**Directions:** This homework is the culminating experience for Learning Module 5: Logistic Regression. This homework is to be completed individually. While you may discuss problems with each other, the work you turn in must be your own. You will turn in this cover page with your assignment. Note: Any problem that begins CH Problem is a problem from the textbook.

By signing your name below, you verify that the work done on this homework is your own.

**Print Name:** Monica Klosin

**Submitting Your Assignment:** The following steps describe how to access the homework assignment and then submit it via BB.

1. Access the Word file homework assignment saved to BB. If you are reading this, then you have done Step 1 successfully!
2. Complete your homework. You may write your answers by hand or you may type your answers into this document. Whichever you choose, be sure to do the following:

* Write in dark ink or dark pencil if you are writing answers by hand.
* Show your work on any required calculations.
* Use the snipping tool to embed any R output you are asked to produce directly into your homework paper. If you choose to write out answers by hand, then print the necessary R output and place it in the appropriate spot on the homework. Please do not place all R output at the end.

1. Scan your homework into a single PDF. I do not want separate PDFs for each page! The Math and Stats (Tutoring) Center has two videos that show how to use a smartphone to do this:

iPhone – A short video named, “Scan With iPhone” that describes how to scan your homework papers with an iPhone is saved to BB under the Technology Menu item.

Android – A short video named, “Scan With Android” that describes how to scan your homework papers with an Android is saved to BB under the Technology Menu item.

1. Upload your PDF to BB – A short video named “Uploading PDF to BB” that describes how to upload your homework PDF is saved to BB under the Technology Menu item.

**Scoring:**

* Any student who chooses to turn in the assignment on or before the due date given in class will be eligible to score 100 points on the assignment.
* Any student who chooses to turn in the assignment by midnight on Tuesday will be eligible to score 80 points on the assignment.
* Any student who chooses to turn in the assignment after midnight on Tuesday will have the assignment reviewed and commented on but the assignment will not be eligible to score any points.

**There are three pages to this homework assignment including the cover page.**

**Software Investigation**

Since we are on homework 5 and the semester is winding down, I am going to give you a break on this Software Investigation. Instead of writing code, I am going to give you R code and output that you can use to answer questions.

We use a data set on 1012 people connected with the People’s Temple of Jonestown on November 18, 1978. The data comes from the Alternative Considerations of Jonestown and Peoples Temple website from San Diego State University (<https://jonestown.sdsu.edu/>). Quoting from the website.

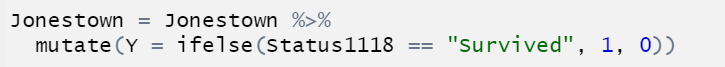
*“Jonestown” is a word with several meanings. First, it refers to an agricultural project established by the Peoples Temple, a religious group based in California which immigrated to Guyana in the mid-1970s to establish an agricultural utopia. Second, it refers to the events of 18 November 1978 in which a U.S. Congressman was assassinated, along with four other individuals, at a jungle airstrip in South America. These tragic killings were followed by the mass murders and suicides of 900 men, women, and children by ingesting potassium cyanide mixed into a vat of fruit punch and tranquilizers. Finally, the word “Jonestown” has been used to describe any New Religious Movement which may or may not have the potential for violence, as in “Heaven’s Gate was another Jonestown.”*

The variables in the dataset are:

* ID – Observation Identifier
* DOB – Date of Birth
* PlaceOB – Place of Birth
* BirthState – Birth State
* BirthCountry – Birth Country
* GuyanaEntry – Date Person Entered Guyana
* Status1118 – Status on 11/18/1978
* Age1118 – Age on 11/18/1978
* Gender – m = male, f = female

**We eliminated any observations where the age and/or gender was unknown. This left us with a data set of 1000 observations. All summaries and models are based on these observations.**

1. The program begins with the following code:



1. What is a success in the binary response variable Y?

If there is a success in the binary response variable Y, Y=1, this means the individual survived on or before November 18th, 1978. If Y = 0, the individual died on or before November 18th, 1978.

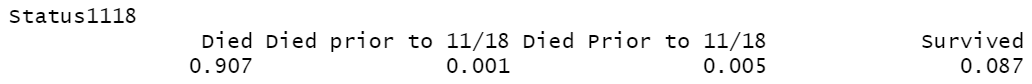
1. True/False: After running this code the data frame Jonestown will have more rows?

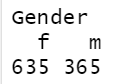
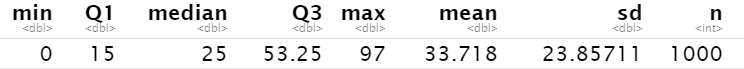
False. mutate created another column Y, not another row.

1. True/False: After running this code the data frame Jonestown will have more columns?

True. mutate created another column Y.

1. The output below is a simple exploratory data analysis of Status1118, Gender, and Age1118.



1. What proportion survived?

8.7% of the population survived, the proportion is 0.087.

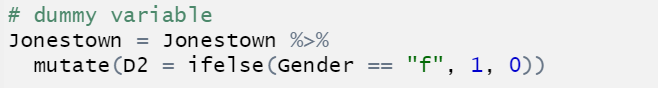
1. How many females are included in the data set?

635 females are included in the data set.

1. What do the mean and median suggest about the shape of the distribution of ages?

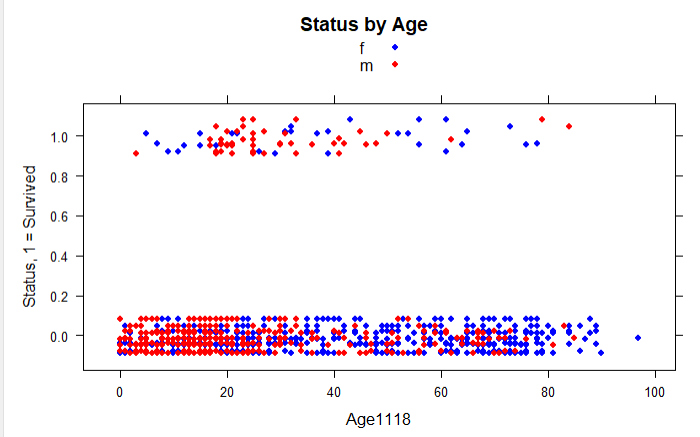
The mean is 33.718, and the median is 25. Since the median is less than the mean, this means the distribution of ages is skewed right.

1. True/False: The code below creates a dummy variable named D2 for X2 where D2 = 1 if Gender = f, and D2 = 0 otherwise.



True, if we consider X2 the variable Gender from the dataset.

1. The following is a jittered scatterplot of Y versus Age1118.



1. Write a sentence that explains how the scatterplot reflects that the proportion of those who survived is 0.087.

Based on the scatterplot above, there are a lot more dots around the Y = 0, indicating dead, than at Y = 1, which indicates survived. It makes sense that only 8.7% of the population survived based on this graph, since there are so many less dots at Y = 1 than at Y = 0.

1. True/False: Males at Jonestown tend to be younger than females.

True.

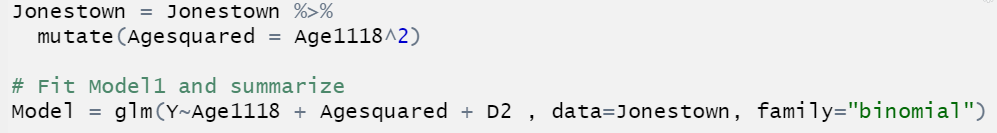
1. It appears that those under 18 and those over 60 are less likely to survive than those between 20 and 60. What variable that is a function of Age1118 would allow us to account for this?

Variable Status1118 along with Age1118 would help indicate if an individual survived or died. We could compare the age range and the average status of the individuals and compare the two groups via a odds ratio.

1. It is difficult to tell from the graph whether there is a difference in survival based on Gender. What tool from STA 215 could be used to determine this? There are several possible answers.

You could write a linear regression model that you would regress survival on Gender. Y = beta0 + beta1(Gender). Coefficents of beta1 tells how much more females are dead compared to males.

1. Consider the code below.

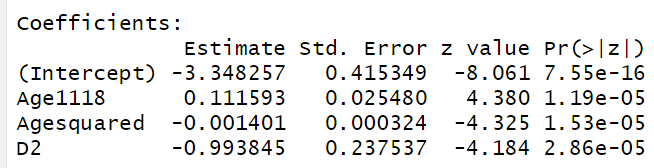


1. What does the mutate code do?

The code creates a new variable, and a new column, named Agesquared which holds values that are the values in Age1118 squared.

1. Write out the model for the probability that a person survives. Use the notation  for the variable from part (a) that was created.

1. The output below is a summary of the R object named Model.



1. True/False: For any age, females are less likely to survive than males?

True.

1. What type of predictor variable could potentially have females less likely to survive for some ages and more likely to survive for other ages?

Age1118^2. A quadratic equation would show a curved nonlinear relationship, it would show higher probability of survival for some ages vs others.

1. Find the probability that a 20-year-old male survived.

beta0 = -3.348257, beta1 = 0.111593

beta12 = -0.001401, beta2 = -0.993845

1. Find the probability that a 20-year-old female survived.
2. Complete the table below with the probabilities of a person surviving.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Gender | Pr ( Y = 1 ) | Age | Gender | Pr ( Y = 1 ) |
| 5 | Female | 0.02147476 | 5 | Male | 0.05597099 |
| 35 | Female | 0.10406663 | 35 | Male | 0.3860397 |
| 50 | Female | 0.09402217 | 50 | Male | 0.21897679 |
| 65 | Female | 0.04703537 | 65 | Male | 0.11765435 |
| 80 | Female | 0.012340936 | 80 | Male | 0.1564741 |

1. Find the odds ratio of a 50-year-old male surviving compared to a 50-year-old female surviving.

beta0 = -3.348257, beta1 = 0.111593

beta12 = -0.001401, beta2 = -0.993845

1. Interpret the odds ratio from part (f).

The odds of a 50-year-old male surviving is 2.7 times more than the odds of a 50-year-old female surviving.

1. Find the relative risk of a 35-year-old male surviving compared to a 5-year-old male surviving.

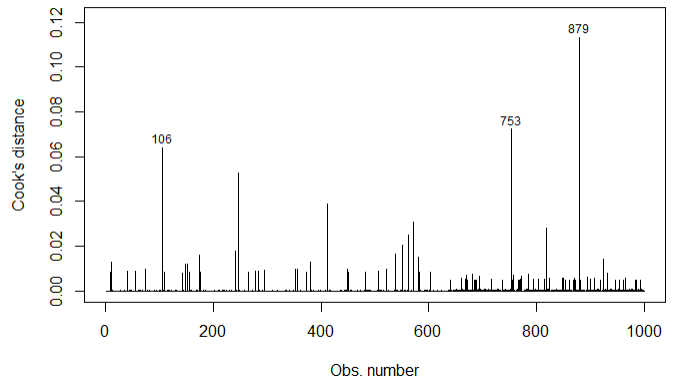
beta0 = -3.348257, beta1 = 0.111593

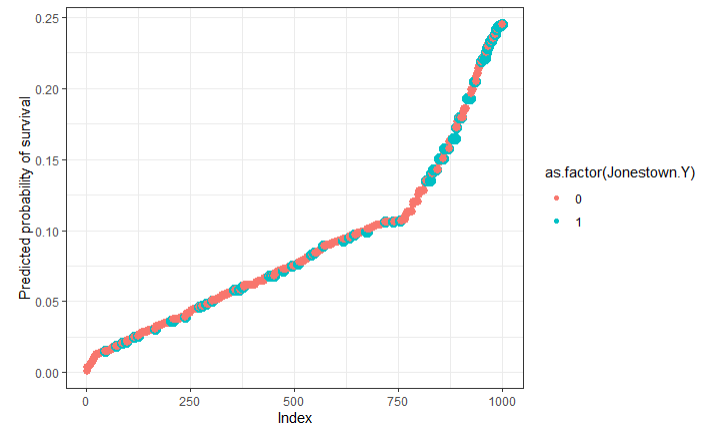
beta12 = -0.001401, beta2 = -0.993845

1. Interpret the relative risk from part (h).

The probability of a 50-year-old male surviving is 6.89 times more probable than the probability of a 5-year-old male surviving.

1. The R output below is diagnostics for the R object Model.

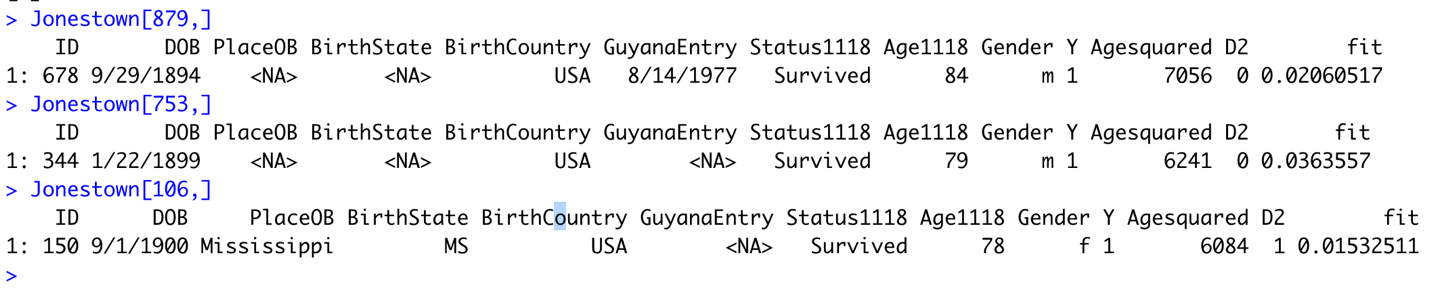




1. Write a sentence that interprets the Cook’s D plot.

The observations with a high value are the ones based on their age, as well as where they are on the graph based on how they can skew the linear regression line.

There are three main outliers shown in the plot above. Using R, we can find more detail about these observations:



Three older individuals, 84, 79, and 79 all survived – while many people around their age did not. This explains why they have a high cooks D, because statistically speaking they should not have survived.

1. Did people with higher probability of survival, survive more often than people with lower probability of survival? Justify your answer by what you see in the plot.

No, we see varying rates of survival/dead (based on the blue vs red dots) along the line in the Model that would be predicting survival or not.

1. Would you say the model is highly proficient at predicting survival? Justify your answer by what you see in the plot.

No. The plot shows the ID of individuals vs their probability of survival. Based on the plot, we know nothing of these individuals, so it is difficult to use it as a way to predict survival.