W271 Group Lab 1

Investigating the 1986 Space Shuttle Challenger Accident

Nathan Martinez, Meera Sharma, Hannah George, and Haile Bizunehe

Contents

Abstract

This report will, indeed, be abstract. No, instead, describe your goals your approach, and what you learn.

1 Introduction

1.1 Research question

Solution:
INSERT WRITUP HERE.

2 Data (20 points)

Complete the following task. In your final submission, please remove this question prompt so that your report reads as a report. The Data Section of this report is worth 20 points.

- Conduct a thorough EDA of the data set.
  - This should include both graphical and tabular analysis as taught in this course.
  - Since the report has a page-limit, you will have to be selective when choosing visuals to illustrate your key points, associated with a concise explanation of the visuals.
- This EDA should begin with an inspection of the given dataset; examination of anomalies, missing values, potential of top and/or bottom code etc.

```
df <- read_csv('../data/raw/challenger.csv')
```

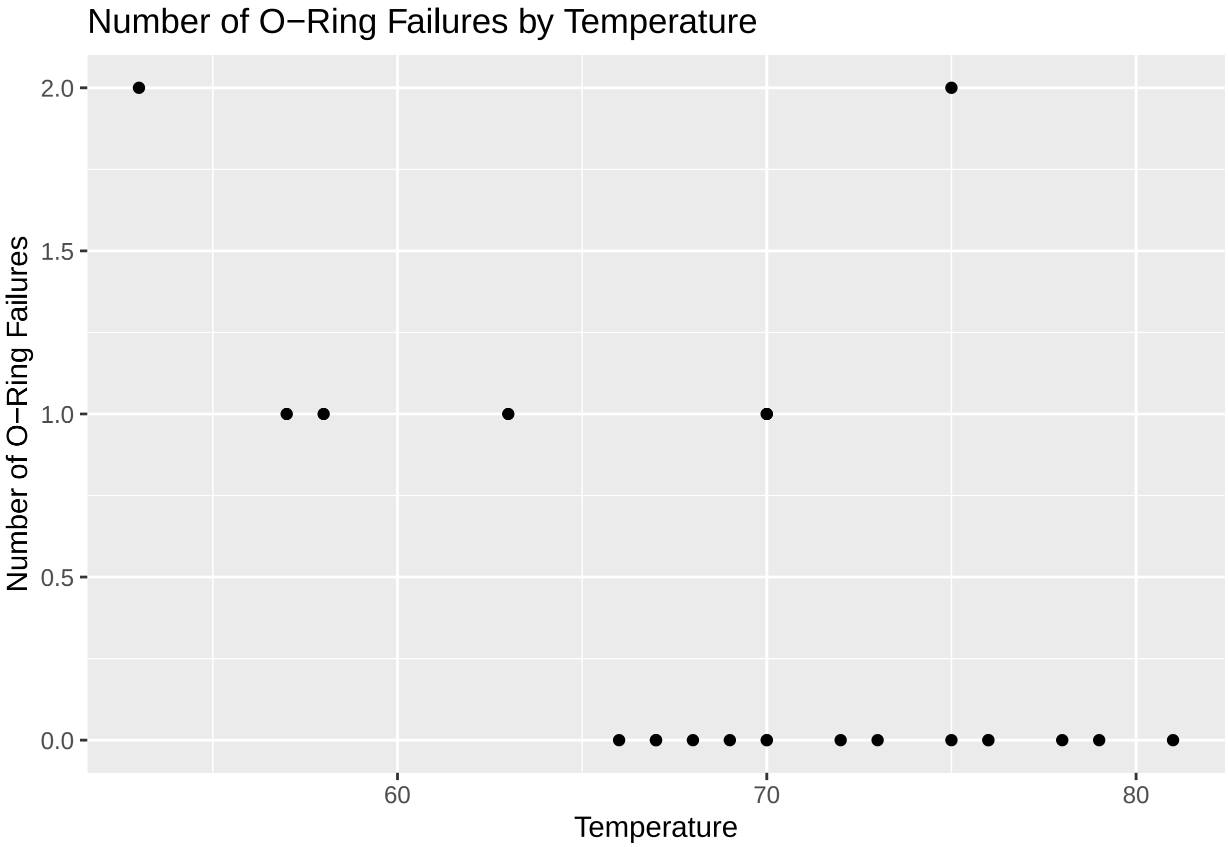
```
## Rows: 23 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): Flight, Temp, Pressure, O.ring, Number
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
summary(df)
```

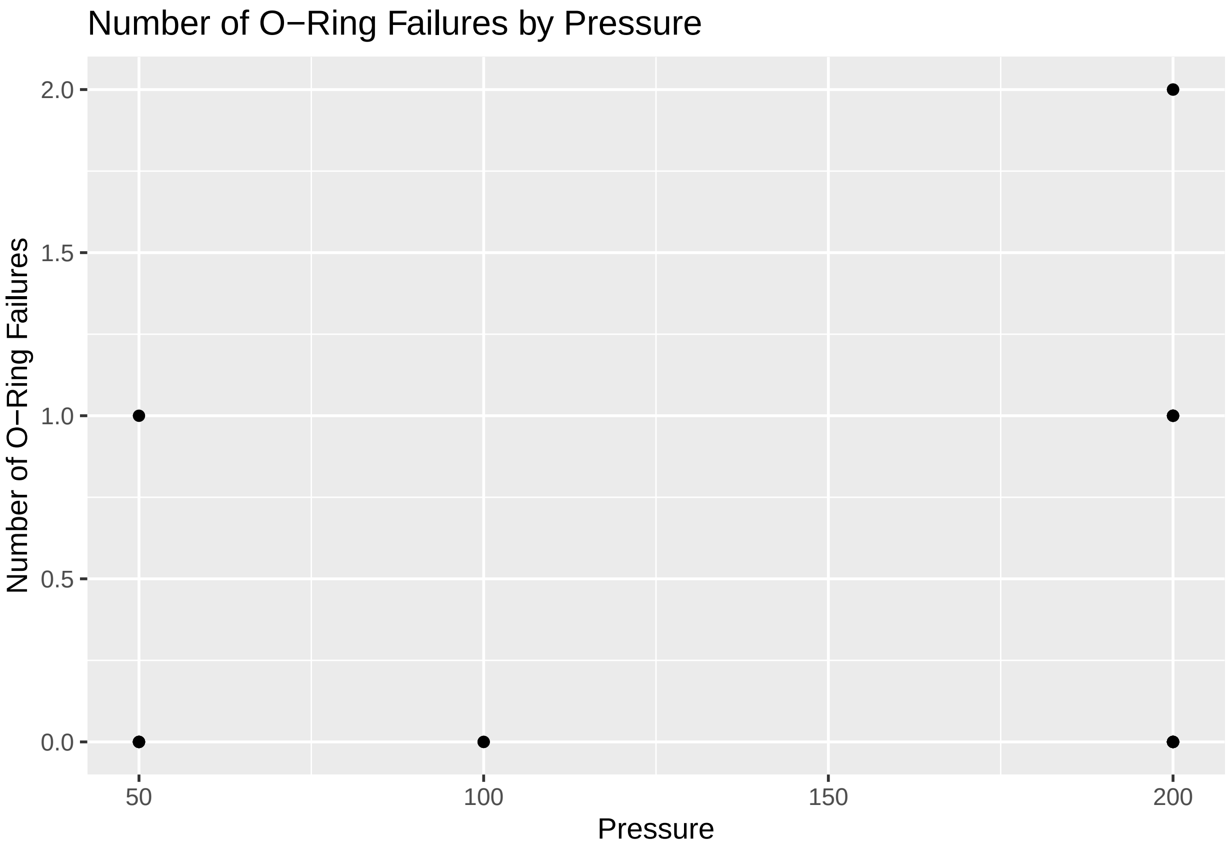
```
##      Flight          Temp          Pressure         O.ring           Number
##  Min.   : 1.0   Min.   :53.00   Min.   : 50.0   Min.   :0.0000   Min.   :6
##  1st Qu.: 6.5   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.:0.0000   1st Qu.:6
##  Median :12.0   Median :70.00   Median :200.0   Median :0.0000   Median :6
##  Mean   :12.0   Mean   :69.57   Mean   :152.2   Mean   :0.3913   Mean   :6
##  3rd Qu.:17.5   3rd Qu.:75.00   3rd Qu.:200.0   3rd Qu.:1.0000   3rd Qu.:6
##  Max.   :23.0   Max.   :81.00   Max.   :200.0   Max.   :2.0000   Max.   :6
```

```
ggplot(df, aes(x = Temp, y = O.ring)) +
  geom_point() +
  labs(title = "Number of O-Ring Failures by Temperature") +
```

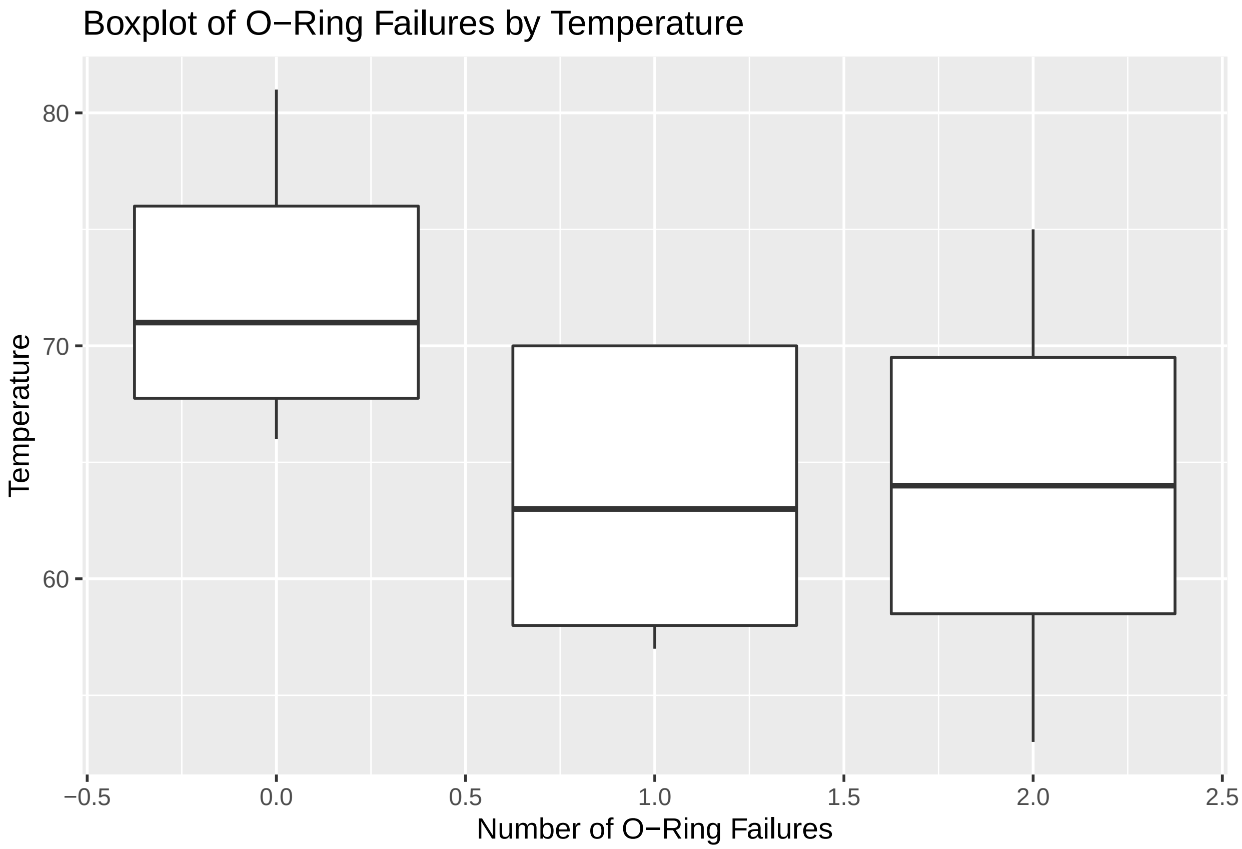
```r
  ylab("Number of O-Ring Failures") +
  xlab("Temperature")
```



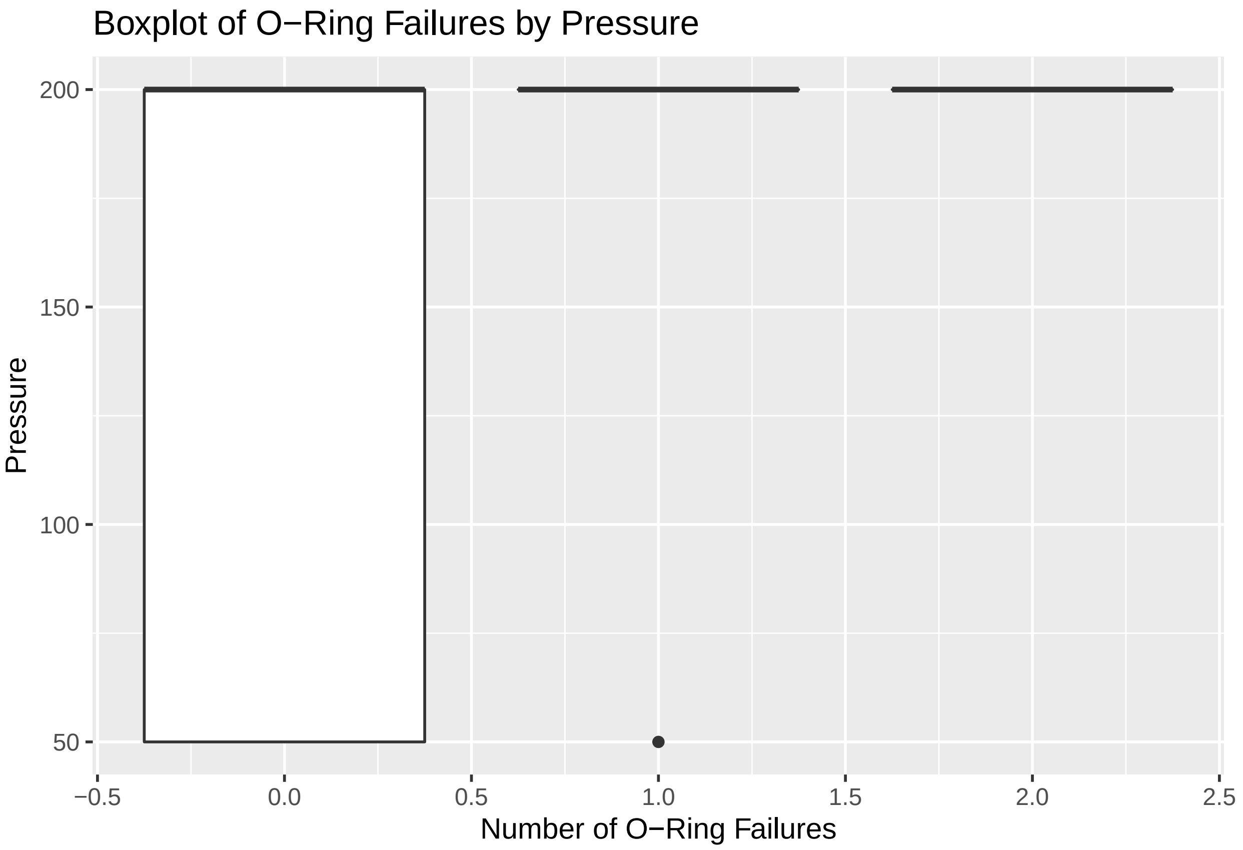
```r
ggplot(df, aes(x = Pressure, y = O.ring)) +
  geom_point() +
  labs(title = "Number of O-Ring Failures by Pressure") +
  ylab("Number of O-Ring Failures") +
  xlab("Pressure")
```



```
ggplot(df, aes(x = O.ring, y = Temp, group = O.ring)) +
  geom_boxplot() +
  labs(title = "Boxplot of O-Ring Failures by Temperature") +
  ylab("Temperature") +
  xlab("Number of O-Ring Failures")
```



```
ggplot(df, aes(x = O.ring, y = Pressure, group = O.ring)) +
  geom_boxplot() +
  labs(title = "Boxplot of O-Ring Failures by Pressure") +
  ylab("Pressure") +
  xlab("Number of O-Ring Failures")
```



Solution:
Based on the scatter plots above, we can see there is a negative relationship between temperature and the number of O-ring failures, that is, as temperature increases the number of O-ring failures appears to decrease. This intuition is confirmed when we look at the box plot, with both the mean and quantiles for temperature decreasing as number of O-ring failures increases.

However, there does not appear to be much of a relationship at all when we look at pressure. The scatter plot appears randomly distributed and the box plot is non descriptive at best.

2.1 Description

Complete the following task. In your final submission, please remove this question prompt so that your report reads as a report.

- Describe the data that you are using. How is this data generated, what is the sampling process that brought it to your availability. If it is helpful, you might describe the population (i.e. the Random Variables) that exist and how samples are produced from these random variables.
- The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

Solution:
INSERT WRITEUP HERE.

2.2 Key Features

Solution:
INSERT WRITUP HERE.

3 Analysis

3.1 Reproducing Previous Analysis (10 points)

Your analysis should address the following two questions. In your final submission, please remove this question prompt so that your report reads as a report.

1. Estimate the logistic regression model that the authors present in their report – include the variables as linear terms in the model. Evaluate, using likelihood ratio tests, the statistical significance of each explanatory variable in the model. Evaluate, using the context and data understanding that you have created in the **Data** section of this report, the practical significance of each explanatory variable in the model.

```r
model_1 <- glm(formula = O.ring / 6 ~ Temp + Pressure, data = df,
               family = 'binomial', weights = Number)

Anova(model_1, test = 'LR')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/6
##          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
```

Solution:
Based on the results of the likelihood ratio test, and the EDA performed above, the pressure variable does not appear to have either a statistically significant or practically significant effect on the number of O-ring failures. However, temperature does appear to be significant and should be used in whichever model we create.

2. Dalal, Fowlkes, and Hoadley (1989) chose to remove `pressure` from the model based on their likelihood ratio tests. Critically evaluate, using your test results and understanding of the question and data, whether `pressure` should be included in the model, or instead, `pressure` should not be included in the model. Your report needs to make a determination, argue why

it is most appropriate choice, and make note of how (if at all) the model results are affected by the choice of including or excluding `pressure`.

Solution:
Based on the results of the likelihood ratio test above, we would also conclude that the pressure variable should not be included in the model. The effect of the variable is not statistically significant, and the scatter plot of pressure versus number of O-ring failures shows no correlation between the two.

3.2 Confidence Intervals (20 points)

No matter what you determined about using or dropping `pressure`, for this section begin by considering the simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$, where π is the probability of an O-ring failure. Complete the following:

1. Estimate the logistic regression model.
2. Determine if a quadratic term is needed in the model for the temperature in this model.
3. Construct two plots:
4. π vs. Temp; and,
5. Expected number of failures vs. Temp.

Specific requirements for these plots:

- Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

- Include the 95% Wald confidence interval bands for π on the plot. Describe, in your analysis of these plots, why the bands much wider for lower temperatures than for higher temperatures?

```r
# 1. Estimate the logistic regression model.
model_2 <- glm(formula = O.ring / 6 ~ Temp, data = df, family = 'binomial',
               weights = Number)
summary(model_2)
```

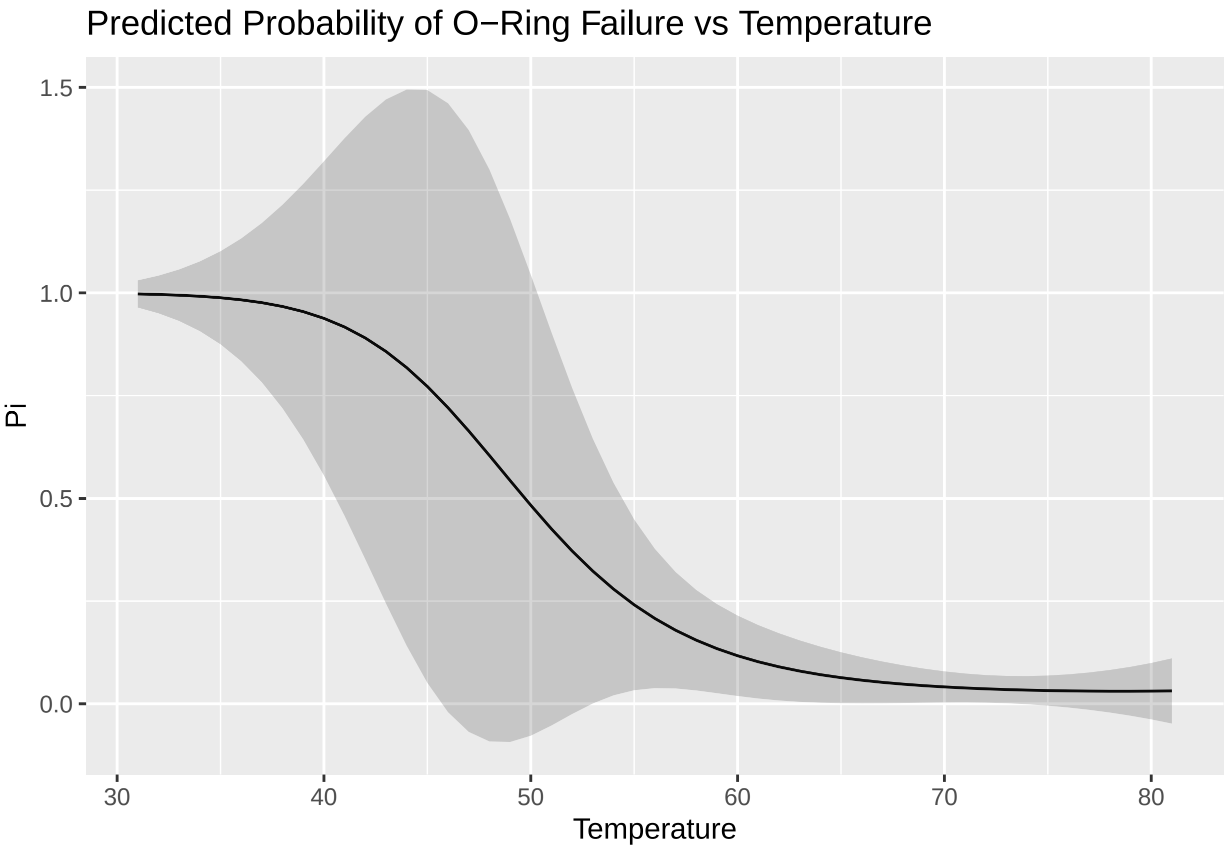
```
##
## Call:
## glm(formula = O.ring/6 ~ Temp, family = "binomial", data = df,
##     weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

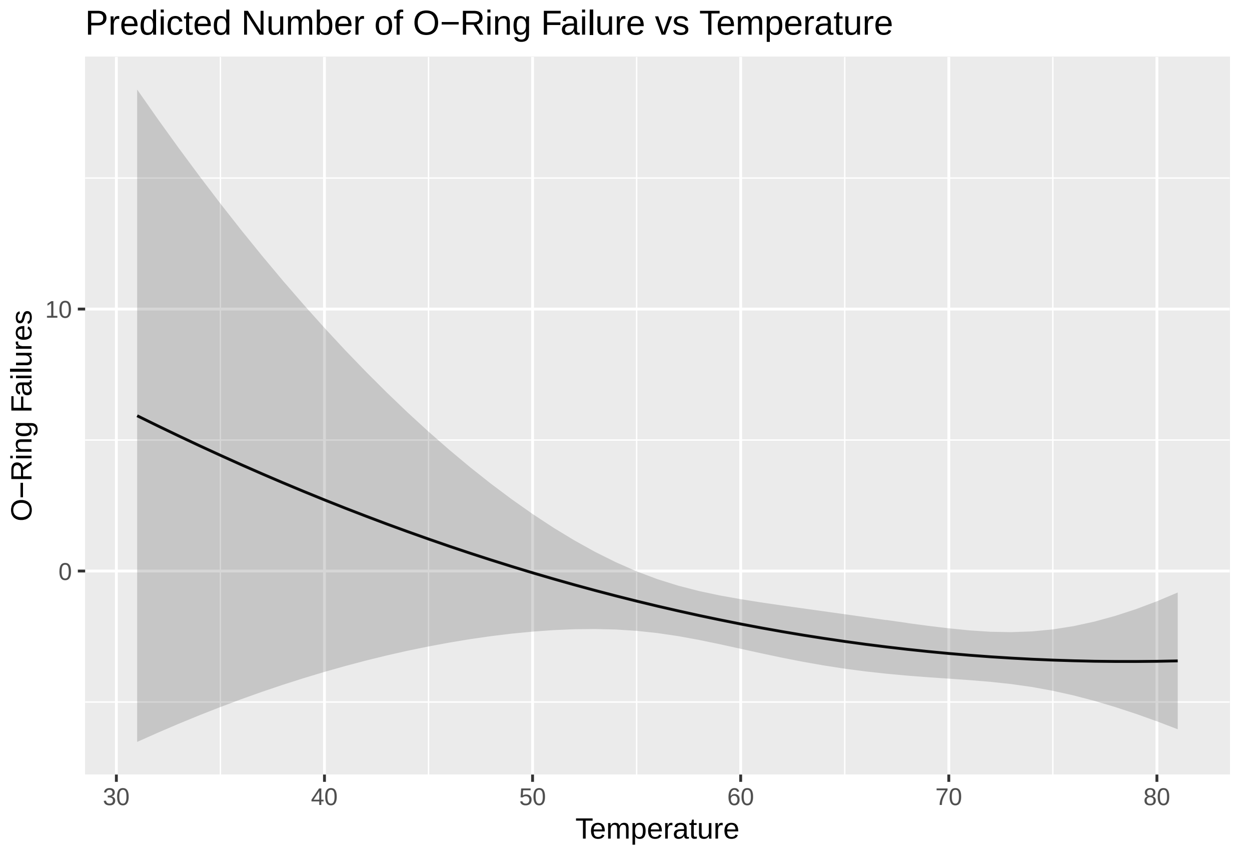
```
# 2. Determine if a quadratic term is needed in the model for the temperature in
#    this model.
model_3 <- glm(formula = O.ring / 6 ~ Temp + I(Temp ^ 2), data = df,
               family = 'binomial', weights = Number)

Anova(model_3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/6
##           LR Chisq Df Pr(>Chisq)
## Temp       0.71878  1     0.3965
## I(Temp^2)  0.49470  1     0.4818
```

```
# Pi vs Temp plot.
res <- data.frame(Temp = 31:81)
pred_prob <- predict(model_3, res, type="response", se = TRUE)
res['pi_hat'] <- pred_prob$fit
res['lower'] <- pred_prob$fit - qnorm(1 - 0.05 / 2) * pred_prob$se.fit
res['upper'] <- pred_prob$fit + qnorm(1 - 0.05 / 2) * pred_prob$se.fit

ggplot(res, aes(x = Temp, y = pi_hat)) +
  geom_line() +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  labs(title = "Predicted Probability of O-Ring Failure vs Temperature") +
  ylab("Pi") +
  xlab("Temperature")
```



```r
# Expected number of failures vs Temp plot.
pred <- predict(model_3, res, se = TRUE)
res['pred_failures'] <- pred$fit
res['failures_lower'] <- pred$fit - qnorm(1 - 0.05 / 2) * pred$se.fit
res['failures_upper'] <- pred$fit + qnorm(1 - 0.05 / 2) * pred$se.fit

ggplot(res, aes(x = Temp, y = pred_failures)) +
  geom_line() +
  geom_ribbon(aes(ymin = failures_lower, ymax = failures_upper), alpha = 0.2) +
  labs(title = "Predicted Number of O-Ring Failure vs Temperature") +
  ylab("O-Ring Failures") +
  xlab("Temperature")
```



Solution:
After looking at the likelihood ratio tests for a quadratic term, we would conclude that it is not necessary to include in the model.

3. The temperature was 31° at launch for the Challenger in 1986. Estimate the probability of an O-ring failure using this temperature, and compute a corresponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

Solution:
When the temperature is 31°, the estimated probability of O-ring failure is 0.9973 (0.9643, 1.0303).

3.3 Bootstrap Confidence Intervals (30 points)

Rather than relying on asymptotic properties, consider using a parametric bootstrap, as did Dalal, Fowlkes and Hoadley. To do this:

1. Simulate a large number of data sets (n = 23 for each) by re-sampling with replacement from the data.
2. Estimate a model for each dataset.
3. Compute the effect at a specific temperature of interest.

To produce a confidence interval, the authors used the 0.05 and 0.95 observed quantiles from the simulated distribution as their 90% confidence interval limits.

Using the parametric bootstrap, compute 90% confidence intervals separately at each integer temperature between 10° and 100° Fahrenheit.

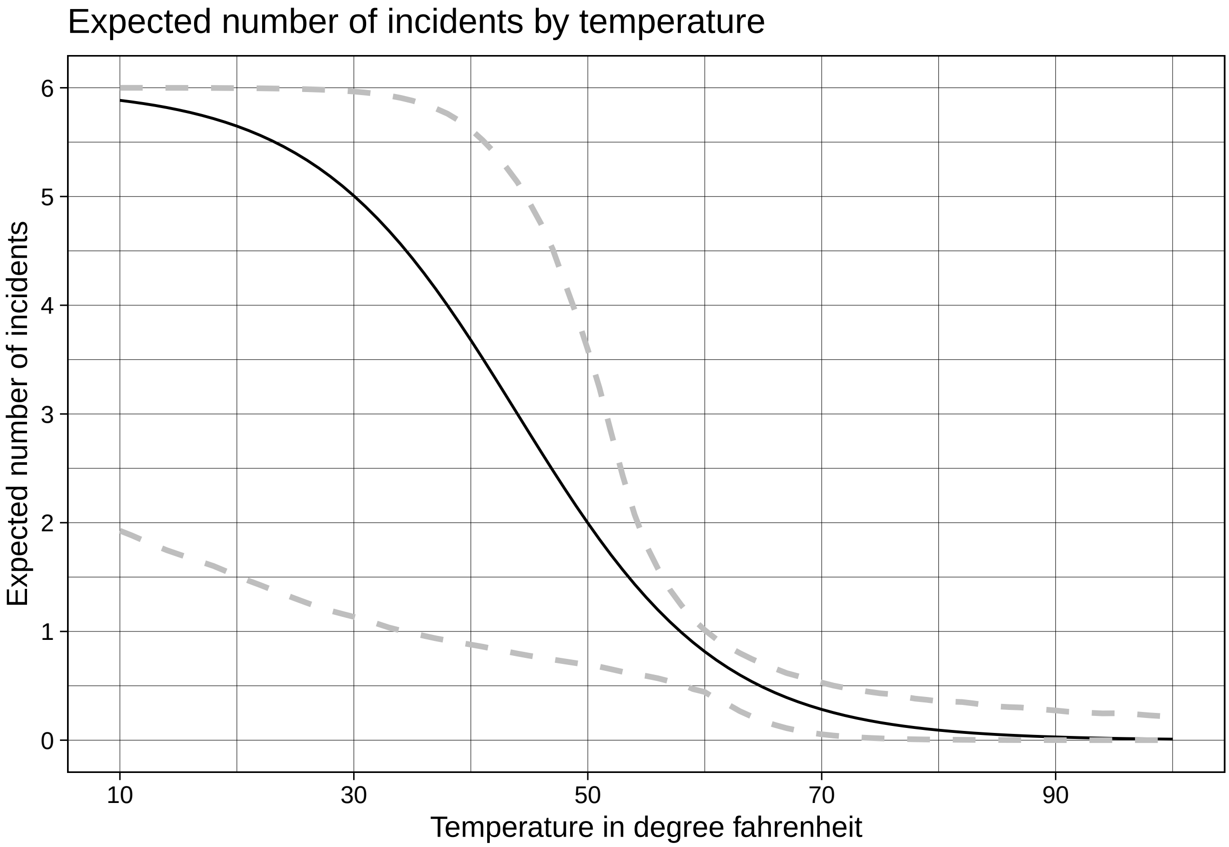
In this section, you should describe your process, justify such a process, and present your results in a way that is compelling for your reader.

```r
incidents.estimator <- function(b, seed=42, number=6) {
  # Setting seed to make the plot exactly reproduced
  set.seed(seed)
  n <- 23 # Resampled size
  estimation <- array(data = NA, dim=c(91, b))

  # Performing re-sampling repeatedly "b" times
  for(i in 1:b) {
    samples <- sample(x = 1:n, size = n, replace = TRUE)
    model   <- glm(formula = O.ring / Number ~ Temp, data = df[samples,],
                   weights = Number, family = 'binomial')
    # Performing prediction for 10-100° Fahrenheit using each fitted model
    for(t in 10:100) {
      p <- predict(object = model, newdata = data.frame(Temp = t),
                   type = "response")
      estimation[t-9, i] <- p
    }
  }

  # Fitting a model with the original data for prediction
  model <- glm(formula = O.ring/6 ~ Temp, data = df,
               weights = Number, family = 'binomial')
  # Estimating Expected number of incidents along with 90% CI
  incidents <- data.frame(Temperature = integer(),
                          Incidents = double(),
                          CI.Upper = double(),
                          CI.Lower = double())
  for(t in 10:100) {
    i <- t-9 # Index to locate data in the data.frame
    p <- predict(object = model, newdata = data.frame(Temp = t),
                 type = "response")
    # Getting the 90% CI and multiply by 6 to get estimation for a single flight
    ci <- quantile(estimation[i,], probs = c(0.05, 0.95), names=FALSE)
    incidents[i, 1] <- t
    incidents[i, 2] <- p*number
    incidents[i, 3] <- ci[1]*number
    incidents[i, 4] <- ci[2]*number
  }

  incidents
}
```

```r
# Performing predictions for 10°-100° temperature and plotting the result
predicted.incidents <- incidents.estimator(1000)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
ggplot(predicted.incidents, aes(x=Temperature)) +
  geom_line(aes(y = Incidents)) +
  geom_line(aes(y = CI.Upper), color = "gray", linetype="dashed", size = 1) +
  geom_line(aes(y = CI.Lower), color = "gray", linetype="dashed", size = 1) +
  scale_y_continuous(breaks = seq(0, 6, by = 1)) +
  scale_x_continuous(breaks = seq(10, 100, by = 20)) +
  labs(title = "Expected number of incidents by temperature") +
  xlab("Temperature in degree fahrenheit") +
  ylab("Expected number of incidents") +
  theme_linedraw()
```



3.4 Alternative Specification (10 points)

With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case? Explain why.

Solution:
INSERT WRITEUP HERE.

4 Conclusions (10 points)

Interpret the main result of your preferred model in terms of both odds and probability of failure. Summarize this result with respect to the question(s) being asked and key takeaways from the analysis.

Solution:
INSERT WRITEUP HERE.