DONNER PARTY SURVIVAL: A REASSESSMENT WITH NONLINEAR REGRESSION WITH R

Eugene D. Gallagher
School for the Environment
UMass Boston
Boston MA 02125
8/13/23

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 analysis of binary response data. 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 sample size limitations, the rcs and GAM analyses reveal striking curvilinear patterns in survivorship, patterns not previously noted for these data. Survivorship was nearly perfect for females aged 5 to 40, whereas males had low survival decreasing linearly with age. Of the 79 snow-trapped travelers no females and only one male past age 40 survived. 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 teamsters and servants who died earlier and at much higher rates than females and travelers with family ties.

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 after most resorted to cannibalism.

Grayson (1990, 1994, 1997) analyzed the demographics of Donner Party survival, concluding "...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 1990, p 223)." The same conclusion was reached by Rarek (2008), who also stressed the low rates of survival of the teamsters and servants who worked for the families.

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 statistics 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 data 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, as in Ramsey & Schafer (2013), nonlinear regression using restricted cubic splines and generalized additive models (GAMs) and survival analysis. The following analyses are in the spirit of Harrell (2015) and Andrews (2021) who teach binary logistic regression and 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, plotly, 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 below). The demographic data were mostly adapted from Table 1 in Grayson (1990), which was largely based on Stewart's (1960) roster. Age, occupation, cause of death, and date of death were obtained from Stewart (1960), Brown (2009), and Rarek (2008). Rarek's (2009, p. x-xi) ages and occupations largely match Grayson (1990), but when they conflicted Grayson's dates were used. Grayson (1990) did not contain ages or first names for the Wolfingers, but Rarek (2008, p xi) lists Doris Wolfinger's age as 20. Mr. Wolfinger was dropped since he was murdered by Reinhardt & Spitzer for his money and rifle prior to the Donner Party entering the mountains (Stewart 1960, p 166).

I'll test combinations of four covariates—Age, Sex, Family Group Size and Teamster & Servant status— on pared data (all ages, n=79) from which the 8 travelers who died or crossed the Sierra Nevada mountains before the first major snowfall on 28 October 1846 are 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. Stewart's (1960 p 363-364) Roster is the source of Grayson's (1990) Table 1, and Stewart's Roster provides the source of the differences between family names and Family Group Sizes. 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 was a high priority, so 3 knots, the minimum allowed in rcs regression, was used if its model likelihood was within 4 AIC 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) presented rules-of-thumb 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. The appropriate *k*, the basis function parameter controlling the smoothing of the GAMs, was determined by *k*-fold cross-validation using the caret package, choosing the k basis function or smoothing parameter that produced the lowest Root Mean Square Error.

Tests of null hypotheses use Wald Chi-square tests and Wilks' Chi-square drop-in-deviance test, described in Ramsey & Schafer (2013). The curved surface plot (Figure 7) was generated using the plotly package.

Cox proportional hazards models and Kaplan-Meier survivorship curves were generated using the survival and survminer packages. Gallagher 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. 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 results are shown in Table 1 and Figure 1. The regression coefficients are not provided as they have no simple interpretation for restricted cubic spline regression (Harrell, 2021).

Table 1. Wald statistics and effect sizes for the rcs(Age, 5 knot) x Sex interaction model.

1.1 Wald statistics

Factor	Chi-Square	d.f.	P
Age (Factor+Higher Order Factors)	9.98	4	0.0408
All Interactions	5.54	2	0.0626
Nonlinear (Factor+Higher Order Factors)	8.67	2	0.0131
Sex (Factor+Higher Order Factors)	7.67	3	0.0535
All Interactions	5.54	2	0.0626
Age * Sex (Factor+Higher Order Factors)	5.54	2	0.0626
Nonlinear	5.46	1	0.0195
Nonlinear Interaction : f(A,B) vs. AB	5.46	1	0.0195
TOTAL NONLINEAR	8.67	2	0.0131
TOTAL NONLINEAR + INTERACTION	8.84	3	0.0315
TOTAL	11.13	5	0.0489

1.2 Effect sizes A 14-y old female's odds of survival were 40 times higher (exp(3.68)) than a 14-y old male (95% CI: 2.7 to 570 times)

Factor	Effect	S.E.	Lower 0.95	Upper 0.95
Age	0.31576	0.63733	-1.5859	0.95444
Sex - Female:Male	3.68050	1.33500	1.0199	6.34110

Adjusted to: Age=14 Sex=Male

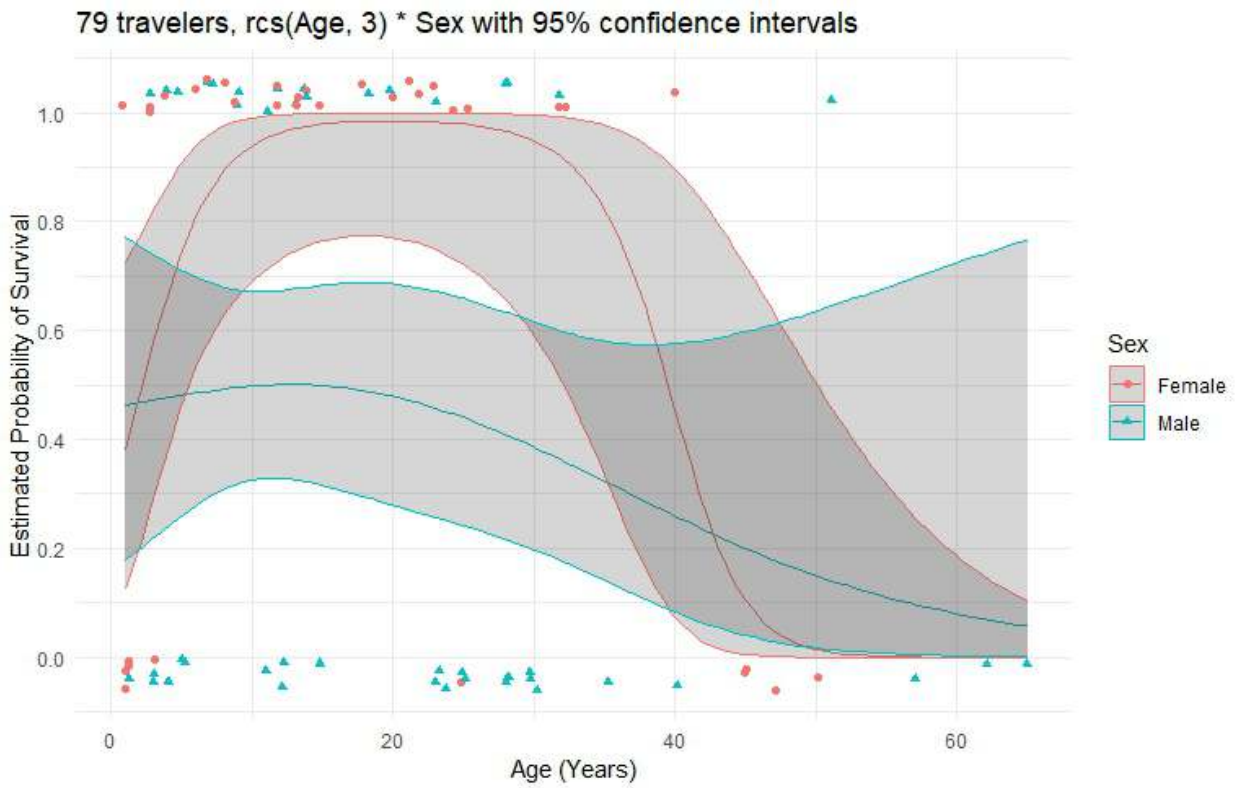

Figure 1. Display of the restricted 3-knot cubic spline model of the interaction of Age and Sex. Between the ages of 4 and 42 only one Female died (Eleanor Eddy, Age 25). Only two travelers survived past age 40: Mary Breen (40) and Patrick Breen (51).

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 are shown in Table 2 and Figure 2.

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

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       10.545      3.554   2.967  0.00301

Approximate significance of smooth terms:
                                       edf Ref.df Chi.sq p-value
s(Age):as.numeric(Sex == "Male")     2.000  2.000 11.152 0.00379
s(Age):as.numeric(Sex == "Female") 1.909  1.992  7.976 0.01711

Rank: 6/7
R-sq.(adj) =  0.258   Deviance explained = 24.1%
UBRE = 0.17052  Scale est. = 1         n = 79
```

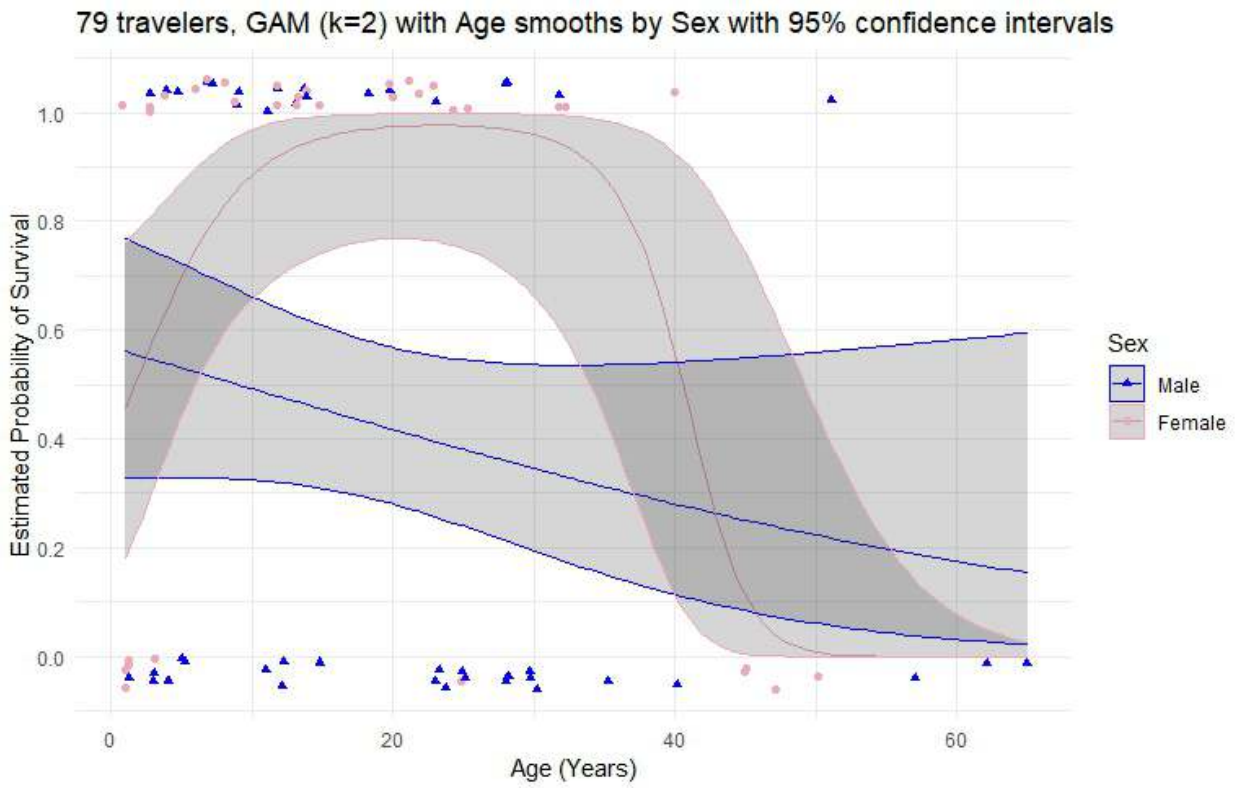

Figure 2. Display of the GAM model of Sex and Gender, with the model basis parameter k determined to be 2 by a k-fold cross-validation analysis. The curvilinear relationship is similar to that shown in Figure 1 using a restricted cubic spline regression.

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 Wald tests (Table 3), but the relationship was not linear as shown in Figure 3. The AIC for knots 3 through 7 were 104.14, 101.39, **100.65**, 99.11 and 97.11 making 5 knots the appropriate choice based on an AIC threshold difference of 4 AIC units.

Table 3. Wald statistics for Family Group Size. The restricted cubic spline regression with 5 knots for Family Group Size provides strong evidence for a curvilinear survivorship pattern.

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

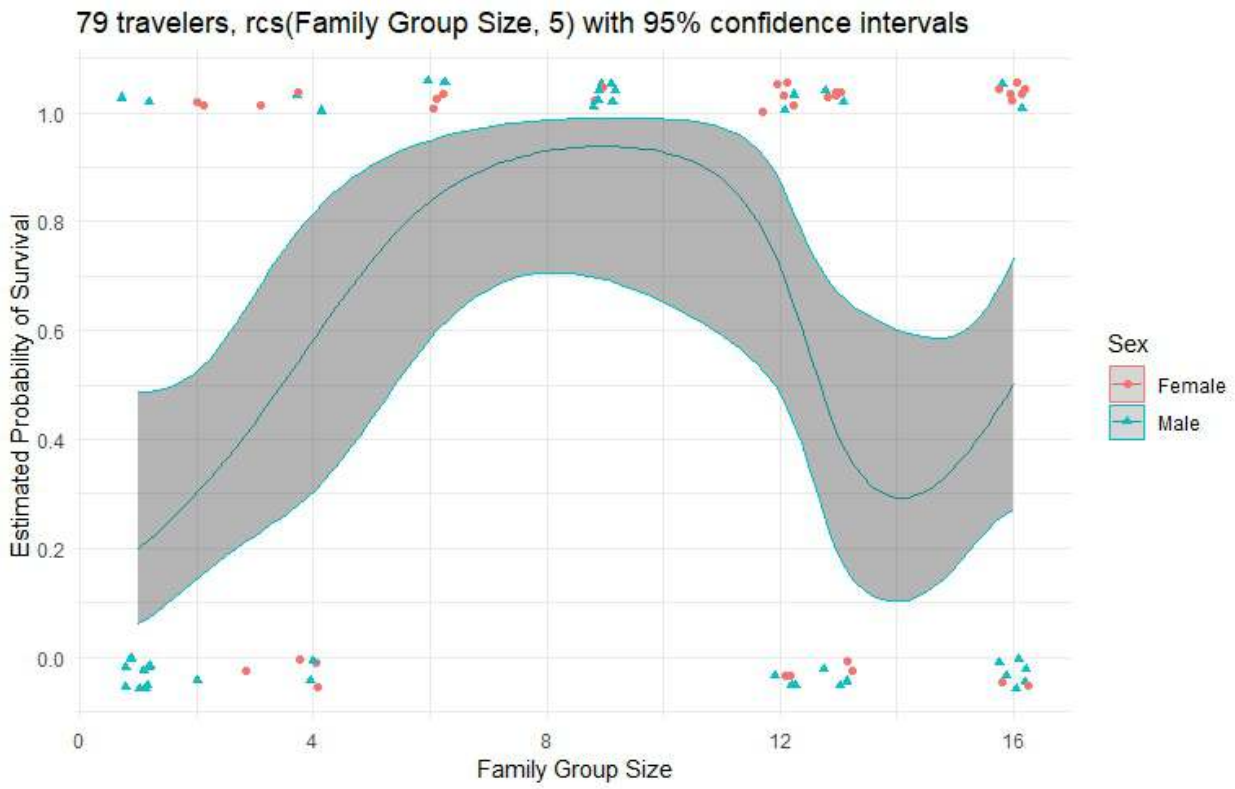

Figure 3. Effect of Family Group Size on Survival, modeled with a restricted cubic spline with 5 knots. In Family Group Sizes of six (the Reed Family) and nine (the Breen family), everyone survived.

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* =3.

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.1921     0.2423   0.793    0.428

Approximate significance of smooth terms:
                           edf Ref.df Chi.sq p-value
s(Family_Group_Size) 1.905  1.991  9.709 0.00651
R-sq.(adj) =  0.122   Deviance explained = 11.2%
UBRE = 0.2974  Scale est. = 1         n = 79
```

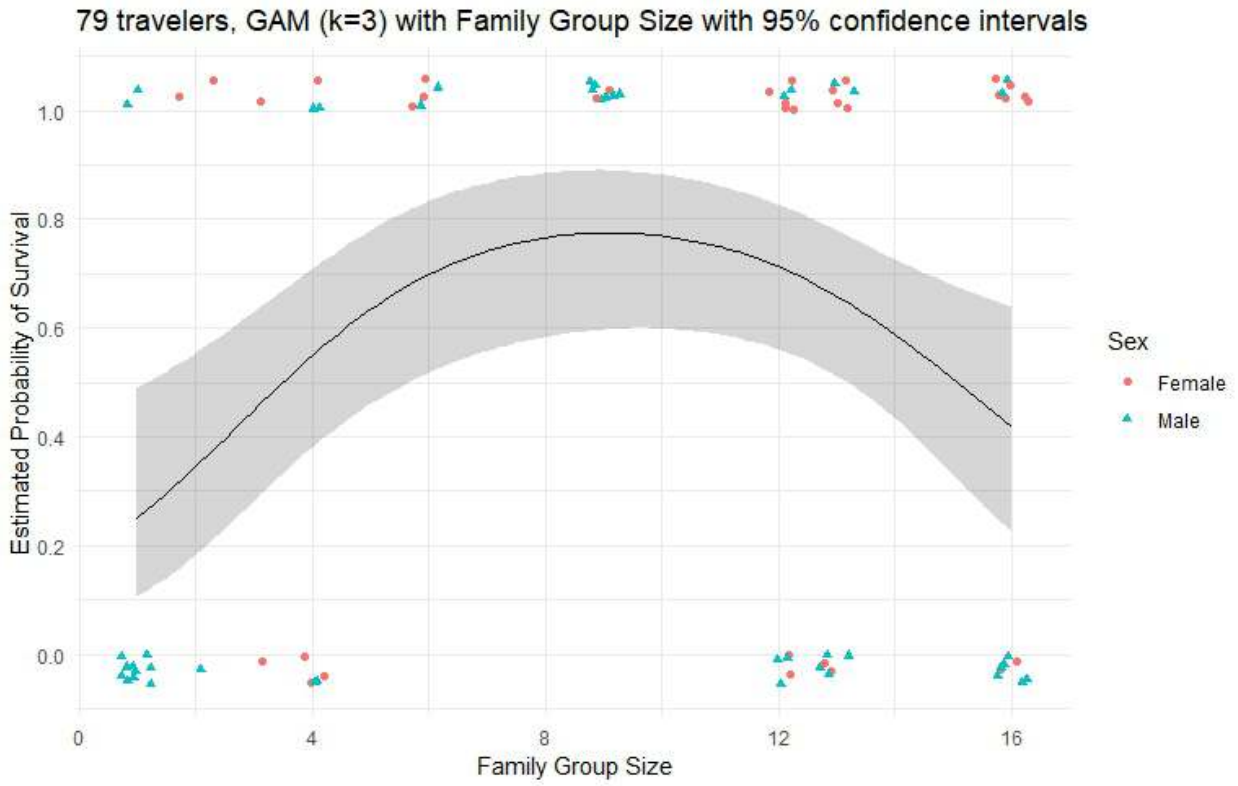

Figure 4. Effect of Family Group Size on Survival, modeled with a GAM with *k* = 3, chosen using *k*-fold cross validation. Note that the reduced number of knots produced a smoother function than restricted cubic spline regression with 5 knots (Figure 3).

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 3 knots should be used for Age and 5 knots for Family Group Size. Table 5 displays the Wald Chi-square tests and Figure 5 shows a three-dimensional view of survivorship.

Table 5. Wald statistics for the rcs(Age, 3) *Sex + rcs(Family Group Size, 5) The odds of a 14-year old female surviving were 30 times higher (exp(3.3963) than a 14-y old male (95% CI: 1.8 to 490 times).

5.1 Effect Sizes

Factor	Low	High	Diff.	Effect	SE	Lower 0.95	Upper 0.95
Age	6.5	28	21.5	2.3705	1.1623	0.051764	4.6892
Family_Group_Size	4.0	13	9.0	-1.0217	1.0496	-3.115500	1.0721
Sex - Female:Male	2.0	1	NA	3.3963	1.4062	0.591090	6.2015

Adjusted to: Age=14 Sex=Male

5.2 Wald Statistics

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.3291	1.9934	-2.673	0.00751
rcs(Age, 3)Age	0.4884	0.1641	2.977	0.00291
rcs(Age, 3)Age'	-1.3871	0.4987	-2.782	0.00541
SexMale	-1.6188	1.5400	-1.051	0.29318
rcs(Family_Group_Size, 5)Family_Group_Size	0.8454	0.5523	1.531	0.12581
rcs(Family_Group_Size, 5)Family_Group_Size'	2.3890	3.4037	0.702	0.48275
rcs(Family_Group_Size, 5)Family_Group_Size''	-8.4285	6.7530	-1.248	0.21199
rcs(Family_Group_Size, 5)Family_Group_Size'''	138.8141	54.4293	2.550	0.01076
rcs(Age, 3)Age:SexMale	-0.1740	0.1868	-0.931	0.35181
rcs(Age, 3)Age':SexMale	0.7191	0.5170	1.391	0.16425

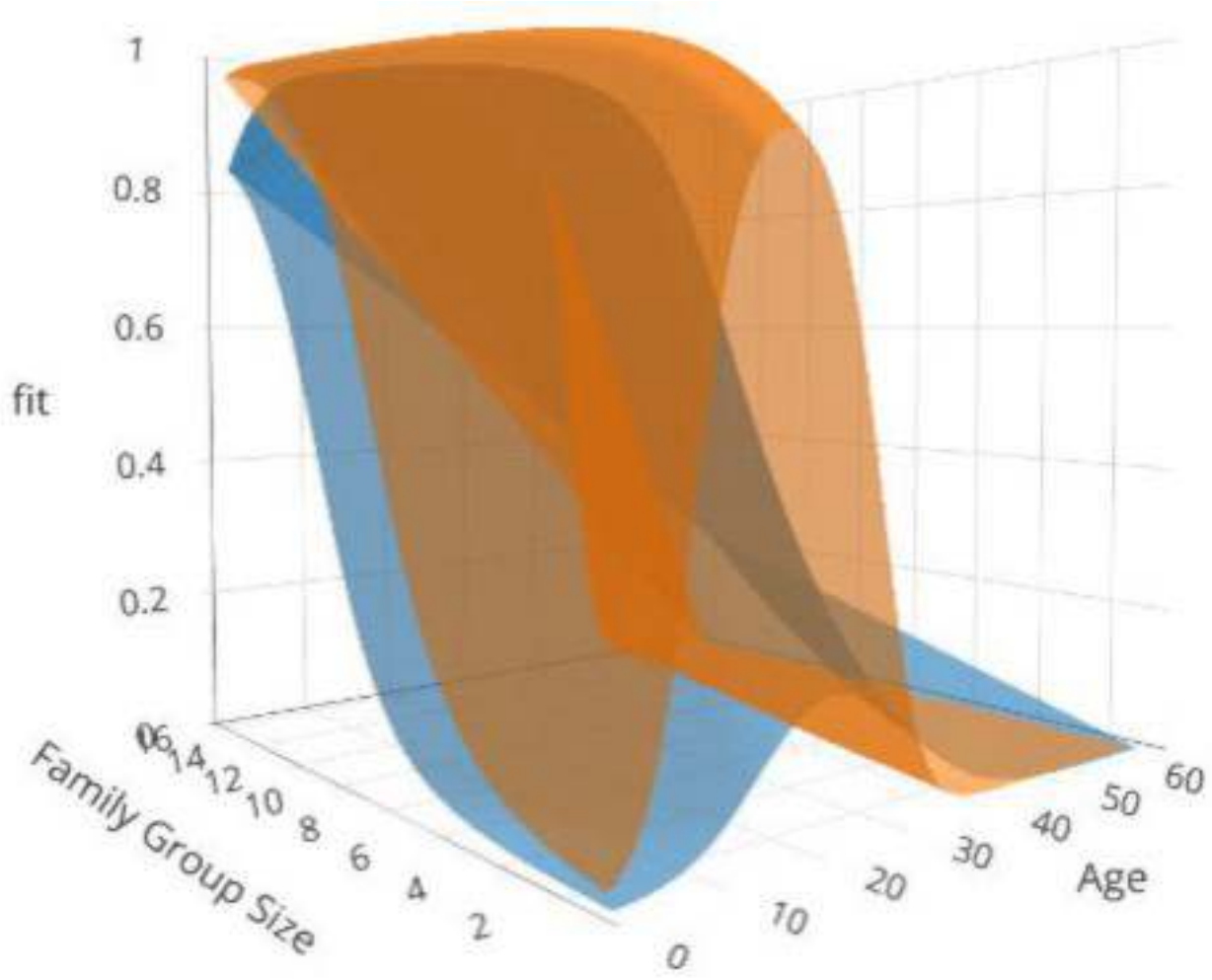

Figure 5. Three dimensional perspective of the probability of survival in the Donner Party as a function of Age and Family Group Size, modeled with restricted cubic splines with 3 and 5 knots, respectively. The Female fitted surface is orange, and the Male surface is blue. The striking Survival of all but one female between the ages of 5 and 41 and the high death rates of Males combined with the complete survival of the Reed (size 6) and Breen (size 9) families cause the age-related humps in Male and Female survival.

3.2 Survival analysis

3.2.1 Effects of Gender

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 storm (Harriet McCutcheon, age 1) with 11 males dying in the first 15-d interval.

Table 6. Cox proportional hazard model The effects of Sex on Survival Time were analyzed. There was a pronounced gender effect on survival time (p = 0.01), with males being 2.6 times more likely to die (1.2 to 5.5 95% CI). A cox.zph test in the survival package indicated no violation of the equal 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

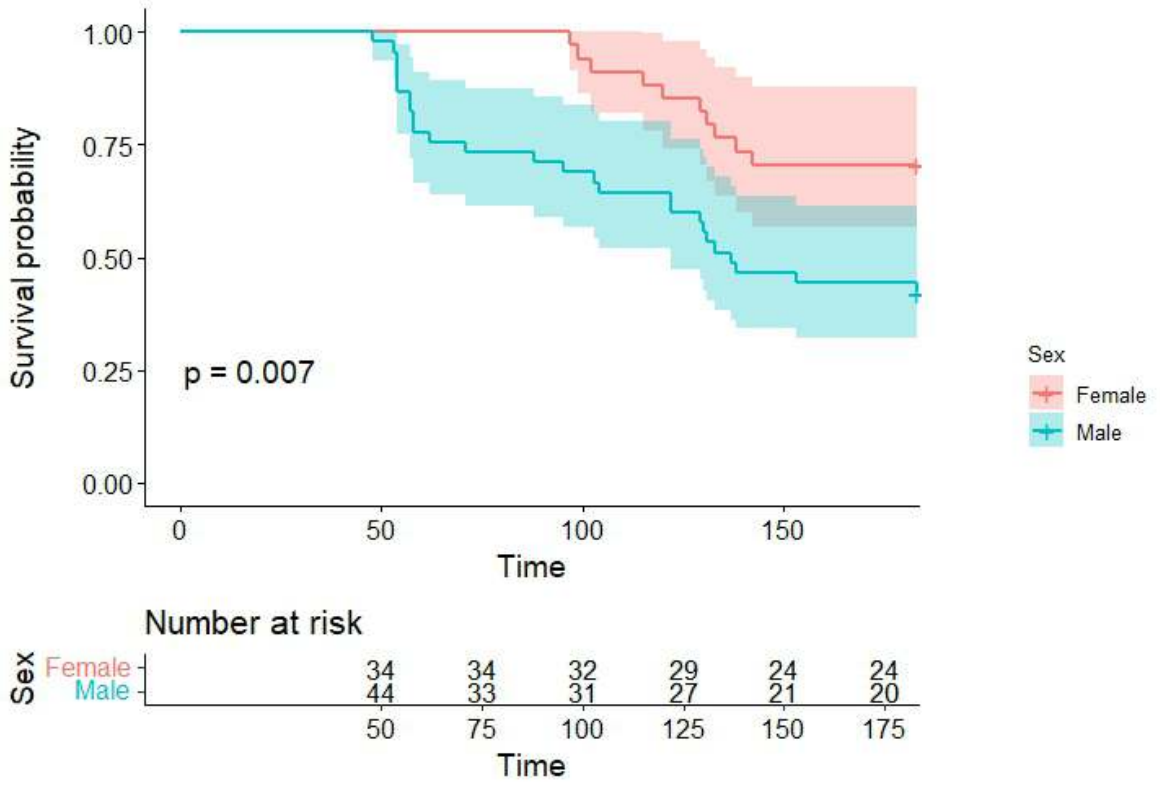

Figure 6. Kaplan-Meier Survivorship curve as a function of Gender. Males are 2.6 times more likely to die than females (Table 6). The equal proportion assumption is tenable.

3.2.2 Effects of working class

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; the others are not included in the analyses. The survivorship analysis indicated that these 9 employees died more rapidly than family members or non-employee bachelors (Table 9, Figure 7).

Table 9. Cox proportional hazard model Analysis of Teamster/Servant vs. Family member survival. Teamsters died at 4.5 times the rate 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

                     chisq df     p
Teamster_Hired_Hands  6.45  1 0.011
GLOBAL                6.45  1 0.011
```

The Cox proportional hazards test (survival::cox.zph) indicated that the hazard ratio of Teamsters & Servants to Family members was not constant with time, violating the Cox equal proportion assumption (p = 0.011), so an additional covariate, Survival_Time, was added to the model. A plot of Schonfield residuals indicated a reasonable fit to the proportional hazards model with the relative risk ratios dropping with time. The results are shown in Table 10 and Figure 8.

Table 10. 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: coxph(formula = Surv(Survival_Time, Death) ~ Teamster_Hired_Hands + SurvTime_Teamster, data = Donner)

```
  n= 79, number of events= 36

                          coef exp(coef)  se(coef)      z Pr(>|z|)
Teamster_Hired_Hands   6.17773 481.89610   1.27357  4.851 1.23e-06 ***
SurvTime_Teamster     -0.04490   0.95609   0.01401 -3.205  0.00135 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                     exp(coef) exp(-coef) lower .95 upper .95
Teamster_Hired_Hands  481.8961   0.002075   39.7090 5848.1386
SurvTime_Teamster       0.9561   1.045927    0.9302    0.9827

Concordance= 0.622  (se = 0.035 )
Likelihood ratio test= 26.59  on 2 df,   p=2e-06
Wald test            = 30.87  on 2 df,   p=2e-07
Score (logrank) test = 62.76  on 2 df,   p=2e-14
```

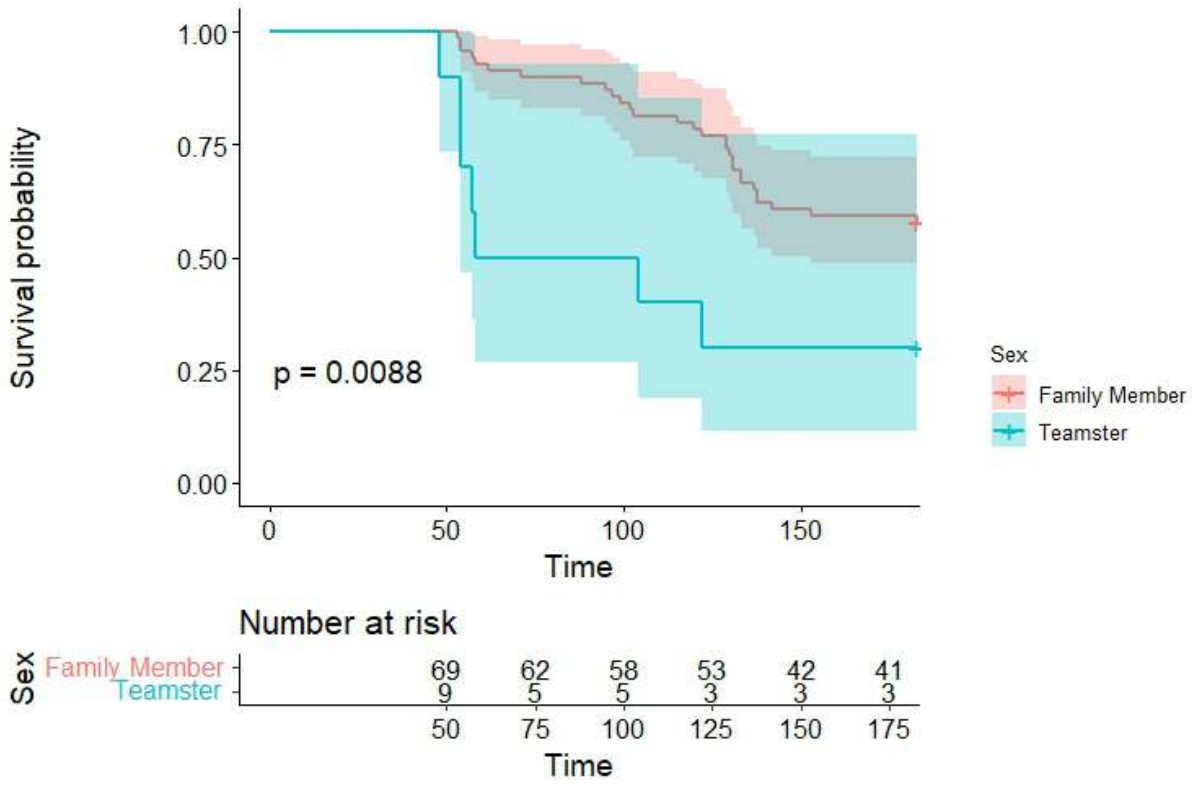

Figure 7. Kaplan-Meier Survivorship curves as a function of Teamster or Servant vs. Family Member or non-employed bachelor. Teamsters & Servants are initially 480 times more likely to die than family members, but the relative risk ratio declines 4% per day (Table 10).

4 Discussion

Why present these analyses? 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 17 Donner Party members ventured over the Sierras with only seven surviving the trip after resorting cannibalism. Cannibalism is the key feature in most tales of the Donner Party, but not all families were cannibals. The Breen and Reed families didn’t cannibalize, and both families survived intact. Rarek (2008, p 239) attributes the complete survival of the 9-member Breen family to their more abundant supply of beef. The only two Donner survivors 40 or older were Mary (40) and Patrick (51) Breen. Rarek

(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.

Second, this study confirms the conclusions of Grayson (1990, 1994, 1997) and Rarek (2008) that three factors largely control the survivorship of the Donner Party: Age, Sex and Family Group Size. Note that family group size is strongly confounded with employee status since those workers were classified as singletons in the Family Group Size analyses. As noted perceptively by Rarek (2008), those teamsters and servants died earlier and at a higher rate than family members, a conclusion stoutly confirmed by the survivorship analyses. The Kaplan-Meier survivorship curves (Figures 6 and 7) clearly indicate that starvation created striking increases in male mortality after about 50 days, with teamsters dying at a high rate more than a month before notable female mortality.

Whereas Grayson (1990, 1994, 1997) presented statistics and graphical displays to assess the roles of gender, age and family size on survivorship, he neither used nonlinear regression nor formal survivorship analyses to reveal the curvilinear patterns which appear to be the keys to the Donner survivorship data.

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. For example, Nathaniel Philbrick (2000, p 167) in describing the tragedy of the Whaleship Essex, which also involved cannibalism, cites the Donner Party to explain why white Nantucket whalers on the Essex with higher fat content outlived the Black Essex whalers from Boston and New Bedford who may have had a lower fat content. 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 and he did not.

The survival of every woman aged 4 to 40 years but one is perhaps the most striking pattern in the data. Eleanor Eddy, the only women in that age span to die, died aged 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 Forlorn Hope rescue group over the Sierras, leaving her alone with two children. Only 7 of the 17 Forlorn Hope group survived their snow-shoed trek to California, but all 5 women in the Forlorn Hope group survived.

While the effects of family group size and kinship interactions are important, the results 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 Rarek (2008) noted that they were more likely to die from starvation due to their greater exertion climbing the Wasatch mountains, the trek across the Great Salt Lake desert, and the ascension up the lower Eastern slope of the Sierra Nevada mountains. Both Rarek (2008) and Brown (2009) argue that their larger body mass 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 10).

Finally and most importantly, 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 datasets, where each datum tells a story.

Only a few of the methods described here are presented in introductory statistics books, like Ramsey & Schafer (2013). Dalgaard (2008, p 251-258) provides an introductory level treatment of Cox and Kaplan-Meier survival analysis 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 function. 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).

Much of the R coding in this paper was aided by OpenAI’s GPT-4. GPT-4 through its pretraining through September 2021 is fully aware of how to program using Harrell’s R rms package, in particular his Glm and rcs 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 (Wood 2017, 2019), including *k*-fold cross-validation to select the appropriate value of the *k*, the basis function or “smoothing” parameter. I was unaware of how to plot 3-dimensional surface plots in R, so I prompted GPT-4 to provide the code to analyze the response variables I specified, and it wrote 20 lines of R code using the plotly package to do so. GPT-4 also wrote the code for the Cox and Kaplan-Meier survivorship

functions and prompted me to ask for analyses to check the proportional hazard assumption, which I did.

Supplementary Material

To demonstrate another of GPT-4's skills, I prompted GPT-4 to write a villanelle summarizing the results from this paper:

Endurance of the Fairer

In days of old, the data tells the tale,
Of men and teamsters, lost before their prime.
While women strong, through youth and storm, prevail.

The splines and GAMs trace curves that never fail,
To show the young and aged fall in time.
In days of old, the data tells the tale.

The children's cries, the elderly turn pale,
But women midst of age, they climb and climb.
While women strong, through youth and storm, prevail.

The families of middle size unveil,
A thriving pulse, a rhythm in their rhyme.
In days of old, the data tells the tale.

The men who worked, their lives so often frail,
Before the age of forty met their chime,
While women strong, through youth and storm, prevail.

One woman lost, her story leaves no trail,
Yet all her sisters through the toughest clime,
In days of old, the data tells the tale,
While women strong, through youth and storm, prevail.

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.

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. (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.

Rarek, 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