# Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 1

Jan Forslow, Kari Ross, Jason Xie

February 12th, 2017

## Report Purpose

This report provides analysis of the data available before the 1989 Space Shuttle Challenger Accident and creates models on the available data to predicit the liklihood of an O-ring failure.

The analysis includes:

1. A comphrehensive Exploratory Data Analysis with a focus on
   - Detecting anomalies in the data, including missing values
   - The potential of top and/or bottom code
   - Explanation of the EDA findings
2. Answering question 4 in Chapter 2 of Bilder and Loughin's "Analysis of Categorical Data with R", which includes exploration of the Logicstic Regression as performed by Dala et al (1989) paper including
   - Discussion of authors assumptions in LRT
   - Estimate of Logistic Regression Model using explanatory variables in a linear form
   - Model analysis and subsequent judgement of the explanatory variables in the LRT
   - Discussion and issues of author(s) removal of Pressure variable and any problems with removal
3. Answering question 5 in Chapter 2 of Bilder and Loughin's "Analysis of Categorical Data with R", which involves using a simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$ where $\pi$ is the probablity of an O-ring failure in order to:
   - Estimate the simplified model
   - Construct plots of $\pi$ versus *Temperature*, using temperature ranges of $31^o$ to $81^o$ as well as Expected Failures versus Temperature
   - Construct the 95% Wald Confidence interval bands for $\pi$
   - Estimate the probabilty of an O-ring failure at $31^o$ with the corresponding confidence interval. The assumptions of this inference procedure are discussed.
   - Compute the 90% confidence intervals using a bootstrap method, for the temperatures of $31^o$ and $81^o$
   - Explore if a quadratic term is needed in the model for temperatures

4. Investigates other models besides the logit moce, including probit and complementary log-log regression.
5. Compares the models of logit, probit, and log-log regression and then selects a Final Model.
6. Plots the main effect with the y-axis being the probabilty of failure and the x-axis being *temperature*

## Section 1: EDA of Challenger Data

As we are considering to use a binomial probability distribution for the model, it is of particular interest to validate the conditions required for using a binomial probablity model, which are:

1. There are n identical trials.
2. Each trial has two possible outcomes, typically referred to as a success or failure.
3. The trials are independent of each other.
4. The probability of success, denoted by $\pi$, remains constant for each trial. The probability of a failure is 1-$\pi$. In the case of the Challenger data, a success is an incident of an O-ring failure.
5. The random variable, w, represents the number of successes. Not interested in the order in which they occur.

### Examine the Validity of the Data

This section examines the structure and the integrity of the data (including the number of observations, number of variables, types of the variables, number of missing values (or oddly coded values) in each of the variables, descriptive statistics of each of the variables, etc).

The challenger.csv file holds 23 observations of 5 variables, two of which are categorical (Pressure and O.ring) that are stored as integer type (instead of as factors).

Flight: Flight number is an integer and starts with 1 and increments up to 23.

Temp: Temperature (F) at launch is an integer with range 53 to 81 and median 70 (and mean 69.57 is almost the same). The occurrences are fairly evenly distributed across the range with a somwhat negative skew.

Pressure: Combustion pressure (psi) in an integer with range 50 to 200 and median 200 and mean 152.2, which means that we have a negatively skewed distribution. There is only three values in sample: 50 (6 occurences), 100 (only 2 occurences) and 200 (15 occurences). We will therefore consider to bin the two 100 psi occurences with 50 psi in a model later on.

O.ring: Number of primary field O-ring failures is an integer 0 to 2. Add one for each failure meaning theoretical max is actually 6, but from the data it varies only from 0 (16

occurences), 1 (5 occurences) and 2 (2 occurences). Mean is 0.3913, while median is 0.000, which means that we have a positively skewed distribution.

Number: Total number of primary field O-rings (six total, three each for the two booster rockets). All rockets have 6 primary field o-rings in the sample. This variable is as such a constant for the sample. Can be used as n for probability assessments.

According to the describe() function, there are no missing values and no values that are beyond the boundaries of the variables.

Our analysis is provided below.

```
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 55), tidy = TRUE)

rm(list = ls())

# Load libraries
require(stargazer)
require(sandwich)
require(lmtest)
require(car)
require(Hmisc)
require(psych)
require(ggplot2)
require(mcprofile)
require(gmodels)
require(rgl)
require(effects)

# Set working directory wd <- 'C:/Users/Jan
# Forslow/Documents/Berkeley/W271_Statistical_Methods/W271_2017Spring_Lab1'
# wd <- '~/Downloads/W271-Lab1/data/'
wd <- "C:/Users/Kari/source_code/W271-Lab1/data"
setwd(wd)

# Load the data and save it as a data frame in our
# current work space
df <- read.csv("challenger.csv", stringsAsFactors = F)

# Examine the structure of the data A basic description
# of the variables is provided in Bilder and Loughlin
# page 129 and is extended here.

str(df)

## 'data.frame':    23 obs. of  5 variables:
##  $ Flight  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Temp    : int  66 70 69 68 67 72 73 70 57 63 ...
##  $ Pressure: int  50 50 50 50 50 50 100 100 200 200 ...
```

```
##  $ O.ring  : int  0 1 0 0 0 0 0 0 1 1 ...
##  $ Number  : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
summary(df)
```

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

```
# For categorical variables it doesn't make a lot of
# sense to display the median and other quantiles.  For
# categorical variables, we use frequency tables to
# describe them. Using the describe() function we can
# see if any missing values.
```

```
library("psych")  #Added as required for describe()
describe(df)
```

```
##          vars  n   mean    sd median trimmed  mad min max range  skew
## Flight      1 23  12.00  6.78     12   12.00 8.90   1  23    22  0.00
## Temp        2 23  69.57  7.06     70   70.00 5.93  53  81    28 -0.57
## Pressure    3 23 152.17 68.22    200  157.89 0.00  50 200   150 -0.69
## O.ring      4 23   0.39  0.66      0    0.26 0.00   0   2     2  1.31
## Number      5 23   6.00  0.00      6    6.00 0.00   6   6     0   NaN
##          kurtosis    se
## Flight      -1.36  1.41
## Temp        -0.27  1.47
## Pressure    -1.50 14.23
## O.ring       0.39  0.14
## Number        NaN  0.00
```

```
table(df$Flight)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

```
table(df$Temp)
```

```
##
## 53 57 58 63 66 67 68 69 70 72 73 75 76 78 79 81
##  1  1  1  1  1  3  1  1  4  1  1  2  2  1  1  1
```

```
table(df$Pressure)
```

```
##
##  50 100 200
##   6   2  15
```

```
table(df$O.ring)
```

```
##
##  0  1  2
## 16  5  2
```

```
table(df$Number)
```

```
##
##  6
## 23
```

## Conduct EDA of Explanatory Variables

a. Examine the distribution for the explanatory variables *Temp* and *Pressure*.

b. Examine the relationship between *Temp* and *Pressure*.

We will first summarize our analysis and then provide the actually analysis following the summary.

The stem-and-leaf graph shows that we have a positive curtosis and negative skew for the Temp score distributon. This is confirmed also in the histogram. The Shapiro-Wilk normality test provides a p-value of 0.34 > 0.05 and as such we cannot reject the null hypothesis that the distribution for temperature is in the form of a normal distribution. This indicates that we would not need to do transformations.

For the Pressure variable, we have only three categorical values present with six samples for 50 psi, two samples for 100 psi and fifteen samples for 200 psi. The histogram as such shows a quite non-linear graph with heavy negative skew and the normal distribution test indicates that we can reject the null hypothesis that the Pressure scores are normally distributed.

A scatter plot of Temperature on x-axis vs. Pressure on y-Axis shows that for the sample that low pressure measurements coinsided with medium temperature range. We do a Pearson's R test for independence and that shows weak correlation as less than 0.1 [Should be at least 0.5 or greater for strong correlation]. We also have a p-value = 0.857 > 0.05 so we cannot reject the null hypothesis that the two variables are independent. We factorize Pressure given that we in our sample only have three levels. The associated boxplot shows that we have very few scores (2) in the middle range of Pressure (100 psi) and that we

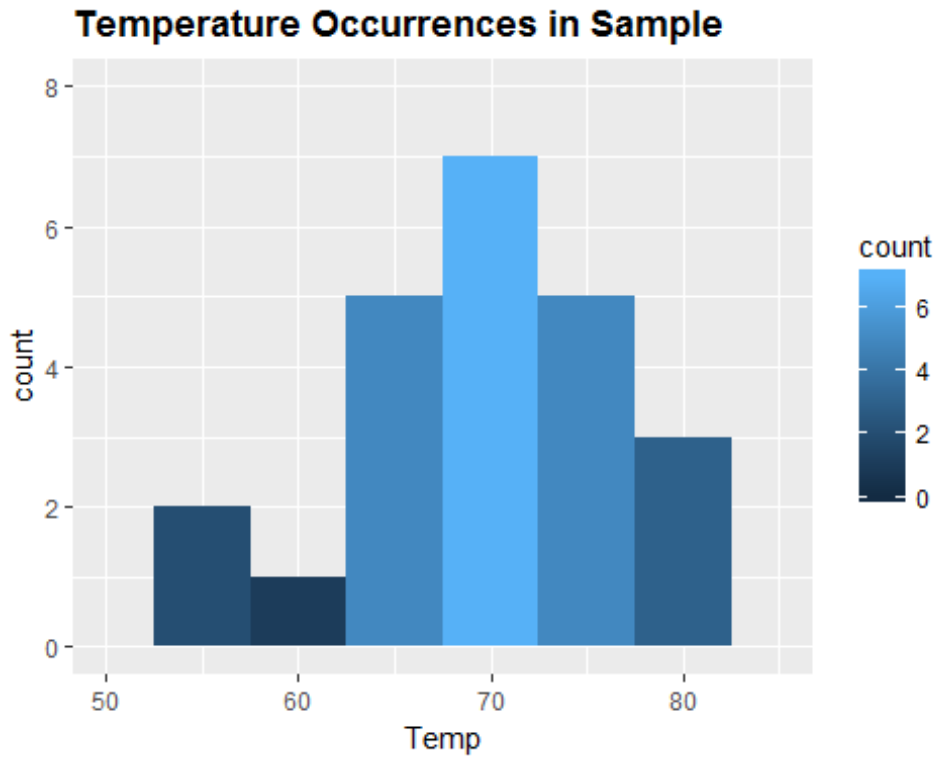
have a wide spread for Pressure = 200 psi (where we also have most data points). This also indicates small dependency between the two explanatory variables.

The scatterplots of Temperature and Pressure over time (Flight) shows that the Pressure was gradually changed from 50 psi for the first flights to 100 psi for two Flights and then to 200 psi for all the flights. This means that unless we take Pressure into account in the model, the condition that each trial is identical does not hold unless Pressure has no impact on O-ring failure. We will consider this further in our modelling. Temperature, on the other hand, does not show any systematic or linear relationship to Flight (time) and as such the trials looks independent of each other from Temp perspective.

```
# Analysis of the Temp variable

summStat <- function(x = Temp, y = first.quartile, z = fourth.quartile) {
    descStat <- describe(x)
    tbl1 <- table(x[x <= y])  # First quartile
    tbl2 <- table(x[x >= z])  # Fourth quartile
    count1 <- length(x[x <= y])
    count2 <- length(x[x >= z])
    result <- list(result1 = descStat, result2 = tbl1, result3 = tbl2,
        result4 = count1, result5 = count2)
    return(result)
}

summStat(df$Temp, 67, 75)

## $result1
##    vars  n  mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 23 69.57 7.06     70      70 5.93  53  81    28 -0.57    -0.27
##      se
## X1 1.47
##
## $result2
##
## 53 57 58 63 66 67
##  1  1  1  1  1  3
##
## $result3
##
## 75 76 78 79 81
##  2  2  1  1  1
##
## $result4
## [1] 8
##
## $result5
## [1] 7

# A traditional stem-and-leaf graph;
stem(df$Temp)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 378
##   6 | 3677789
##   7 | 000023556689
##   8 | 1
```

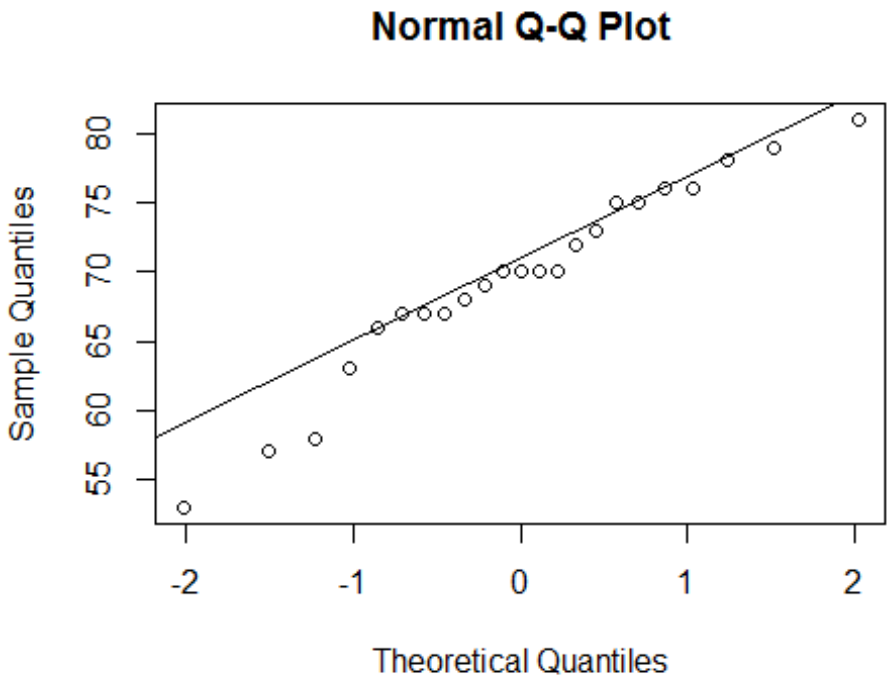
```r
# Histogram for Temperature scores:
library("ggplot2")  # Added for plots
ggplot(df, aes(x = Temp, fill = ..count..)) + xlim(50, 85) +
    ylim(0, 8) + ggtitle("Temperature Occurrences in Sample") +
    theme(plot.title = element_text(lineheight = 1, face = "bold")) +
    geom_histogram(binwidth = 5)
```




```r
# Another way to review binning:
Temp_bin <- cut(df$Temp, c(50, 55, 60, 65, 70, 75, 80, 85))
table(Temp_bin)
```

```
## Temp_bin
## (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85]
##       1       2       1      10       4       4       1
```

```r
# Examining the normality of the distribution
qqnorm(df$Temp)
qqline(df$Temp)
```



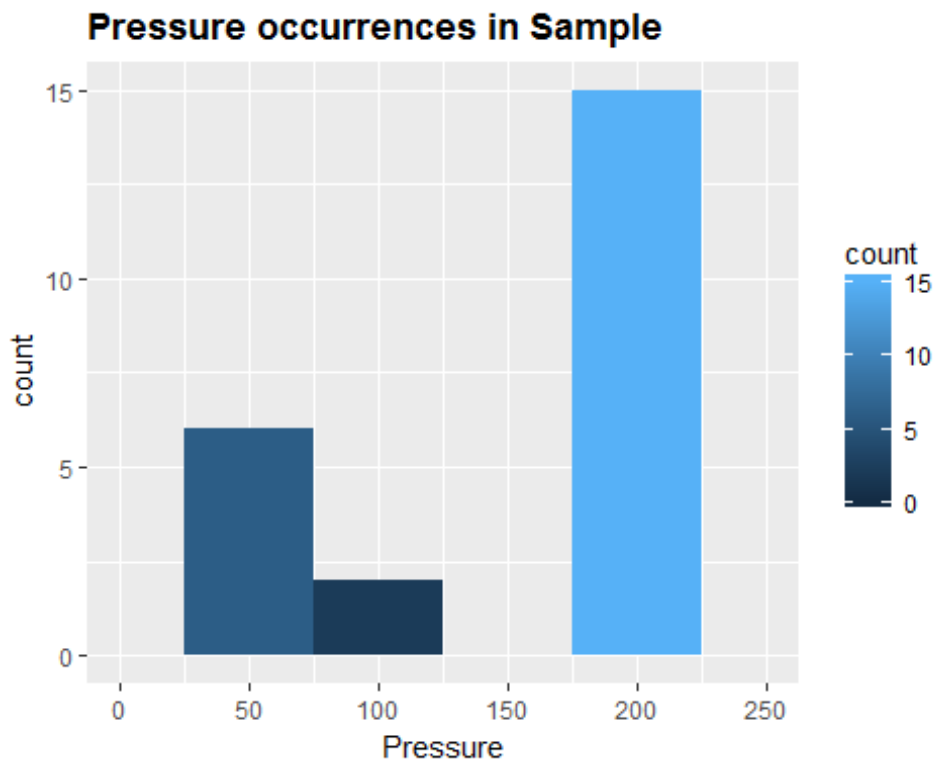

```
shapiro.test(df$Temp)

##
##  Shapiro-Wilk normality test
##
## data:  df$Temp
## W = 0.95368, p-value = 0.3482

# Analysis of the Pressure variable in the same way
summStat(df$Pressure, 75, 200)

## $result1
##    vars  n   mean    sd median trimmed mad min max range  skew kurtosis
## X1    1 23 152.17 68.22    200  157.89   0  50 200   150 -0.69     -1.5
##       se
## X1 14.23
##
## $result2
##
## 50
##  6
##
## $result3
##
## 200
##  15
##
```

```
## $result4
## [1] 6
##
## $result5
## [1] 15
```

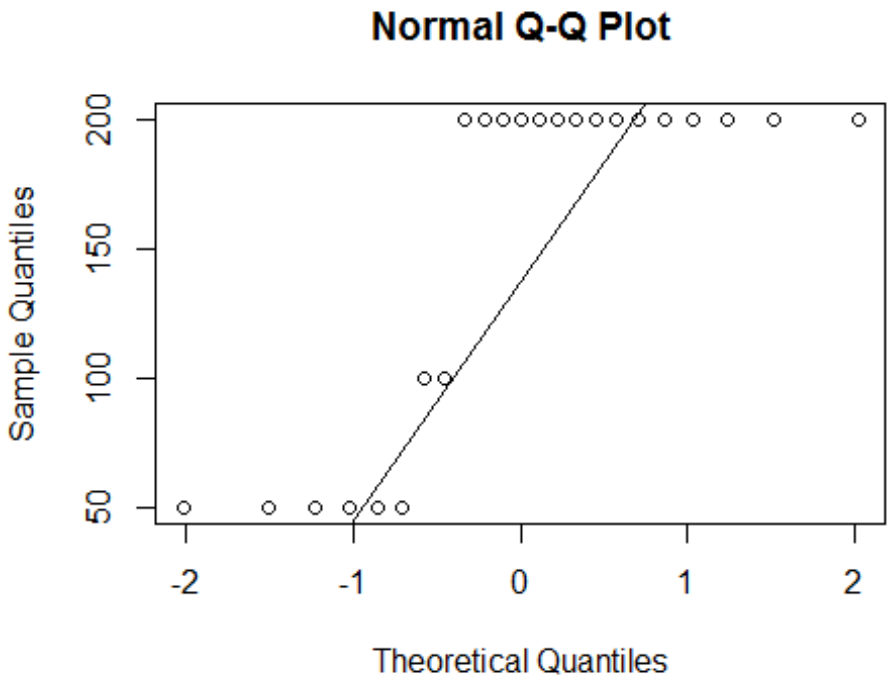
```
# A traditional stem-and-leaf graph
stem(df$Pressure)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 555555
##   1 | 00
##   1 |
##   2 | 000000000000000
```

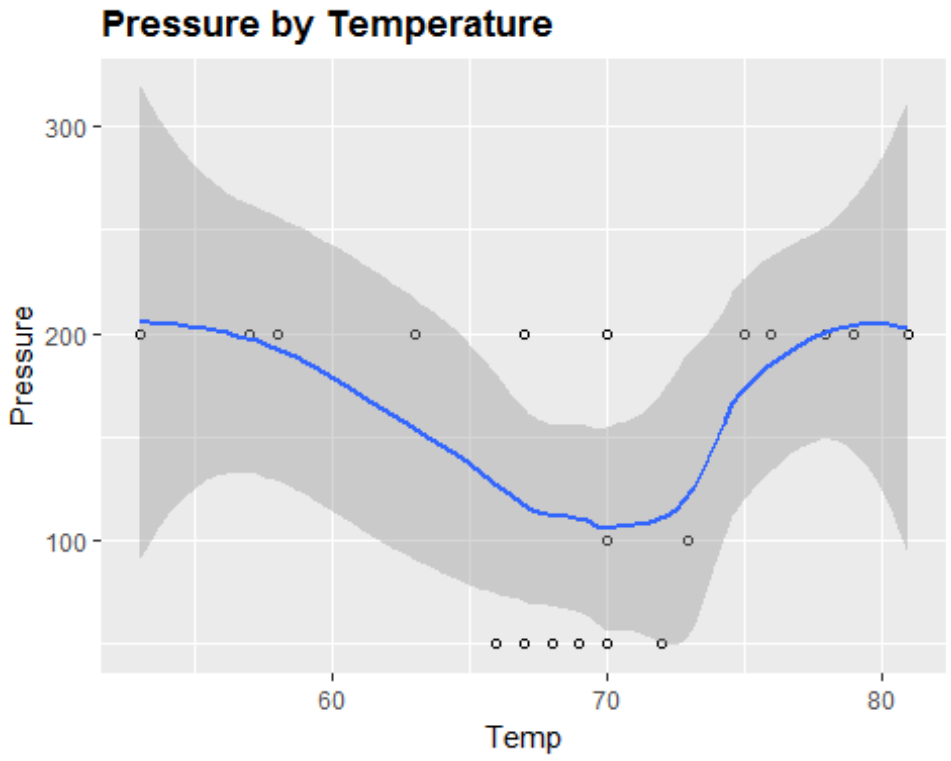
```
# Histogram for Pressure scores
ggplot(df, aes(x = Pressure, fill = ..count..)) + xlim(0,
    250) + ylim(0, 15) + ggtitle("Pressure occurrences in Sample") +
    theme(plot.title = element_text(lineheight = 1, face = "bold")) +
    geom_histogram(binwidth = 50)
```



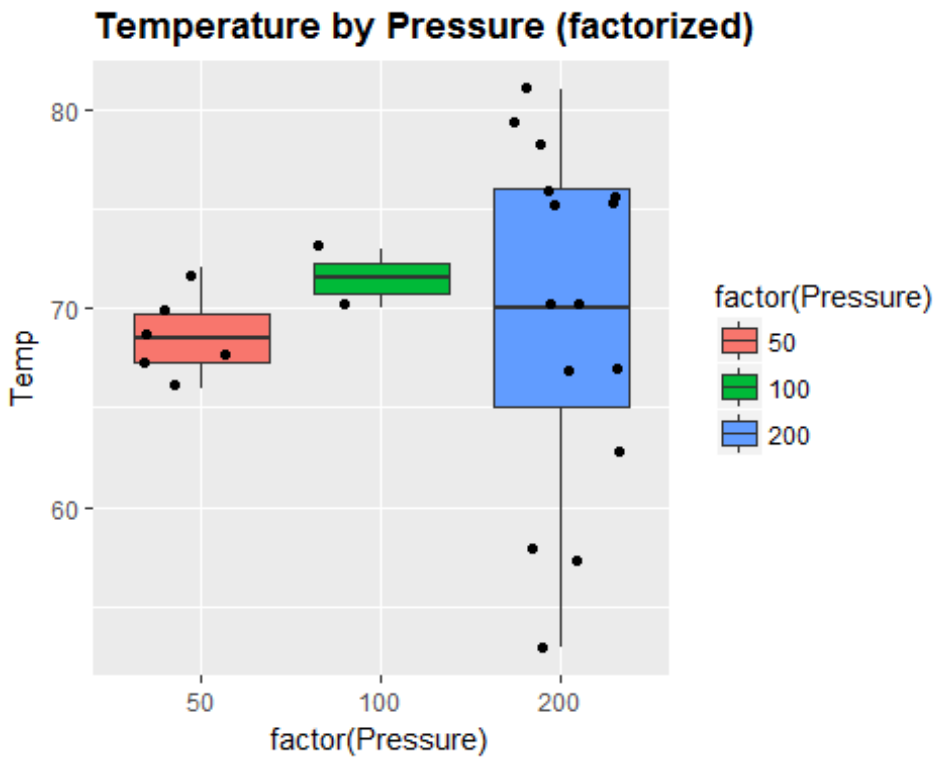

```
# Examining the normality of the distribution
qqnorm(df$Pressure)

qqline(df$Pressure)
```



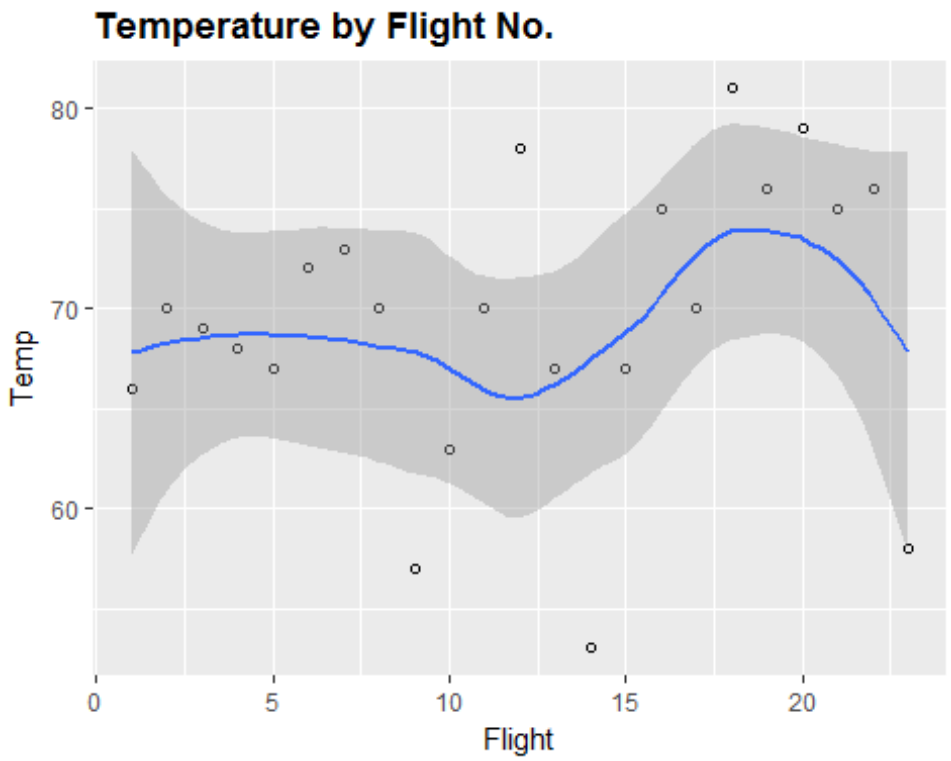

```
shapiro.test(df$Pressure)

##
##  Shapiro-Wilk normality test
##
## data:  df$Pressure
## W = 0.63952, p-value = 2.526e-06

# Scatterplot: Pressure by Temperature
ggplot(df, aes(x = Temp, y = Pressure)) + geom_point(shape = 1) +
    ggtitle("Pressure by Temperature") + theme(plot.title =
element_text(lineheight = 1,
    face = "bold")) + geom_smooth()
```



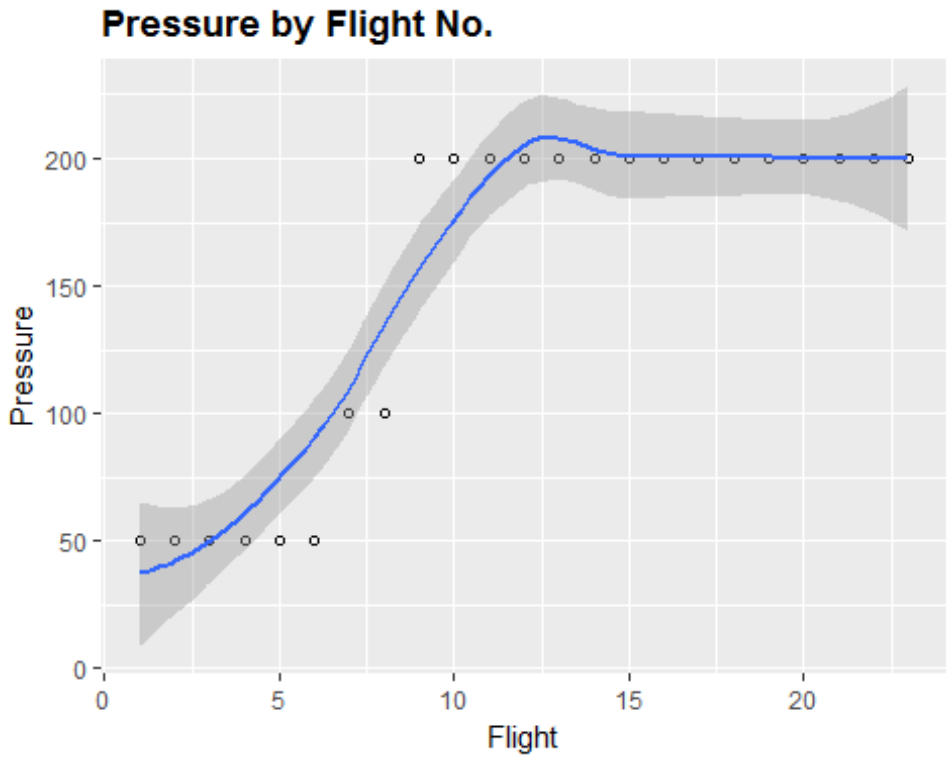

```
# Test for independence using Pearson's R:
cor.test(df$Temp, df$Pressure)

##
##  Pearson's product-moment correlation
##
## data:  df$Temp and df$Pressure
## t = 0.18261, df = 21, p-value = 0.8569
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3785984  0.4447207
## sample estimates:
##       cor
## 0.03981769

# Boxplot: Temperature by Pressure (factor)
ggplot(df, aes(x = factor(Pressure), y = Temp)) + geom_boxplot(aes(fill =
factor(Pressure))) +
    geom_jitter() + ggtitle("Temperature by Pressure (factorized)") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
# Scatterplot: Temperature by Flight No. (over time)
ggplot(df, aes(x = Flight, y = Temp)) + geom_point(shape = 1) +
    ggtitle("Temperature by Flight No.") + theme(plot.title =
element_text(lineheight = 1,
    face = "bold")) + geom_smooth()
```




```
# Scatterplot: Pressure by Flight No. (over time)
ggplot(df, aes(x = Flight, y = Pressure)) + geom_point(shape = 1) +
    ggtitle("Pressure by Flight No.") + theme(plot.title =
element_text(lineheight = 1,
    face = "bold")) + geom_smooth()
```




## Conduct EDA of Response Variable

We visually inspect and comment on the distribution of the variable, O.ring. The O.ring variable has strong positive skew with most scores for zero (no failure). We have only two occurences of two or more failures for a primary field o-ring and as such, it seems from this perspective reasonable to bin these with the one failure occurences into a binary response variable. We then satisfy the condition for binomial mode and probability distribution of only two outcomes (success and failure).

When attempting a t-test for the two groups of no-failure (0) and at least one failure (1) vs. Temperature, we get a p-value = 0.035 < 0.05 and a 95% confidence interval above zero. As such we can reject the null hypothesis that the means for the two groups are equal, which is a good starting point in creating a binary regression model with this parameter.

The same test for Pressure, does not give any significance though and we will need more analysis during the model state to see if Pressure is valuable as an explanatory variable.

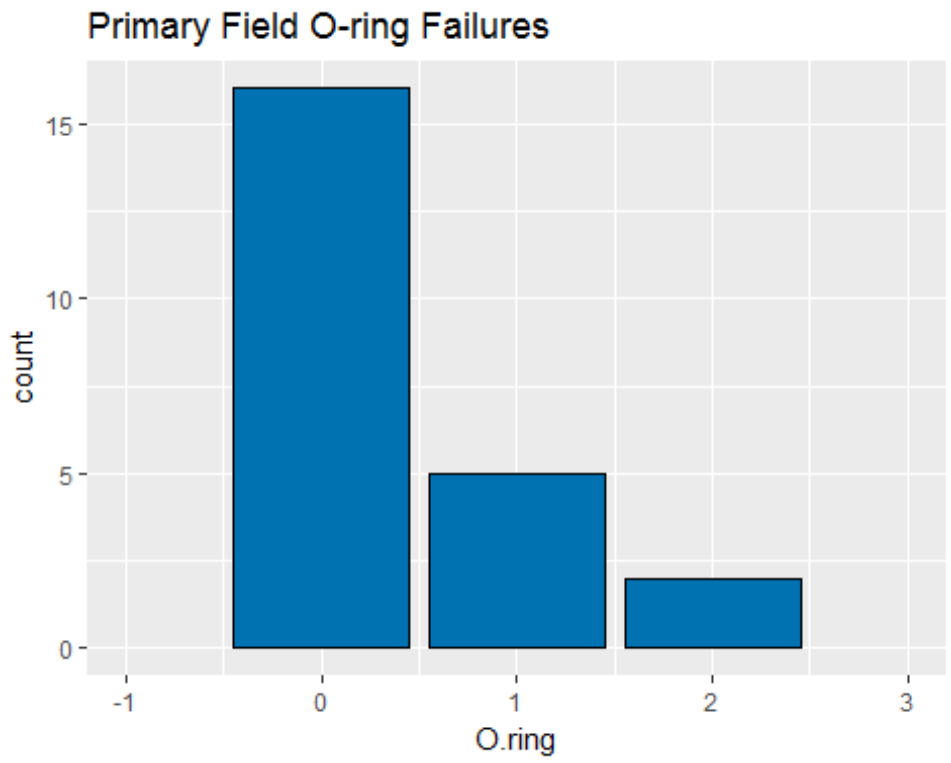
The scatterplot of primary field o-ring failures over time (Flight), does not show any systematic or linear behavior. It is an early indication that even if Pressure changed for the latter Flights it did not look to have a strong influence on o-ring failures. To make sure, we perform also a t-test for group with and without at least one o-ring failure. The p-value for this t-test is 0.7 (> 0.05) and the 95% confidence interval is -8.33 to 5.87 (spans zero) and as such we cannot reject the null hypothesis that the means are equal. We also review in a boxplot Flight with O-ring failures as factor and this shows also no systematic relationship (quite a wide spread). We continue with the assumption that the individual Flight

observations are independent of each other. For our continued analysis, we do not care of the order in which the successes and failures occur across the Flight trials.

The scatterplot of primary O-ring failure to Temperature does show a downward trend for O.ring failures as the Temperature increases. However, there is an outlier that raises the curve at (Temp = 75, O.ring = 2). The associated boxplot with number of O-ring failures as factor shows that we have higher Temperature mean and lower spread between first and fourth quartile for no O-ring failures than in case of one or two O-ring failures. Again, this indicates that Temperature should be in the model.

A boxplot of Pressure by O-ring failures as a factor shows that we have a large spread for O-ring failures = 0, while the medians are the same (200 psi), which is natural as we have most observations in this region. Reviewing as a table instead as both variables can be considered categorical variables. This shows more clearly that we have the main number of O-ring failures (1x4 + 2x2 = 8) and also total observations (15) at 200 psi. Other Pressure levels have very little information (only 1 O-ring failure in 8 observations). This indicates that the sample in itself does not have data to help indicate strong dependency between O-ring failures and Pressure. Still we will take Pressure into account in at least one model creation to verify this further as we have seen the gradual change in Pressure over Flight Numbers from before. A Pearson's Chisquare dependency test gives p-value = 0.67 > 0.05, and as small sample we also review Fisher's test with p = 1, which means that we cannot reject the null hypothesis that they are independent.

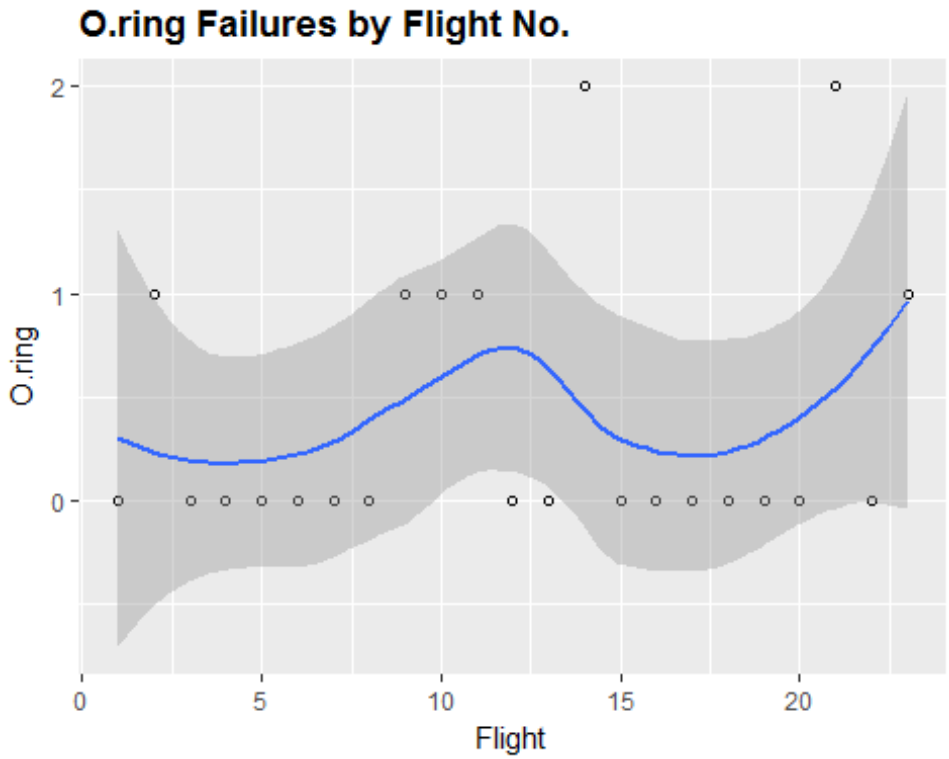
```
describe(df$O.ring)

##    vars  n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 23 0.39 0.66      0    0.26   0   0   2     2 1.31     0.39 0.14

# Another way to display a frequency table
table(df$O.ring, useNA = "always")

##
##    0    1    2 <NA>
##   16    5    2    0

# Distribution of O.ring
ggplot(df, aes(x = O.ring)) + xlim(-1, 3) + geom_bar(stat = "count",
    fill = "#0072B2", colour = "black") + ggtitle("Primary Field O-ring
Failures")
```



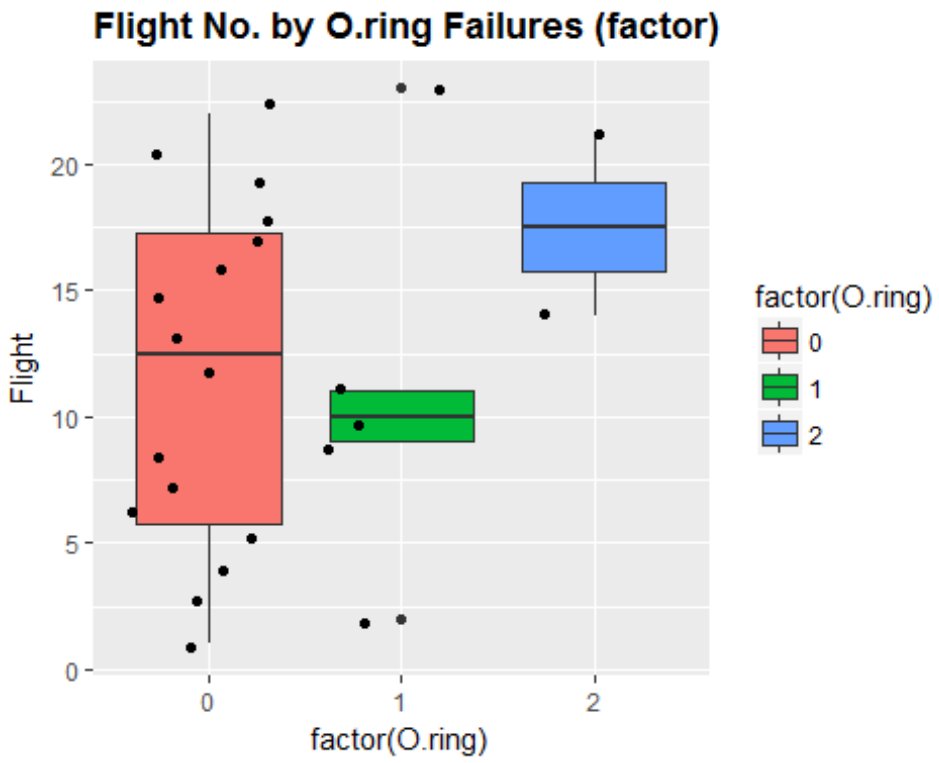

```
# Create a binary variable out of o-ring failure (at
# least one):
df$O.ring.failure <- ifelse(df$O.ring == 0, 0, 1)
df$O.ring.fail.factor <- factor(df$O.ring.failure)
# Perform t-test for the two groups of zero and at least
# one primary field o-ring failure.  FIXME disuss
# results of the t-tests
t.test(df$Temp ~ df$O.ring.failure)

##
##  Welch Two Sample t-test
##
## data:  df$Temp by df$O.ring.failure
## t = 2.5387, df = 7.9166, p-value = 0.03507
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.7569431 16.0644855
## sample estimates:
## mean in group 0 mean in group 1
##        72.12500        63.71429

t.test(df$Pressure ~ df$O.ring.failure)

##
##  Welch Two Sample t-test
##
## data:  df$Pressure by df$O.ring.failure
```

```
## t = -1.362, df = 14.4, p-value = 0.1941
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -97.54714  21.65428
## sample estimates:
## mean in group 0 mean in group 1
##        140.6250        178.5714
```

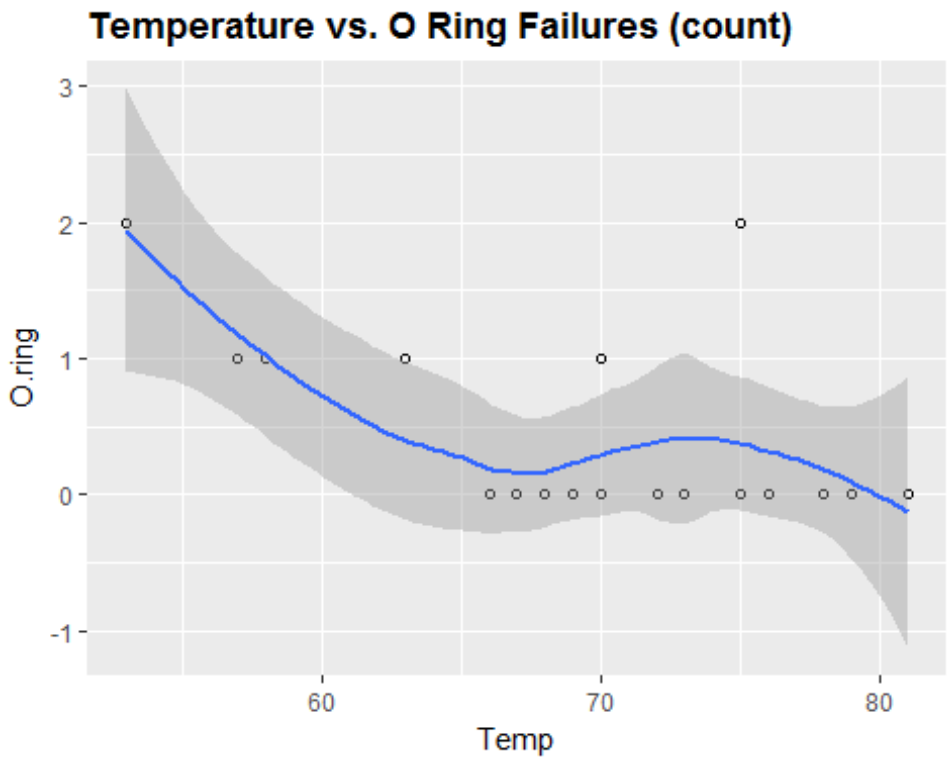
```r
# Scatterplot: O.ring Failure by Flight No. (over time)
ggplot(df, aes(x = Flight, y = O.ring)) + geom_point(shape = 1) +
    ggtitle("O.ring Failures by Flight No.") + theme(plot.title =
element_text(lineheight = 1,
    face = "bold")) + geom_smooth()
```



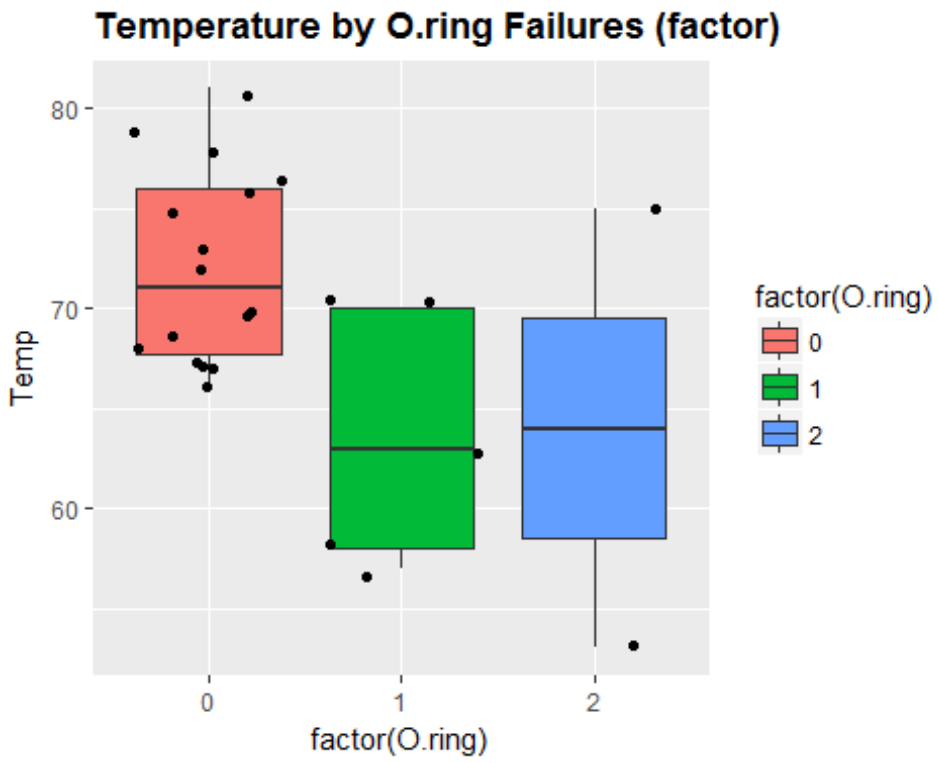

```r
# Boxplot Flight No. by O.ring Failures (factor)
ggplot(df, aes(x = factor(O.ring), y = Flight)) + geom_boxplot(aes(fill =
factor(O.ring))) +
    geom_jitter() + ggtitle("Flight No. by O.ring Failures (factor)") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



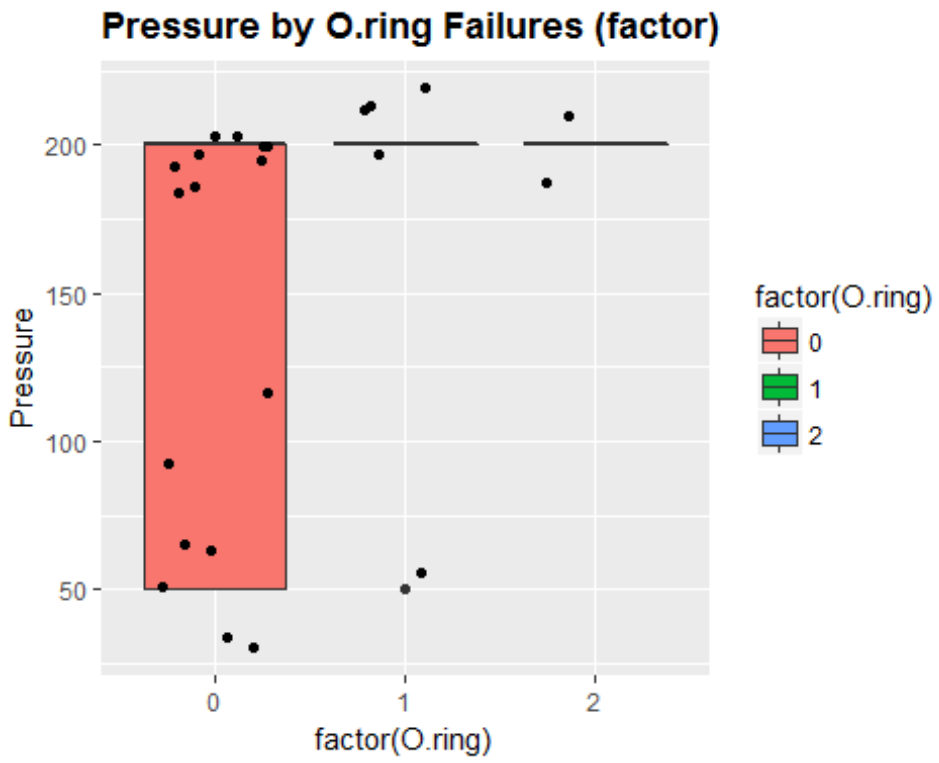

```
# Perform t-test for the two groups of zero and at least
# one primary field o-ring failure related to Flight
# number (time).
t.test(df$Flight ~ df$O.ring.failure)

##
##  Welch Two Sample t-test
##
## data:  df$Flight by df$O.ring.failure
## t = -0.38261, df = 10.845, p-value = 0.7094
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.332455  5.868170
## sample estimates:
## mean in group 0 mean in group 1
##        11.62500        12.85714

# Scatterplot: O.ring Failures by Temperature
ggplot(df, aes(x = Temp, y = O.ring)) + geom_point(shape = 1) +
    ggtitle("Temperature vs. O Ring Failures (count)") +
    theme(plot.title = element_text(lineheight = 1, face = "bold")) +
    geom_smooth()
```




```r
# Boxplot: Temperature by O.ring Failures (factor)
ggplot(df, aes(x = factor(O.ring), y = Temp)) + geom_boxplot(aes(fill =
factor(O.ring))) +
    geom_jitter() + ggtitle("Temperature by O.ring Failures (factor)") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
# Boxplot: Pressure by O.ring Failures (factor)
ggplot(df, aes(x = factor(O.ring), y = Pressure)) + geom_boxplot(aes(fill =
factor(O.ring))) +
    geom_jitter() + ggtitle("Pressure by O.ring Failures (factor)") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
# Table: Pressure (factor) by O.ring Failures (factor)
tab <- table(df$Pressure, df$O.ring)
addmargins(tab)
```

```
##
##         0  1  2 Sum
##   50    5  1  0   6
##   100   2  0  0   2
##   200   9  4  2  15
##   Sum 16  5  2  23
```

```
# Test dependency between categorical variables using
# Pearson's Chisquare test and also Fisher's exact test
# as small sample.  install.packages('gmodels')
library(gmodels)
CrossTable(df$Pressure, df$O.ring, fisher = TRUE, chisq = TRUE,
    expected = TRUE, sresid = TRUE, format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
```

```
## |             Std Residual |
## |--------------------------|
##
## Total Observations in Table:  23
##
##              | df$O.ring
##  df$Pressure |         0  |         1  |         2  | Row Total  |
## -------------|------------|------------|------------|------------|
##           50 |         5  |         1  |         0  |          6  |
##              |     4.174  |     1.304  |     0.522  |             |
##              |     0.163  |     0.071  |     0.522  |             |
##              |    83.333% |    16.667% |     0.000% |    26.087% |
##              |    31.250% |    20.000% |     0.000% |             |
##              |    21.739% |     4.348% |     0.000% |             |
##              |     0.404  |    -0.266  |    -0.722  |             |
## -------------|------------|------------|------------|------------|
##          100 |         2  |         0  |         0  |          2  |
##              |     1.391  |     0.435  |     0.174  |             |
##              |     0.266  |     0.435  |     0.174  |             |
##              |   100.000% |     0.000% |     0.000% |     8.696% |
##              |    12.500% |     0.000% |     0.000% |             |
##              |     8.696% |     0.000% |     0.000% |             |
##              |     0.516  |    -0.659  |    -0.417  |             |
## -------------|------------|------------|------------|------------|
##          200 |         9  |         4  |         2  |         15  |
##              |    10.435  |     3.261  |     1.304  |             |
##              |     0.197  |     0.168  |     0.371  |             |
##              |    60.000% |    26.667% |    13.333% |    65.217% |
##              |    56.250% |    80.000% |   100.000% |             |
##              |    39.130% |    17.391% |     8.696% |             |
##              |    -0.444  |     0.409  |     0.609  |             |
## -------------|------------|------------|------------|------------|
## Column Total |        16  |         5  |         2  |         23  |
##              |    69.565% |    21.739% |     8.696% |             |
## -------------|------------|------------|------------|------------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  2.367083     d.f. =  4     p =  0.6685838
##
##
##
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Alternative hypothesis: two.sided
## p =  1
```

```
##
##
##         Minimum expected frequency: 0.173913
## Cells with Expected Frequency < 5: 8 of 9 (88.88889%)
```

## Section 2: Answers to Question 4 in Chapter 2 of Bilder and Loughin's "Analysis of Categorical Data with R"

### Question 4a: Author's Assumptions in using a logistic regression to estimate the probability of an O-ring failure

The models assumes that at temperature and pressure of each of the six O-rings would suffer damage independently with the same probability. The logistical model employed. The logistical model relaxes several assumptions, including that there are no distributional assumptions about the predictor variables.

### Question 4b: Estimate the logistic regression model using the explanatory variables in a linear form.

A Logistical Regression model was deemed an appropiate model because a binary outcome (O-ring failure) is being predicted from a set of continuous variable, Temperature and Pressure. The Logistical regression has the following form:

$$logit(\pi) = \beta_0 + \beta_1 Temp + \beta_2 Pressure$$

where $\pi$ is the conditional mean of Failure (the probability that Failure = 1 given a set of Pressure and Temperature Values.)

The original data set had a numerical value for the O.ring failure per shuttle launch. This O.ring value was converted to a dichotomous variable with a zero to be "No Failure" and anything greater than 0, to be a "Failure"

After the O.ring.failure counts have been transformed to a dichotomous variable, the O.ring.fail.factor was then able to be used as the Outcome Variable in a logistical regression model.

```
df$O.ring.failure <- ifelse(df$O.ring == 0, 0, 1)
df$O.ring.fail.factor <- factor(df$O.ring.failure, levels = c(0,
    1), labels = c("No Failure", "Failure"))
table(df$O.ring.fail.factor)
```

```
##
## No Failure    Failure
##         16          7
```

```
glm_temp_pressure <- glm(formula = O.ring.fail.factor ~
    Temp + Pressure, family = binomial(link = logit), data = df)
round(coef(glm_temp_pressure), 3)

## (Intercept)        Temp    Pressure
##      13.292      -0.229       0.010

summary(glm_temp_pressure)

##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp + Pressure, family = binomial(link
= logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1993  -0.5778  -0.4247   0.3523   2.1449
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.292360   7.663968   1.734   0.0828 .
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## Pressure     0.010400   0.008979   1.158   0.2468
## ---
## 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: 18.782  on 20  degrees of freedom
## AIC: 24.782
##
## Number of Fisher Scoring iterations: 5
```

The resulting prediction equation is:

$$logit(\pi) = 13.292 - 0.229Temp + 0.010Pressure$$

## Question 4c: Perform LRTs to judge the importance of Pressure and Temperature as Predictors

An Anova function of Type 2 was used to generate and then examine Likelihood ratio test (LRT)

```
Anova(glm_temp_pressure, test = "LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring.fail.factor
##          LR Chisq Df Pr(>Chisq)
```

```
## Temp        7.7542  1   0.005359 **
## Pressure    1.5331  1   0.215648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The function Anova() computes a type 2 (partial) test, where each null-hypothesis model consists of all the other variables listed in the right side of the equation: in this case, Pressure and Temperature.

LRTs and the Anova() tests was used to test the hypothesis of Temperature with

$$H_0: \beta_1 = 0 \; vs. \; H_a: \beta_1 \neq 0$$

From the Anova() results, $-2log(\Lambda) = 7.754$ with a p-value of P( A > 7.7542) = 0.005359. Thus there is significant evidence that Temperature is important to include in the model.

Performing a similar analysis for Pressure,

$$H_0: \beta_2 = 0 \; vs. \; H_a: \beta_2 \neq 0$$

and from the Anova() results, $-2log(\Lambda) = 1.5331$ with a p-value of P( A > 1.5331) = 0.215648. This shows there is evidence to not reject the null hypothesis that $\beta_2 = 0$.

## Question 4d: The authors chose to remove Pressure from the model based on LRTs. A discussion on why this was done, based on our results, follows.

LRTs and the Anova() tests was used to test the hypothesis of Temperature with

$$H_0: \beta_1 = 0 \; vs. \; H_a: \beta_1 \neq 0$$

From the Anova() results, $-2log(\Lambda) = 7.754$ with a p-value of P( A > 7.7542) = 0.005359. Thus there is significant evidence that Temperature is important to include in the model.

Performing a similar analysis for Pressure,

$$H_0: \beta_2 = 0 \; vs. \; H_a: \beta_2 \neq 0$$

and from the Anova() results, $-2log(\Lambda) = 1.5331$ with a p-value of P( A > 1.5331) = 0.215648. This shows there is evidence to not reject the null hypothesis that $\beta_2 = 0$.

The result of Anova() test suggests that when Temperature and Pressure are both included in the model, that Pressure is not a significant predictor.

However, before removing Pressure, it is important to compare the models and use an Type I anova test of these models

$$Model \;\; 1: \;\; logit(\pi) = \beta_0 + \beta_1 Temp$$

$$Model \;\; 2: \;\; logit(\pi) = \beta_0 + \beta_1 Temp + \beta_2 Pressure$$

reduced model with only Temperature for the predictor fits as well as the model with Temperature and Pressure as predictors. Therefore it is reasonable to base the model off of the simplified model using only Temperature as a Predictor.

The anova() function in R performs a Type I test with Model 1 and Model 2. The value of p-value of the result is 0.2156, showing that Pressure does not add significance when compared to Model 1. Therefore Pressure may be removed from the model.

```
glm_temp <- glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link =
logit),
    data = df)
anova(glm_temp, glm_temp_pressure, test = "Chisq")
```

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

```
qchisq(0.95, df = 19)
```

```
## [1] 30.14353
```

## Section 3: Answering question 5 in Chapter 2 of Bilder and Loughin's "Analysis of Categorical Data with R" using a simplified logit model.

### Question 5a) Estimate the simplified model

The simplified model was created using a generalized linear model of the family binomial logit. As noted in the previous section, Pressure was dropped from the model so this leaves only Temperature as the explanatory variable. The simplified model has the following form:

$$Simplified\ Model\text{:}\ \ logit(\pi) = \beta_0 - \beta_1 Temp$$

The values for $\beta_0$ and $\beta_1$ were determined from the output of glm() function, as shown below.

```
glm_temp <- glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link =
logit),
    data = df)
summary(glm_temp)
```

```
##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link = logit),
##     data = df)
```

```
##
## 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 *
## ---
## 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
```

The model shows that Temperature is slightly statistically significant and produces the following model:

$$Simplified\ Model:\ \ logit(\pi) = 15.04 - 0.23Temp$$

Odds Ratio for Logistic Regression Model with Temperature only as Explanatory Variable. The odds of a failure is 1.261 times as large for a one degree decrease in Temperature.

```
exp(glm_temp$coefficients[2])

##      Temp
## 0.7928171

1/exp(glm_temp$coefficients[2])

##     Temp
## 1.261325
```

Confidence Interval for Logistic Regression Model with Temperature only as Explanatory Variable. The 95% profile likelihood ratio confidence interval for the Temperature coefficient is -0.515 < $\beta_1$ < -0.061. The profile likelihood ratio confidence ratio confidence interavl for the odds ratio is 1.063 < OR < 1.674.

```
beta.ci <- confint(object = glm_temp, parm = "Temp", level = 0.95)
beta.ci

##       2.5 %      97.5 %
## -0.51547175 -0.06082076

as.numeric(rev(1/exp(beta.ci)))

## [1] 1.062708 1.674428
```

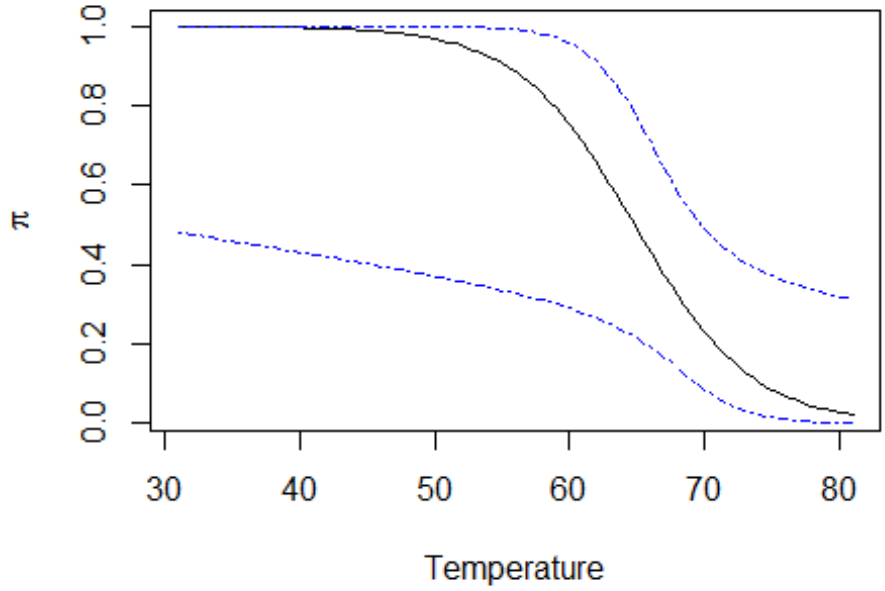
## Question 5b-c) Construct plots of the simplifed model with 95% Wald Confidence Intervals

A plot of $\pi$ versus Temp of the model of $logit(\pi) = \beta_0 + \beta_1 Temp$ was constructed for a temperature range of $31^0$ to $81^0$ on the x-axis, where

$$\pi = \frac{e^{15.04-0.231Temp}}{1 + e^{15.04-0.231Temp}}$$

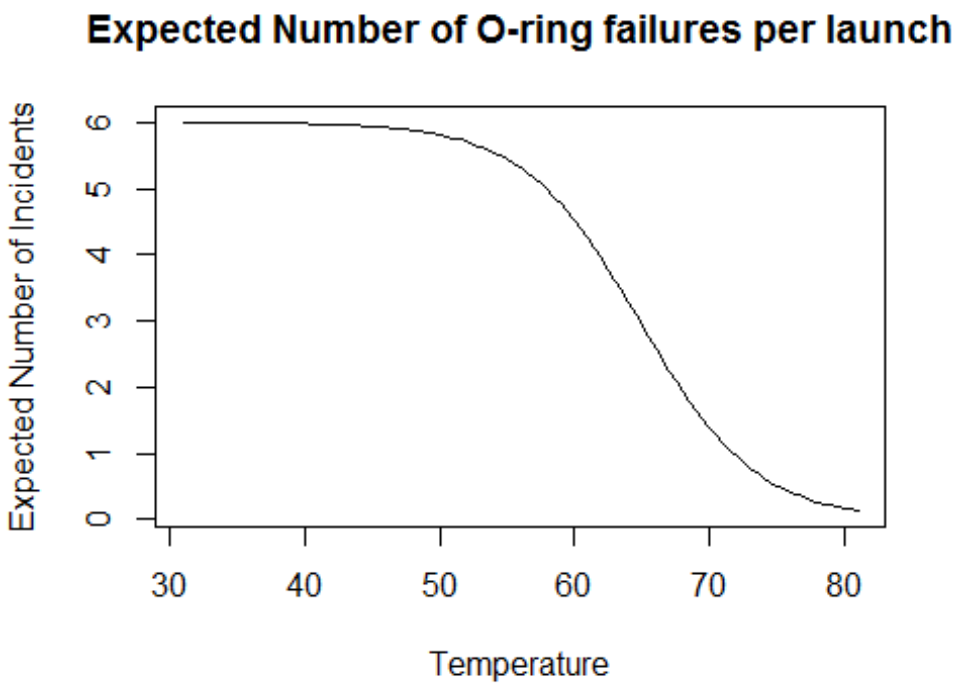
Plot of $\pi$ $vs$ $Temp$

```
par(mfrow = c(1, 1))
beta0 <- glm_temp$coefficients[1]
beta1 <- glm_temp$coefficients[2]
alpha <- 0.5

curve(expr = exp(beta0 + beta1 * x)/(1 + exp(beta0 + beta1 *
    x)), xlim = c(31, 81), col = "black", xlab = "Temperature",
    ylab = expression(pi))

ci.pi <- function(newdata, mod.fit.obj, alpha) {
    linear.pred <- predict(object = mod.fit.obj, newdata = newdata,
        type = "link", se = TRUE)
    CI.lin.pred.lower <- linear.pred$fit - qnorm(p = 1 -
        alpha/2) * linear.pred$se
    CI.lin.pred.upper <- linear.pred$fit + qnorm(p = 1 -
        alpha/2) * linear.pred$se
    CI.pi.lower <- exp(CI.lin.pred.lower)/(1 + exp(CI.lin.pred.lower))
    CI.pi.upper <- exp(CI.lin.pred.upper)/(1 + exp(CI.lin.pred.upper))
    list(lower = CI.pi.lower, upper = CI.pi.upper)
}

curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = glm_temp,
    alpha = 0.05)$upper, col = "blue", lty = "dotdash",
    add = TRUE, xlim = c(31, 81))
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = glm_temp,
    alpha = 0.05)$lower, col = "blue", lty = "dotdash",
    add = TRUE, xlim = c(31, 81))
```




A plot for the Expected number of failures vs. Temp was also constructed using n = 6 for the number of O-rings in a shuttle. The plot was constructed with the assumption that each O-ring has the same probabilty of failure and the failure is independent of other O-ring failures.

```
curve(expr = 6 * exp(beta0 + beta1 * x)/(1 + exp(beta0 +
    beta1 * x)), xlim = c(31, 81), col = "black", main = "Expected Number of
O-ring failures per launch",
    xlab = "Temperature", ylab = "Expected Number of Incidents")
```




## Question 5d) Estimate the probability of an O-ring failure using Temperature = $31^0$ at the shuttle launch and compute the corresponding confidence intervals

A 95% confidence level was used to estimate the probablity of an O-ring failure at the temperature of $31^0$ .

```
predict.glm(glm_temp, data.frame(Temp = 31), se.fit = TRUE,
    type = "response")

## $fit
##         1
## 0.9996088
##
## $se.fit
##           1
## 0.001580137
##
## $residual.scale
## [1] 1

ci.pi(newdata = data.frame(Temp = 31), mod.fit.obj = glm_temp,
    alpha = 0.05)

## $lower
##         1
```

```
## 0.4816106
##
## $upper
##         1
## 0.9999999
```

The results show that the probablity of a failure of an O-ring was 0.9996 at $31^0$. The confidence intervals for

## Question 5e) Bootstrapping

A 90% confidence interval can also be obtained.

For 31 degrees, our bootstrap CI is (0.967, 1).

For 81 degrees, our bootstrap CI is (9.39e-05, 0.117)

```
library(boot)

bootstrapPredict <- function(temp_of_interest, data, indices) {
    d <- data[indices, ]
    bootstrap_fit <- glm(formula = O.ring.fail.factor ~
        Temp, family = binomial(link = logit), data = d)
    return(predict.glm(bootstrap_fit, data.frame(Temp = temp_of_interest),
        type = "response"))
}

# the bootstrap quantile for 31 degrees
results_31 <- boot(data = df, statistic = bootstrapPredict,
    R = 1000, temp_of_interest = 31)
quantile(results_31$t, c(0.05, 0.95))

##        5%       95%
## 0.9569149 1.0000000

# the bootstrap quantile for 81 degrees
results_81 <- boot(data = df, statistic = bootstrapPredict,
    R = 1000, temp_of_interest = 81)
quantile(results_81$t, c(0.05, 0.95))

##           5%          95%
## 2.220446e-16 1.293973e-01
```

## Question 5f) Addition of a Quadtratic Term to the Simplified Model

A quadratic term of Temperature was expored to determine if this would add prediction to the simplifed model for the temperature predictor.

$$Quadratic\ Model\colon\ \ logit(\pi) = \beta_0 + \beta_1 Temp + \beta_2 Temp^2$$

```
glm_temp_sq <- glm(formula = O.ring.fail.factor ~ Temp +
    I(Temp^2), family = binomial(link = logit), data = df)
summary(glm_temp_sq)
```

```
##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp + I(Temp^2), family =
binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9614  -0.6608  -0.4735   0.1781   2.1185
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 78.48317   76.67353   1.024    0.306
## Temp        -2.09430    2.18479  -0.959    0.338
## I(Temp^2)    0.01359    0.01553   0.875    0.381
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 19.389  on 20  degrees of freedom
## AIC: 25.389
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(glm_temp_sq)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring.fail.factor
##           LR Chisq Df Pr(>Chisq)
## Temp       1.19322  1     0.2747
## I(Temp^2)  0.92649  1     0.3358
```

Examining the results of the output, it is shown that the p-value for the Wald test of

$$H_0\colon \beta_2 = 0 \; vs. \; H_a\colon \beta_2 \neq 0$$

is 0.381, suggesting that there is not evidence of a quadratic relationship between Temperature and O-ring Failure. The quadratic model was dropped from any further analysis.

## Section 4: Alternative Model Exploration

Several different models were explored, including:

- Logistic model with Temperature and Pressure Interacting

- Probit Model
- Complementary Log-Log Regression Model

## Logistic Model with Temperature and Pressure Interaction

As we saw in the EDA there is a marginal dependency between Temperature and Pressure. For this purpose we will also test a model with an interaction between the two. The interaction coefficient is small and negative (-0.003) which means that the estimated probability for a O.ring failure will slowly descrease twith increased product of Temperature and Pressure. The AIC value for the interaction model (Ha) is 26.075, which is higher than the original model without iteraction (Ho), AIC = 24.315. From the anova Chi-square test we also get that $-2log(\Lambda) = 0.707$ with a p-value of P( A > 18.07) = 0.401, where A ~ Chi-square with df of 19 (30.14). Thus there is no evidence of a Temperature and Pressure interaction. We will not go forward with this interaction model.

```
glm_temp_pressure_interact <- glm(formula = O.ring.fail.factor ~
    Temp + Pressure + Temp:Pressure, family = binomial(link = logit),
    data = df)
summary(glm_temp_pressure_interact)

##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp + Pressure + Temp:Pressure,
##     family = binomial(link = logit), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2083  -0.5879  -0.4178   0.3049   2.0406
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -27.217458  50.238881  -0.542    0.588
## Temp            0.358212   0.720644   0.497    0.619
## Pressure        0.221088   0.258398   0.856    0.392
## Temp:Pressure  -0.003054   0.003711  -0.823    0.410
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 18.075  on 19  degrees of freedom
## AIC: 26.075
##
## Number of Fisher Scoring iterations: 5

# Likelihood-Ratio Test of the two models Ho and Ha
anova(glm_temp_pressure, glm_temp_pressure_interact, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: O.ring.fail.factor ~ Temp + Pressure
```

```
## Model 2: O.ring.fail.factor ~ Temp + Pressure + Temp:Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        20     18.782
## 2        19     18.075  1  0.70689   0.4005
```

```
qchisq(0.95, df = 19)
```

```
## [1] 30.14353
```

## Probit Regression Model

The probit model's link function is the inverse CDF.

The simple probit model with just temperature as the predictor has a slightly higher AIC than the logit counterpart.

The probit model coefficient is hard to interpret, because the increase in probability of failure for a single unit increase of temp depends on the starting value of the temp.

Model Form: $probit(\pi) = \beta_0 + \beta_1 Temp$

```
glm_probit <- glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link
= "probit"),
    data = df)
summary(glm_probit)
```

```
##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link =
"probit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0831  -0.7930  -0.3747   0.4413   2.2081
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.77490    3.87231   2.266   0.0234 *
## Temp        -0.13510    0.05646  -2.393   0.0167 *
## ---
## 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: 20.378  on 21  degrees of freedom
## AIC: 24.378
##
## Number of Fisher Scoring iterations: 6
```

## Complementary Log-Log Regression Model

The log-log model's link function is $ln(-ln(1-\pi))$.

The simple loglog model with just temperature as the predictor has a slightly lower AIC than the logit counterpart.

It's also important to note that the complementary log-log model is asymmetrical and used when the probability of event occuring tends to be very large or small.

Model Form: $probit(\pi) = \beta_0 + \beta_1 Temp$

```
glm_loglog <- glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link
= "cloglog"),
    data = df)
summary(glm_loglog)
```

```
##
## Call:
## glm(formula = O.ring.fail.factor ~ Temp, family = binomial(link =
"cloglog"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0358  -0.7361  -0.3891   0.1729   2.2050
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 12.30215    5.19483   2.368   0.0179 *
## Temp        -0.19583    0.07809  -2.508   0.0122 *
## ---
## 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: 19.531  on 21  degrees of freedom
## AIC: 23.531
##
## Number of Fisher Scoring iterations: 8
```

## Section 5: Model Analysis and comparision

The selection of the final model includes tradeoffs between predicitve accuracy (including deviance statistics and their associated chi-square statistic) and parsimony (using AIC).

There were several models investigated in this report

- Logistic Regression with Temperature and Pressure as predictors

- Logistic Regression with only Temperature as a Predictor
- Logistic Regression with Temperature as a predictor, with a quadtratic term
- Logistic Regression model with Temperature and Pressure Interacting
- Probit Model
- Complementary Log-Log Model

The first four models are all logistic regression models and we compared them against each other to determine which one exhibited the best fit.

```
anova(glm_temp, glm_temp_pressure, test = "Chisq")
```

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

The nonsignificant chi-square value (p=0.2156 ) suggests that the model using only temperature fits as well as the model using temperature and pressure. Therefore the linear model using only temperature can be used instead of the more complex model of tempertaure and pressure. The deviance of the logistic model with temperature and pressure was slightly smaller than the logistic model (18.78 versus 20.31) with just temperature but the lack of signficance in the parameters excluded it from consideration.

The results of the quadratic model showed that the p-value for the Wald test of

$$H_0\colon \beta_2 = 0 \; vs. \; H_a\colon \beta_2 \neq 0$$

is 0.381, suggesting that there is not evidence of a quadratic relationship between Temperature and O-ring Failure.

The AIC value for the interaction model (Ha) is 26.075, which is higher than the original model without iteraction (Ho), AIC = 24.315. From the anova Chi-square test we also get that $-2log(\Lambda) = 0.707$ with a p-value of P( A > 18.07) = 0.401, where A ~ Chi-square with df of 19 (30.14). Thus there is no evidence of a Temperature and Pressure interaction. We will not go forward with this interaction model.

The comparison of these four models left the model with Temperature as a Predictor as the best of the logistical regression models.

## Comparing the Logistic, Probit, and Complementary Log-log Models

We are comparing Logistic, Probit, and Complementary Log-log Regression Models with only Temperature as explanatory variable.

From the individual model creation summary statements earlier on we have the AIC information Criteria ratings of each model. The Logistic model has an AIC = 24.315, the

Probit model has AIC = 24.378 and the Complementary Log-Log model has an AIC = 23.531. From this perspective, the Log-Log model is preferred. When reviewing the coefficients we can also see a lower p-value for Log-Log on both the intercept and the Temperature coefficients.

When comparing the estimated probability for the three models at different Temperatures, we see that the Log-Log model has a slightly higher estimated probability at low and high Temperature values. This is per say not an indication of being a better model, but explains how it behaves vs. the other models.

We see the same in a bubble plot of the estimated probability of failure based on Temperature. The observed probability of failure at each Temperature is plotted with the plotting point size being proportional to the number of observations at a Temperature. The estimated logistic, probit, and complementary log-log regression models are all given in the plot. We see little difference between the models except for at low Temmperatures, where the Log-Log model has higher probability.

We calculate the odds ratios between three different Temperatures (55 to 70 and 70 to 85) as both Probit and Log-Log model odds ratios are dependent on (Temperature in this case). As expected, the odds ratios are exactly the same for the logistic regression model (33), but are different for the Probit and Complementary Log-Log models. Specifically for the preferred Log-Log model we have that the odds of a failure changes by 363 times for an increase in Temperature from 55 degrees to 70 degrees. On the other hand, the odds of a failure changes by 21 times for an increase in Temperature from 70 degrees to 85 degrees.

```
# Estimate the logistic, probit, and complementary
# log-log models

# Logistic
round(summary(glm_temp)$coefficients, 4)

##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786  2.0387   0.0415
## Temp         -0.2322     0.1082 -2.1450   0.0320

# Probit
round(summary(glm_probit)$coefficients, 4)

##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.7749     3.8723  2.2661   0.0234
## Temp         -0.1351     0.0565 -2.3926   0.0167

# Complementary log-log
round(summary(glm_loglog)$coefficients, 4)

##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  12.3022     5.1948  2.3682   0.0179
## Temp         -0.1958     0.0781 -2.5077   0.0122
```

```r
# Compare pi_hat values with predict()
predict.data <- data.frame(Temp = c(55, 70, 85))
logistic.pi <- predict(object = glm_temp, newdata = predict.data,
    type = "response")
probit.pi <- predict(object = glm_probit, newdata = predict.data,
    type = "response")
cloglog.pi <- predict(object = glm_loglog, newdata = predict.data,
    type = "response")
round(data.frame(predict.data, logistic.pi, probit.pi, cloglog.pi),
    4)
```

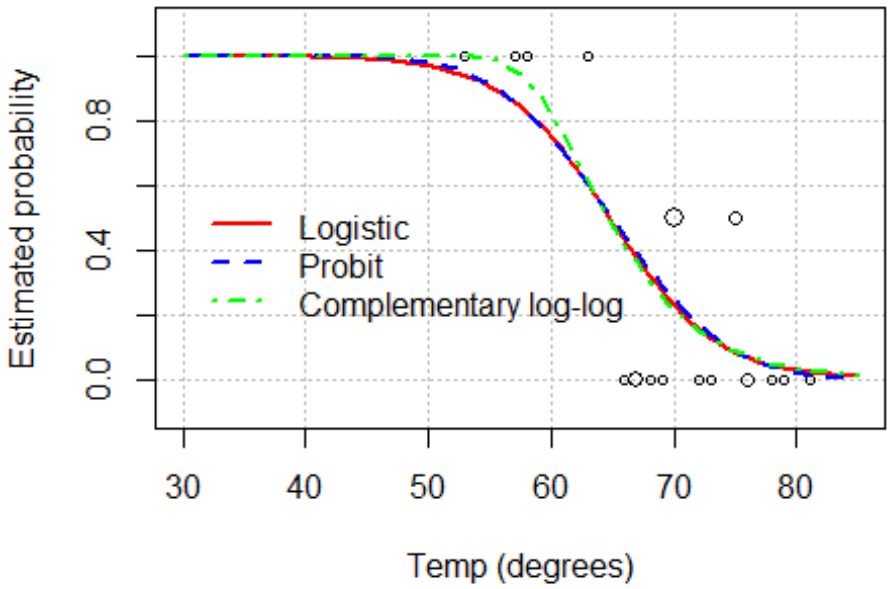
```
##   Temp logistic.pi probit.pi cloglog.pi
## 1   55      0.9067    0.9106     0.9902
## 2   70      0.2300    0.2477     0.2174
## 3   85      0.0091    0.0034     0.0129
```

```r
# Bubble plot
head(df)
```

```
##   Flight Temp Pressure O.ring Number O.ring.failure O.ring.fail.factor
## 1      1   66       50      0      6              0         No Failure
## 2      2   70       50      1      6              1            Failure
## 3      3   69       50      0      6              0         No Failure
## 4      4   68       50      0      6              0         No Failure
## 5      5   67       50      0      6              0         No Failure
## 6      6   72       50      0      6              0         No Failure
```

```r
w <- aggregate(formula = O.ring.failure ~ Temp, data = df,
    FUN = sum)
n <- aggregate(formula = O.ring.failure ~ Temp, data = df,
    FUN = length)
w.n <- data.frame(Temp = w$Temp, Failure = w$O.ring.failure,
    trials = n$O.ring.failure, proportion =
round(w$O.ring.failure/n$O.ring.failure,
        4))

# Plot of the observed proportions with logistic
# regression model
symbols(x = w$Temp, y = w$O.ring.failure/n$O.ring.failure,
    circles = sqrt(as.numeric(n$O.ring.failure)), inches = 0.05,
    xlim = c(30, 85), xlab = "Temp (degrees)", ylab = "Estimated
probability",
    panel.first = grid(col = "gray", lty = "dotted"))
# Estimated logistic regression model
curve(expr = predict(object = glm_temp, newdata = data.frame(Temp = x),
    type = "response"), col = "red", lwd = 2, add = TRUE,
    lty = 1, xlim = c(30, 85))
# Estimated probit model
curve(expr = predict(object = glm_probit, newdata = data.frame(Temp = x),
    type = "response"), col = "blue", lwd = 2, add = TRUE,
    lty = 2, xlim = c(30, 85))
```

```
# Estimated complementary log-log model
curve(expr = predict(object = glm_loglog, newdata = data.frame(Temp = x),
    type = "response"), col = "green", lwd = 2, add = TRUE,
    lty = 4, xlim = c(30, 85))
# Legend
legend(x = 30, y = 0.6, legend = c("Logistic", "Probit",
    "Complementary log-log"), lty = c(1, 2, 4), lwd = c(2,
    2, 2), bty = "n", col = c("red", "blue", "green"), cex = 1)
```




```
# ORs

pi.hat <- data.frame(predict.data, logistic.pi, probit.pi,
    cloglog.pi)
odds.x55 <- pi.hat[1, 2:4]/(1 - pi.hat[1, 2:4])
odds.x70 <- pi.hat[2, 2:4]/(1 - pi.hat[2, 2:4])
odds.x85 <- pi.hat[3, 2:4]/(1 - pi.hat[3, 2:4])

OR.55.70 <- odds.x55/odds.x70
OR.70.85 <- odds.x70/odds.x85

data.frame(OR = c("55 vs. 70", "70 vs. 85"), round(rbind(OR.55.70,
    OR.70.85), 2))

##          OR logistic.pi probit.pi cloglog.pi
## 1 55 vs. 70       32.54     30.95     363.43
## 2 70 vs. 85       32.54     97.01      21.24
```

## Section 6: Plots of Final Model Selection

Below are plots for the Simplified Logistic Regression Model with Temperature as only Explanatory Variable:

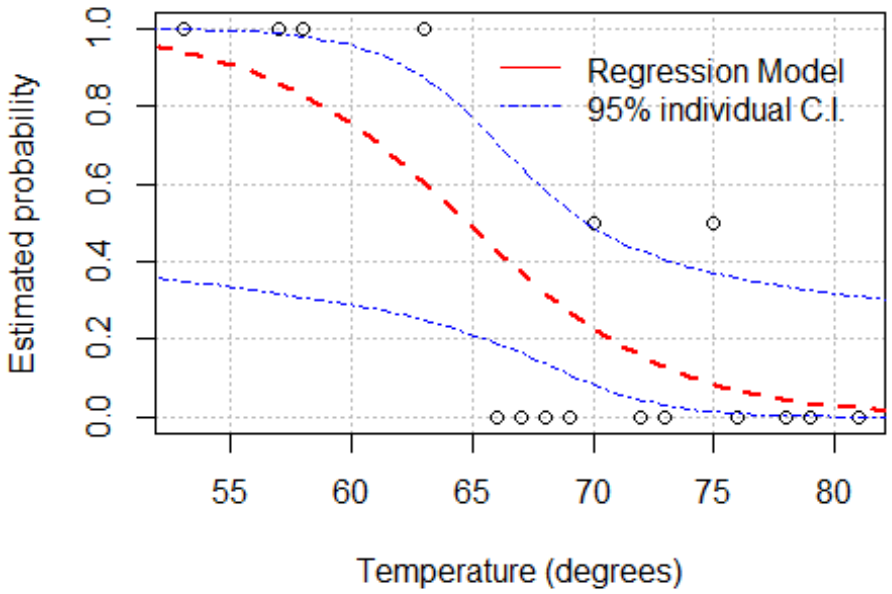
$$SimplifiedLogisticModel: \ \ logit(\pi) = 15.04 - 0.23Temp$$

The reason we are not showing plots of the Complementary Log-Log model even if it has a better fit is that we had few references in the text book and most of the below plot functions were not available for the Complementary Log-Log Regression Model.

We start to show an estimated probability plot, which shows the observed proportion of failures at each Temperature and then overlaying the estimated regression model and associated CI. We see how the estimated probability for O-ring failure decreases with higher temperature and that the confidence intervall is higher at low and high temperatures respectively.

```
# Plot of the observed proportions with regression model
plot(x = w$Temp, y = w$O.ring.failure/n$O.ring.failure,
    xlab = "Temperature (degrees)", ylab = "Estimated probability",
    panel.first = grid(col = "gray", lty = "dotted"))
# Put estimated regression model on the plot
curve(expr = predict(object = glm_temp, newdata = data.frame(Temp = x),
    type = "response"), col = "red", lwd = 2, add = TRUE,
    lty = 2, xlim = c(30, 85))
# Function for C.I.s - need in order to use with curve
# function
ci_cloglog.pi <- function(newdata, mod.fit.obj, alpha) {
    linear.pred <- predict(object = mod.fit.obj, newdata = newdata,
        type = "link", se = TRUE)
    CI.lin.pred.lower <- linear.pred$fit - qnorm(p = 1 -
        alpha/2) * linear.pred$se
    CI.lin.pred.upper <- linear.pred$fit + qnorm(p = 1 -
        alpha/2) * linear.pred$se
    CI.pi.lower <- exp(CI.lin.pred.lower)/(1 + exp(CI.lin.pred.lower))
    CI.pi.upper <- exp(CI.lin.pred.upper)/(1 + exp(CI.lin.pred.upper))
    list(lower = CI.pi.lower, upper = CI.pi.upper)
}

# Test cases
ci_cloglog.pi(newdata = data.frame(Temp = 55), mod.fit.obj = glm_temp,
    alpha = 0.05)

## $lower
##         1
## 0.3350234
##
## $upper
##         1
## 0.9946932
```

```
ci_cloglog.pi(newdata = data.frame(Temp = 70), mod.fit.obj = glm_temp,
    alpha = 0.05)

## $lower
##         1
## 0.085086
##
## $upper
##          1
## 0.4895482

# Plot C.I. bands
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = glm_temp,
    alpha = 0.05)$lower, col = "blue", lty = "dotdash",
    add = TRUE, xlim = c(30, 85))
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = glm_temp,
    alpha = 0.05)$upper, col = "blue", lty = "dotdash",
    add = TRUE, xlim = c(30, 85))

# Legend
legend(x = 65, y = 1, legend = c("Regression Model", "95% individual C.I."),
    lty = c("solid", "dotdash"), col = c("red", "blue"),
    bty = "n")
```

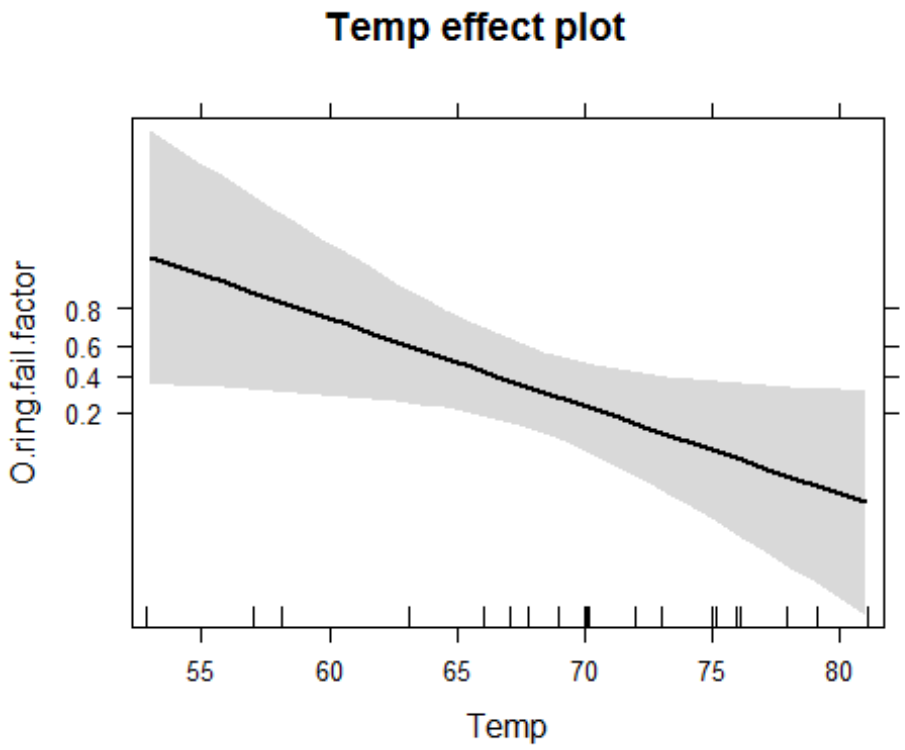



Next we show a three-dimensional plot of the log-likelihood function as it changes for different beta coefficents. As expected, the plot shows the log-likelihood estimate to be top at $\beta_0$ = 15 and $\beta_1$ = 0.2, which is aligned with the Regression Model.

```
# 3D plot of the log-likelihood function

# Evaluate the log-likelihood function at a lot of
# different values for beta0 and beta1
logL <- function(beta, x, Y) {
    pi <- exp(beta[1] + beta[2] * x)/(1 + exp(beta[1] +
        beta[2] * x))
    sum(Y * log(pi) + (1 - Y) * log(1 - pi))
}
beta0.values <- seq(from = 2, to = 25, by = 0.1)
beta1.values <- seq(from = -0.65, to = -0.05, by = 0.01)
count <- 1
save.logL <- numeric(length(beta0.values) * length(beta1.values))

for (beta0 in beta0.values) {
    for (beta1 in beta1.values) {
        save.logL[count] <- logL(beta = c(beta0, beta1),
            x = df$Temp, Y = df$O.ring.failure)
        count <- count + 1
    }
}
max(save.logL)

## [1] -10.15782

library(package = rgl)  # Package that does 3D interactive plots
open3d()  # Open plot window

## wgl
##   1

# 3D plot with gridlines
persp3d(x = beta1.values, y = beta0.values, z = save.logL,
    xlab = "beta1", ylab = "beta0", zlab = "log(L)", ticktype = "detailed",
    col = "red")
grid3d(c("x", "y+", "z"))
```

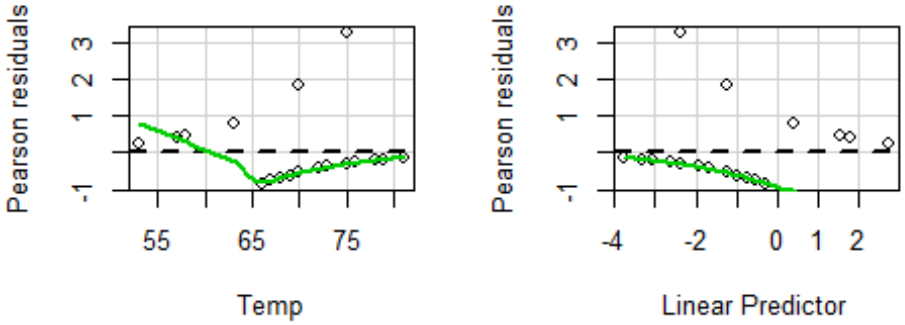
A temp effect plot is shown next and displays the relationship between increasing O-ring.failure probability and decreasing temperature. It also shows the higher uncertainty at low and high temperature ranges. This time the effect plot has no real new information as compared to the first plot.

```
library(effects)
plot(allEffects(glm_temp, default.levels = 10))
```




Finally a residual plot is used to see that there are no systematic patterns between residuals and the explanatory variables. The smoother line shows an upwards curve from -1 to 0 residuals for Tempr = 65 to 85 degrees but is based on a subset of values and ther are still a lot of outliers.

```
residualPlots(glm_temp, layout = c(2, 2))
```




```
##      Test stat Pr(>|t|)
## Temp     0.926    0.336
```