W271 Group Lab 1

Investigating the 1986 Space Shuttle Challenger Accident

Please fill in with your names.

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

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.

```r
summary(data)
```

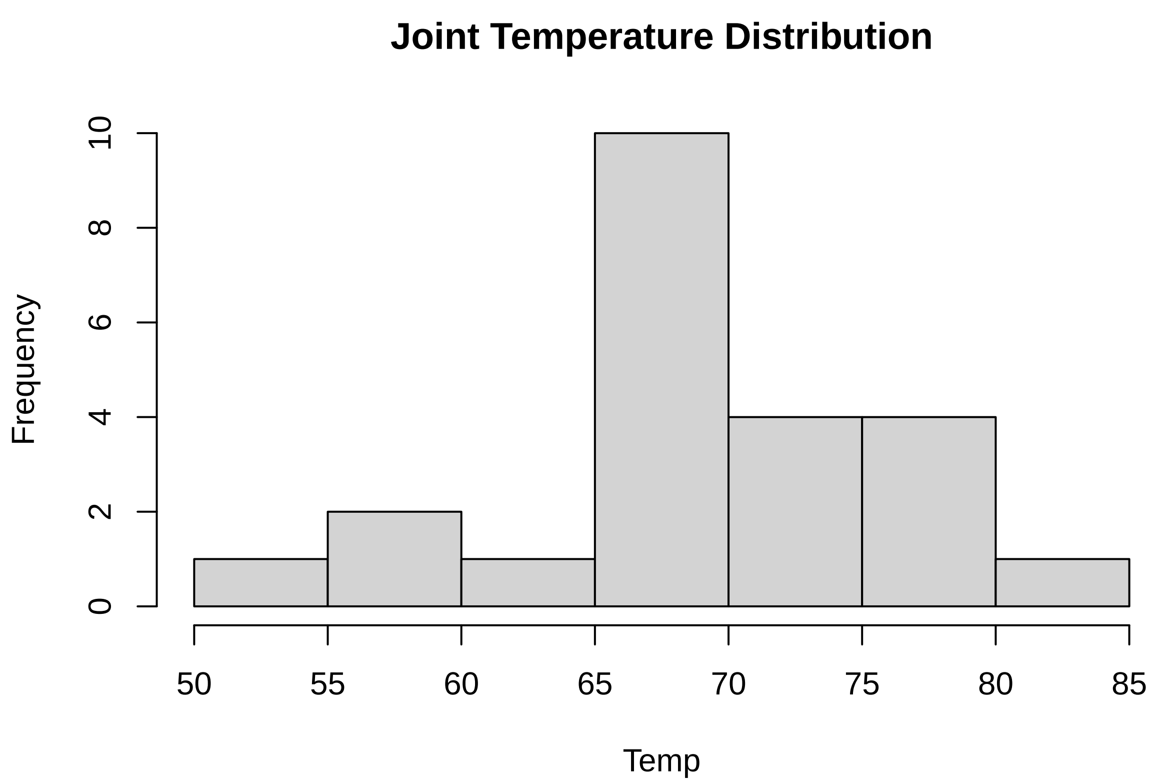
```
##      Flight          Temp          Pressure         O.ring           Number
##  Min.   : 1.0   Min.   :53.00   Min.   : 50.0   Min.   :0.0000   Min.   :6
##  1st Qu.: 6.5   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.:0.0000   1st Qu.:6
##  Median :12.0   Median :70.00   Median :200.0   Median :0.0000   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
```

```r
describe(data)
```

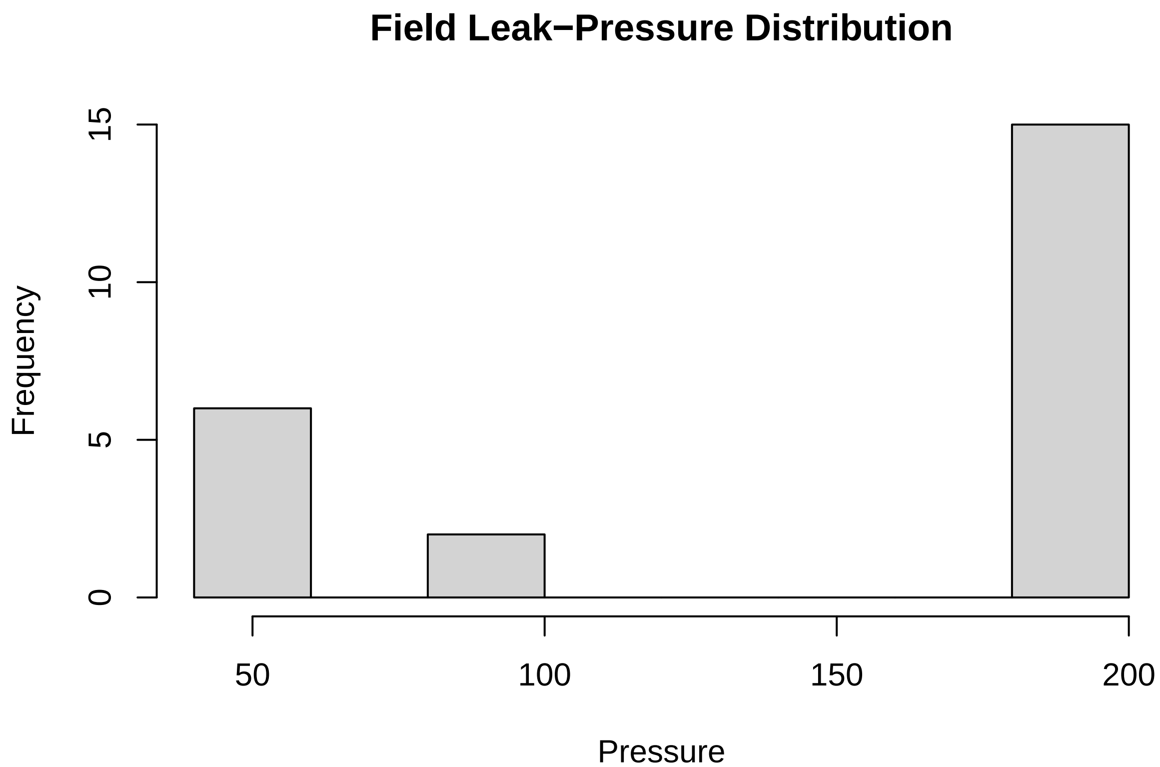
```
## data
##
##  5  Variables      23  Observations
## --------------------------------------------------------------------------------
## Flight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       23        0       23        1       12        8      2.1      3.2
##      .25      .50      .75      .90      .95
##      6.5     12.0     17.5     20.8     21.9
##
## lowest :  1  2  3  4  5, highest: 19 20 21 22 23
## --------------------------------------------------------------------------------
## Temp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       23        0       16    0.992    69.57    7.968     57.1     59.0
##      .25      .50      .75      .90      .95
##     67.0     70.0     75.0     77.6     78.9
##
## Value      53.00 56.92 57.76 62.80 65.88 67.00 67.84 68.96 69.80 71.76 72.88
## Frequency      1     1     1     1     1     3     1     1     4     1     1
## Proportion 0.043 0.043 0.043 0.043 0.043 0.130 0.043 0.043 0.174 0.043 0.043
##
## Value      74.84 75.96 77.92 78.76 81.00
## Frequency      2     2     1     1     1
## Proportion 0.087 0.087 0.043 0.043 0.043
##
## For the frequency table, variable is rounded to the nearest 0.28
## --------------------------------------------------------------------------------
## Pressure
##        n  missing distinct     Info     Mean      Gmd
##       23        0        3    0.706    152.2    67.59
##
## Value       50.0  99.5 200.0
## Frequency      6     2    15
## Proportion 0.261 0.087 0.652
##
```

```
## For the frequency table, variable is rounded to the nearest 1.5
## --------------------------------------------------------------------------------
## O.ring
##        n  missing distinct     Info     Mean      Gmd
##       23        0        3    0.654   0.3913   0.6087
##
## Value          0     1     2
## Frequency     16     5     2
## Proportion 0.696 0.217 0.087
##
## For the frequency table, variable is rounded to the nearest 0.02
## --------------------------------------------------------------------------------
## Number
##        n  missing distinct     Info     Mean      Gmd
##       23        0        1        0        6        0
##
## Value       6
## Frequency  23
## Proportion  1
## --------------------------------------------------------------------------------
```

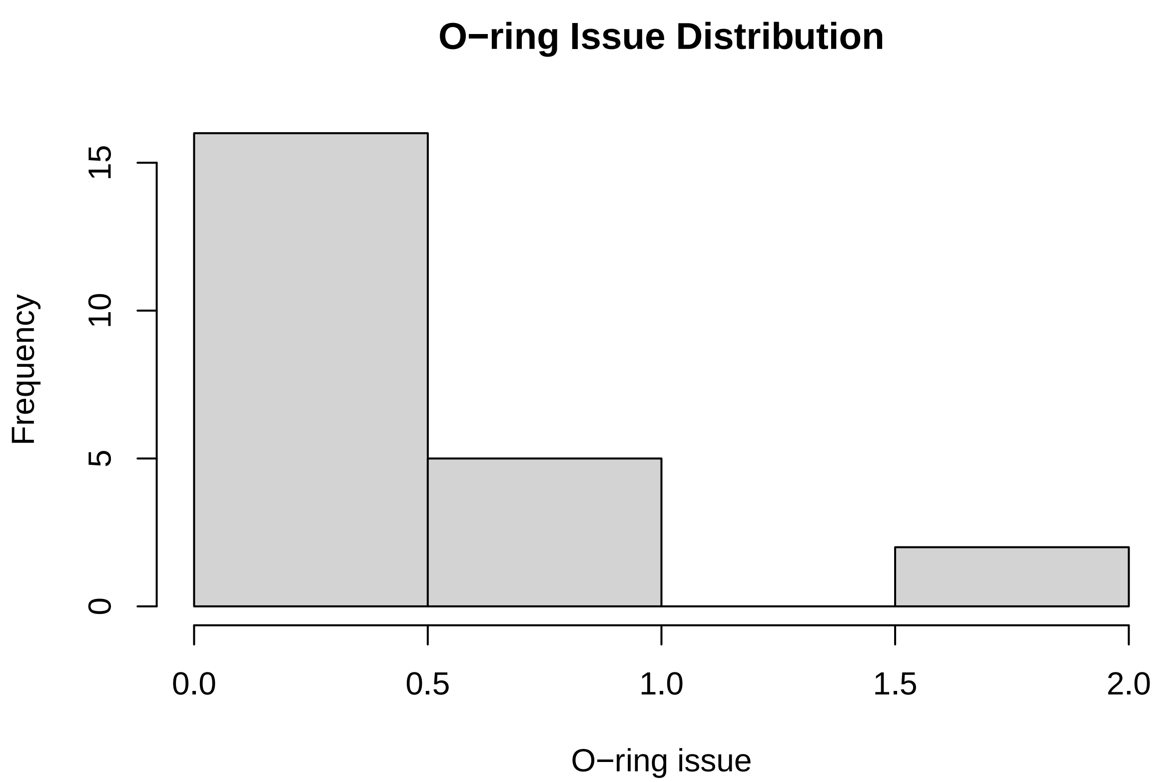
```r
# Histograms

hist(data$Temp, main = "Joint Temperature Distribution", xlab = "Temp")
```



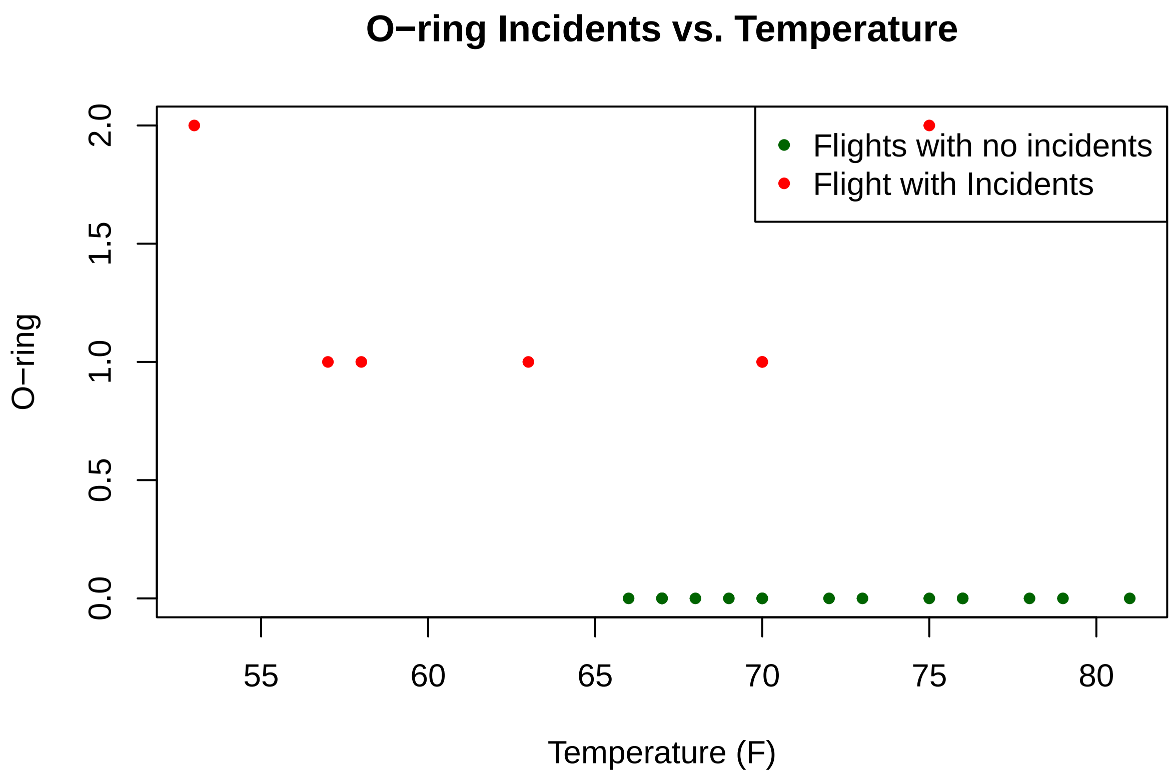
```r
hist(data$Pressure, main = "Field Leak-Pressure Distribution", xlab = "Pressure")
```




```r
hist(data$O.ring, main = "O-ring Issue Distribution", xlab = "O-ring issue")
```

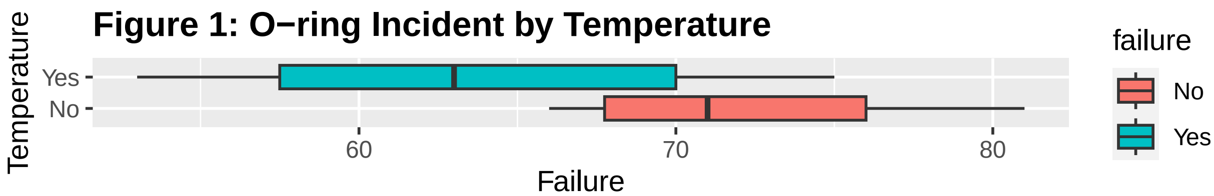


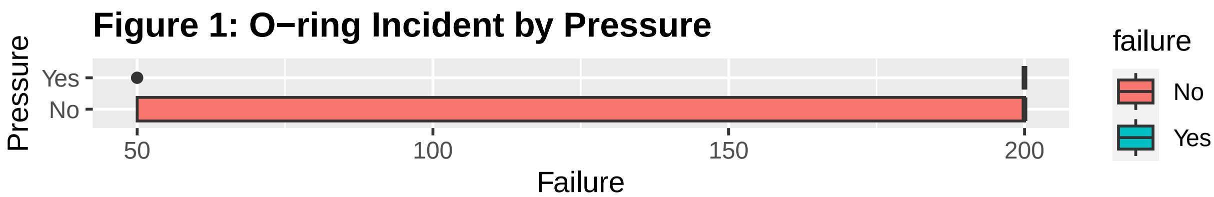
```r
# Define colors based on O.ring values
color <- ifelse(data$O.ring > 0, "red", "dark green")

# Create the scatter plot
plot(data$Temp, data$O.ring, xlab = "Temperature (F)", ylab = "O-ring", main = "O-ring Inciden
# Create a legend
legend("topright", legend = c("Flights with no incidents", "Flight with Incidents"), col = c("
```

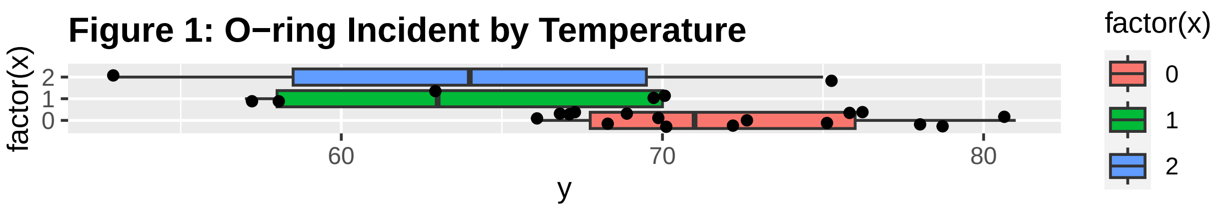


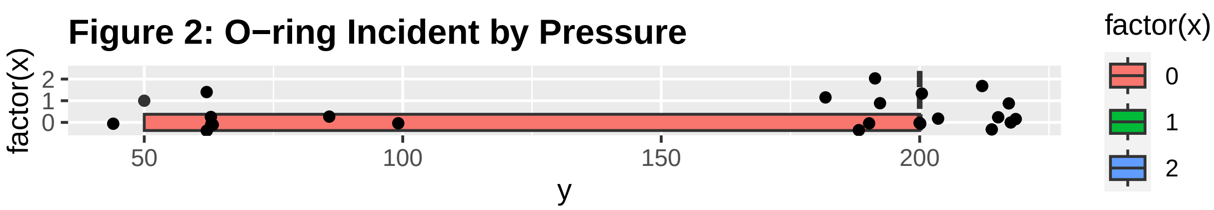

```r
data %>%
  count(O.ring) %>%
  mutate(prop = round(prop.table(n), 2)) %>%
  kable(col.names = c("Incidents", "Number", "Proportion"))
```

Incidents	Number	Proportion
0	16	0.70
1	5	0.22
2	2	0.09









From Data EDA, we don't see any anomalous observations. The launch temperature is approximately normal distributed with a mean of 65-70F. In the sample, approximately 70% of the launches had zero O-ring incidents. Preliminary graphs show that temperature has an effect on O-ring incidents. Lower temperature might increase the probability of O-ring incidents.

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.

The data used in this analysis pertains to the Challenger Space Shuttle disaster and was generated by recording observations of major parameters such as temperature, pressure and O-ring incidents across different shuttle launches. It is sampled from a population of all shuttle launches. The random variables are temperature, pressure, and O-ring failure status. The specific values of these random variables are produced from the actual launches. The sampling process is not a random sampling process because it doesn't involve randomly selecting elements from a larger population. Rather, it involves recording observations from a sequence of events (the shuttle launches) that actually took place. Each launch can be thought of as an experiment where the temperature and pressure are somewhat uncontrollable variables and the O-ring failure status is the outcome.

The assumption of independence in the context of logistic regression means that we assume the outcome (O-ring failure) of one launch doesn't affect the outcome of another launch. This is necessary because if the outcomes were dependent, then conditional probability would need to be accounted for. The dependence of one outcome on another would violate the independence assumption of the logistic regression model. As a result, the estimation of the parameters from the usual logistic regression model would not have statistical significance. The potential problem with this assumption in the current context is that the launches are not truly independent events. Factors like technological advancements, changes in maintenance practices, and learning from past failures could cause later launches to be dependent on earlier ones. Due to technological advancement, more recent launches might see fewer O-ring incidents, skewing the proportion of incidents lower. On the other hand, the physical conditions of the O-rings might not be independent, for example, if the same set of O-rings were used in multiple launches. The metal fatigue as a result of repeated usage of the same set of O-rings could skew the incident proportion higher.

2.2 Key Features

Data Source and Context: The data set is derived from the Challenger Space Shuttle missions, detailing the conditions of each launch, particularly focusing on temperature, pressure, and the crucial aspect of O-ring failure. The observations represent specific shuttle launches under various conditions.

Variables: The primary variables in this data set include temperature (Temp), pressure (Pressure), and the status of the O-ring (O.ring). Each variable provides critical information about the conditions during each space shuttle launch.

Target Variable: The target variable in the analysis is O-ring failure, a binary variable that indicates whether an O-ring failure occurred during a launch (1 = failure, 0 = no failure).

Explanatory Variables: The temperature and pressure during each launch are the explanatory variables. These continuous variables provide context for the conditions that may influence O-ring failure.

Data Structure: The data set is structured as a cross-sectional data set, with each row representing a different space shuttle launch. Random variables “temperature” and “pressure” are column variables “Temp” and “Pressure”, respectively.

3 Analysis

Temperature (Temp):

Distribution: The temperature during the launches varied from 53°F to 81°F. The distribution of temperature values was fairly normal but slightly skewed towards the higher end.

Relationship with O-ring failure: The exploratory data analysis and logistic regression model indicated a significant relationship between temperature and O-ring failure. Specifically, lower temperatures were associated with a higher risk of O-ring failure. This relationship was found to be statistically significant, making temperature a key variable to consider when predicting O-ring failure.

Insights: The fact that colder temperatures increase the risk of O-ring failure can have important practical implications. For example, it could inform decisions about the optimal conditions for launching a space shuttle to minimize the risk of O-ring failure.

Pressure (Pressure):

Distribution: The pressure during the launches was mostly around 200 psi, with a few instances of lower pressure at 50 psi and 100 psi. The distribution was therefore skewed towards the higher end.

Relationship with O-ring failure: The exploratory data analysis and logistic regression model did not indicate a statistically significant relationship between pressure and O-ring failure. Although pressure was initially included as an explanatory variable in the model, the likelihood ratio test suggested that it did not significantly improve the model's fit to the data.

Insights: While it might seem logical that higher pressure could increase the risk of O-ring failure, the data did not provide statistical evidence for this. Therefore, based on this data set, pressure does not appear to be a key variable to consider when predicting O-ring failure.

In conclusion, while both temperature and pressure are important factors in the overall operating conditions during a space shuttle launch, temperature appears to be the more crucial variable when it comes to predicting O-ring failure.

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
# Check if O.ring is a factor, if not convert it
if (!is.factor(data$O.ring)) {
  data$O.ring <- as.factor(data$O.ring)
}

# Fit the logistic regression model
model <- glm(O.ring ~ Temp + Pressure, data = data, family = binomial(link = "logit"))

# Fit reduced models
model_temp <- glm(O.ring ~ Temp, data = data, family = binomial(link = "logit"))
model_pressure <- glm(O.ring ~ Pressure, data = data, family = binomial(link = "logit"))

summary(model)
```

The coefficient of -0.228671 indicates that for each additional degree in temperature, the log-odds of O-ring failure decrease by about 0.23, holding pressure constant. This would suggest that higher temperatures are associated with a lower risk of O-ring failure. Because the p-value (0.0376) is below 0.05, this result is statistically significant at the 5% level.

The coefficient of 0.010400 suggests that for each additional unit increase in pressure, the log-odds of O-ring failure increase by about 0.01, holding temperature constant. However, the p-value (0.2468) is greater than 0.05, suggesting that this result is not statistically significant at the 5% level. Therefore, based on this model, pressure does not have a significant effect on the log-odds of O-ring failure.

From a practical perspective, temperature seems to have a significant effect on O-ring failure, with higher temperatures reducing the likelihood of failure. However, pressure doesn't seem to significantly affect the likelihood of failure.

```r
# Perform likelihood ratio tests
anova(model_temp, model, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Temp
## Model 2: O.ring ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     18.782  1   1.5331   0.2156
```

The test compares the deviances of the two models to assess whether adding Pressure to the model significantly improves the model fit. The deviance difference (1.5331) is not statistically significant (p=0.2156), as indicated by the Pr(>Chi) value. This means that, based on this test, adding Pressure as a predictor does not significantly improve the model.

```r
anova(model_pressure, model, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Pressure
## Model 2: O.ring ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     26.536
## 2        20     18.782  1   7.7542 0.005359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test compares the deviances of the two models to assess whether adding Temp to the model significantly improves the model fit. The deviance difference (7.7542) is statistically significant (p=0.005359), as indicated by the Pr(>Chi) value. This means that adding Temp as a predictor significantly improves the model.

In summary, these results suggest that Temp is a significant predictor of O.ring failure, while Pressure is not, at least not when added to a model that already includes Temp.

```r
exp(coef(model))
```

```
##  (Intercept)         Temp     Pressure
## 5.926501e+05 7.955903e-01 1.010454e+00
```

The exponentiated coefficient for Temp is 0.7955903. This means that for each unit increase in Temp, the odds of O-ring failure are multiplied by approximately 0.80, or in other words, they decrease by about 20%, assuming all other variables are held constant. This matches with our previous interpretation that an increase in temperature decreases the likelihood of O-ring failure.

The exponentiated coefficient for Pressure is 1.010454. This means that for each unit increase in Pressure, the odds of O-ring failure are multiplied by approximately 1.01, or they increase by about 1%, assuming all other variables are held constant. However, remember from the earlier analysis that this effect was not statistically significant.

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`.

Based on the likelihood ratio tests that we have conducted, we found that the inclusion of Pressure in the model does not significantly improve the model's fit when Temp is already in the model. This is indicated by a non-significant p-value ($p = 0.2156$) when we compared the model with Temp and Pressure to the model with Temp alone.

On the contrary, when we compared the model with Pressure alone to the model with Temp and Pressure, we observed a significant improvement in the model fit with the inclusion of Temp ($p = 0.005359$).

Taking these results together, it would be reasonable to exclude Pressure from the model. By doing so, we would have a simpler model that is easier to interpret, without a significant loss in predictive power.

Excluding Pressure from the model does not appear to negatively affect the model's ability to predict O-ring failure. The significant predictor, Temp, remains in the model and its coefficient remains statistically significant, indicating that it is still able to predict O-ring failure effectively.

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.

```
model_temp <- glm(O.ring ~ Temp, family = binomial(link = "logit"), data = data)
coef(model_temp)
```

```
## (Intercept)        Temp
##  15.0429016  -0.2321627
```

2. Determine if a quadratic term is needed in the model for the temperature in this model.

```
model_temp_quad <- glm(O.ring ~ Temp + I(Temp^2), family = binomial(link = "logit"), data = da
anova(model_temp, model_temp_quad, test = "Chisq")
```

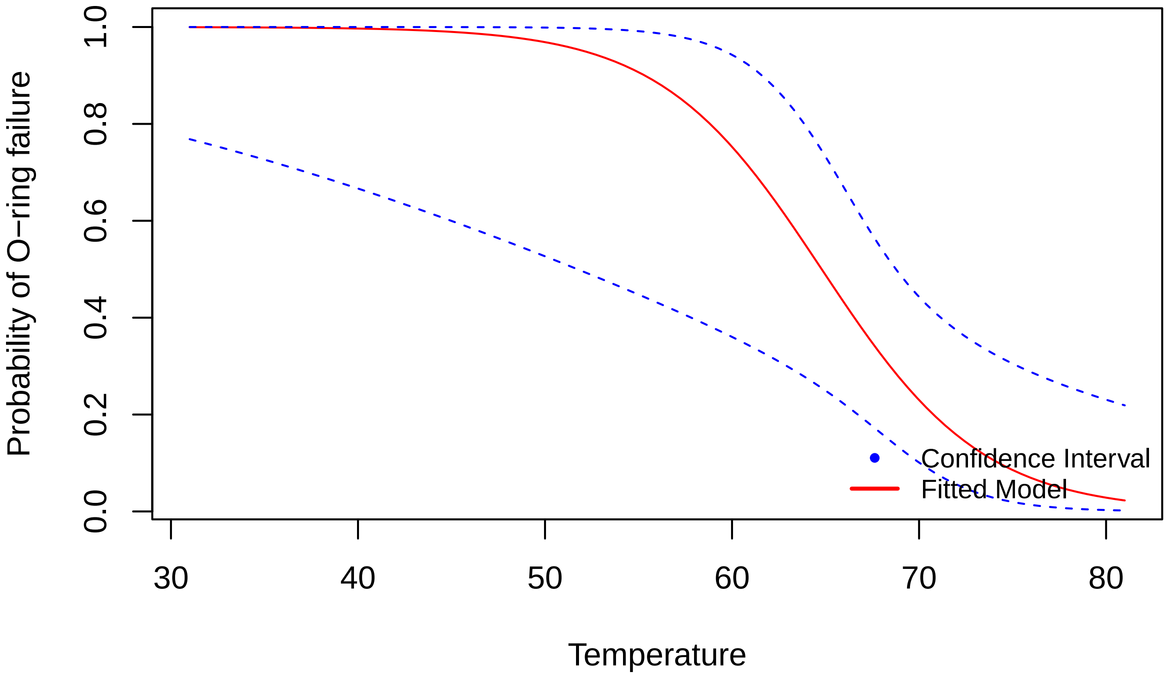
```
## Analysis of Deviance Table
##
```

```
## Model 1: O.ring ~ Temp
## Model 2: O.ring ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     19.389  1  0.92649   0.3358
```

Not needed.

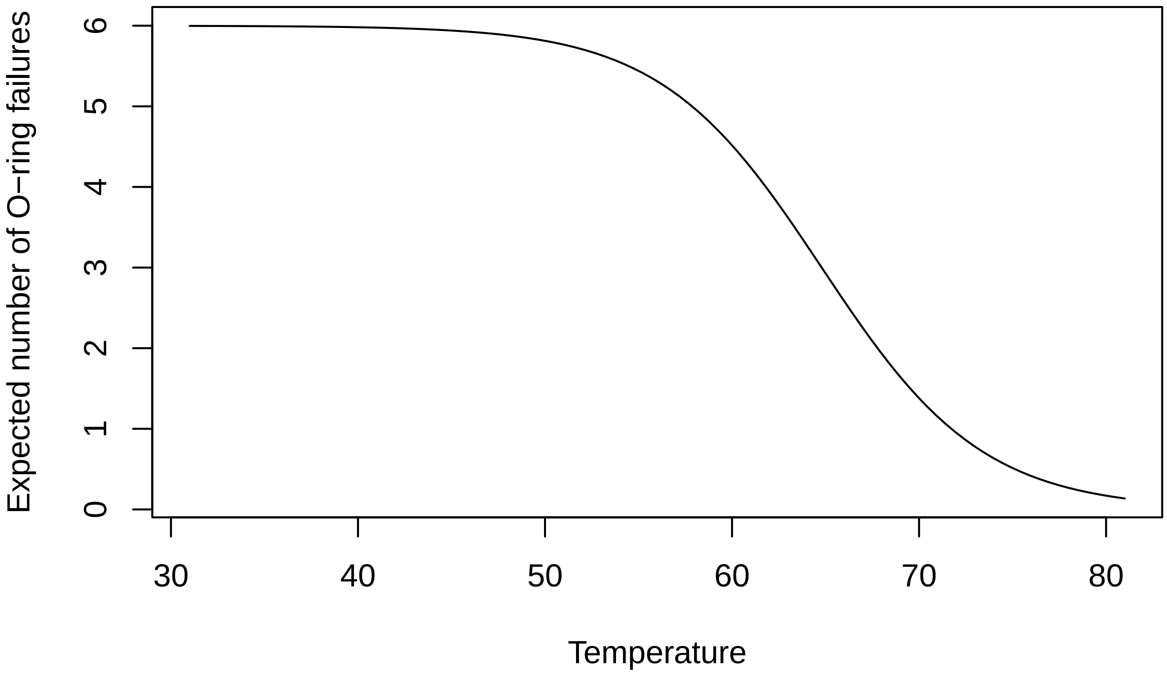
3. Construct two plots:
4. π vs. Temp; and,

```r
temp_range <- seq(31, 81, length.out = 100)

prob_est <- predict(model_temp, newdata = data.frame(Temp = temp_range), type = "response")
conf_int <- predict(model_temp, newdata = data.frame(Temp = temp_range), type = "link", se.fit

alpha = 0.05
CI.pred.upper = conf_int$fit + qnorm(0.95)*conf_int$se.fit
CI.pred.lower = conf_int$fit + qnorm(0.05)*conf_int$se.fit
CI.pi.upper = exp(CI.pred.upper)/(1+exp(CI.pred.upper))
CI.pi.lower = exp(CI.pred.lower)/(1+exp(CI.pred.lower))

plot(temp_range, prob_est, type = "l", xlab = "Temperature", ylab = "Probability of O-ring fai
lines(temp_range, CI.pi.upper, lty = 2, col='blue')
lines(temp_range, CI.pi.lower, lty = 2, col='blue')
legend("bottomright",
       legend = c("Confidence Interval", "Fitted Model"),
       col = c("blue", "red"),
       pch = c(20, NA),
       lty = c(NA, 1),
       cex = 0.8,
       bty = "n",
       lwd = c(NA, 2))
```



5. Expected number of failures vs. Temp.

```
expected_failures <- prob_est * 6 # Assuming there are 6 O-rings
plot(temp_range, expected_failures, type = "l", xlab = "Temperature", ylab = "Expected number o
```



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?

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.

```
prob_31 <- predict(model_temp, newdata = data.frame(Temp = 31), type = "response")
conf_int_31 <- predict(model_temp, newdata = data.frame(Temp = 31), type = "link", se.fit = TRU
CI_31.pred.upper = conf_int_31$fit + qnorm(0.95)*conf_int_31$se.fit
CI_31.pred.lower = conf_int_31$fit + qnorm(0.05)*conf_int_31$se.fit
CI_31.pi.upper = exp(CI_31.pred.upper)/(1+exp(CI_31.pred.upper))
CI_31.pi.lower = exp(CI_31.pred.lower)/(1+exp(CI_31.pred.lower))

prob_31
```

```
##         1
## 0.9996088
```

```
CI_31.pi.upper
```

```
##         1
## 0.9999995
```

```
CI_31.pi.lower
```

```
##         1
## 0.7684619
```

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.

```
# Load necessary library



# Set seed for reproducibility
set.seed(123)
```

```r
# Fit the initial model to the data
fit <- glm(O.ring ~ Temp, data = data, family = binomial)

# Extract the fitted parameters
alpha <- coef(fit)[1]
beta <- coef(fit)[2]

# Number of bootstrap resamples
R <- 1000

# Preallocate a matrix to store the bootstrap coefficients
bootstrap_coefs <- matrix(NA, nrow = R, ncol = 2,
                          dimnames = list(1:R, names(coef(fit))))

i <- 0
# Run the bootstrap
while (i < R) {
  # Generate new response values from the fitted model
  y_star <- rbinom(n = nrow(data), size = 1, prob = plogis(alpha + beta * data$Temp))

  # Fit the model to the new data
  fit_star <- glm(y_star ~ data$Temp, family = binomial)

  if (fit_star$deviance >0.1) {
  # Record the estimated parameters
  i = i+1
  bootstrap_coefs[i,] <- coef(fit_star)
  }
}

# Compute the 90% confidence interval for alpha
alpha_CI <- quantile(bootstrap_coefs[, "(Intercept)"], probs = c(0.05, 0.95))

# Preallocate a matrix for the beta confidence intervals
beta_CI <- matrix(NA, nrow = 91, ncol = 2,
                  dimnames = list(10:100, c("Lower", "Upper")))

# Compute the 90% confidence interval for beta at each temperature
for (temp in 10:100) {
  # Calculate the beta values at this temperature
  beta_at_temp <- alpha_CI + bootstrap_coefs[, "Temp"] * temp
  # Compute the confidence interval
  beta_CI[temp - 9,] <- quantile(beta_at_temp, probs = c(0.05, 0.95))
}
```

```r
beta_CI
```

```
##           Lower    Upper
## 10    0.3123808 56.12224
## 11   -0.3075079 55.98768
## 12   -0.9273967 55.85312
## 13   -1.5472855 55.71856
## 14   -2.1671743 55.58401
## 15   -2.7870631 55.44945
## 16   -3.4069519 55.31489
## 17   -4.0268407 55.18033
## 18   -4.6467295 55.04577
## 19   -5.2666182 54.91121
## 20   -5.8865070 54.77665
## 21   -6.5063958 54.64210
## 22   -7.1262846 54.50754
## 23   -7.7461734 54.37298
## 24   -8.3660622 54.23842
## 25   -8.9859510 54.10386
## 26   -9.6058398 53.96930
## 27  -10.2257285 53.83474
## 28  -10.8456173 53.70019
## 29  -11.4655061 53.56563
## 30  -12.0853949 53.43107
## 31  -12.7052837 53.29651
## 32  -13.3251725 53.16195
## 33  -13.9450613 53.02739
## 34  -14.5649500 52.89283
## 35  -15.1848388 52.75828
## 36  -15.8047276 52.62372
## 37  -16.4246164 52.48916
## 38  -17.0445052 52.35460
## 39  -17.6643940 52.22004
## 40  -18.2842828 52.08548
## 41  -18.9041716 51.95092
## 42  -19.5240603 51.81637
## 43  -20.1439491 51.68181
## 44  -20.7638379 51.54725
## 45  -21.3837267 51.41269
## 46  -22.0036155 51.27813
## 47  -22.6235043 51.14357
## 48  -23.6936744 51.00901
## 49  -24.3760238 50.87446
## 50  -25.0532437 50.73990
## 51  -25.6845339 50.60534
## 52  -26.3158242 50.47078
## 53  -26.9471144 50.33622
## 54  -27.5784047 50.20166
## 55  -28.2096949 50.06710
## 56  -28.8409852 49.93255
```

```
## 57  -29.4722754 49.79799
## 58  -30.1035657 49.66343
## 59  -30.7348559 49.52887
## 60  -31.3661462 49.39431
## 61  -31.9974364 49.25975
## 62  -32.6287266 49.12519
## 63  -33.2600169 48.99064
## 64  -33.8913071 48.85608
## 65  -34.5225974 48.72152
## 66  -35.1538876 48.58696
## 67  -35.7851779 48.45240
## 68  -36.4164681 48.31784
## 69  -37.0477584 48.18329
## 70  -37.6790486 48.04873
## 71  -38.3103389 47.91417
## 72  -38.9416291 47.77961
## 73  -39.5729194 47.64505
## 74  -40.2042096 47.51049
## 75  -40.8354999 47.37593
## 76  -41.4667901 47.24138
## 77  -42.0980804 47.10682
## 78  -42.7293706 46.97226
## 79  -43.3606609 46.83770
## 80  -43.9919511 46.70314
## 81  -44.6232414 46.56858
## 82  -45.2545316 46.43402
## 83  -45.8858219 46.29947
## 84  -46.5171121 46.16491
## 85  -47.1484023 46.03035
## 86  -47.7796926 45.89579
## 87  -48.5146327 45.76123
## 88  -49.7035365 45.62667
## 89  -50.8924403 45.49211
## 90  -52.0813441 45.35756
## 91  -53.2702479 45.22300
## 92  -54.4591517 45.08844
## 93  -55.6480555 44.95388
## 94  -56.8369593 44.81932
## 95  -57.8774636 44.68476
## 96  -58.5552397 44.55020
## 97  -59.2330159 44.41565
## 98  -59.9107920 44.28109
## 99  -60.5885681 44.14653
## 100 -61.2663442 44.01197
```

```
# Convert log-odds to probabilities
beta_CI_prob <- plogis(beta_CI)
```

```r
# Print the probability confidence intervals
print(beta_CI_prob)
```

```
##             Lower Upper
## 10  5.774663e-01     1
## 11  4.237231e-01     1
## 12  2.834532e-01     1
## 13  1.754787e-01     1
## 14  1.027372e-01     1
## 15  5.802728e-02     1
## 16  3.207891e-02     1
## 17  1.751821e-02     1
## 18  9.501775e-03     1
## 19  5.134535e-03     1
## 20  2.768970e-03     1
## 21  1.491626e-03     1
## 22  8.030545e-04     1
## 23  4.322071e-04     1
## 24  2.325757e-04     1
## 25  1.251402e-04     1
## 26  6.732984e-05     1
## 27  3.622487e-05     1
## 28  1.948946e-05     1
## 29  1.048551e-05     1
## 30  5.641275e-06     1
## 31  3.035038e-06     1
## 32  1.632866e-06     1
## 33  8.784893e-07     1
## 34  4.726312e-07     1
## 35  2.542776e-07     1
## 36  1.368025e-07     1
## 37  7.360032e-08     1
## 38  3.959729e-08     1
## 39  2.130351e-08     1
## 40  1.146138e-08     1
## 41  6.166271e-09     1
## 42  3.317480e-09     1
## 43  1.784818e-09     1
## 44  9.602400e-10     1
## 45  5.166132e-10     1
## 46  2.779401e-10     1
## 47  1.495330e-10     1
## 48  5.128236e-11     1
## 49  2.591955e-11     1
## 50  1.316784e-11     1
## 51  7.004040e-12     1
## 52  3.725485e-12     1
```

```
## 53  1.981604e-12     1
## 54  1.054025e-12     1
## 55  5.606414e-13     1
## 56  2.982080e-13     1
## 57  1.586183e-13     1
## 58  8.436990e-14     1
## 59  4.487678e-14     1
## 60  2.387019e-14     1
## 61  1.269667e-14     1
## 62  6.753425e-15     1
## 63  3.592181e-15     1
## 64  1.910699e-15     1
## 65  1.016311e-15     1
## 66  5.405808e-16     1
## 67  2.875376e-16     1
## 68  1.529427e-16     1
## 69  8.135101e-17     1
## 70  4.327102e-17     1
## 71  2.301607e-17     1
## 72  1.224237e-17     1
## 73  6.511776e-18     1
## 74  3.463647e-18     1
## 75  1.842331e-18     1
## 76  9.799454e-19     1
## 77  5.212379e-19     1
## 78  2.772491e-19     1
## 79  1.474702e-19     1
## 80  7.844014e-20     1
## 81  4.172271e-20     1
## 82  2.219252e-20     1
## 83  1.180431e-20     1
## 84  6.278775e-21     1
## 85  3.339712e-21     1
## 86  1.776410e-21     1
## 87  8.518492e-22     1
## 88  2.594349e-22     1
## 89  7.901217e-23     1
## 90  2.406355e-23     1
## 91  7.328672e-24     1
## 92  2.231983e-24     1
## 93  6.797615e-25     1
## 94  2.070247e-25     1
## 95  7.313696e-26     1
## 96  3.713492e-26     1
## 97  1.885507e-26     1
## 98  9.573564e-27     1
## 99  4.860928e-27     1
## 100 2.468111e-27     1
```

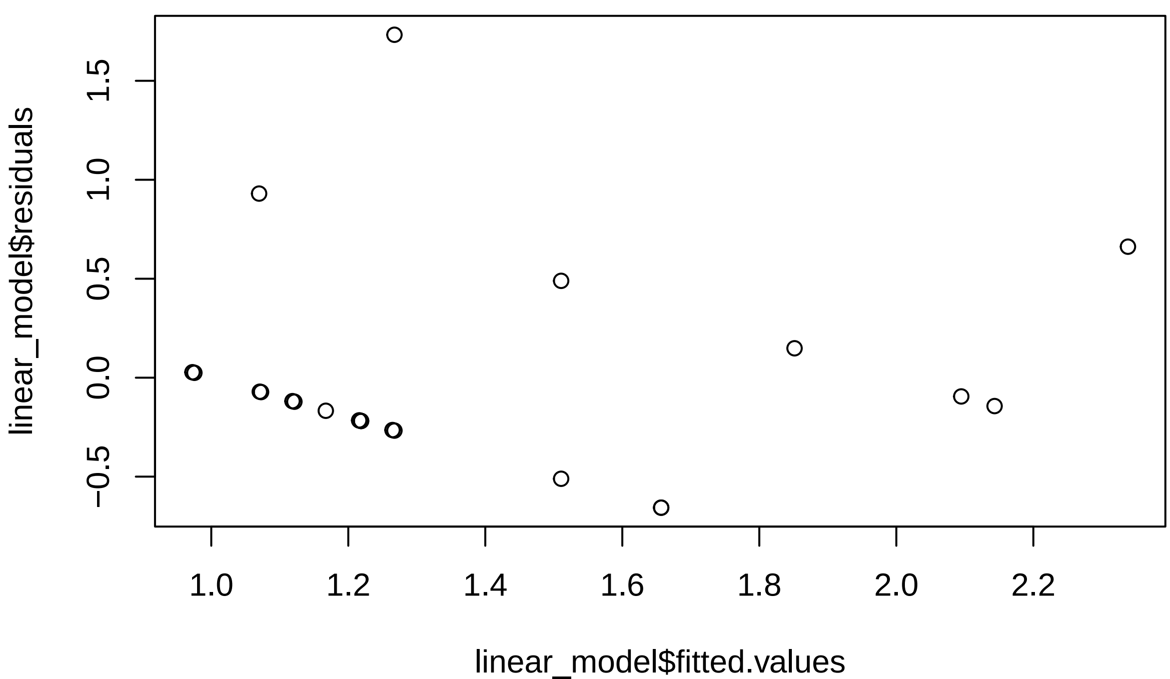
3.4 Alternative Specification (10 points)

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.

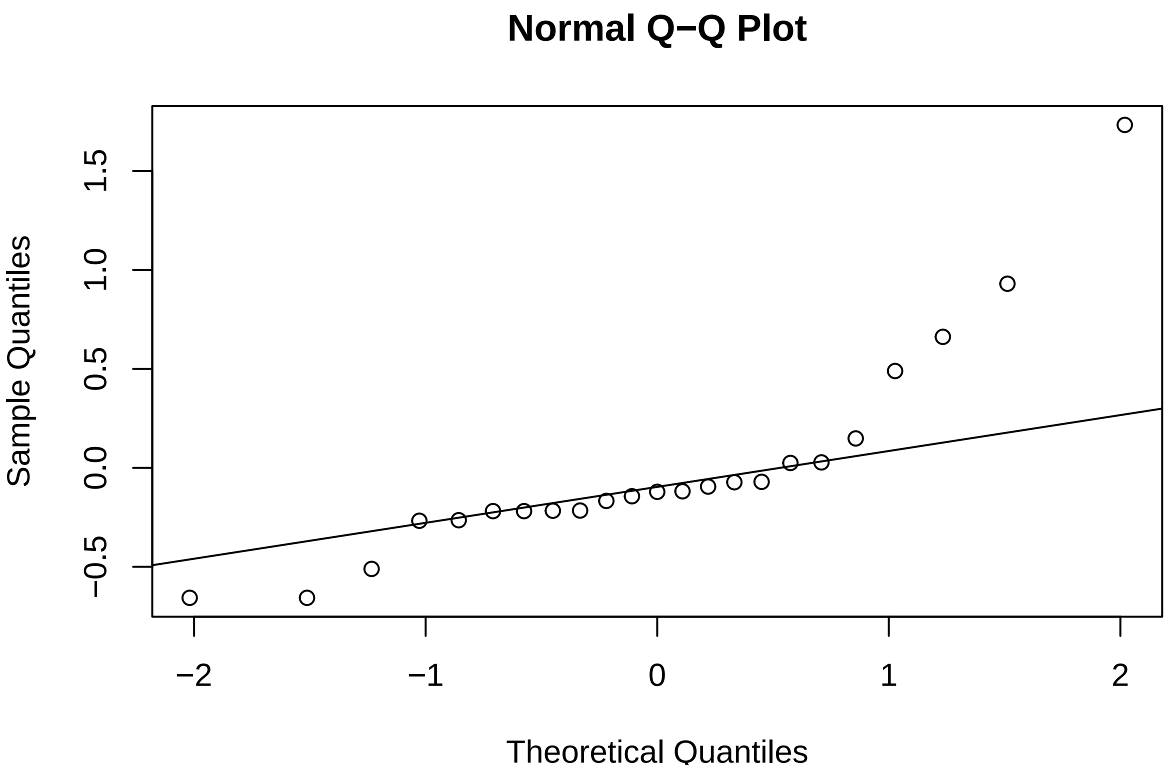
```r
linear_model <- lm(as.numeric(O.ring) ~ Temp + Pressure, data = data)
summary(linear_model)
```

```
##
## Call:
## lm(formula = as.numeric(O.ring) ~ Temp + Pressure, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65673 -0.21869 -0.12135  0.02609  1.73264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.329831   1.188132   3.644  0.00161 **
## Temp        -0.048671   0.016724  -2.910  0.00865 **
## Pressure     0.002939   0.001730   1.699  0.10481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5531 on 20 degrees of freedom
## Multiple R-squared:  0.3544, Adjusted R-squared:  0.2899
## F-statistic:  5.49 on 2 and 20 DF,  p-value: 0.01257
```

```r
plot(linear_model$fitted.values, linear_model$residuals)
```



```r
qqnorm(linear_model$residuals)
qqline(linear_model$residuals)
```



In this case, since the dependent variable is binary (0 or 1), it might not be appropriate to use a linear regression model. Linear regression may predict values outside the range [0,1] which doesn't make sense for a binary outcome. Instead, logistic regression, which predicts the log-odds of the outcome, might be more suitable.

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.