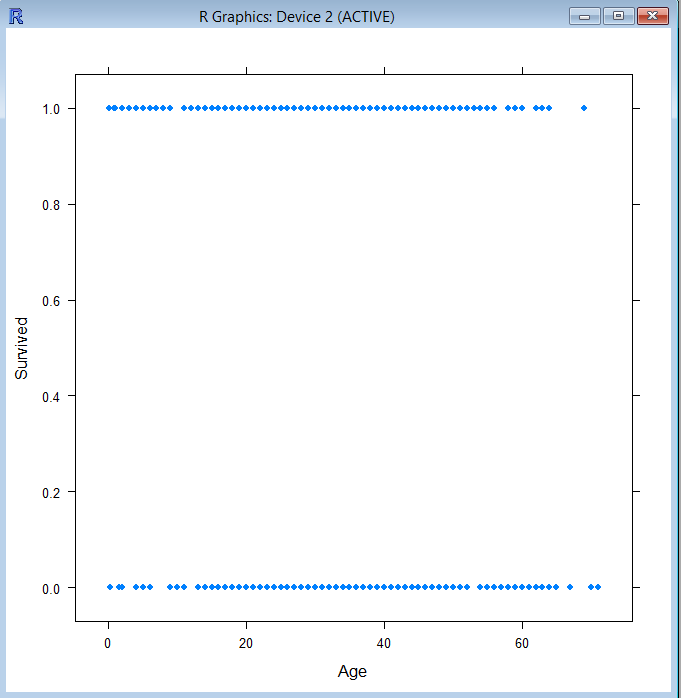
Timothy Goonan

Math 256 Project #2

The sinking of the Titanic was a major tragedy in the 20th century. Many of the people on the boat did not survive, but a lot the passengers did due to lifeboats. Who was able to get on the lifeboats and who was not may have been determined by a set of characteristics. These characteristics such as age and sex will be explored using logistic regression models, two-way tables, and graphical plots. A logistic regression is the statistical model used when the response is binary such as Survival. In the exploration of the effects age and sex had on survival, first let us look at the relationship between age and survival.

First we will use a scatterplot to explore the relationship between age and survival. If there is a significant effect on how old you were and whether you lived, the lines should look different at certain ages.



The plots above depicting the relationship between age and survival indicate that there is not a drastic effect on whether a person will survive or not depending on just their age. You can see that there isn’t a huge effect because the line representing survival is almost the same as the line representing death. The only thing a little different is that survival is denser towards the younger ages and death is a little denser towards the older ages. It is hard to tell how significant of a difference it is just by observing the plots. We have to take a closer look at this significance by fitting a logistic regression with survival being the response and age being the predictor.

The probability form is logit(pi) = (e^Beta0 + Beta1 \* X) so if we apply the form to the Titanic problem we get: logit(Survival) = e^Beta0 + Beta1 \* Age. Using software technology, we discover the logistic model is:

logit(Survival) = -.0814 - .008795(Age).

Now that we have the estimated logistic model we can test if there is a significant relationship between age and survival.

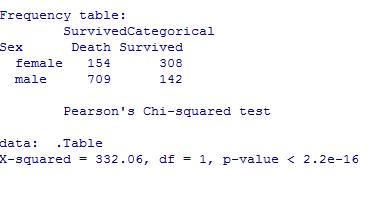
The test:

Null Hypothesis = H0 : Beta1 = 0

Alternative Hypothesis = Ha : Beta1 != 0 “does not equal zero”

The test statistic is z = Beta1hat / the standard error of Beta1hat, this is the z-statistic or the Wald statistic. The Wald statistic for Age is -1.67 and the p-value is 0.0928, which is not less than a significance level of .001, so we fail to reject the null. It should be noted that the p-value is less than a significance level of 0.1 therefore the relationship between age and survival is not significant, but may have a weak relationship. Let us also calculate the G test statistic, which equals the null deviance minus the residual deviance and compare that to a chi-square distribution, to double check the significance. G = 1025.6 – 1022.7 = 2.9. When that is compared to a chi square distribution with a degree of freedom of one, the p-value equals 0.088, which is very close to the previous p-value. This means the conclusion of the relationship not being significant stays the same. We have already determined the magnitude to be weak, but since the slope is negative, the direction is also negative. This means the older you were, the less likely you were of surviving, but it isn’t a significant statement to be made.

Now that the relationship between age and survival has been examined we have to take a look at the relationship between sex and survival. We will first explore this relationship by looking at a two way table.



It is evident from the two way table that there will be a significant difference between being male and female and whether you survive or not. The percentage of females that survived is 308 / 462, which equals a 67% survival rate. The percentage of males that survived is 142 / 851, which equals a 17% survival rate. Without even having to calculate the percentages there are about twice as many men on the ship, but over twice as many women survived versus men. We still have to prove statistically that there is a significant relationship between sex and survival and we will do so by fitting a logistic model. The model is :

logit(Survival) = -1.608 + 2.301(SexCode)

Now that we have the estimated logistic model we can test if there is a significant relationship between age and survival.

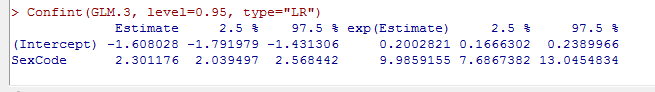
The test:

Null Hypothesis = H0 : Beta1 = 0

Alternative Hypothesis = Ha : Beta1 != 0 “does not equal zero”

The z-statistic equals 17.06 and the p-value is less than 2e^-16. Since the p-value is less than a significance level of 0.001, we reject the null, therefore the relationship between sex and survival is significant. Let us also calculate the G-statistic, the difference in null deviance and residual deviance. G equals 1688.1 – 1355.5 = 332.6, if we compare that to a chi-square distribution with degrees of freedom of one, the p-value is 2.6e^-74, which is a more accurate version of our p-value from earlier. Therefore the conclusion will stay the same that the relationship between sex and survival is significant. If you plug in a 1 (female) into the probability model: (e^-1.608 + 2.301(SexCode) / 1 + e^-1.608 + 2.301(SexCode)) for the SexCode the probability you get is 67% and if you plug in a zero (male) the probability is 17%. If you remember from earlier, these are the same exact probabilities discovered using the two-way table confirming our results from the descriptive analysis. The nature of the model is binary since SexCode can only be two numbers (67% or 17%) and the magnitude is large. There is a substantial difference between 67% and 17% and it weighs heavily in the females favor in this case.

The estimated slope coefficient of SexCode in the model is 2.301. The odds ratio equals the slope coefficient exponentiated so e^Beta1 or e^2.301 = 9.986. The odds ratio in context to the Titanic tragedy means that females are 9.986 times more likely to survive then males.



As seen from the picture above, the 95% confidence interval for the odds ratio is (7.69, 13.05). This interval means that we can confidently predict that 95% of the time the odds ratio will be within this interval or that females will 95% of the time be this much more times likely to survive. We are also confident that 5% of the time the odds ratio will not be within the interval.

Using the two-way table from earlier we found out the probability of a female surviving was 67% so the odds are 67:33 and for males it was 17% so the odds are 17:83. The odds ratio would then be (67/33) / (17/83) = (67 x 83) / (17 x 33) = 5561 / 561 = 9.913, which is very close to the odds ratio we found earlier at 9.986. The difference between the two is that I used rounding to the nearest integer for my probabilities in the two-way table. We can then use that odds ratio of 9.986, the more accurate one, and take the log of it, log(9.986) = 2.301, the estimated coefficient.

Using the model we created early where logit(Survival) = -1.608 + 2.301(SexCode), we can then use that to create the probability model. The probability model or P(Survival) = (e^-1.608 + 2.301(SexCode) / 1 + e^-1.608 + 2.301(SexCode)). When we plug 1 into the SexCode for females, the probability we get is 0.667.

When fitting any type of model we always have to check the conditions beforehand to make sure the model is suitable to use. The conditions for a logistic regression are that it has to be linear, random, and independent. To check for linearity, we use an empirical logit plot. Luckily for our case the predictor (SexCode) is binary. That means linearity is automatic because the plot can only have two points. The next conditions that have to be checked are randomness and independence. Both of these conditions cannot be checked using plots, but instead we have to think about the process that produced the data. The data is historical in nature so there wasn’t any random assignment. Independence is also not met because the death or survival of a parent or sibling could affect the death or survival of their children or family members. Therefore randomness and independence could be problems when interpreting the p-values and intervals as trustful. However, even if conditions are not met, the p-values and confidence intervals can still be used as useful rough guides.

After multiple plots, models, and intervals have been created, the relationships age and sex had with the survival of the passengers on the titanic have become much clearer. Using a scatterplot to guide my intuition when fitting a model for age and survival, I was able to realize that the relationship was not going to be strong. The model backed up my initial thoughts and proved that the relationship was not significant. I then moved to examine the relationship between sex and survival. I used a two-way table to get an understanding of the percentages of males and females that survived. I didn’t realize how useful the information was until I fit a model and the significance of the relationship was clear. Females were 9.986 times more likely to survive than their male counterpart.

Overall the statistical analysis of the Titanic tragedy has taught me the real life applications of statistical modeling. It has also given me a deeper understanding of how a historical moment truly transpired. Although we already knew the outcome of the Titanic, in other situations statistical modeling could be used to predict the outcomes. You use the same pattern that we used here of assessing the conditions, choosing a model, fitting the model, and then testing it. For the Titanic the conditions weren’t exactly met, we chose a logistic regression since we needed a binary response, we created our model, and used two tests to check for significance. The analysis was still successful even though the conditions weren’t met.