Exam#3

Sunny Lee

2021-04-14

1) Data set Donner: The Donner data set has 87 observations and 5 variables. The variables are . Name: Name of individual in the group . Gender: int 1 male and 0=female . Age: The age of the people . Survived: categorical variable 1=survived and 0=dead . Family group size: the size of the family.

In 1846, a group of 87 people (called the Donner Party) were heading west from Springfield, Illinois, for California. The leaders attempted a new route through the Sierra Nevada and were stranded there throughout the winter. The harish wather conditions and lack of food resulted in the death of many people within the group. Social scientists have used the data to study the theory that females are better able than men to survive harsh conditions.

a) Create a logistic regression model (M1) using Gender to estimate the probability of survival.

```
M1 <- glm(Survived~Gender, family = "binomial")
summary(M1)
```

```
##
## Call:
## glm(formula = Survived ~ Gender, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6304  -1.0669   0.7842   1.2921   1.2921
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0217     0.3887   2.628  0.00858 **
## Gender       -1.2874     0.4774  -2.697  0.00701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 119.67  on 86  degrees of freedom
## Residual deviance: 111.85  on 85  degrees of freedom
## AIC: 115.85
##
## Number of Fisher Scoring iterations: 4
```

b. Submit the logistic model $P(Y = 1|x)$.

Using the coefficients found above, we obtain the model: $P(Y = 1|x) = \frac{e^{1.0217-1.2874(Gender)}}{1+e^{1.0217-1.2874(Gender)}}$

b) Interpret the model in terms of the odds ratio. Use the Wald statistic to create a 95% confidence interval for the odds ratio.

Calculating the odds ratio of the model: $\frac{P(Y=1|x=1)}{P(Y=1|x=0)} = \frac{e^{1.0217-1.2874(1)}}{e^{1.0217-1.2874(0)}} = e^{-1.2874}$ which tells us that it is

$e^{-1.2874} = .276$ times more likely to survive if you are a male than a female or $e^{1.2874} = 3.62$ times more likely to survive if you are a female. Using the Wald statistic, we can create a confidence interval for our logistic regression model: $(e^{-1.2874-1.96 \cdot 0.4774}, e^{-1.2874+1.96 \cdot 0.4774}) = (0.2513, 0.30306)$.

c) Test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using the Wald's test.

Our 95% Odds Ratio confidence interval does not include zero. Since zero is not in our 95% confidence interval, we reject the null hypothesis and conclude $\beta_1 \neq 0$ and thus β_1 is significant to our model.

d) Discuss the overall significance of the model. Clearly define the null and alternative hypothesis and write your conclusion.

Overall, since our Wald Statistic does not include zero, we can reject the null hypothesis $H_0 : \beta_1 = 0$ and accept the alternative hypothesis $H_a : \beta_1 \neq 0$.

e) Create another model (M2) using bothe Gender and FamilyGroupSize to estimate the probability of survival and sumbit the logistic model.

```
M2 <- glm(Survived~Gender+FamilyGroupSize, family = "binomial")
summary(M2)
```

```
##
## Call:
## glm(formula = Survived ~ Gender + FamilyGroupSize, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7455  -1.0748   0.7013   1.1357   1.4197
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.58141    0.56898   1.022   0.3068
## Gender          -1.17851    0.48867  -2.412   0.0159 *
## FamilyGroupSize  0.04350    0.04185   1.040   0.2985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 119.67  on 86  degrees of freedom
## Residual deviance: 110.76  on 84  degrees of freedom
## AIC: 116.76
##
## Number of Fisher Scoring iterations: 4
```

From the above, we obtain the model: $P(Y = 1|x) = \frac{e^{.58141-1.17851(Gender)+0.04350(FamilyGroupSize)}}{1+e^{.58141-1.17851(Gender)+0.04350(FamilyGroupSize)}}$

f) Calculate the probability of survival for a male familygroupsize of 3.

```
exp(M2$coefficients[1] + M2$coefficients[2]*0 + M2$coefficients[3]*3) / (1 + exp(M2$coefficients[1] + M
```

```
## (Intercept)
##   0.6708266
```

The probability of a male with a famliygroupsize of 3 surviving is $\frac{e^{.58141-1.17851(0)+0.04350(3)}}{1+e^{.58141-1.17851(0)+0.04350(3)}} = .6708 = 67.08\%$.

g) Calculate the 95% confidence interval for β_2.

The 95% confidence interval for β_2 is: $(e^{0.04350-1.96 \cdot 0.04185}, e^{0.04350+1.96 \cdot 0.04185}) = (.9622, 1.1337)$

h) Check if there is a statistical significance difference between M1 and M2. Clearly define hypothesis test and make a conclusion.

To check if there is a statistical significant difference between M1 and M2, we use the drop-in deviance test with null hypothesis $H_0 : \beta_2 = 0$ and alternative hypothesis: $H_a : \beta_2 \neq 0$.

```
G <- M1$deviance - M2$deviance
G
```

```
## [1] 1.087363
```

```
pchisq(G, 1, lower.tail = F)
```

```
## [1] 0.2970561
```

Since our test statistic is 1.08, we use this number and a chi-sqare distribution with degrees of freedom 1 to obtain the p value. Since our p value is 0.2970561 which is greater than .05, we fail to reject the null hypothesis and conclude our original model with only β_0 and β_1 is better.

2) Assume that for a particular temperature $x_i = 70^0F$ the true probability of success is $\pi_i = 0.75$. If there are four launches made at $x_i = 70^0F$, Find the probability that all four launches are successful, $P(Y_i = 4|X_i = 70^0F)$. Also find the probability that one is successful.

Since the probability of a successful launch at $70^oF = .75$, the probability of having four successful launches is $(.75)^4 = 0.3164$ or 31.64%. The probability that exactly one is successful is found by $\binom{4}{1}(.75)(1 - .75)^3 = 0.04687 = 4.687\%$.

3)Consider the lung cancer data study with 120 observations

```
Lung<-cbind(c(41,19,60),c(28,32,60),c(69,51,120))
rownames(Lung)<-c("Smoker","Nonsmoker","Total")
colnames(Lung)<-c("Yes","No","Total")
Lung
```

```
##            Yes No Total
## Smoker      41 28    69
## Nonsmoker   19 32    51
## Total       60 60   120
```

a) What is the proportion of females in the study who have lung cancer?

The proportion of the females in the study who have lung cancer is $\frac{60}{120}$ or 50%.

b)What is the propotion of smokers who got lung cancer?

The proportion of smokers who have lung cancer is $\frac{41}{69}$ or 59.42%

c)What is the proportion of nonsmokers who got lung cancer?

The proportion of nonsmokers who have lung cancer is $\frac{19}{51}$ or 37.25%

d) What is the relative risk when having lung cancer is defined as a success?

The relative risk is found by the probability of lung cancer in the smoker group over the probability of lung cancer in the nonsmoker group: $\frac{\frac{41}{69}}{\frac{19}{51}} = 1.59$. Thus the risk of having lung cancer is 1.59 times higher for smokers than nonsmokers.

e) what is the odds ratio when having a lung cancer is defined as a success?

The odds of having lung cancer in the smoker group is $\frac{41}{28}$ and the odds of having lung cancer in the nonsmoker $\frac{19}{32}$. Thus the odds ratio is $\frac{\frac{41}{28}}{\frac{19}{32}} = 2.47$. Thus the odds of having lung cancer is 2.47 times higher for smokers than nonsmokers.