Class Activity#13

Write Your Name

2021-03-26

Logistic Regression: This exercise is a continuation of the previous class activity#12. We used the maximum likelihood estimator to estimate the coefficients

$$odds = \frac{\pi_i}{1-\pi_i} = e^{-15.043+0.232xi}$$

1) When x_i increases by 10, state in terms of e^{b_1} how much you would expect the odds to change.

```
Shuttle<-read.csv("C7 Shuttle.csv")
Logit<-glm(Shuttle$Success~Shuttle$Temperature,family = "binomial")
summary(Logit)
```

```
##
## Call:
## glm(formula = Shuttle$Success ~ Shuttle$Temperature, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2175  -0.4524   0.3783   0.7613   1.0611
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -15.0429     7.3786  -2.039   0.0415 *
## Shuttle$Temperature   0.2322     0.1082   2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

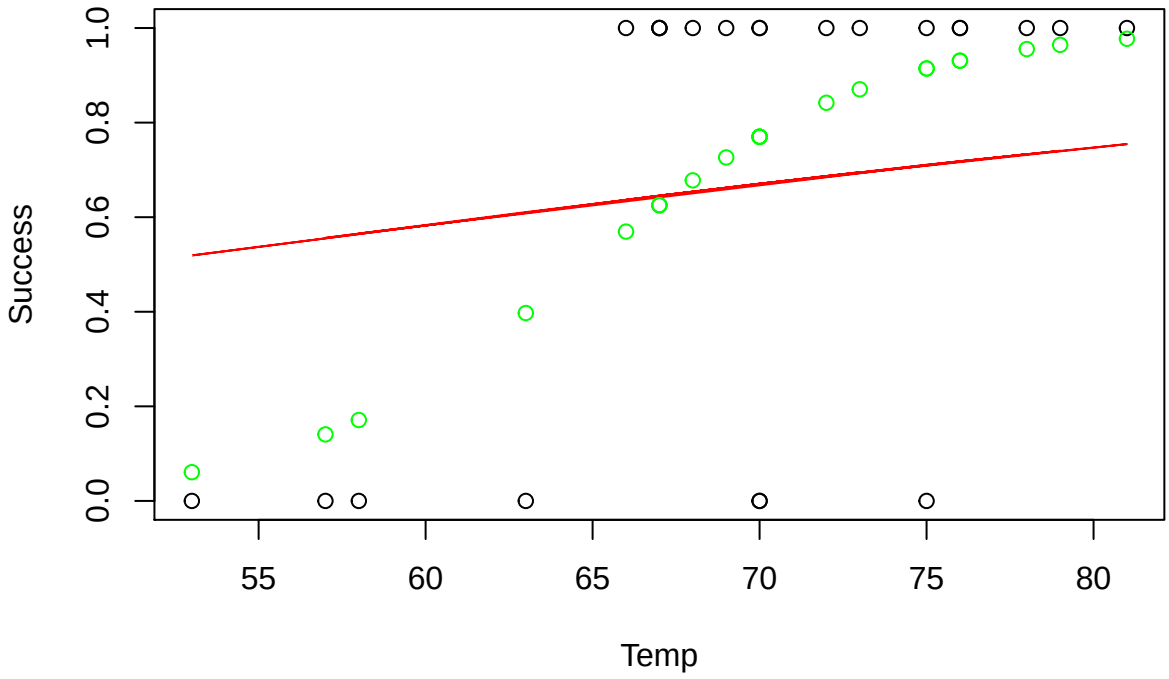
If x_i increases by 10, our odds would increase by a factor of $e^{2.32}$.

2) The difference between the odds of success at 60^oF and 59^oF is about $0.3285 - 0.2605 = 0.068$. Would you expect the difference between the odds at 52^oF and 51^oF to also be about 0.068? Explain why or why not.

We would not expect our difference in odds to be the same, as we are dealing with an exponential function and although we keep the temperature interval the same, our success interval would differ.

3) Plot temperature versus the estimated probability using maximum likelihood estimates and temperature versus the estimated probability using the least squares estimates.

```
Coef<-Logit$coefficients
MLE_y<-(exp(Coef[1]+Coef[2]*Shuttle$Temperature))/(1+(exp(Coef[1]+Coef[2]*Shuttle$Temperature)))
Lmodel<-lm(Shuttle$Success~Shuttle$Temperature)
Coef_L<-Lmodel$coefficients
LSE_y<-(exp(Coef_L[1]+Coef_L[2]*Shuttle$Temperature))/(1+(exp(Coef_L[1]+Coef_L[2]*Shuttle$Temperature)))
plot(Shuttle$Success~Shuttle$Temperature,ylab="Success",xlab="Temp")
points(LSE_y~Shuttle$Temperature,type="l",col="red")
points(MLE_y~Shuttle$Temperature,col="green")
```



We see from the graph that the logistic model does a much better job of predicting our success than our linear model.

4) Calculate the odds ratio of a successful launch between 31^oF and 60^oF. Provide a confidence interval of this odds and interpret your result.

```
Shuttle<-read.csv("C7 Shuttle.csv")
Logit<-glm(Shuttle$Success~Shuttle$Temperature,family = "binomial")
summary(Logit)
```

```
##
## Call:
## glm(formula = Shuttle$Success ~ Shuttle$Temperature, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2175  -0.4524   0.3783   0.7613   1.0611
##
## Coefficients:
```

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -15.0429     7.3786  -2.039   0.0415 *
## Shuttle$Temperature   0.2322     0.1082   2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

```
exp(0.2322 - 1.96* 0.1082)
```

```
## [1] 1.020332
```

```
exp(0.2322 + 1.96* 0.1082)
```

```
## [1] 1.559355
```

To calculate the odds ratio: $OR = \frac{e^{\beta_0+\beta_1(31)}}{e^{\beta_0+\beta_1(60)}} = e^{-\beta_1(29)}$. The confidence interval will take the form $(e^{\beta_1 - z \cdot Se(\beta_1)}, e^{\beta_1 + z \cdot Se(\beta_1)}) = (e^{0.2322 - 1.96 \cdot 0.1082}, e^{0.2322 + 1.96 \cdot 0.1082}) = (1.020332, 1.559355)$

5) The first model were calculated when a successful launch was given a value of 1. Conduct a logistic regression analysis where 1 indicates an O-ring failure and 0 represents a successful launch.

```
Shuttle$success0 <- 1 - Shuttle$Success
logit_0 <- glm(success0~Temperature, data = Shuttle, family = "binomial")
summary(logit_0)
```

```
##
## Call:
## glm(formula = success0 ~ Temperature, family = "binomial", data = Shuttle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0611  -0.7613  -0.3783   0.4524   2.2175
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## Temperature  -0.2322     0.1082  -2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

```
exp(-0.2322 - 1.96* 0.1082)
```

```
## [1] 0.641291
```

```
exp(-0.2322 + 1.96* 0.1082)
```

```
## [1] 0.9800732
```

a) Explain any relationship between the model shown from the previous activity.

The AIC, null and residual deviance even the pvalues are the same. What is different are the signs of the β_i.

b) How did the regression coefficients change.

The regression coefficients flipped signs.

c) How did the odds ratio change.

The odds ratio has flipped. Our first odds ratio was $e^{.2322} : 1 = 1.26 : 1$ and so it was 1.26 times more likely for $y = 1$ than $y = 0$. Our second odds ratio is $e^{-.2322} : 1 = .79 : 1$ and so it was .79 times more likely for $y = 0$ than $y = 1$. We see from the two ratios that one is the reciprocal of the other.

d) Create a 95% Wald confidence interval for the new ratio and interpret the results.

The confidence interval is calculated by $(e^{-0.2322-1.96\cdot 0.1082}, e^{-0.2322+1.96\cdot 0.1082}) = (0.641291, 0.9800732)$. Since our 95% confidence interval does not include zero, we can conclude that our $\beta_1 \neq 0$ and thus reject the null hypothesis, meaning our model is significant.

6) Use statistical software to calculate the LRT for the space shuttle data. Submit the p-value and state your conclusions.

```
-2*sum(Shuttle$success0 * log(logit_0$fitted.values) + (1-Shuttle$success0)*log(1-logit_0$fitted.values
```

```
## [1] 20.31519
```

```
28.267 - 20.315
```

```
## [1] 7.952
```

```
pchisq(7.952, df = 1, lower.tail = F)
```

```
## [1] 0.004803426
```

To calculate the LRT, we can use the fitted values of our logistic regression and we get 20.31519 which is precisely our null deviance. To get our test statistic, we subtract the residual deviance from the null deviance, from which we get 7.952. Since our test statistic is a Chi Sqaure distribution, we can find the pvalue with the pchisq function and we get: 0.004803426. Thus, we can reject the null hypothesis and conclude that our full model is significant.

Data set: Cancer2

7) Create a logistic regression model using Radius and Concave as explanatory variables to estimate the probability that a mass is malignant.

a) Using Radius as the first explanatory variable, x_1, and Concave as the second explanatory variable, x_2, submit the logistic regression model. In other words, find the coefficients for the model

$$y_i = \frac{e^{\beta_o+\beta_1 x_{1i}+\beta_2 x_{2i}}}{1 + e^{\beta_o+\beta_1 x_{1i}+\beta_2 x_{2i}}}, \quad i = 1, 2, \cdots, n.$$

```
cancer2 <- read.csv("C7 Cancer2.csv")

m7 <- glm(Malignant~radius+concavity, data = cancer2, family = "binomial")
summary(m7)
```

```
##
## Call:
## glm(formula = Malignant ~ radius + concavity, family = "binomial",
```

```
##     data = cancer2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3698  -0.2839  -0.1327   0.1160   2.8343
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.1320     1.4932  -8.795  < 2e-16 ***
## radius        2.7175     0.3663   7.418 1.19e-13 ***
## concavity     3.3192     0.3545   9.362  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 224.02  on 566  degrees of freedom
## AIC: 230.02
##
## Number of Fisher Scoring iterations: 7
```

```
pchisq(527.42, df = 2, lower.tail = F)
```

```
## [1] 2.966212e-115
```

From our glm, we obtain the function $\frac{e^{-13.1320+2.7175x_{1i}+3.3192x_{2i}}}{1+e^{-13.1320+2.7175x_{1i}+3.3192x_{2i}}}$.

b) Submit the likelihood ratio test results, including the log-likelihood (or deviance) values.

```
-2*sum(cancer2$Malignant * log(m7$fitted.values) + (1-cancer2$Malignant)*log(1-m7$fitted.values))
```

```
## [1] 224.016
```

We find the likelihood ratio test to be 224.016 and a null deviance of 751.44 and a residual deviance of 224.02.

c) Concave=0 represents round cells and concave=1 represents concave cells. Calculate the event probability when Radius=4 and the cells are concave. Also calculate the event probability when Radius=4 and the cells are not concave.

```
exp(m7$coefficients[1] + m7$coefficients[2]*4 + m7$coefficients[3]*1) / (1 + exp(m7$coefficients[1] + m
```

```
## (Intercept)
##   0.7421557
```

```
exp(m7$coefficients[1] + m7$coefficients[2]*4 + m7$coefficients[3]*0) / (1 + exp(m7$coefficients[1] + m
```

```
## (Intercept)
##  0.09432166
```

The probability that the cells are malignant when the Radius is 4 and the cells are concave is 74.22% while the probability when the radius is 4 and the cell is round is 9.43%.

8) Create a logistic regression model only Radius as an explanatory variable to estimate the probability that a mass is malignant.

```
radius <- glm(Malignant~radius, data = cancer2, family = "binomial")
summary(radius)
```

```
##
```

```
## Call:
## glm(formula = Malignant ~ radius, family = "binomial", data = cancer2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5470  -0.4694  -0.1746   0.1513   2.8098
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.2459     1.3246  -11.51   <2e-16 ***
## radius        3.6165     0.3258   11.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 330.01  on 567  degrees of freedom
## AIC: 334.01
##
## Number of Fisher Scoring iterations: 6
```

```
pchisq(105.99, df = 1, lower.tail = F)
```

```
## [1] 7.410824e-25
```

a) Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values.

```
-2*sum(cancer2$Malignant * log(radius$fitted.values) + (1-cancer2$Malignant)*log(1-radius$fitted.values)
```

```
## [1] 330.0108
```

From the above, the likelihood ratio test result is 330.0108 and the null deviance is 751.44 and the residual deviance is 330.01.

b) Calculate the event probability when Radius= 4.

```
exp(radius$coefficients[1] + radius$coefficients[2] * 4) / (1 + exp(radius$coefficients[1] + radius$coe
```

```
## (Intercept)
##   0.3143713
```

Using our coefficients from our model, we would predict a cell with radius 4 to have a 31.43% chance of being malignant.