Logistic Model for predicting the failure of the O-ring of the challenger

In order to predict the failure of the O-Ring of the challenger in the morning of January 28, 1986 we used a data set of 23 past launches conducted by NASA that included two variables: the temperature measured in the morning of the launch (in Fahrenheit) and the binary outcome which indicates if there was a crack in the O-Ring on that flight. Our attempt includes using logistic regression to predict the failure of the O-Ring structure based on the temperature on the morning of the takeoff. We could not use linear regression for this model since one of our variables had a binary value and the scatter plot of the two variables 'temp' and 'failure' showed a polar distribution due to the binary value.

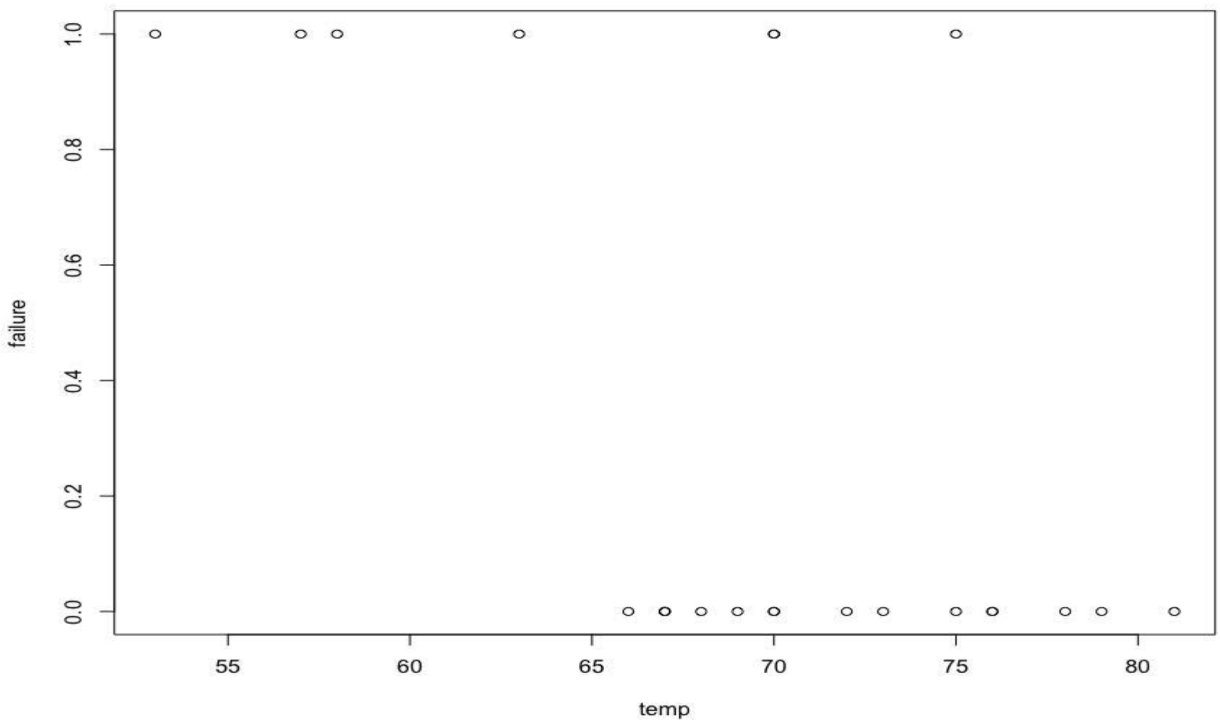

Figure – (i)
Shows the scatter plot between temperature and failure.

We started our project by importing the data into R-Studio, eliminated unnecessary columns, named the two variables ('temp' and 'failure') and sorted the data rows by the temperature in an increasing order. Then we ran a logistic regression using the 'glm' function where 'failure' was dependent on 'temp'. The summary is as follows:

```
Call:
glm(formula = failure ~ temp, family = binomial())

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 *
temp          -0.2322     0.1082  -2.145   0.0320 *
```

Since the z values of the estimators are statistically significant, we decided to proceed by using our estimators: β_0= 15.0429 and β_1=-0.2322 (we denoted β_0 as 'intercept' and β_1 as 'temp_coefficients' for an easier future use). Then we created the probably of failure function 'predicted_probability' which returns the probability 'prob' that included the aforementioned estimators and the predictive value 'temp' (x_i) as illustrated in the following formula:

$$\frac{\Sigma \quad {}_{\Sigma}}{\qquad\qquad}$$

Then we created a vector 'P' that holds the predicted probabilities of failure for all the 23 temperature values. We plotted a graph with x as 'Temperature' and Y as 'Probability of the failure':

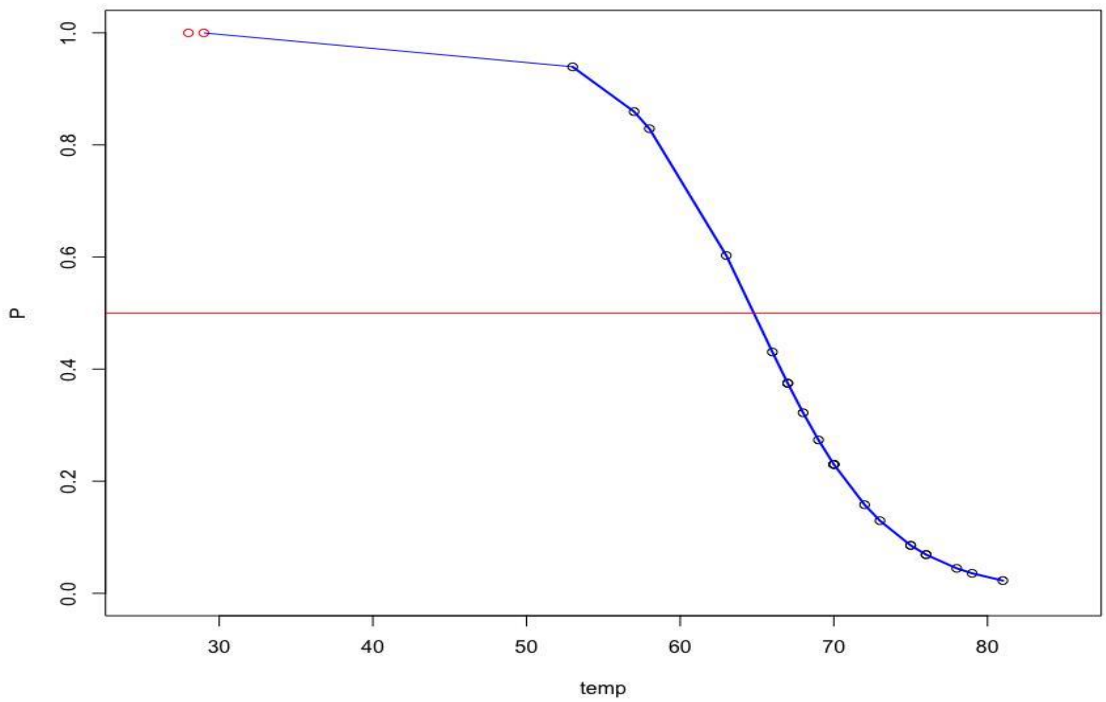

Figure – (ii)

Shows the estimated probability of failure for all temperatures on the day of launches. The red points indicate the probability of failure for the temperatures $28F^{0}$ and $29F^{0}$.

According to the assignment details, the temperature on the day of the launch was between $28\text{-}29F^{0}$. Thus, our next step was to find the probability of failure with the given temperature. The probability of failure for $29F^{0}$ was 0.9997541 or **99.97%** and the probability of failure for $28F^{0}$ was 0.999805 or **99.98%**. Therefore, we can conclude that the launch had a very high likelihood of failing in the morning of January 28, 1986 according to the past data that we used.

While we were considering the scatter plot in Figure (i), we noticed 3 possible outliers that indicated a failure in a relatively high temperature (>=70 F^0). As was mentioned in classes during the description of the assignment, the breaks in the O-Rings that were found in the ocean could have been caused by the Solid Rocket Motor's impact with the water after its fall from the sky. Thus, the temperature did not affect that failure. Thus, we one might suggest eliminating one or all three of the optional outliers. That approach would lead to a higher probability of failing for the launch on the January 28^{th}. Yet, removing all three optional outliers would simply create a z-shaped line that gives all temperatures lower than the highest observed failure a value of 1, a value approaching 0 to all temperatures higher than the lowest observed success, and a negatively steeped linear function for all the temperatures that are in between the highest observed failure and lowest observed success.

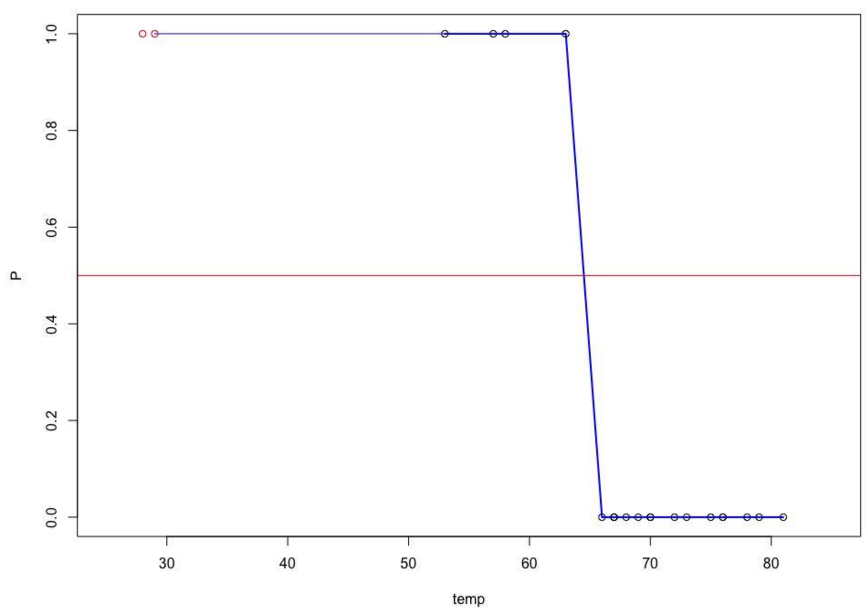

Figure – (iii)

Shows the estimated probability of failure for all temperatures on the day of launches after removing the optional three outliers mentioned above from our data set.