w271 Lab 1: Investigation of the 1989 Space Shuttle Challenger Accident

Ahmad Azizi, Oswaldo Olivo, George Jiang

6/9/2022

Contents

Abstract

In this report, we are going to analyze the data set for the Space Shuttle Challenger 1989. We will be using statistical analyses to investigate the events surrounding the space shuttle Challenger accident that occurred on January 27, 1989. Our task is to test various models to find a good predictor for O-ring failure.

Introduction

In this report, we are going to analyze the data set for the Space Shuttle Challenger 1989. We will be using statistical analyses to investigate the events surrounding the space shuttle Challenger accident that occurred on January 27, 1989. Our task is to test various models to find a good predictor for O-ring failure. The data set has four explanatory variables that consist of Flight, Temperature, Pressure, and Number. The outcome variable is O.ring. We will be conducting a full explanatory analysis to get a better understanding of the data set. We will explore various logistic regression models to find the best predictive model. Following the logistic regression analysis, we will be using a linear regression to analyze the data and compare the results with the logistic regression models.

Research question

What are the effects of the given explanatory variables on the probability of O.ring failure?

Data (20 points)

We are dealing with a very small data set. We have a total of 23 observations across 5 columns. ***Table 1*** shows the first few data points. There are no missing values in the data set. We also do not see any anomalies. Flight is an integer value that indicates flight number. Number variable shows the number of O.rings and is a constant value of 6. The outcome variable O.ring takes values of 0, 1, and 2. ***Table 2*** shows how the frequency of each outcome variable given our data set. It appears that outcome variable 0 is the most frequent in our data set. Pressure has three values: 50, 100, 200.

Table 1: First few datapoints

Flight	Temp	Pressure	O.ring	Number
1	66	50	0	6
2	70	50	1	6
3	69	50	0	6
4	68	50	0	6
5	67	50	0	6

Table 2: Number of observation outcome variable

Var1	Freq
0	16
1	5
2	2

Table 3 shows the distribution of the outcome variable with Temperature and Pressure variables. We can see that when pressure is 200, we have the most number of 0 outcome variables. We can further see that the highest temperature is associated with outcome variable 0. However this may be misleading, since there are only two observations for outcome variable 2. Let's try to visualize these relationships to get a better

understanding of what's happening. ***Figure 1*** confirms our observations from ***table 3***. Basically, pressure with value of 200 has the most number of observations. Due to low number of observations for outcome variable 2 and 1, the density graphs of temperature are overlapping.

Table 3: Distribution of outcome variable with Temperature and Pressure variables.

Dependent: O.ring		0	1	2
Temp	Mean (SD)	72.1 (4.8)	63.6 (6.3)	64.0 (15.6)
Pressure	50	5 (31.2)	1 (20.0)	0 (0.0)
	100	2 (12.5)	0 (0.0)	0 (0.0)
	200	9 (56.2)	4 (80.0)	2 (100.0)

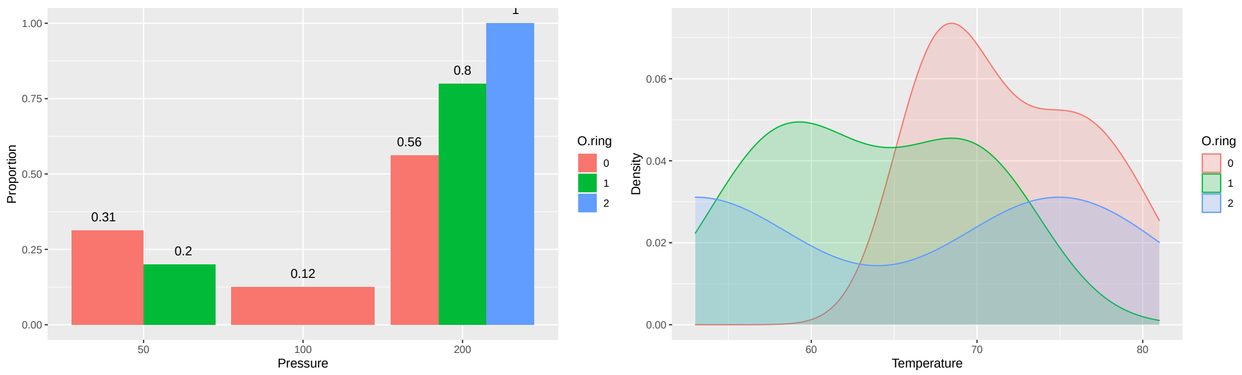


Figure 1: Density plot for Temperature with respect to Outcome variable and a bar plot of Pressure with respect to Outcome variable

Let's try to do a box plot of Temperature vs outcome variable to see if we can find any other insights. ***Figure 2*** shows that most observations occur within 0 and 1 O.ring incidents and we can see that 0 O.ring incidents generally has higher temperature than O.ring incidents 1 and 2.

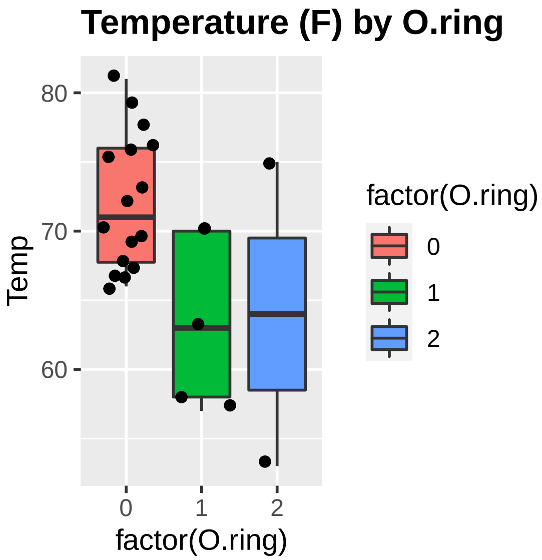


Figure 2: Shows boxplot of temperature vs Outcome variable

Next, let's try to see the relationship between Temperature and Pressure variables to gain some insights. ***Figure 3*** shows the this relationship and we can see that the majority of observations occurred under 200 psi and there is the most variation in the temperature curve for these values.

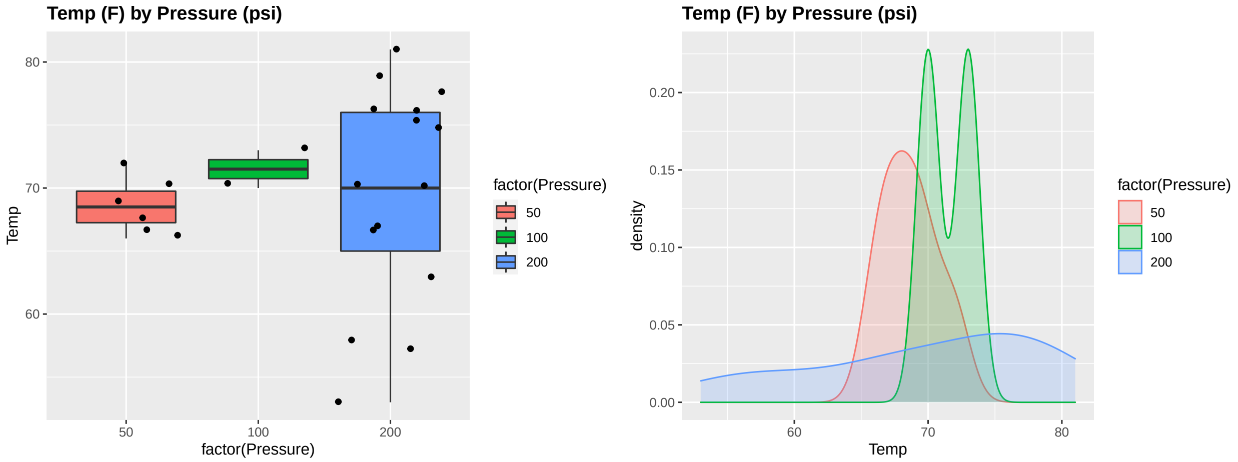


Figure 3: Shows the relationship between Temperature and Pressure

Description

Each launch is an observations with a sample size of six. In binary terms, each trial has two possible outcomes, a success or a failure. The assumption is made that each trial is independent of each other. It is assumed that each O.ring would suffer damage independent of each other given designated pressure and temperature. This is a very strong assumption because it is possible that that if one O-ring fails, another is more likely to fail due to stress being put on the other O-rings. This may result in creating the potential for a cascade of failures invalidating the independence assumption. Furthermore, there may be omitted variables that influence O-ring quality across each launch. However, for logistic regression to be applicable, this assumption has to be made.

Analysis

Reproducing Previous Analysis (10 points)

The logistic regression model is as follows.

```
logistc_temp_pressure <- glm(formula = O.ring/ Number ~ Temp + Pressure,
                             weights = Number, family =
              binomial (link = logit), data = data)
```

```
##
## =====================================
##                   Dependent variable:
##                   -------------------
##                      O.ring/Number
## -------------------------------------
## Temp                   -0.098**
##                         (0.045)
## Pressure                 0.008
##                         (0.008)
## Constant                 2.520
##                            (3.487)
## -------------------------------------
## Observations                  23
## Log Likelihood              -15.053
## Akaike Inf. Crit.           36.106
## =====================================
```

We can see that only the Temp variable is statistically significant with a coefficient of (β_1)= -0.0982968 and a final model of $logit(\pi) = 2.52 - 0.098 Ttemp + 0.008 Pressure$

Since pressure is not significant based on the above model. Let us run a likelihood ratio test as follows.

```
Anova(logistc_temp_pressure, test='LR')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/Number
##          LR Chisq Df Pr(>Chisq)
## Temp       5.1838  1     0.0228 *
## Pressure   1.5407  1     0.2145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$, we get value of $-2log(\Lambda) = 5.18$ and a p-value of 0.0228 which means Temp variable is significant. For $H_0 : \beta_2 = 0$ and $H_a : \beta_2 \neq 0$, we get value of $-2log(\Lambda) = 1.54$ and a p-value of 0.2145 which means Pressure variable is not significant.

Based on the above model and likelihood ratio test, we can see that Pressure variable can indeed be dropped from the model. We concur with the authors. However, we do have to note that the authors have assumed that the relationship between Temp and Pressure variables is linear. More exploration would be needed to determine if pressure truly has no impact on the outcome variable and that we can leave it out of our model.

Confidence Intervals (20 points)

We explore a simple logistic regression model for predicting the probability of failure of an O-ring based on environment temperature. Specifically, we compute a model for the following problem $logit(\pi) = \beta_0 + \beta_1 Temp$.

```
model_logistic_regression_temp <- glm(formula = O.ring/Number ~ Temp,
                                      weights = Number, family = binomial (link = logit),
                                      data = data)
```

We create another model that includes temperature as a quadratic term, and apply an ANOVA test to see if it improves our predictive power.

```
model_logistic_regression_temp_with_square <- glm(formula =
                                                  O.ring/Number ~ Temp + I(Temp^2),
                                                  weights = Number,
                                                  family = binomial (link = logit),
                                                  data = data)
```

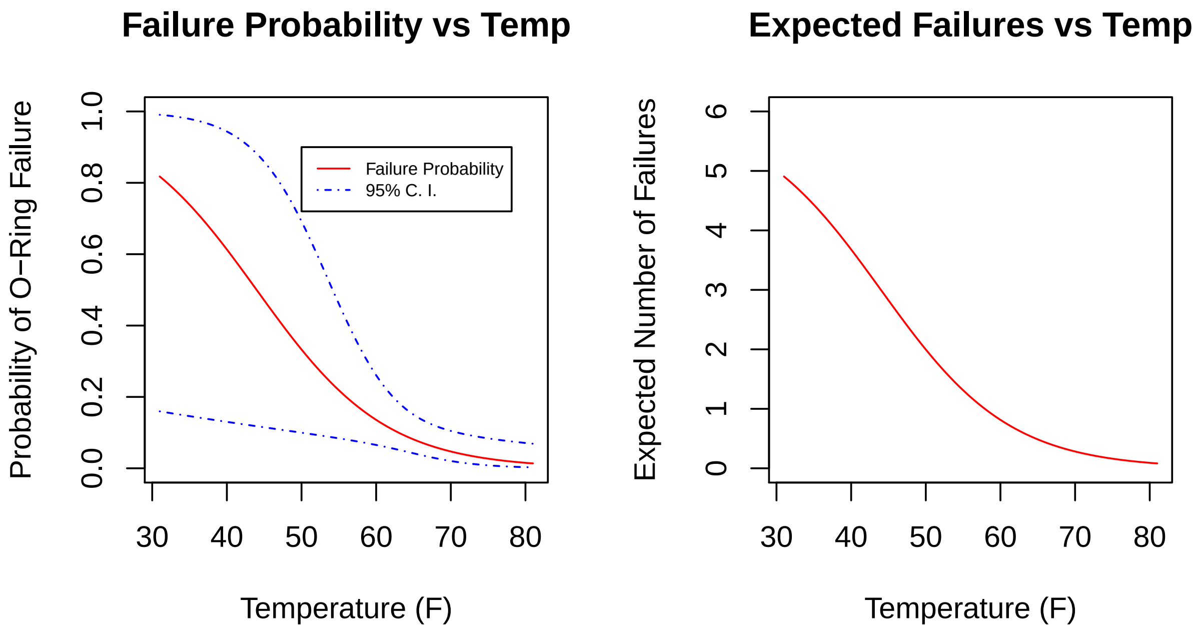
```
##
## =====================================
##                  Dependent variable:
##                  -------------------
##                     O.ring/Number
## -------------------------------------
## Temp                       -0.651
##                            (0.741)
## I(Temp2)                    0.004
##                            (0.006)
## Constant                   22.126
##                           (23.794)
## -------------------------------------
## Observations                  23
## Log Likelihood             -15.576
## Akaike Inf. Crit.          37.152
## =====================================
```

```
anova(model_logistic_regression_temp, model_logistic_regression_temp_with_square,
      test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Temp
## Model 2: O.ring/Number ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     18.086
## 2        20     17.592  1   0.4947   0.4818
```

We can drop the quadratic temperature term from our models, given the p-value of 0.4818 (> 0.05) in the ANOVA test.

Below is a visualization of the predicted failure probabilities of O-rings for temperatures ranging from 31 F to 81 F, along with lower and upper bands of corresponding to 95% Wald confidence intervals. The second visualization is the expected number of failures for each temperature.



We can see a clear decline of failure probabilities as temperatures increase, from roughly 80% failures at 31 F, down to close to zero at 81 F. Interestingly, the confidence intervals are very wide for low temperatures and narrow down as temperatures increase. The confidence intervals are wider for lower temperatures due to the lack of data points for low temperatures and bigger standard deviation. we expect a larger number of failures for lower temperatures (between 4 to 5 failures for 31 F and close to zero failures at 80 F).

We will compute the failure probability and confidence interval for the conditions of the Challenger launch. Our model relies on the assumption of i.i.d., meaning that all O-rings in a flight fail independently from each other, and have the same underlying distribution of failure probability.

The data set used to train the model doesn't have any datapoints with temperature below 53 F, while the temperature on the Challenger launch day was 31 F. Therefore, we're relying on our model to extrapolate the correlation between temperature and failure probability to predict datapoints very different from the training data set, such as the Challenger's launch day.

```
linear.pred = predict(object = model_logistic_regression_temp,
                      newdata = data.frame(Temp = 31),
                      type = "link", se = TRUE)
pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit))
CI.lin.pred = linear.pred$fit + qnorm(p = c(0.05/2, 1-0.05/2))*linear.pred$se.fit
CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred))
round(data.frame(pi.hat, lower=CI.pi[1], upper=CI.pi[2]),2)
```

```
##   pi.hat lower upper
## 1   0.82  0.16  0.99
```

Our model predicts a probability of O-ring failure of 82% for the Challenger launch, which is substantial. Notice that our confidence intervals are very wide, as we saw on the previous confidence interval bands, due to lack of data points with temperature below 53 F.

Bootstrap Confidence Intervals (30 points)

In this section, we use parametric bootstrap to compute confidence intervals, rather than relying in asymptotic properties. We compute confidence intervals for each temperature between 10 F and 100 F. For each temperature, we create 1000 data sets, each data set constructed by sampling 23 entries with replacement from the original data set. With this approach we are bootstrapping many data sets from our original data set.

For each data set, we fit a logistic regression model that correlates temperature to O-ring failures, and perform a prediction over the temperature that we're currently analyzing.

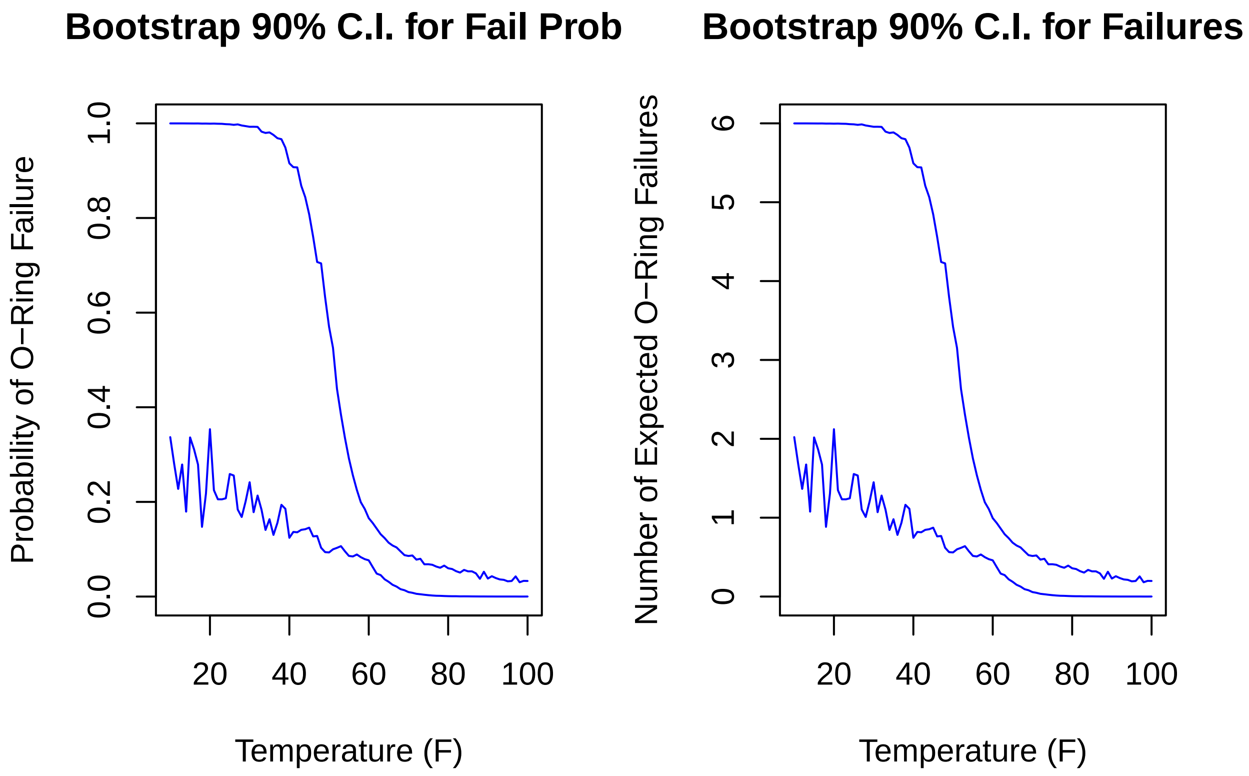
After computing a prediction from each model at a given temperature, we've produced a list of probability failures for a given temperature, which we can order and compute quantiles at different probability failures. In our case, we compute 90% confidence intervals by extracting quantiles at probabilities 5% and 95%.

```
lowerBound = c()
upperBound = c()

for (currentTemp in 10:100) {
  currentTempDataset = data.frame(Temp = currentTemp)
  currentPredictions = c()

  for (currentDataset in 1:1000) {
    sampleDataset = data[sample(nrow(data), 23, replace=TRUE),]
    model = glm(O.ring / Number ~ Temp, data = sampleDataset,
                family = binomial(link = "logit"))
    prediction = predict(object = model, newdata = currentTempDataset,
                         type = "link", se = TRUE)
    failureProbability = exp(prediction$fit)/(1+exp(prediction$fit))
    currentPredictions = append(currentPredictions, failureProbability)
  }
  quantiles = quantile(currentPredictions, probs = c(.05, .95), na.rm = TRUE)
  lowerBound = append(lowerBound, quantiles[1])
  upperBound = append(upperBound, quantiles[2])
}

confidenceIntervalsDatasets = data.frame(Temp = 10:100, LowerBound = lowerBound,
                                         UpperBound = upperBound)
```



We can see a similar trend with the bootstrapped confidence intervals: probability of failures and number of expected failures tend to decrease with higher temperatures. For very low temperatures we expect failure probabilities of at least 20%, while the lower bounds and upper bounds for the failure probabilities are very close to zero for temperatures above 80 F.

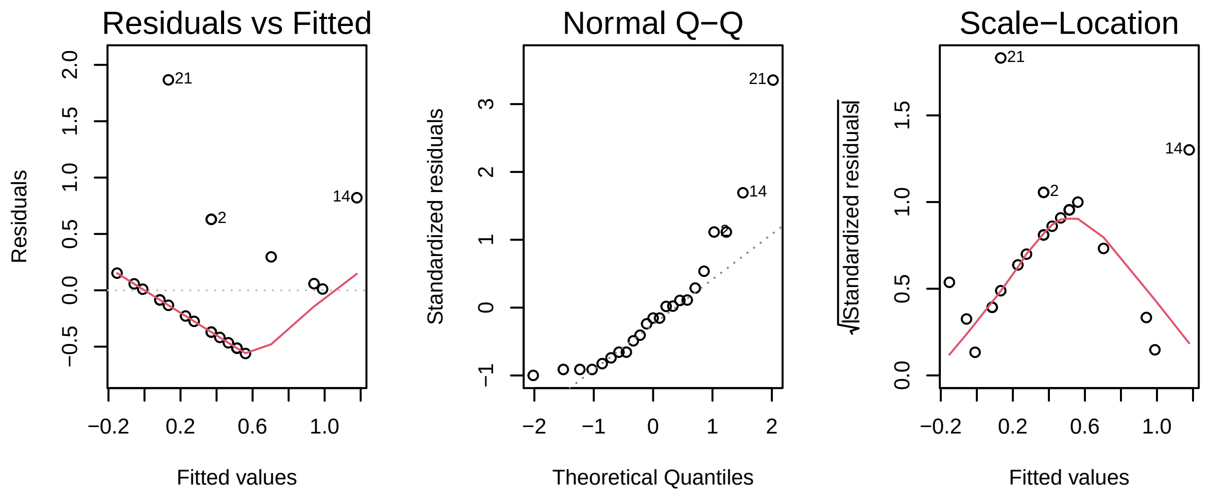
Alternative Specification (10 points)

To further analyze the data set, we will now create a linear model as follows. Point to note here is that we are not using O.ring/Number proportions because we would like to directly measure the linear relationship between temperature and O.ring failure and not a proportion. $O.ring = \beta_0 + \beta_1 Temp$

```
mod_linear <- lm(O.ring ~ Temp, data = data)
```

```
##
## ================================
##              Dependent variable:
##              -------------------
##                     O.ring
## --------------------------------
## Temp                -0.048**
##                         (0.017)
## Constant               3.698***
##                         (1.220)
## --------------------------------
## Observations              23
## R2                       0.261
## Adjusted R2              0.226
## ================================
```

As we can see above, Temperature variable is significant with p value of 0.013. Next let's talk about Central Limit Theorem Assumptions. For a linear model to be applicable, we need to have random sampling (i.i.d), no perfect multi-colinearity, zero-conditional mean, homoskedasticity, and normality error. For our data set, we do not have random sampling as explained in previous sections, however we made the assumption that the data is i.i.d. There is no perfect multi-colinearity since there is only one explanatory variable in our model. As we can see below in the Scale-Location plot, the errors do not have a uniform distributions so there is no homoskedasticty. And the QQ-plot shows deviation from normality. Residual vs Fitted plot shows a deviation from zero-conditional mean. Furthermore, AIC for the linear model with a value of 43.9 is higher than the AIC for the logistic model than a value of 35.65



```
AIC(mod_linear)
```

```
## [1] 43.91752
```

Conclusions (10 points)

Our preferred model is the following model: $logit(\pi) = \beta_0 + \beta_1 Temp$ We would like to estimate the effects on the odds of O.ring failure based on a c unit change in temperature. We can use the following formula to calculate this. $\widehat{OR} = exp(c\hat{\beta}_1)$

```
c = seq(from = 1, to = 10, by = 3)
OR.hat = round(exp(c * model_logistic_regression_temp$coefficients[2]),2)
probability_failure = round(
  exp(
    c * model_logistic_regression_temp$coefficients[2])/
    (1+exp(c * model_logistic_regression_temp$coefficients[2])),
  2)
```

Table 4: ODDs ratio and Probabality of failure for c unit change in Temperature

change.in.temperature	OR.hat	probability_of_failure
1	0.89	0.47
4	0.63	0.39
7	0.45	0.31
10	0.31	0.24

For explanation purposes, let us look at odds ratio and probability of failure when *c* is 1. We can interpret as follows. The odds of an O.ring failure change by 0.89 with an increase of 1 unit for the temperature variable. Furthermore, the probability of failure is 0.47 when c = 1.

All in all, as described in the previous sections, logistic regression seems to be best model for the given data set. Below is the summary table of the three models that we used above. Three logistic regressions and one linear model was used in our analysis. Based on the results, we can see that the logistic regression with only temperature variable as explanatory variables gives us the best results. Furthermore, when comparing model_logistic_regression_temp to mod_linear, we can clearly see that the AIC is smaller for the logistic regression.

```
##
## Comparing GLM and Linear Models
## =====================================
##                 Dependent variable:
##                ----------------------
##                O.ring/Number  O.ring
##                  logistic      OLS
##                    (1)         (2)
## -------------------------------------
## Temp             -0.116**    -0.048**
##                   (0.047)    (0.017)
## Constant          5.085*     3.698***
##                   (3.052)    (1.220)
## -------------------------------------
## AIC               35.647      43.918
## Observations        23          23
## R2                           0.261
## Adjusted R2                  0.226
## Log Likelihood   -15.823
## =====================================
```