W271 Lab 1

Gurdit Chahal
Zach Day
Victoria Eastman

October 1, 2018

I. Introduction

Following the Challenger Space Shuttle's destruction in 1986, a commission appointed by President Reagan determined the cause to be a gas leak through a field joint. This problem was well-known to NASA and is frequently referred to as an O-ring failure. In 1989, Dalal et al. collected data from previous space shuttle launches to study the probability of an O-ring failure under conditions similar to those that occured during the Challenger launch in 1986. In this analysis, we used their dataset to echo the purpose of their study and attempt to determine the effect of key explanatory variables (temperature and pressure) using LRT's on O-ring failure. We also explore the possibility of using a linear model and assess it's validity while comparing to the logistic model. In the end we specified a logistic regression model on temperature with the following formula:

$$\text{logit}(\hat{\pi}) = \frac{\hat{\pi}}{1-\hat{\pi}} = 5.08498 - 0.11560\text{Temp}$$

The probability of an O-ring failure at temperature 31°F, the temperature when Challenger was launched, is 0.818 with a 95% confidence interval of about (0.16,0.991) using profile LR intervals, which would correspond to about five out of the six primary field O-rings failing.

We first begin our analysis with a thorough exploratory data analysis in order to understand the variables we are working with. Then, we estimate a series of models that we use to predict O-ring failure.

```
# Import libraries
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(Hmisc))
suppressPackageStartupMessages(library(car))

#setwd("/Users/gurditchahal/w271_lab1")
#setwd("/home/victoriaeastman/berkeley/w271/w271_lab1")
df <- read.csv("challenger.csv")
```

II. EDA

II (a) Data Summary and Overview

```
# Start with basic looks at the data
glimpse(df)
```

```
## Observations: 23
## Variables: 5
## $ Flight   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Temp     <int> 66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 5...
## $ Pressure <int> 50, 50, 50, 50, 50, 50, 100, 100, 200, 200, 200, 200,...
## $ O.ring   <int> 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 2, 0, 0, 0, 0,...
```

```
## $ Number   <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6,...
```

```r
describe(df[,c("Temp", "Pressure", "O.ring")])
```

```
## df[, c("Temp", "Pressure", "O.ring")]
##
##  3  Variables      23  Observations
## ---------------------------------------------------------------------------
## 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    57    58    63    66    67    68    69    70    72
## Frequency      1     1     1     1     1     3     1     1     4     1
## Proportion 0.043 0.043 0.043 0.043 0.043 0.130 0.043 0.043 0.174 0.043
##
## Value         73    75    76    78    79    81
## Frequency      1     2     2     1     1     1
## Proportion 0.043 0.087 0.087 0.043 0.043 0.043
## ---------------------------------------------------------------------------
## Pressure
##        n  missing distinct     Info     Mean      Gmd
##       23        0        3    0.706    152.2    67.59
##
## Value         50   100   200
## Frequency      6     2    15
## Proportion 0.261 0.087 0.652
## ---------------------------------------------------------------------------
## 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
## ---------------------------------------------------------------------------
```

A glimpse of the data shows we have 23 observations and 5 variables in our dataset. There are also no missing values. Again, we are primarily interested in the effects of temperature and pressure on o-ring failure. The dependent variable : the percentage of O-ring failures is *O.ring/Number* where *1) O.ring* is the number of primary field O-ring failures and *2)Number* is the total number of primary field O-rings (six total, three for each for two booster rockets). Our two potential explanatory variables are *1)Temp* is temperature at launch in degrees Fahrenheit and *2)Pressure* is combustion pressure (psi).

We can see that the `O.ring` variable has 3 distinct values: 0, 1, and 2 but only 2 observations for the value 2. To deal with this concern in representation we might want to consider formulating our problem in terms of having *at least one* O-ring failure.This might also help with assumptions of independence between O-rings (discussed later). Therefore, for one version of our model we will binarize our outcomes and transform those flights with 2 failures as having 1 failure for the purposes of this study.We will later compare this to the binomial model when finalzing our model selection.

```r
# Keep raw counts as new variable of total
df$O.ring.total = df$O.ring
# Binary version
```

```r
df$O.ring[df$O.ring > 1] = 1
```

In addition, we see that the dataset contains 23 data points from other shuttle launches and none of the variables are missing any entries. Interestingly, pressure is generally considered to be a continuous variable, however, we see three distinct values of 50, 100, and 200. We could potentially see reason to use this as a categorical variable in the regression estimation below.

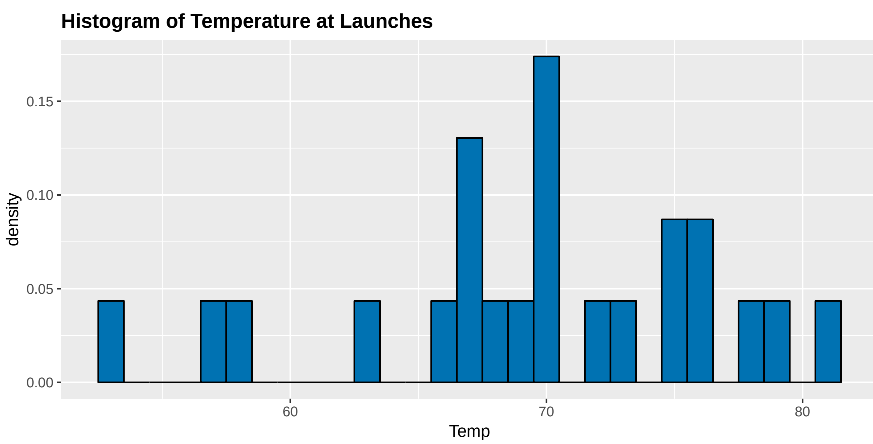
```r
#quick check for potential collinearity as well as surface level relations
cor(df[,c("Temp", "Pressure", "O.ring")])
```

```
##                 Temp   Pressure     O.ring
## Temp      1.00000000 0.03981769 -0.5607143
## Pressure  0.03981769 1.00000000  0.2616884
## O.ring   -0.56071429 0.26168839  1.0000000
```

We take a quick look at correlation between variables to assess wether there might be collinearity as well as a rough gauge of predictive power between these variables prior to any transformations. We see that there is no perfect collinearity. We see a moderate negative correlation between O-ring failures and temperature and a weakly positive correlation between pressure and failures- leading us to our first consideration of these variables as the explanatory variables in our model. We see negligible positive correlation between temperature and pressure and thus don't worry about collinieարity.

II (b) Temperature

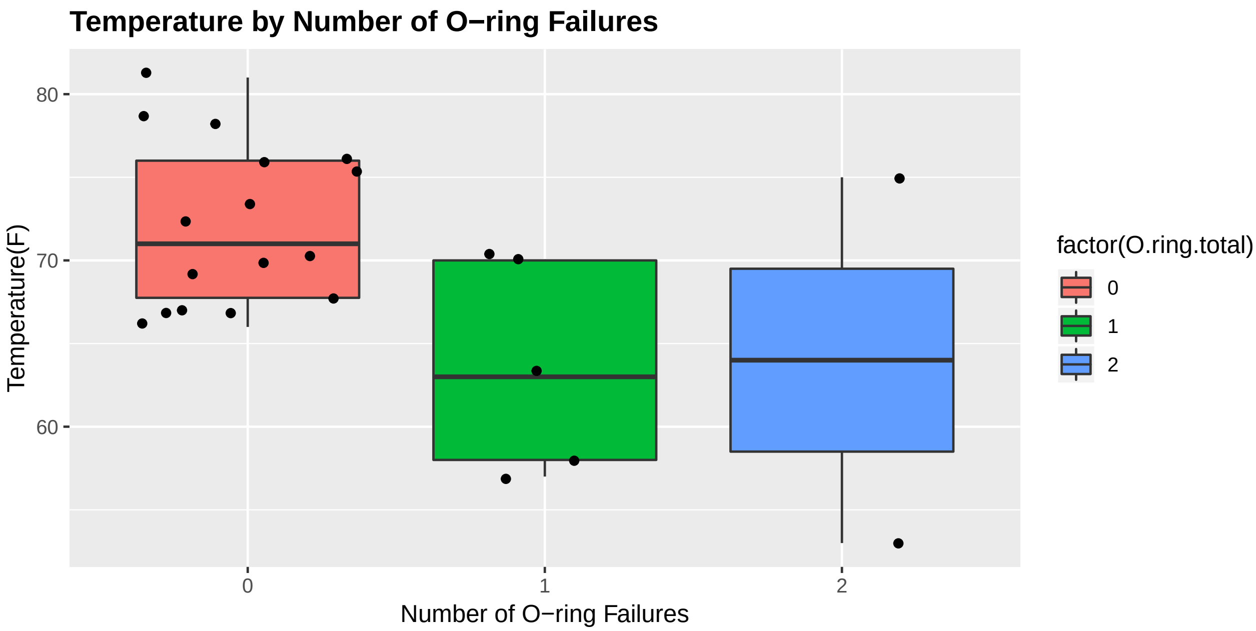
```r
#What does the distribution of temperatures look like?
ggplot(df, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Histogram of Temperature at Launches") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



The distribution of the temperature explanatory variable has a slight negative skew. The mode of temperatures is 70°F and the range is 53, 81. This is noteworthy because the temperature when Challenger was launched is 31°F and there are no observations of how many O-ring failures occur at that temperature. Moreover,there are fewer observations at lower temperatures as we look to the left of 67°F (everything below has no more than one count) and is worth noting in terms of precision of prediction concerns once we reach the modeling stage.

```r
#group temperature variable by O.ring failure
  print(ggplot(df, aes(factor(O.ring.total), Temp)) +
          geom_boxplot(aes(fill = factor(O.ring.total))) +
          geom_jitter() +
```

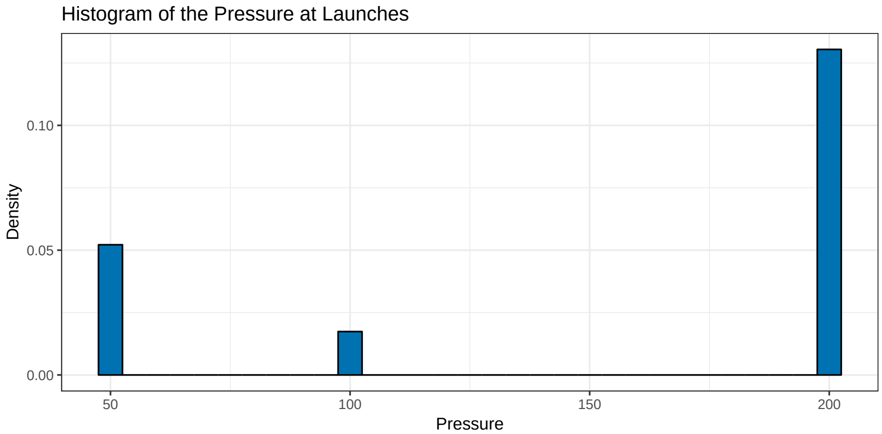
```r
        ggtitle("Temperature by Number of O-ring Failures") + ylab('Temperature(F)')
        + xlab('Number of O-ring Failures')+
        theme(plot.title = element_text(lineheight=1, face="bold")))
```



The boxplot above of temperature grouped by o-ring failure shows the average temperature when O-rings failed was lower than when they did not. What's interesting is we can see from the chart that when the temperature is below 65°F, at least one O-ring fails. Therefore, it lends some credibility to a belief that O-rings are more likely to fail when temperature is lower. Another interesting observation that jumps out is that there is one observation of 2 O-ring failures at 75°F which goes against the grain of the general observation between temperature and failures. This outcome could be suspect to having some influence by our other explanatory variable of *Pressure*.

II (c) Pressure

```r
ggplot(df, aes(x = Pressure)) + geom_histogram(aes(y = ..density..),
binwidth = 5, fill = "#0072B2", colour = "black") +
        ggtitle("Histogram of the Pressure at Launches")+
        theme(plot.title = element_text(lineheight = 1, face = "bold")) +
        theme_bw() + labs(x = "Pressure", y = "Density")
```

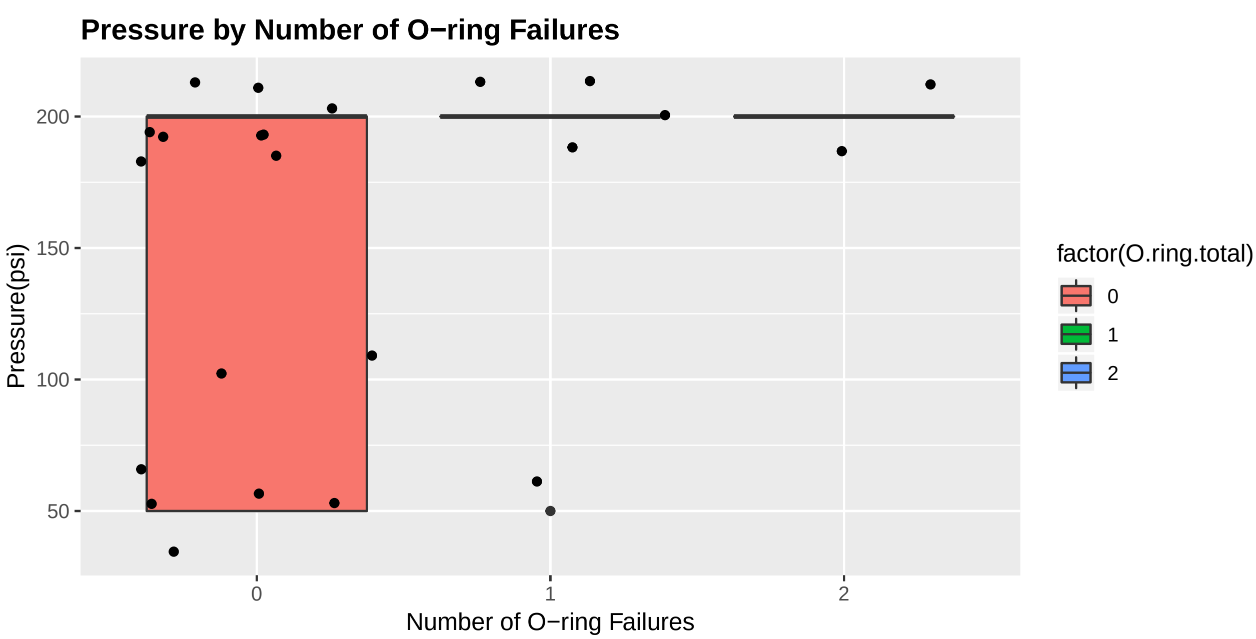



```
table(df$Pressure)
```

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

As noted in the table summary, although pressure in psi is usually considered continuous, the manner in which it's been observed at these three distinct values suggests a natural discretization at these key values. The test pressure from the observations are 50psi, 100psi and 200psi. Moreover, we notice only two observations at 100. Depending on where the seemingly abnormal failure point at higher temperature is with relation to pressure as well as the relationship of pressure to the failures, we could consider a further discretiziation like binarization (e.g. at 50 =1 >50=0).

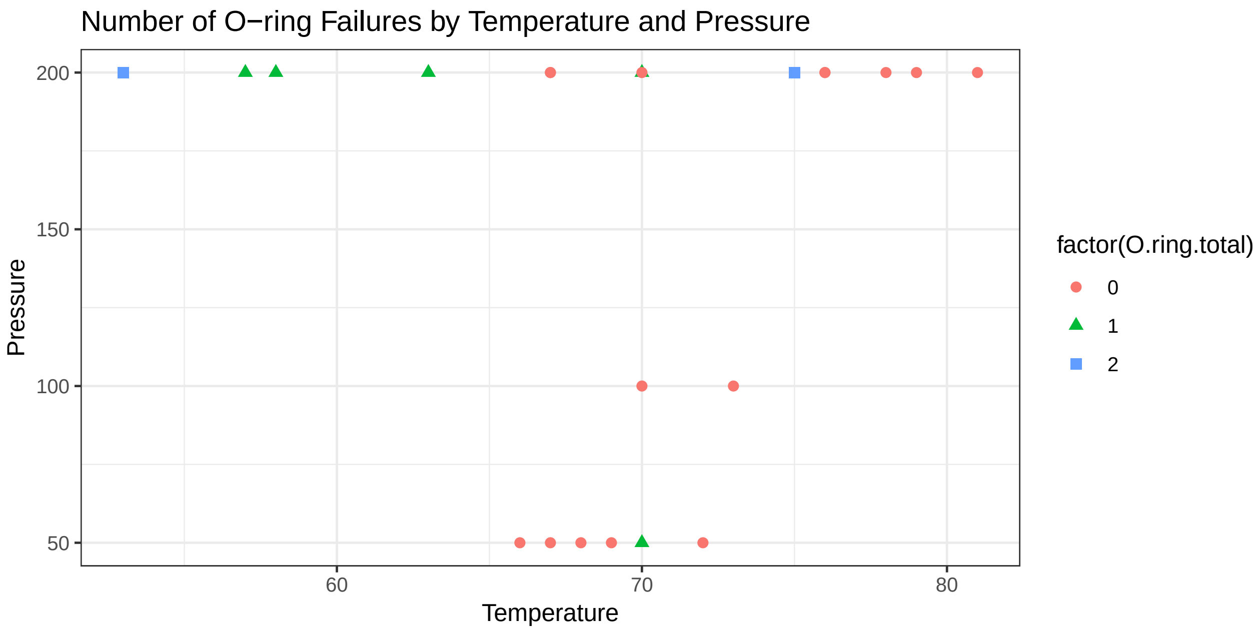
```
print(ggplot(df, aes(factor(O.ring.total), Pressure)) +
          geom_boxplot(aes(fill = factor(O.ring.total))) +
          geom_jitter() +
          ggtitle("Pressure by Number of O-ring Failures") + ylab('Pressure(psi)')
                +xlab('Number of O-ring Failures') +
          theme(plot.title = element_text(lineheight=1, face="bold")))
```



From the Pressure by Number of O-ring Failures chart, there are no O-ring failures in observations occuring at a pressure of 100psi, while the observations at a pressure 50psi or 200psi have a mix of both O-ring failures reported and no failures. Thus, O-rings may be more likely to fail at either low pressure (50psi) or high pressure (200psi) but we would need more observations at 100 psi to be more confident in the trend.

II (c) Pressure and Temperature

```
#slicing by pressure, any distinct relations between failure and temperature?
ggplot(df, aes(Temp, Pressure, shape = factor(O.ring.total),
color = factor(O.ring.total))) + geom_point(size = 2) +
ggtitle("Number of O-ring Failures by Temperature and Pressure") +
theme(plot.title = element_text(lineheight = 2, face = "bold")) + theme_bw() +
        labs(x = "Temperature")
```



Combining all three variables into the plot, we can see that almost all o-ring failures occured at 200 psi with much lower temps than those that did not fail. The "special" failure point at 75 noted earlier also occurs at 200 psi. From the chart above, we can also see that below a temperature of 65, all observations contain at least one O-ring failure. Additionally, at high pressure values (200(psi)) there is a larger proportion of observations that contain at least 1 incident of an O-ring failure than the low pressure observations. This helps in suggesting that both temperature and pressure have some effects on the probability of O-ring failure.

```
#bin pressure to 200 or not
df$P200=df$Pressure
df$P200[df$P200!=200]=0
df$P200[df$P200==200]=1
df$P200=factor(df$P200)
```

Since 200 psi had some interesting distinguishing qualities in relation to the other variables such as proportion of failures as well as the 75 failure point,we create a binarized version of our pressure variable where we distinguish by wether or not the the pressure observed is 200 psi or not. We will analyze the impact of this at our model selection stage.

III. Modeling

Question 4 From Book

First, we will assume that the failure of each O-ring is independent of the others and the probability of each O-ring failure is the same so that we can employ the binomial distribution to predict the number of O-ring failures. It is necessary for deriving the likelihood-based solution as we can take products of the probabilities.

4 (a) Why is independence of each observation necessary?

Potential issues is that the quality/durability of the O-ring might be dependent on the factory or even batch that it came from (clustering) and could interfere with producing a more accurate estimate when left unaccounted for. The Dalal et. al. paper notes that O-ring failure is common when joint rotation occurs in the rocket, and the failure of one O-ring could cause a change in the probability of another O-ring failure on the same rocket. This kind of damage could affect the probability of damage in subsequent O-rings and violates our assumption of fixed, equal probabilities of failure for each O-ring. Moreover, O-rings are known to degrade over time, thus the O-rings in different shuttle launches may not have equal probabilities of failure if the O-ring ages are not equivalent. We will revisit this assumption later in our modeling stage.

III.I Modeling O-ring Failures with Temperature and Pressure

4 (b) Estimate logistic regression model

Based on our exploratory data analysis, we believe that both temperature and pressure have some effect on the probability of O-ring failure. We create a logistic regression model using both variables:

$$\text{logit}(\hat{\pi}) = \beta_0 + \beta_1 \text{Pressure} + \beta_2 \text{Temp}$$

What's critical is that we have *two working versions of our model where in one, Pressure is treated as continuous and in the other, a binary variable.* We will use likelihood ratio tests to explore these choices.

```r
# Initial model of pressure as continuous
mod_binomial.fit1<-glm(formula=O.ring.total/Number~Pressure+Temp,data=df,
                       family=binomial(link = logit),weights=Number)
summary(mod_binomial.fit1)
```

```
##
## Call:
## glm(formula = O.ring.total/Number ~ Pressure + Temp, family = binomial(link = logit),
##     data = df, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## Pressure     0.008484   0.007677   1.105   0.2691
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

Our first estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = 2.520195 + 0.008484\text{Pressure} - 0.098297\text{Temp}$$

```r
#Initial model of pressure as binary variable
mod_binomial.fit2<-glm(formula=O.ring.total/Number~P200+Temp,data=df,
                       family=binomial(link = logit),weights=Number)
summary(mod_binomial.fit2)
```

```
##
## Call:
## glm(formula = O.ring.total/Number ~ P200 + Temp, family = binomial(link = logit),
##     data = df, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0459  -0.6545  -0.5310  -0.1675   2.3178
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.84181    3.24061   0.877   0.3805
## P2001        1.29316    1.10356   1.172   0.2413
## Temp        -0.09678    0.04467  -2.166   0.0303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.332  on 20  degrees of freedom
## AIC: 35.892
##
## Number of Fisher Scoring iterations: 5
```

Our second estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = 2.84181 + 1.29316\text{Pressure} - 0.09678\text{Temp}$$

We quickly note that the coefficient in front of pressure in the first model describes the change in odds-ratio for incremental changes in pressure whereas the second model's pressure coefficient describes change in odds-ratio compared to the baseline of when the pressure is not 200 psi (from our observations, below 200).

4 (c) Perform LRTs to judge the importance of the explanatory variables in the model.

To investigate the important of the explanatory variables, likelihood ratio tests were performed.

```r
Anova(mod_binomial.fit1,Test='LRT')
```

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

```r
Anova(mod_binomial.fit2,Test='LRT')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring.total/Number
##      LR Chisq Df Pr(>Chisq)
## P200   1.7546  1    0.18530
## Temp   5.0724  1    0.02431 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test evaluates the importance of each explanatory variable . For the first explanatory variable, pressure, we test the hypothesis: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. The test statistic for the continuous pressure variable model is $-2log(\Lambda) = 1.5407$ and the p-value is 0.2145 so we fail to reject the null hypothesis that pressure has no effect on o-ring failure at $\alpha = 0.05$. Similarly, we fail to reject the null hypothesis that pressure has no effect on o-ring failure at $\alpha = 0.05$ in the binary case with $-2log(\Lambda) = 1.7546$ and the p-value is 0.18530.

For the second explanatory variable, temperature, we test the hypothesis: $H_0 : \beta_2 = 0$ vs $H_a : \beta_2 \neq 0$. The first model test statistic is $-2log(\Lambda) = 7.7542$ and the p-value is 0.0228 so we reject the null hypothesis and are encouraged that evidence suggests of an effect of temperature on o-ring failure. We reach to the same conclusion in the second model.

4 (d) Why did they remove pressure? Why could this be a problem?

In terms of statistical significance, pressure wasn't found to be statistically significant in either continuous nor the binary case. Even though our exploratory data analysis shows that pressure may affect the probability of O-ring failure, the likelihood ratio test shows that the effect is not statisctically significant and pressure does not need to be included in the model. Potential problems could be losing precision in probability, especially for edge case observations.

III.II Modeling O-ring Failures with Temperature

Question 5 (book)

Next we estimate the model with the only one regressor found significant: temperature.

5 (a) Estimate the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{Temp}$$

```r
mod_binomial.fit_temp<-glm(formula=O.ring.total/Number~Temp,data=df,
                           family=binomial(link = logit),weights=Number)
summary(mod_binomial.fit_temp)
```

```
##
## Call:
## glm(formula = O.ring.total/Number ~ Temp, family = binomial(link = logit),
##     data = df, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

The estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = 5.08498 - 0.11560\text{Temp}$$

Note that the new coefficients are different from the models that we estimated in our first versions. The new model also shows that using

$$\alpha = 0.05$$

, we still reject the null hypothesis that the coefficient of Temp is 0 because the p-value = 0.014.

III.III Checking the Independence Assumption

To help validate the independence assumption of the failure of individual O-rings, we compare our binomial model with a binary model. As we mentioned early in our EDA, the *O.ring* variable groups flights where at least one failure occured. Unlike the binomial model, this binary model does not require the statistical independence of the failure of each O-ring.

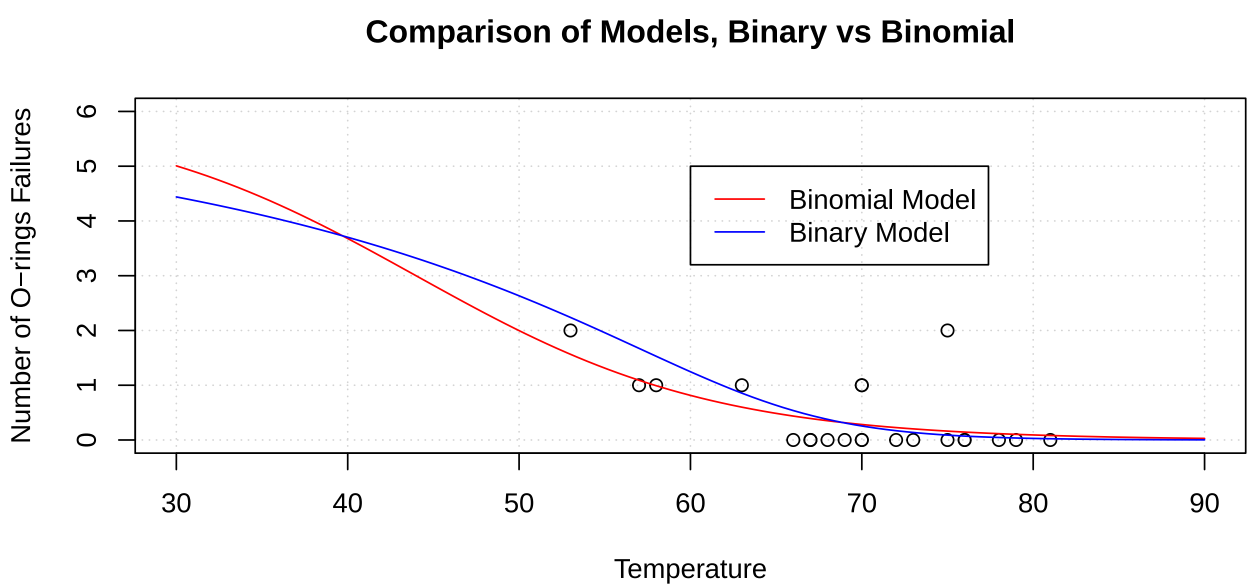
```r
binary.fit<-glm(O.ring ~ Temp, family = binomial(link = "logit"), data = df)
summary(binary.fit)
```

```
##
## Call:
## glm(formula = O.ring ~ 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 binary model also shows that the Temp variable is statiscally significant and should be included in the model. We look at the binary and binomial models using logistic regression side-by-side to compare.

```r
plot(df$Temp, df$O.ring.total, ylim = c(0, 6), xlim = c(30,
90), ylab = "Number of O-rings Failures", xlab = "Temperature",
main = "Comparison of Models, Binary vs Binomial", panel.first = grid())
# Binomial Model
curve(expr = predict(object = mod_binomial.fit_temp, newdata = data.frame(Temp = x),
                     type = "response") * 6, col = "red", add = TRUE,
                        xlim = c(30,90))
# Binary Model
curve(expr = (1 - (1 - predict(object = binary.fit, newdata = data.frame(Temp = x),
                               type = "response"))^(1/6)) * 6, col = "blue",
                                add = TRUE,
xlim = c(30, 90))
legend(60, 5, c("Binomial Model", "Binary Model"), lwd = c(1, 1), col = c(2, 4))
```

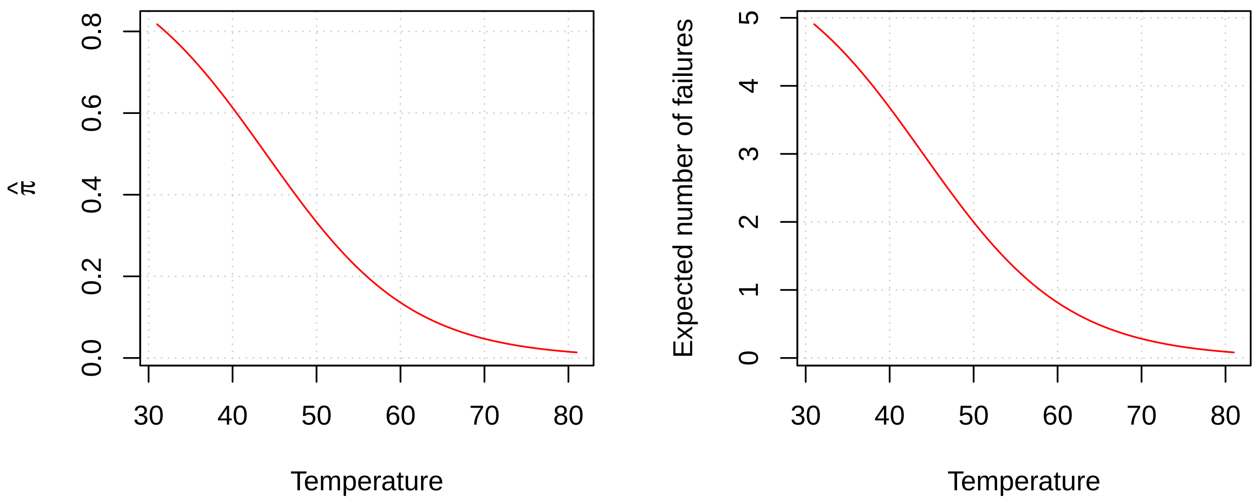


Although not a surefire validation, we are further assured in supporting the claim that the O-ring failures can be treated independently and we can use the binomial model. This validation check is on top of the subsequent analysis that Dalal et. all use to support using the binomial model.

5 (b) Construct two plots: (1)π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31 to 81 on the x-axis even though the minimum temperature in the

data set was 53.

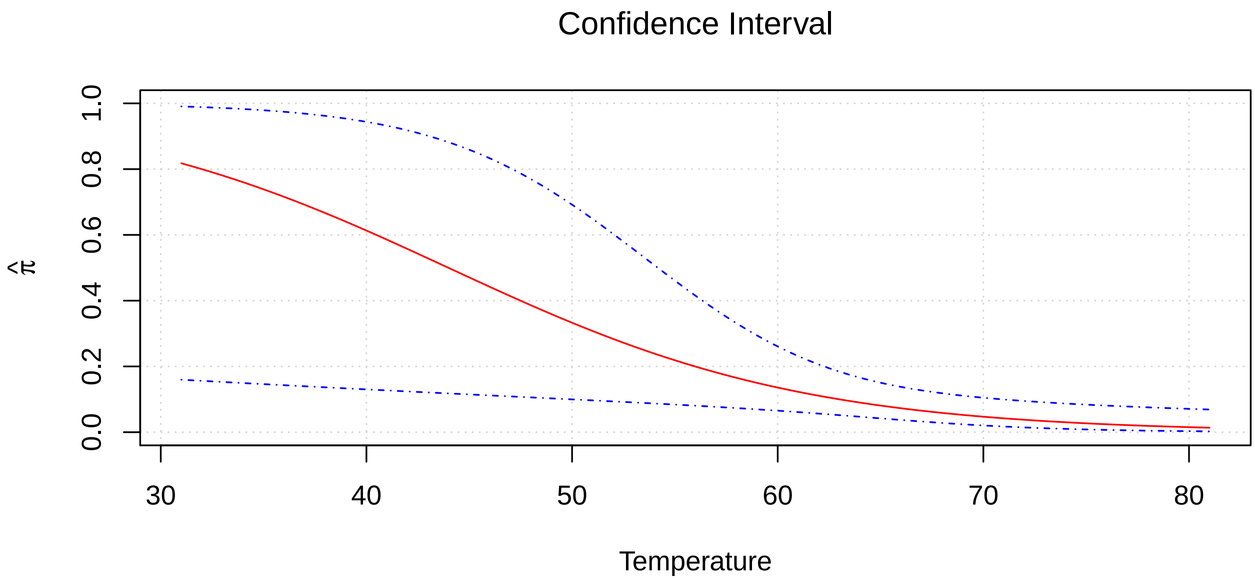
```
#
# (Question 5b)
par(mfrow = c(1, 2))
# pi vs. temperature
curve(expr = exp(mod_binomial.fit_temp$coefficients[1] +
                         mod_binomial.fit_temp$coefficients[2] *x)/
                        (1 + exp(mod_binomial.fit_temp$coefficients[1] +
                                        mod_binomial.fit_temp$coefficients[2] * x)),
      xlim = c(31, 81), col = "red", xlab = "Temperature",
      ylab = expression(hat(pi)), panel.first = grid())

# expected number of failures vs. temperature
curve(expr = predict(object = mod_binomial.fit_temp, newdata = data.frame(Temp = x),
type = "response") * 6, col = "red", xlim = c(31, 81), xlab = "Temperature",
ylab = "Expected number of failures", panel.first = grid())
```



5 (c) Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

```
curve(expr = exp(mod_binomial.fit_temp$coefficients[1] +
                         mod_binomial.fit_temp$coefficients[2] *x)/
                        (1 + exp(mod_binomial.fit_temp$coefficients[1] +
                                        mod_binomial.fit_temp$coefficients[2] *x)),
      ylim = c(0, 1), xlim = c(31, 81), col = "red",
      main = expression("Confidence Interval", hat(pi)), xlab = "Temperature",
      ylab = expression(hat(pi)),
panel.first = grid())


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 = mod_binomial.fit_temp,
                 alpha = 0.05)$lower, col = "blue", lty="dotdash",add=TRUE,xlim=c(31,81))

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



The dashed lines on this plot help to visualize the 95% confidence intervals for the estimated probability of O-ring failure. Above 60 degrees, the confidence interval is comparedly narrow, but at lower temperatures the confidence interval covers a much larger range. This larger range is due to the lack of data points at the lower temperatures.The model is extrapolating from the data points at higher temperatures and must account for the diversity of models that could be fitted to this range of observations.

5 (d) 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.

The challenger disaster occured at a temperature of 31 degrees, using the binomial model, we can estimate the probability of O-ring failures at this temperature from our model. This inference assumes that our observations are independent, there is no perfect collinearity, there is no complete separation and that there is linearity of the independent variables.

```
alpha=0.05
# Set data to predict on
predict.data <- data.frame(Temp=31)
# Linear part of model
linear.pred=predict(object = mod_binomial.fit_temp, newdata = predict.data,
                    type = "link", se = TRUE)
# Estimate probability
pi.hat31 = exp(linear.pred$fit)/(1+exp(linear.pred$fit))
number_fail=6*pi.hat31
# Confidence interval
CI.pi=ci.pi(newdata = predict.data, mod.fit.obj = mod_binomial.fit_temp,
            alpha = alpha)
```

The estimated probability of o-ring failure at 31°F is 0.8178 with and 4.9066 of the 6 O-rings failing. The corresponding 95% confidence interval is 0.1596025, 0.9906582.

5 (e) Bootstrap Estimates

```
# First, fit on observed
out <- glm(formula=O.ring.total/Number~Temp, family = binomial(link = logit),
           data=df, weights=Number)


# Estimated probabilities from observation-fitted model
df$pred <- predict(out,data=df$Temp, type = "response")

n <- length(df$Temp) #sample size of 23 , like original data
nboot <- 1000 #number of bootstrap samples
pi.star <- double(nboot) #array to store probability estimates

for (i in 1:nboot) { #for each bootstrap sample
    samp_df<-df[sample(nrow(df),size=n,replace=TRUE),]
    # Generate outcome for each temperature with the estimated probability
    samp_df$O.star <- rbinom(n, 6, samp_df$pred)
    # Fit new model on these generated outcomes
    out.star <- glm(O.star/Number ~Temp, family = binomial(link = logit),
                    data=samp_df,weights=Number)
    test=data.frame(Temp<-72.27) #test temperature
    # Predict probability of  O-ring failures for test temp.
    pi.star[i] <- predict(object=out.star,newdata=test, type = "response")

}
pi.star.72 <- mean(pi.star) #bootstrapped estimate of probability
ci.72 <- quantile(pi.star,c(.05,.95)) #90% confidence interval from bootstrap simulation

for (i in 1:nboot) { #for each bootstrap sample
    samp_df<-df[sample(nrow(df),size=n,replace=TRUE),]
    # Generate outcome for each temperature with the estimated probability
    samp_df$O.star <- rbinom(n, 6, samp_df$pred)
    # Fit new model on these generated outcomes
    out.star <- glm(O.star/Number ~Temp, family = binomial(link = logit),
                    data=samp_df,weights=Number)
    test=data.frame(Temp<-31) #test temperature
    # Predict probability of  O-ring failures for test temp.
    pi.star[i] <- predict(object=out.star,newdata=test, type = "response")

}
pi.star.31 <- mean(pi.star) #bootstrapped estimate of probability
ci.31 <- quantile(pi.star,c(.05,.95)) #90% confidence interval from bootstrap simulation
```

The 90% confidence interval was calculated using temperatures of 31°F and 72°F via bootstrap simulation. At each tempera- ture, the confidence interval is determined by choosing the simulated values that correspond to the 5th percentile and 95th percentile.

At 31°F, the 90% confidence interval of

$$\hat{\pi}*$$

from the simluated model is 0.1318595, 0.9928616 with an estimate of 0.7167264 . At 72°F, the estimate is 0.0364587 and the 90% confidence interval of

$$\hat{\pi}*$$

from the simluated model is 0.0107101, 0.0673745. We see a wide interval for 31°F, likely due to the lack of data points at the lower temperatures. For 72°F, the confidence interval is narrower in reflection to the data we have for higher temperatures.

5 (f) Determine if a quadratic term is needed in the model for the temperature.

As one last step to finalizing the model specification, we check if higher orders of temperature, particularly quadratic, contribute to the failure of O-rings.

```
mod.fit.Ha<-glm(formula=O.ring.total/Number~Temp+I(Temp^2),data=df,
                family=binomial(link = logit),weights = Number)
anova(mod_binomial.fit_temp,mod.fit.Ha,test="Chisq")
```

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

The Quadratic term fails to produce significant effect in change in residual deviance and so we fail to reject that the coefficient is actually 0 for the quadratic term. Hence we do not include it in our final logistic model.

IV.Model Selection: Fitting and Assessing a Linear Regression Model

Estimate Linear Regression Model and Diagnose/Interpret (3b)

3 (b). 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. Please explain.

```
mod.lm <- lm(O.ring.total/Number ~ Temp, data = df)
summary(mod.lm)
```

```
##
## Call:
## lm(formula = O.ring.total/Number ~ Temp, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09347 -0.06573 -0.01423  0.01760  0.31118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.616402   0.203252   3.033  0.00633 **
## Temp        -0.007923   0.002907  -2.725  0.01268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.09624 on 21 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.2261
## F-statistic: 7.426 on 1 and 21 DF,  p-value: 0.01268
```

The estimated linear regression model is

$$\hat{\pi} = 0.616402 - 0.007923\text{Temp}$$

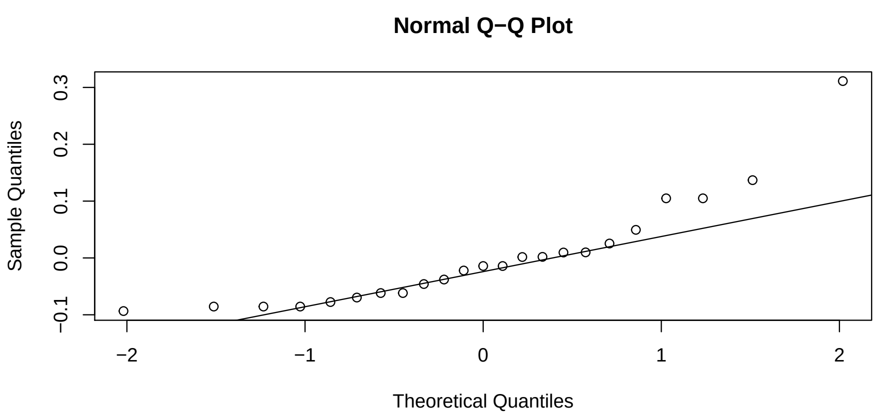
The assumptions we want to test/be on the look out for:

1. The model is linear in it's parameters.
2. The conditional mean of the errors is 0.
3. There is a random sampling of observations. *would need to investiage source since we select for pre-Challenger launches*
4. There is no multi-collinearity/perfect collinearity amongst explanatory variables. *pass:only one explanatory variable*
5. The errors have common constant variance (homoscedasticity).
6. The errors are independent of one another.
7. The errors are normally distributed. *not required for BLUE but do need for reliable inference*

We note since are observations are less than 30, we can't make asymptotic arguments for our parameter estimates. Moreover, for some of our tests, such as Shapiro, we will likely fail to reject due to lack of data and so we must test these assumptions from multiple angles (visualizations, etc.).

7. The errors are normally distributed.

```r
# Justification for these packages: used commonly in 203 to asses LR assumptions
# and found in Week 2 Analysis in 271
suppressPackageStartupMessages(library(lmtest))
suppressPackageStartupMessages(library(plm))


qqnorm(mod.lm$residuals)
qqline(mod.lm$residuals)
```




```r
shapiro.test(mod.lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod.lm$residuals
## W = 0.8233, p-value = 0.0009267
```

```r
coeftest(mod.lm, vcov=vcovHC(mod.lm))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6164021  0.2419719  2.5474  0.01875 *
## Temp        -0.0079233  0.0034482 -2.2978  0.03195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see both from the qq-plot that the distribution deviates from normality at the tail ends. We particularly see a positive skew present in our distribution. With a p-value 0.0009 for the Shapiro-Wilk test we can safely reject the null hypothesis that the residuals follow a normal distribution. Hence, we can't gaurantee precision on our standard errors. We can switch to robust standard errors.

1. The model is linear in it's parameters.

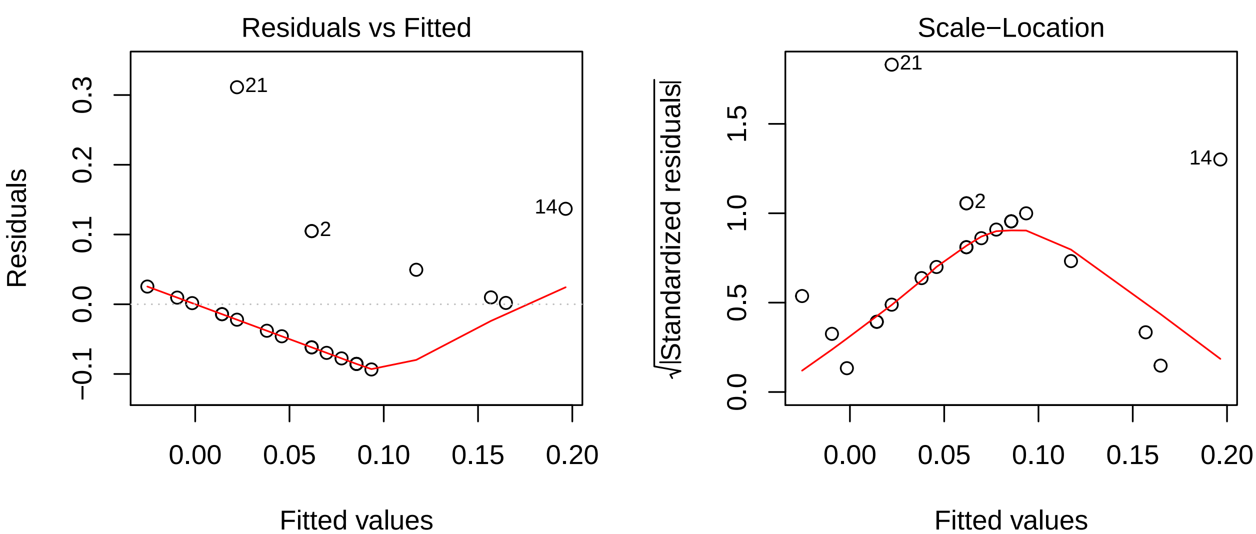
```r
summary(mod.lm$fitted.values)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.02538  0.02216  0.06177  0.06522  0.08554  0.19647
```

We see that the predictions are outside of the range for a probability (can't be negative). This is concerning as we want to assess risk of O-ring failure and probability would be well-suited for that task. Having a range that doesn't correspond gives us little sense of what's going on as the scale becomes relatively arbitrary.

2. The conditional mean of the errors is 0.

```r
par(mfrow=c(1,2))
#plot(mod.lm$fitted.values,mod.lm$residuals,xlab = 'Fitted Values',ylab='Residuals',title='Residuals vs
plot(mod.lm,which = c(1,3))
```



We see from the curvature in the residuals vs. fitted plot that there is cirvature present and that the zero conditional mean assumption is suspect.

5. The errors have common constant variance (homoscedasticity).

```r
ncvTest(mod.lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
```

```
## Chisquare = 0.1173805    Df = 1     p = 0.7318931
```

```r
bptest(mod.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.lm
## BP = 0.04457, df = 1, p-value = 0.8328
```

Both the traditional Breusch-Pagan test and the Studentized Breusch-Pagan test fail to reject the null hypothesis that the variance is homoskedastic. However the lack of even band in residuals vs fitted plot and curvature in scale-location plots suggests violation of this assumption though difficult to say due to sparsity of data. As we noted, at low sample size, these tests might have also lack the power.

6. The errors are independent of one another.

```r
durbinWatsonTest(mod.lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.08055664      2.115704   0.882
##  Alternative hypothesis: rho != 0
```

By the Durbin-Watson test, we fail to reject the null that the residuals are uncorrelated with one another. From the cross-sectional nature of the launch events, we can also feel more comfortable about this result. We also are able to reject the idea that the residuals correlate with our explanatory variables as well (slight but negligble curvature), giving some plausibility to 0 conditional mean of the errors.

After review of both models, we would opt for a logistic regression for several reasons. The linear regression model estimates that the percentage of O-ring failures will decrease by 0.007923 with every 1°F increase in temperature. That means, the probability of O-rings failure will be less than 0 when the temperature is roughly 78°F and above. This does not correspond with the physical picture. Also, the linear regression assumes a linear relationship between the probability of O-rings failure and the temperature which is not realistic. First, the output is more desirable in that we can compare odds of failures as well as compute actual probabilities of failure whereas the linear regression goes out of range. Moreover, the linear regression's questionability in terms of inference (failed normality of errors), likely keeps us restricted to the observed data in terms of predictions.

V. Interpreting the Results of the Final Model (3a)

Having explored the data, tested model specifications with lrt's as well as comparing to other possibilities such as the linear model, our final model is

$$\text{logit}(\hat{\pi}) = 5.08498 - 0.11560\text{Temp}$$

That is to say we picked the model with *Temp* as the only explanatory variable. We use this model to interpret the estimated Odds Ratio for failure of O-rings.

```r
exp(-10*mod_binomial.fit_temp$coefficients[2])
```

```
##     Temp
## 3.177236
```

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

```
## Waiting for profiling to be done...
```

```r
rev(exp(-10*beta.ci))
```

```
##   97.5 %    2.5 %
## 1.277239 8.350003
```

We look at the change in the odds of an O-ring failure when temperature decreases. From our calculations above, we see that for a 10°F decrease in temperature, the odds of an O-ring failure is about 3.18 times as high as the initial odds. With 95% confidence, the odds of an O-ring failure changes by an amount between 1.277 to 8.35 times for every 10°F decrease in temperature.

VI. Conclusion

We analyzed 23 pre-accident launches of the space shuttle and found that the probability of an O-ring failure at 31°F is rather high at about 0.818 with a 95% confidence interval of about (0.16,0.991). This is critical as this is the same temperature at which the Challenger failed and the reason for this investigation.

We made assumptions on independence of O-rings failing independently from one another and a combination of comparing to a binary failure model and the binomial model along with some deference to Dalal's analysis assured us the models are close enough to overcome potential problems. Since we only had 23 observations and the range of temperatures from 53°F to 81°F was relatively far away from the Challenger launch and limited, we obtained a rather wide confidence interval for probability of failure. For future analysis, we hope to analayze a wider range of conditions for failure.

The pressure variable was unable to evidentially produce a statistically significant effect to be included in our final model, even when attention was binarized to the 200 psi level. Again, collections of more observations may change this conclusion in the future. Finally, we find that a logistic regression model performs better than a linear regression model and provides a better rational explanation of the O-ring failures.