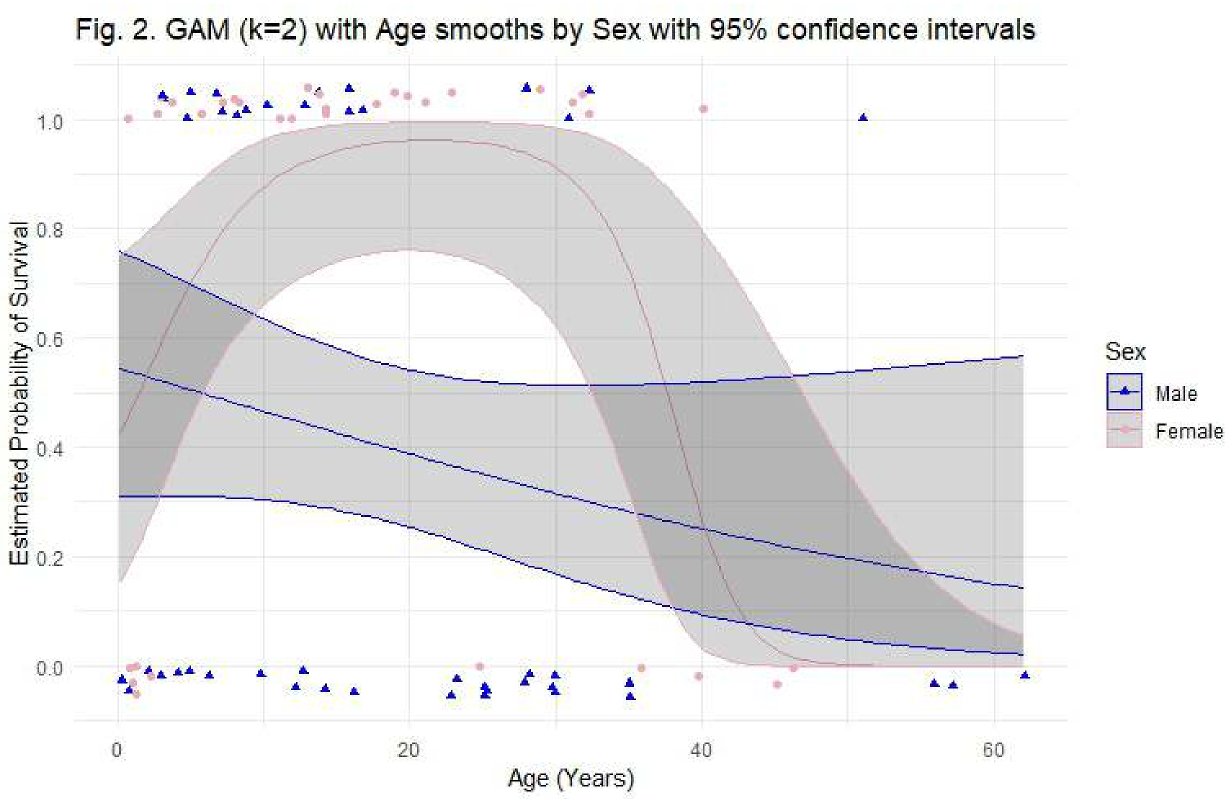

Figure 2. Display of the GAM model of Sex and Age, with the model basis parameter *k* determined to be 2 by a *k*-fold cross-validation analysis. The curvilinear relationship is similar to that shown in Figure 1 using a restricted cubic spline regression.

3.1.3 Effects of Family Group Size analyzed with restricted cubic splines

Family Group Size for the 79-traveler data was analyzed using rcs regression. The rcs regression revealed that Family Group Size was strongly related to Survival as shown by Wilks drop in deviance tests (Table 3), but the relationship was not linear as shown in Figure 3. The AIC for knots 3 through 7 were 103.10, 100.96, **100.414**, 98.8629 and 96.8564 making 5 knots the appropriate choice based on an AIC threshold difference of 4 AIC units.

Table 3. Wald & Wilks statistics and effect sizes for the Family Group Size restricted cubic spline model. Note, that it is not unusual that all 5 coefficients have Wald statistics have $p > 0.05$. The Wilks Drop-in-deviance test and Figure 3 clearly show that the model is explaining more than chance variability in the Donner data.

3.1 Formulae: Null Model: Status ~ 1
Model 3: Status ~ rcs(Family_Group_Size, 5)

3.2 Wald Statistics from glm.

```
Coefficients:                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                      -1.9266     0.9855  -1.955  0.051
rcs(Family_Group_Size, 5)Family_Group_Size        0.5424     0.3839   1.413  0.158
rcs(Family_Group_Size, 5)Family_Group_Size'       0.7826     2.4304   0.322  0.747
rcs(Family_Group_Size, 5)Family_Group_Size''     -3.6399     4.7616  -0.764  0.445
rcs(Family_Group_Size, 5)Family_Group_Size'''    63.8362    36.1297   1.767  0.077
```

3.3 Wilks statistics (Drop-in-Deviance Tests) The 5-knot Family Group Size rcs model had considerably more explanatory value than the null model.

Null deviance: 109.201 on 78 degrees of freedom
Residual deviance: 90.414 on 74 degrees of freedom
Drop in deviance: 18.787 on 4 degrees of freedom, $p < 0.0001$

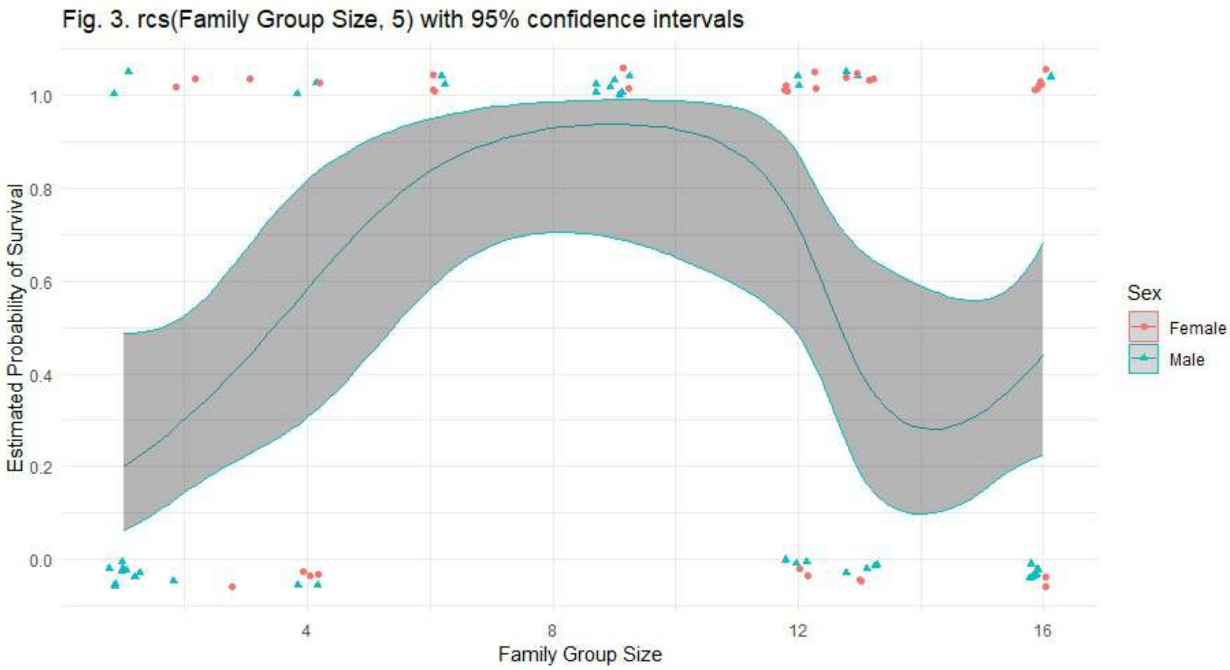

Figure 3. Effect of Family Group Size on Survival, modeled with a restricted cubic spline with 5 knots. In Family Group Sizes of six (the Reed Family) and nine (the Breen family), everyone survived.

3.1.4 Effects of Family Group Size analyzed with a GAM

A *k*-fold cross-validation determined that the GAM smoothing parameter k=3 minimized the RMSE. Similar to the rcs analysis, the GAM revealed that Family Group Size was also strongly related to Survival as shown by Wald tests (Table 4), but the relationship was not linear (Figure 4).

Table 4. Statistics for the binary logistic GAM model of Family Groups Size with the basis function (GAM smoothing parameter) *k* =5.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3082     0.2941   1.048    0.295

Approximate significance of smooth terms:
                        edf Ref.df Chi.sq p-value
s(Family_Group_Size) 3.766  3.957  11.61  0.0195

R-sq.(adj) =  0.173   Deviance explained = 19.2%
UBRE = 0.23719  Scale est. = 1         n = 79
```

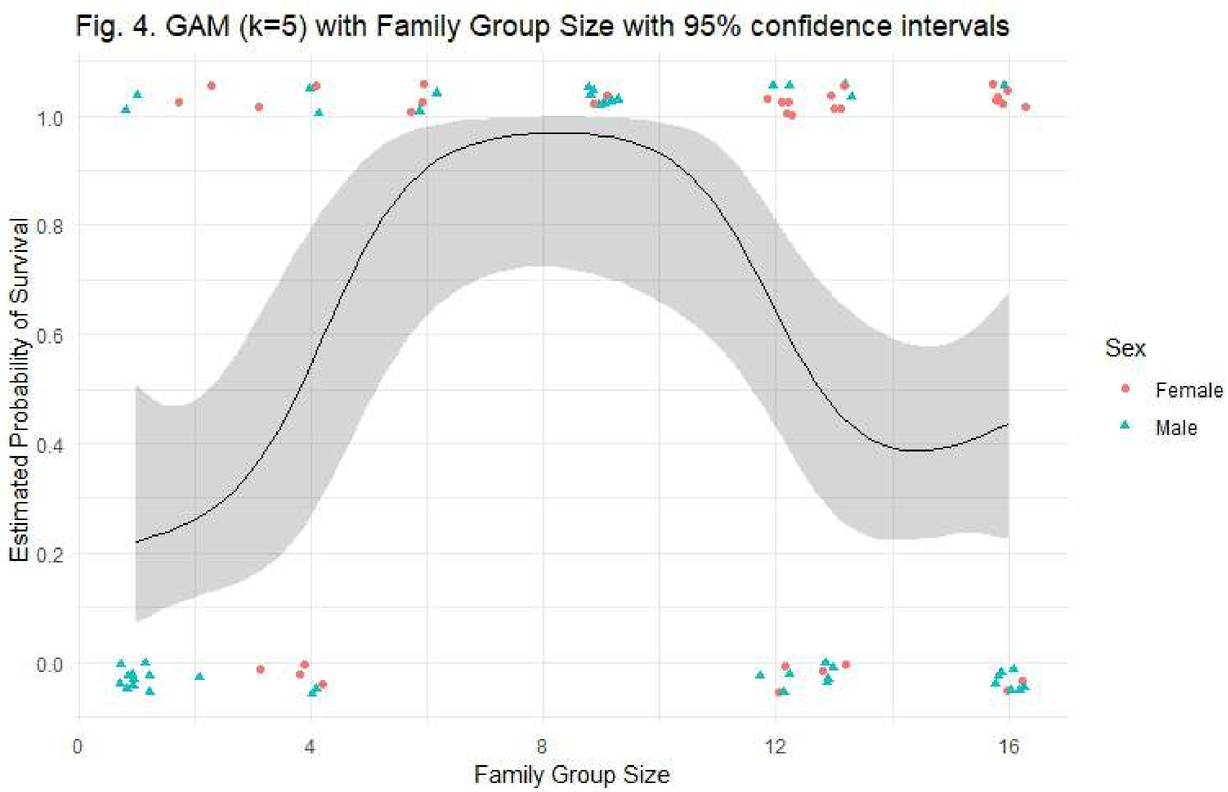

Figure 4. Effect of Family Group Size on Survival, modeled with a GAM with *k* = 5, chosen using *k*-fold cross validation. Most of those in Family Group Size 1 were employees, mainly teamsters, but two servants.

3.1.5 Simultaneous analysis of Sex, Age, and Family Group size with rcs

One way to analyze the joint effects of Sex, Age, and Family Group Size is to model Survival as curved surfaces resulting from the action of all three variables. In this model, an AIC analysis indicated that 6 knots should be used for Age and 6 knots for Family Group Size. The 6,6 model had the lowest AIC value by far (56.328), but the 3-d curved surface plots were full of spikes due to single observations. A *k*-fold cross validation analysis minimizing Root Mean Square Error found that Age with 3 knots and Family Group Size with 5 knots was appropriate even though the AIC was much higher at 78.369. Table 5 displays the Wald Chi-square tests and Figure 5 shows a three-dimensional view of survivorship with those knot sizes.

Table 5. Wald statistics for the rcs(Age, 3) *Sex + rcs(Family Group Size, 5)

5.1 Effect Sizes The odds of a 14-year old female surviving were 88 times higher (exp(4.4726) than a 14-y old male (95% CI: 2.9 to 2700 times).

```
 Factor              Effect   SE      Lower 0.95 Upper 0.95
 Age                  1.0784 1.02410 -0.96465   3.12150
 Family_Group_Size   -1.0421 0.99062 -3.01840   0.93412
 Sex - Female:Male    4.4726 1.70690  1.06740   7.87780
Adjusted to: Age=14 Sex=Male
```

5.2 Wald Statistics

```
Coefficients                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -3.8540     1.8404  -2.094  0.03625
rcs(Age, 3)Age                                     0.5408     0.1931   2.801  0.00510
rcs(Age, 3)Age'                                   -1.3338     0.4800  -2.779  0.00546
SexMale                                           -0.4999     1.4069  -0.355  0.72234
rcs(Family_Group_Size, 5)Family_Group_Size         0.5136     0.5186   0.990  0.32199
rcs(Family_Group_Size, 5)Family_Group_Size'        3.1091     3.1502   0.987  0.32366
rcs(Family_Group_Size, 5)Family_Group_Size''      -9.1724     6.1511  -1.491  0.13591
rcs(Family_Group_Size, 5)Family_Group_Size''' 123.5832    47.0271   2.628  0.00859
rcs(Age, 3)Age:SexMale                            -0.3581     0.2087  -1.716  0.08619
rcs(Age, 3)Age':SexMale                            0.9591     0.4964   1.932  0.05334
```

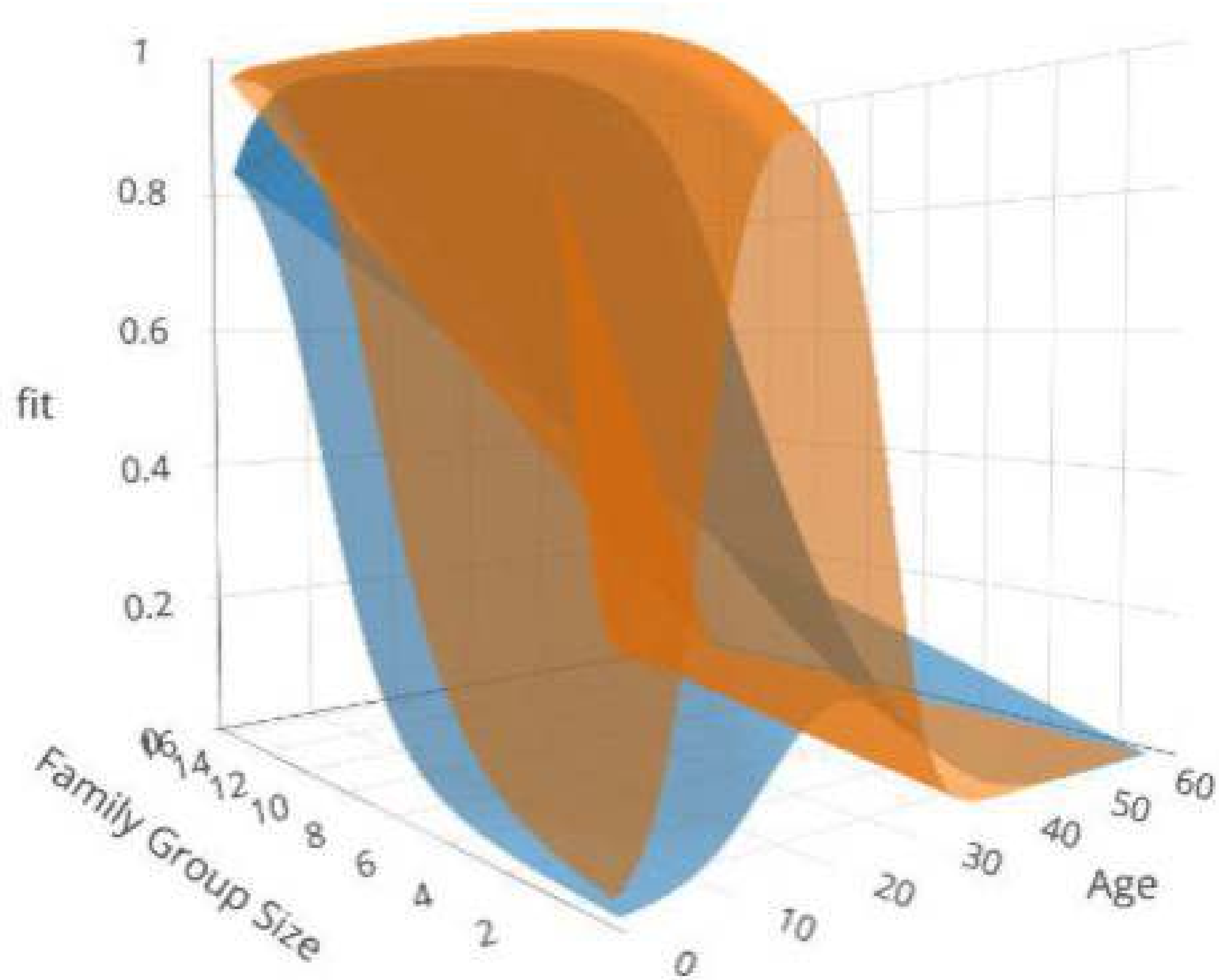

Figure 5. Three dimensional perspective of the probability of survival in the Donner Party as a function of Age and Family Group Size, modeled with restricted cubic splines with 3 and 5 knots, respectively. The Female predicted survival is orange, and the Male is blue. The striking Survival of all but one female between the ages of 5 and 41 and the high death rates of all but one male in their 20s produce the two parabolic patterns along the Age axis. The complete survival of families of size 6 and 9, the very low survival of small families, and the intermediate survival of the largest families produce the nested cowl-like patterns along the Family Group Size axis.

3.2 Survival Analyses

3.2.1 Effects of Age

The 79-traveler data were analyzed with survival analyses, the Cox proportional hazard model (Table 6) and the Kaplan-Meier survivorship curves (Figure 6). The proportional hazard assumption, tested with the survival::cox.zph test, was not violated (Table 6), and the Kaplan-Meier survivorship curve (Figure 6) clearly indicates that males started dying earlier and at a higher rate than females. As noted first by Grayson (1997), 14 males died before the first female death on day 97 after the 28 October storm (Harriet McCutcheon, age 1) with 11 males dying in the first 15-d interval after the first male death. There were four relief parties, called by Donner historians the First Relief to Fourth Relief, and their arrival at the Donner encampments is indicated in Figures 6 and 7. The endpoints of the survivorship curves are at Day 183 (April 29, 1847), when Louis Keseberg the sole survivor in the Donner encampments left for California with Relief Party 4.

Table 6. Cox proportional hazard model The effects of Sex on Survival Time were analyzed. There was a pronounced Sex effect on survival time (p = 0.01), with the odds of male dying being 2.6 times higher than females (1.2 to 5.5 95% CI). A cox.zph test in the survival package indicated only weak evidence against the equal hazard proportion assumption (chisq=2.54, df = 1, p = 0.11).

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
SexMale	0.9683	2.6336	0.3727	2.598	0.00937

Factor	Chi-Square	d.f.	P
Family_Group_Size	12.49	4	0.014
Nonlinear	11.57	3	0.009
TOTAL	12.49	4	0.014

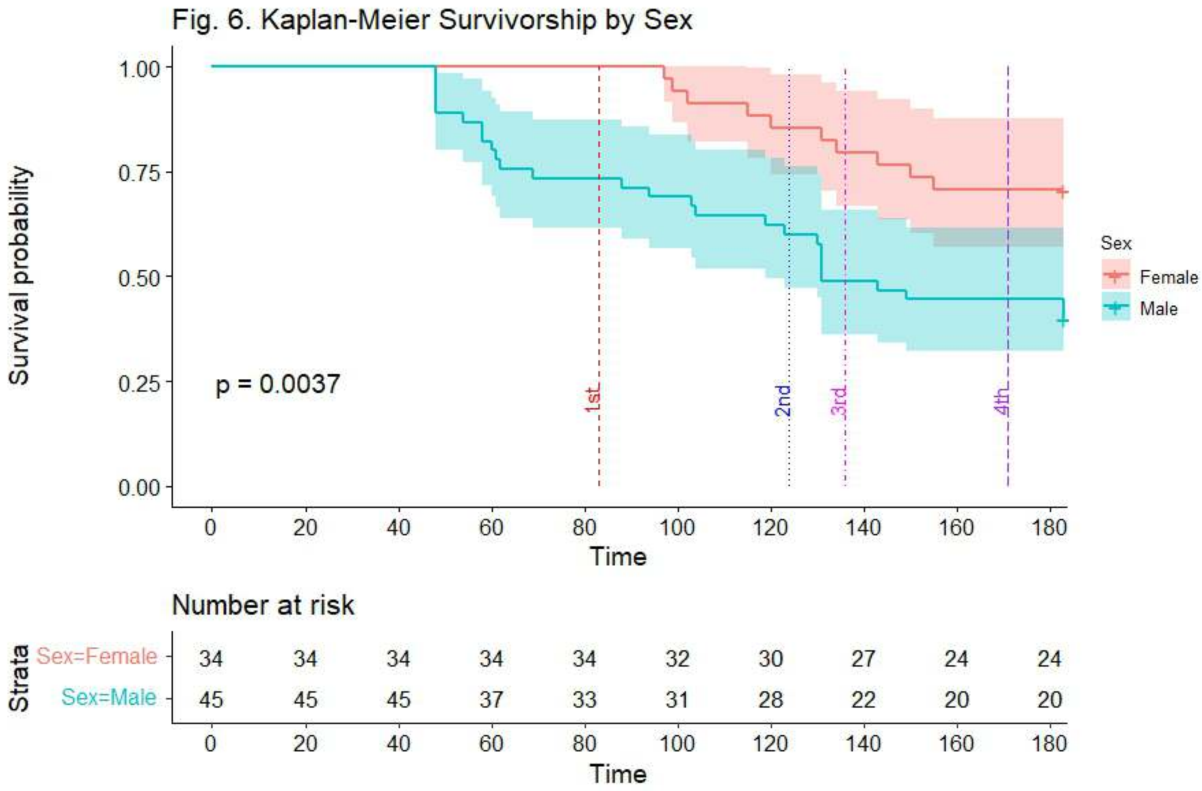

Figure 6. Kaplan-Meier Survivorship curve as a function of Sex. Males are 2.6 times more likely to die than females (Table 6).The equal proportion assumption is tenable (p = 0.08). The dates of arrival at the encampments by the four relief parties (labeled 1st to 4th) are indicated.

3.2.2 Effects of employee status

Thirteen of the 87 travelers in the Donner Party were employed as teamsters or servants. Most were teamsters hired to drive the oxen-powered wagons, but Antonio was a cattle herder and Baylis & Eliza Williams were Reed family servants. Patrick Dolan, Luke Halloran, Joseph Reinhardt, and Charles Stanton were the lone travelers (Family Group Size = 1) who were neither teamsters nor servants. Only 9 of the 13 employees were present after the first major snowfall on 28 October 1846; the others are not included in the analyses. The survivorship analyses indicated that these 9 employees died more rapidly than family members or non-employee bachelors (Tables 7 & 8, Figures 7 & 8).

Table 7. Cox proportional hazard model Analysis of Employee vs. Family member survival. Teamsters died at a rate 4.5 times that of Family members (95% CI: 2.1 to 9.3). A survival::cox.zph test indicated a clear violation of the equal proportion assumption (chisq=6.45, df = 1, p = 0.01).

```
                       coef exp(coef) se(coef)     z Pr(>|z|)
Teamster_Hired_Hands 1.5068    4.5124   0.3679 4.096 4.21e-05 ***

                     exp(coef) exp(-coef) lower .95 upper .95
Teamster_Hired_Hands     4.512     0.2216   2.194     9.281

Concordance= 0.63  (se = 0.038 )
Likelihood ratio test= 13.32  on 1 df,   p=3e-04
Wald test            = 16.77  on 1 df,   p=4e-05
Score (logrank) test = 20.02  on 1 df,   p=8e-06
```

The Cox proportional hazards test (survival::cox.zph) indicated that the hazard ratio of Teamsters & Servants relative to Family members was not constant with time, violating the Cox constant proportional hazard assumption (p = 0.01), so an additional covariate, Survival_Time, was added to the Cox model. A plot of Schoenfield residuals indicated a reasonable fit to the proportional hazards model with the relative risk ratios dropping with time. The results are shown in Table 8 and Figures 7 & 8.

Table 8. Cox time-dependent hazard model. The model determined that employees initial risk of dying was 480 times that of family members, but the relative risk between employees and family members declined by 4% per day as shown by the Kaplan-Meier survivorship curves (Figure 9).

Model:

```
Donner$SurvTime_Employee <- with(Donner, Survival_Time *
Employee)
coxph(formula = Surv(Survival_Time, Death) ~ Employee +
SurvTime_Employee, data = Donner)
  n= 79, number of events= 37
                       coef exp(coef)  se(coef)      z Pr(>|z|)
Employee            6.16800 477.22945   1.26154  4.889 1.01e-06
SurvTime_Employee  -0.04558   0.95545   0.01434 -3.177  0.00149
                  exp(coef) exp(-coef) lower .95 upper .95
Employee           477.2294   0.002095    40.263 5656.5203
SurvTime_Employee    0.9554   1.046632     0.929    0.9827

Concordance= 0.621  (se = 0.035 )
Likelihood ratio test= 27.03  on 2 df,   p=1e-06
Wald test            = 32.13  on 2 df,   p=1e-07
Score (logrank) test = 63.63  on 2 df,   p=2e-14

                  chisq df     p
Employee           4.97  1 0.026
SurvTime_Employee  2.70  1 0.100
GLOBAL             5.14  2 0.077=2e-14
```

Table 8 and Figures 7 & 8 reveal that the proportional hazard ratio between Employees and non-Employees is not constant with time. This is evident in the lack of parallel slopes in the Kaplan-Meier survivorship curves in Figure 7. Kaplan-Meier curves stratified by time period (Figure 8) show the changes through time intervals in the hazard function. While Employees have a lower survival in each time period with no non-employees dying in the first interval, there was a clear convergence of death rates during the last time interval 100 days after the Donner Party was snowed in.

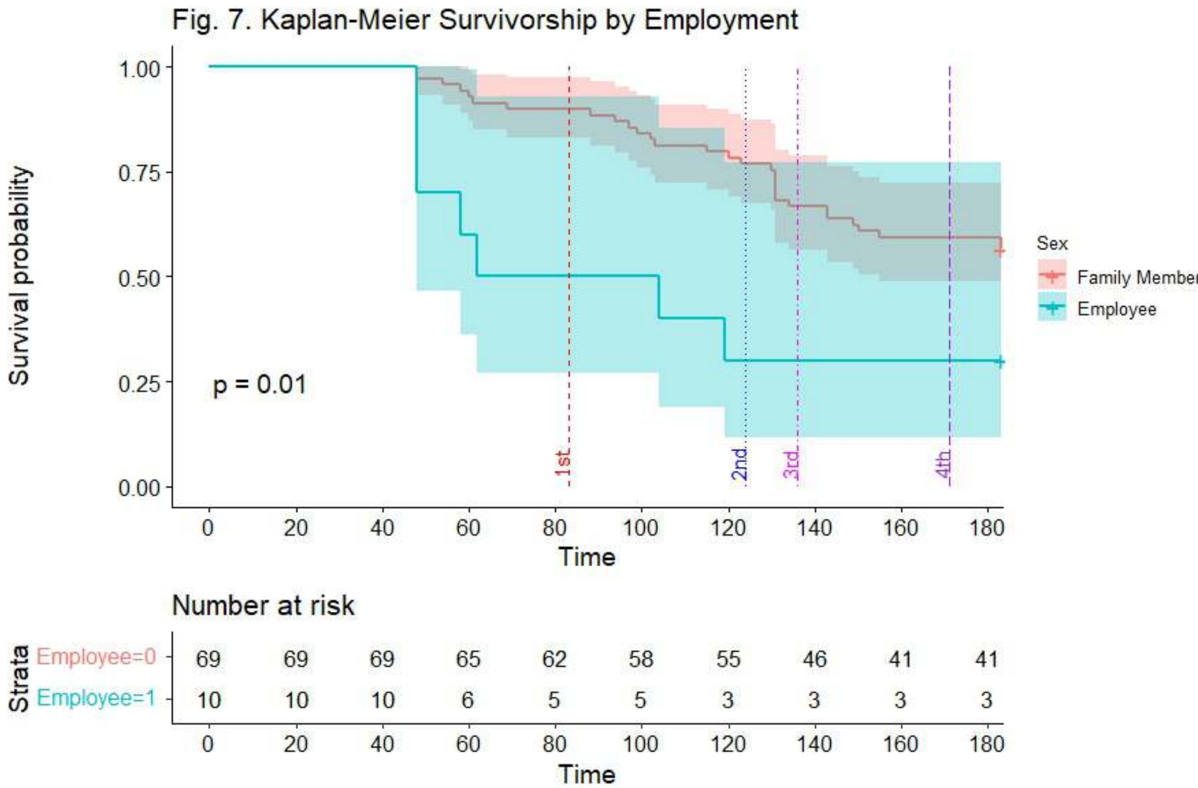

Figure 7. Kaplan-Meier Survivorship curves as a function of Employee vs. Family Member. Employees are initially 480 times more likely to die than family members, but the relative risk ratio declines 4% per day (Table 7).

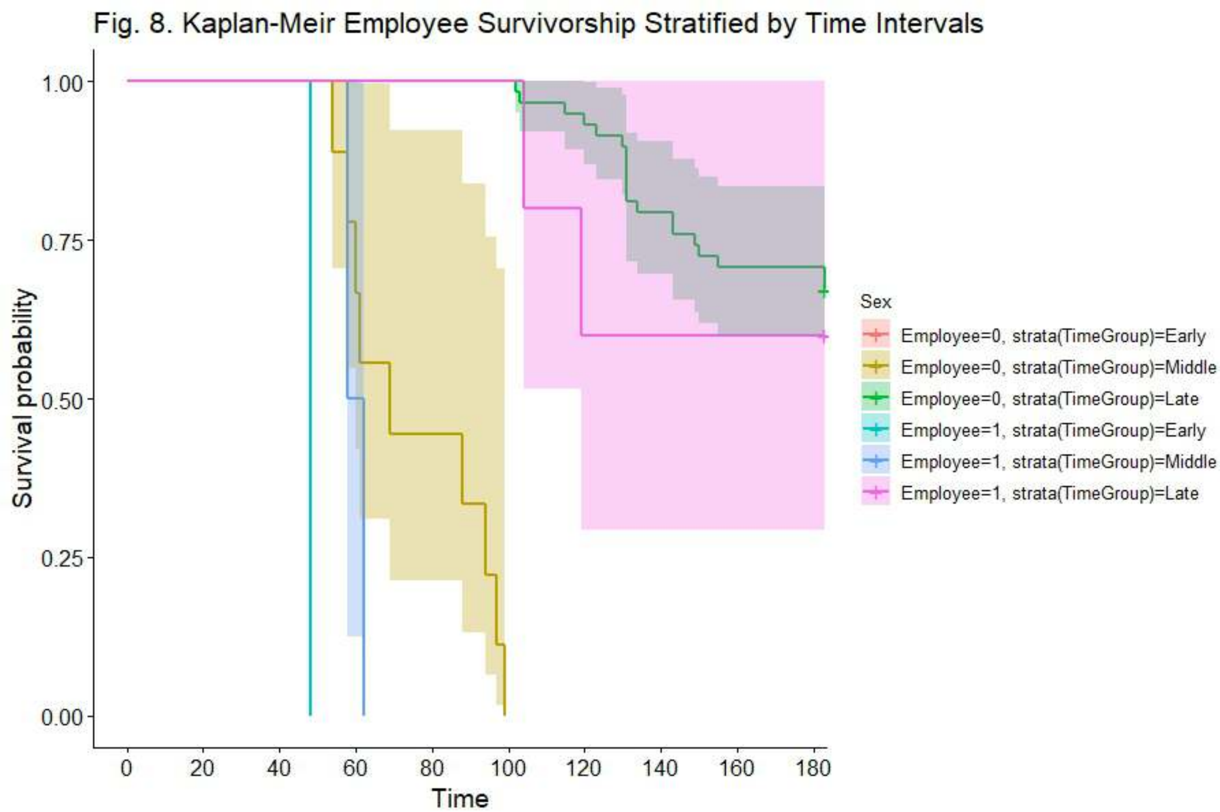

Figure 8. Stratified Kaplan-Meier Survivorship curves as a function of Employee vs. non-Employee status in three time periods with breaks at days 50 and 100. During the Early period only Employees died. In the middle period both groups died at high rates with Employee death rates far exceeding Non-Employees. During the final period, both groups died at a lower rate than the middle period with the Employees dying at a higher rate. During the Late period, there was extensive overlap in 95% confidence intervals.

3.2.2 Effects of Sex on the survivorship of the Forlorn Hope

On 16 December 1846 17 members of the Donner Party left the encampments to cross the crest of the Sierras to get help. Only fourteen had snowshoes, and two without horseshoes, unable to walk in the huge drifts, returned to the encampments on the first day. Because of the dire fate of the fifteen who continued, the group has become known as the Forlorn Hope, a term usually used for troops at the vanguard of a military charge with a low chance of survival. Of the fifteen members of the Forlorn Hope, all 5 women survived but only 2 of the 10 men survived. Of the 8 men who died, 7 were cannibalized.

Using Fisher's exact hypergeometric test, the probability of observing such a Sex-based difference in survivorship (100% to 20%) by chance is just 0.006993. Because of the 100% survival, neither an odds ratio nor 95% confidence limits for the odds ratio can be calculated.

4 Discussion

Why present these analyses? First, the Donner Party has been the subject of more than a dozen books and was the subject of a 1992 Ric Burns PBS documentary. A 2009 movie "Donner Party," was based on the Forlorn Hope rescue mission in which the seven survivors of fifteen resorted to cannibalism.

Cannibalism is the key feature in most tales of the Donner Party, but not all families were cannibals. The Breen and Reed families stated that they did not resort to cannibalism, and both families survived intact. Rarick (2008, p 239) attributes the remarkable survival of the 9-member Breen family to their more abundant supply of beef. The only two Donner survivors 40 or older were Margaret (40) and Patrick (51) Breen. Rarick (2008, p 239) argues that the perfect survival of the 6-member Reed family was not due to food availability but to the indomitable Margaret Reed who had to beg for food after her husband James Reed was driven out of the Donner Party for killing the teamster James Snyder on 5 October 1846.

Second, this study confirms the conclusions of Grayson (1990, 1994, 1997, 2018) and Rarick (2008) that three factors largely control the survivorship of the Donner Party: Age, Sex and Family Group Size. The effects of Family Group Size on survival are strongly confounded with Employee status since the 13 employees were classified as singletons in the Family Group

Size analyses. Employees—teamsters, servants, and a cattle herders—died earlier and at a higher rate than family members (Tables 6 & 7, Figures 6 & 7). Christmas (2008, p 73-74) argues that the high mortality among employees was due to the families becoming more stingy with provisions once starvation and rationing set in. When the food ran out, due to their strenuous exertions in getting the wagons to the encampments 'they were likely the first to go hungry, and the least prepared for it.' Rarick (2008, p 155-156) describes how two employees, Milt Elliot and Eliza Williams, weren't even allowed to die in the same cabin with the Reeds, their former employers who they had considered family.

The Kaplan-Meier survivorship curves (Figure 6) clearly indicate that starvation and cold temperatures created striking increases in male mortality after about 50 days, with employees dying at a high rate (Figures 7 & 8). Males died at a high rate more than a month before notable female mortality (Figure 6). The stratified Kaplan-Meier survivorship curves (Figure 8) reveal that most of the disparity in death rates between Employees and non-Employees occurred before day 100, after which there were just five employees left. The survival of three of the last five employees may have been due to their leaving the encampments. Noah James and Baylis Williams left with the First Relief Group, and Jean Pierre Trudeau left with the Third Relief Group. The first of two remaining employees to die was Milt Elliot, a Reed teamster who tried to walk out with two others on 4 January but returned to the encampments on January 8th and died on 8 February (day 103). The final employee to die was John Denton, a Donner teamster who died on 24 February (day 119) as part of the First Relief journey back to California when the other travelers left him behind on the trail.

Third, the patterns in the data might be of interest to physiologists, anthropologists, archaeologists, or those just curious about how starvation and family ties affect mortality.

Philbrick (2000, p 167) described the tragedy of the Nantucket Whaleship Essex, sunk after being rammed by a sperm whale, the inspiration for Melville's Moby Dick, and cites the Donner Party to explain why 7 of 11 (63.6%) white Nantucket whalers survived on the Essex whaleboats while only 1 of 6 (16.7%) black Essex whalers survived. All but one black whaler were cannibalized. Philbrick argued that the black whalers came from Boston and New Bedford and may have been unhealthier and had a lower fat content than the white Nantucket whalers. Not calculated by Philbrick were the odds of such a racial disparity in survival. The odds of a black whaler dying were 8.75 times that of a white whaler, but the p-value offers only slight evidence to reject the chance null hypothesis (two-sided p = 0.13).

Brown (2009, p 137) calculated the metabolic rate [MR] of Sarah Fosdick (aged 22, MR ~ 3100 kcal/d [cal/d in Brown]) and her father Franklin Graves (aged 57, MR ~ 3600 kcal/d) to explain why she may have survived while he did not. Grayson (2018) presents a thorough review for the possible reasons for the sex-based differential mortality, emphasizing differences in height, weight, and fat content as well as differences in behavior (e.g., sex-based suicide rates). Grayson (1990, 1994, 1997, and 2018) presents statistics and graphical displays to assess the roles of age, sex and family size on survivorship, but used neither nonlinear regression nor formal survivorship analyses to reveal the curvilinear patterns which appear to be the keys to the Donner survivorship patterns.

The survival of every woman but one aged 4 to 40 years is perhaps the most striking pattern in the data. Eleanor Eddy, the only women in that age span to die, died at age 25 on 7 February 1847 a mere 3 days after the death of her 1-y old daughter Margaret and 53 days after Eleanor's husband Edward left her to lead the 17-traveler Forlorn Hope group, leaving her alone with two children and little food.

While the effects of family group size and kinship interactions are important, the Family Group Size effect appear to be strongly influenced by the perfect survival of the Reed and Breen families as shown in Figure 5 and 6, rather than a monotonic pattern in which greater kinship links yield higher survival, one of Grayson's (1990, 1994) conclusions. Teamsters and servants make up the bulk of the single-member families, and Rarick (2008) noted that they were more likely to die from starvation due to their greater exertion reaching the final Donner Pass encampments. Both Rarick (2008) and Brown (2009) argue that the larger body mass of teamsters relative to women and non-employees accounts for the relatively higher male mortality. The Kaplan-Meier survivorship curve (Figure 8) shows that singleton groups, dominated by teamsters and servants, have a tremendously high early mortality, 480 times that of family members (Table 8).

A fourth reason for analyzing the Donner Party data is that these data offer an interesting case study to introduce restricted cubic splines, GAMs, and survivorship analyses to intermediate and upper level statistics classes. Wood (2017, p. 136) commented on his use of a 23-case dataset to introduce the Cox proportional hazard model, "This is clearly a small sample from a statistical point of view, but not from a human point of view, and it is important to try to determine whether there really is evidence for a difference between the treatments." The Donner Party is one of the few historical data, where each datum tells a story.

Only a few of the methods described here are presented in introductory statistics books. Ramsey & Schafer (2013 and two previous editions) use a pared 45-sample Donner dataset to introduce binary logistic regression and Wald and drop-in-deviance tests. Dalgaard's (2008, p 251-258) introductory statistics text introduces Cox and Kaplan-Meier survival analyses and nonlinear curve fitting but covers neither splines nor GAMs. These methods are presented clearly in advanced texts like Harrell (2015), Wood (2017), and Andrews (2021). Harrell (2015) is particularly good on restricted cubic spline regression, for which he wrote the rms::rcs and rms::Predict functions. Wood (2017) provides detailed descriptions of restricted cubic splines, GAMs with his mgcv package, and the Cox proportional hazards model. Zuur et al. (2009) present ecological examples of GAMs based on the first edition of Wood (2017).

Andrews (2021) in a concise chapter on nonlinear regression covers both restricted cubic splines and GAMs. Restricted cubic spline regression permits the user to choose the location of knots or joining points in the curved regressions. In this paper, only the default knot locations provided by Harrell's rms::rcs function were used, e.g., quartiles for 4-knot analyses. Harrell's (2015) book provides examples of selecting specific knot locations. Choosing the appropriate number of knots or the GAM basis function *k* is very important. Simply choosing the model with the lowest AIC appeared to lead to overfitting with too high a number of knots. The 4-AIC threshold used here shifted the optimum knot size to lower values and while the AIC was higher, the fit appeared to be more easily interpreted. The most severe case encountered in this study was fitting both Age and Family Group Size using rcs to model survivorship (Table 5 and Figure 5). Using the AIC criterion with the 4-AIC threshold produced knot sizes of 6 for both parameters with an AIC of 56.33. The Wald statistics were high and the 3-d figure had little predictive value with predicted survival tending to flex around a few individual extreme observations. I then performed a *k*-fold cross validation analyses which found 3 knots for Age and 5 knots for Family Group Size produced the best log loss statistic, used as the optimization parameter for binary logistic regression by the caret function. The results shown in Figure 5 are clearly interpretable. The take-home messages are: 1) Don't use AIC alone to choose knot size for rcs as it tends to select too many knots leading to overfitting, 2) *k*-fold cross validation appears to work well to find the optimal *k*'s for both restricted cubic splines and GAMs, 3) The goal of this study was explanation, not prediction, and the AIC criterion of using AIC with a 4 unit threshold or *k*-fold cross validation produced results which were more easily interpreted.

Much of the R coding for this paper was aided by OpenAI's GPT-4. GPT-4 through its training through September 2021 is fully aware of how to program using Harrell's R rms package, in particular his Glm, rcs and Predict functions. GPT-4 reported on more than one occasion that it hadn't been trained on Harrell's (2015) text but was aware of his R packages. GPT-4 is also adept at coding GAM analyses using Wood's mgcv function (Wood 2017, 2019), including *k*-fold cross-validation to select the appropriate value of the *k*, the GAM basis dimension. GPT-4 also wrote the code for the Cox and Kaplan-Meier survivorship functions and prompted me to ask for analyses to check the Cox proportional hazard assumption, which I did.

Supplementary Material

To demonstrate another of GPT-4's skills, I provided GPT-4 the abstract to this paper and asked it to write a villanelle in iambic pentameter summarizing the results:

The Curves of Survival in Snow's Embrace

In the face of stark Sierra’s bitter chill,
The Donner Party’s tale emerges still,
Where cubic splines and GAMs unveil the case.

With eighty-seven souls, they faced the hill,
Yet causes claimed eight, their numbers to distill,
In the face of stark Sierra's bitter chill.

Their numbers, scarce, yet patterns show the skill,
Of analyses which curves of life distill,
Where cubic splines and GAMs unveil the case.

Young females thrived, while men did face the kill,
Age marked their fate, a grim and steady spill,
In the face of stark Sierra's bitter chill.

Families six and nine, stood firm and still,
Yet lone employees faced a dire ill,
Where cubic splines and GAMs unveil the case.

This tale of snow and strength, of grit and thrill,
Shows man and method, fate and analytic skill,
In the face of stark Sierra's bitter chill,
Where cubic splines and GAMs unveil the case.

Acknowledgments

Many thanks to Donald Grayson whose Donner scholarship inspired this paper and who provided the latest Donner demographic data. Thanks to Frank Harrell whose 2015 book and 2021 4-d course introduced me to restricted cubic spline regression.

References

Andrews, M. (2021), *Doing Data Science in R: An Introduction for Social Scientists*, Los Angeles, CA: Sage.

Bolker, B. M. (2008), *Ecological Models and Data in R*, Princeton, NJ: Princeton University Press.

Brown, D. J. (2009), *The Indifferent Stars Above: the Harrowing Saga of the Donner Party*, Boston: HarperCollins.

Burnham, K. P. and Anderson , D. R. (2004), "Multimodal inference: understanding AIC and BIC in model selection," *Sociological Methods and Research 33*, 261-304.

Christmas, B. K. (2008), *Tragedy in the Sierra Nevada: a Narrative of the Donner Party*. Charleston: CreateSpace.

Grayson, D. K. (1990), "Donner Party Deaths: A Demographic Assessment". *J. Anthropological Research 46*: 223-242.

Grayson, D. K. (1994), "Differential mortality and the Donner Party disaster". *Evolutionary Anthropology 2*: 151-159.

Grayson, D. K. (2018), *Sex and Death on the Western Emigrant Trail: The Biology of Three American Tragedies.* Salt Lake City: University of Utah Press.

Grayson, D. K. (1997) "The timing of Donner Party deaths" Appendix 3 in Hardesty, D. L. (1997) *The Archaeology of the Donner Party*, Reno: University of Nevada Press.

Harrell, F. E. (2015), *Regression Modeling Strategies, 2nd edition*. New York, NY: Springer.

Harrell, F. E. (2021), Regression Modeling Short Course. Tuesday 11 May 2021 notes.

Philbrick, N. (2000), *In the Heart of the Sea: The Tragedy of the Whaleship Essex*. New York, NY: Penguin Books

R Core Team. (2023), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ramsey, F. L. and D. W. Schafer. (2013), *The Statistical Sleuth: a Course in Methods of Data Analysis, 3rd Edition*. Brooks/Cole Cengage Learning, Boston MA, 760 pp.

Rarick, E. (2008), *Desperate Passage: The Donner Party's Perilous Journey West*. Oxford University Press, Oxford. 304 pp.

Stewart, G. E. (1960), *Ordeal by Hunger: the Ordeal of the Donner Party*. Boston: Houghton Mifflin. 392 pp.

Wickham, H. and G. Grolemund. (2017), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Sebastapol CA. 492 p.

Wood, S. N. (2017), *Generalized Additive Models: an Introduction with R*. CRC Press, Boca Raton. 476 p.

Wood, S. (2019), Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothing Estimation. **https://CRAN.R-project.org/package=mgcv**.

Zuur, A. F., E. N. Ieno, M. J. Walker, A. A. Saveliev, and G. M. Smith. (2009), *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York. 574 pp.

Data Availability

Data

https://raw.githubusercontent.com/EugeneGall/donner-data-analysis/main/Donner.csv

R Code

https://raw.githubusercontent.com/EugeneGall/donner-data-analysis/main/Donner_Gallagher_Public.R

This manuscript

https://raw.githubusercontent.com/EugeneGall/donner-data-analysis/main/Donner_Gallagher_MS.pdf