Lab 1 VHE

Victoria Eastman

September 19, 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 will use their dataset to mimic their study and attempt to determine the effect of key explanatory variables (temperature and pressure) on o-ring failure. In the end we specified a logistic regression model on temperature with the following formula:

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) Univariate Analysis

```
# 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,...
```

```
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 5 variables in our dataset:

- Flight: Flight number
- Temp: Temperature in F at launch
- Pressure: combustion pressure in psi at launch
- O.ring: number of primary field o-ring failures
- Number: total number of primary field o-rings

We are primarily interested in the effects of temperature and pressure on o-ring failure. Seeing as the total number of primary field o-rings does not change for our observations and we have no particular reason to see it change, we discard this variable due to lack of immediate use/differentiating behavior between failed and successful launches. Similarly, we discard flight number due to lack of any immediate use.

We can see that the `O.ring` variable has 3 distinct values: 0, 1, and 2. We are going to be nuanced in our analysis and say we want to find the conditions that lead to *at least one* o-ring failure. Therefore, we will recategorize those flights with 2 failures as having 1 failure for the purposes of this study.

```
# We don't want to eliminate raw data
df$O.ring.total = df$O.ring
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.

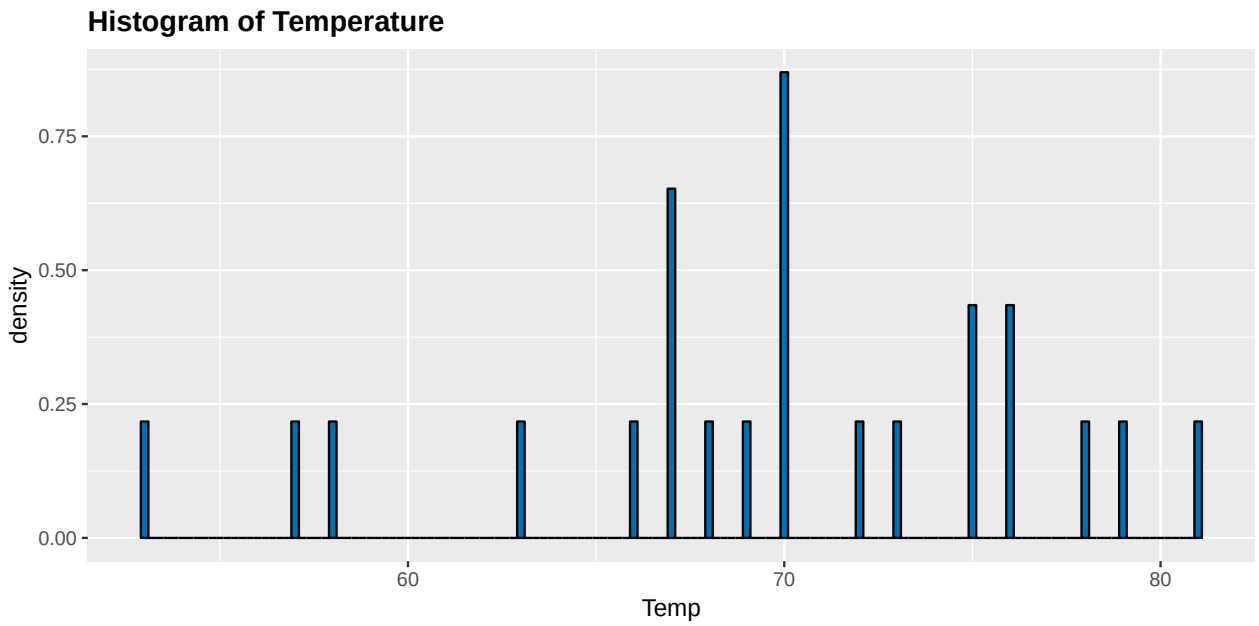
```
#quick check for potential collinearity as well as surface level relations
cor(df[,c("Temp", "Pressure", "O.ring")])
```

```
##                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. We see negligible positive correlation between temperature and pressure and thus don't worry about colliniearity.

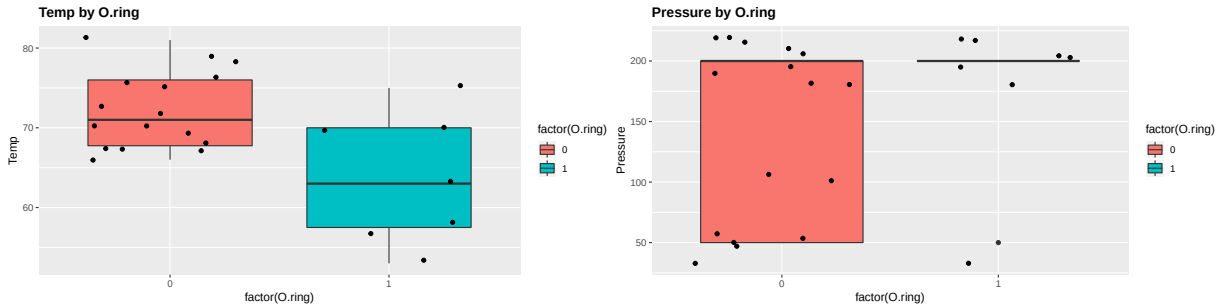
```
#What does the distribution of temperatures look like?
ggplot(df, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill="#0072B2", colour="black") +
  ggtitle("Histogram of Temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



The distribution of the temperature explanatory variable looks to faintly resemble a normal distribution with a very slight negative skew. Due to this, we see no compelling reason to take a log transformation of the variable at this stage. The mode of temperatures is 70 and the range is 53, 81.

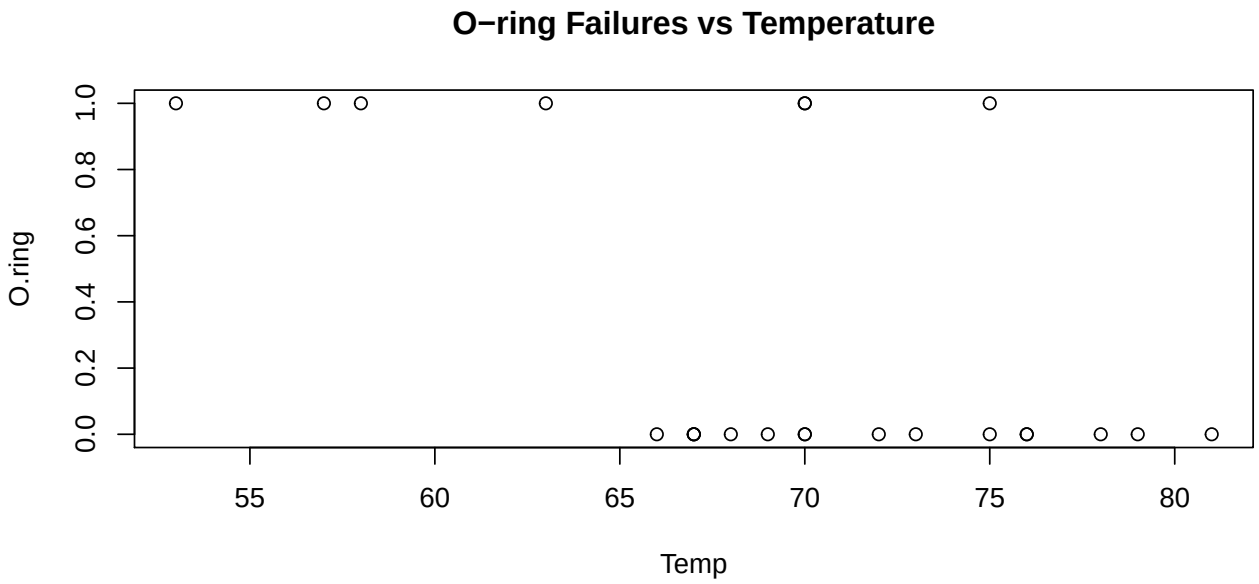
II (b) Bivariate Analysis

```
#group each explanatory variable by O.ring failure
for (i in 2:3){
  print(ggplot(df, aes(factor(O.ring), df[,i])) +
          geom_boxplot(aes(fill = factor(O.ring))) +
          geom_jitter() +
          ggtitle(paste0(colnames(df)[i], " by O.ring")) + ylab(colnames(df)[i]) +
          theme(plot.title = element_text(lineheight=1, face="bold")))
}
```



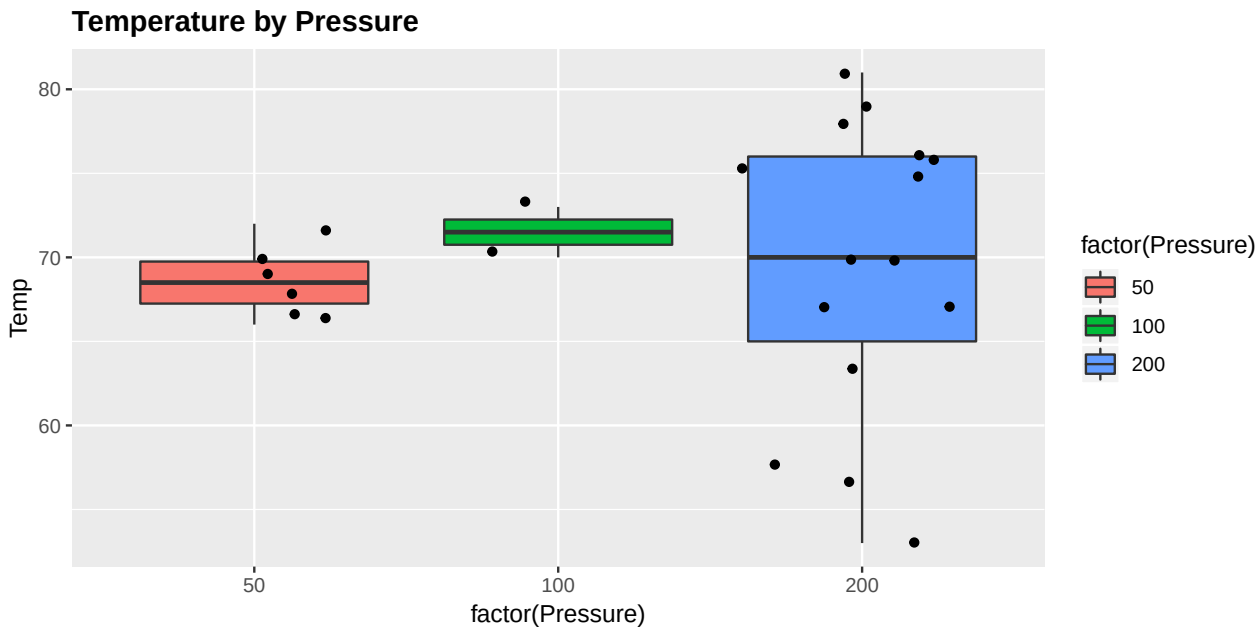
The boxplot above of temperature grouped by o-ring failure shows the average temperature when o-rings failed was lower than when they did not. Also, the overlay scatterplot shows the mismatch of data: there's much more data for non-failures than failures. The second boxplot for pressure shows that all but 2 o-ring failures occured under high pressure conditions. Also, for non-failures, the mean and 75th percentile are overlapping, indicating that the data is skewed.

```
plot(O.ring~Temp,data=df, main="O-ring Failures vs Temperature")
```

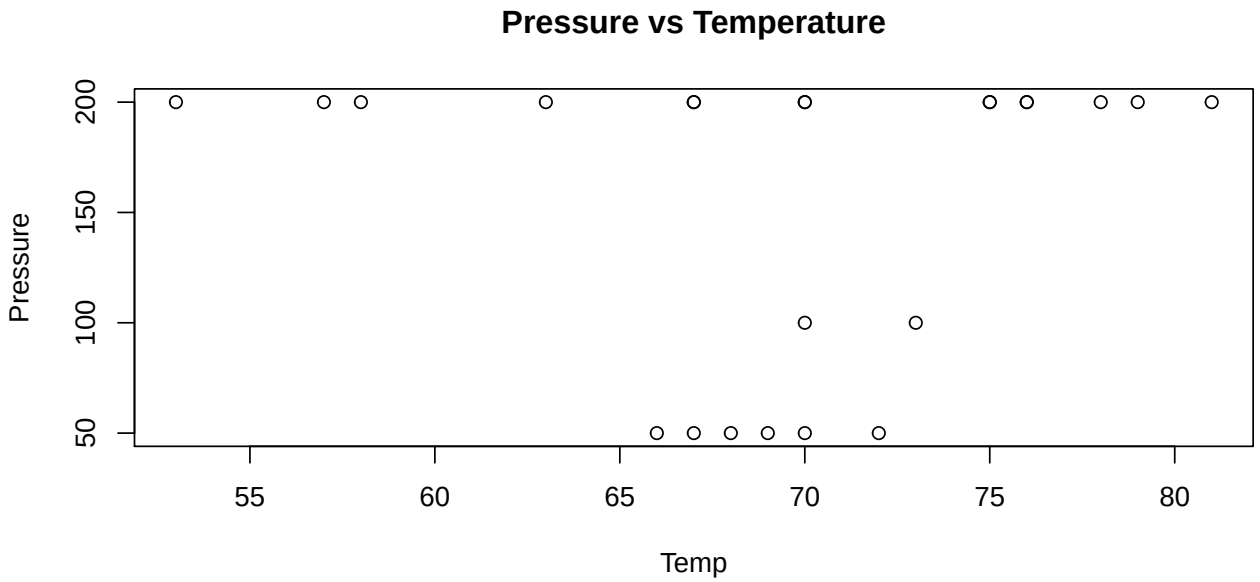


Directly looking at the distribution of o-ring failures to temperature, we can determine our data is not completely separated (hence traditional logistic regression is still a valid possibility for model selection).

```
#how much does temperature vary by each pressure stage
ggplot(df, aes(factor(Pressure), Temp)) +
          geom_boxplot(aes(fill = factor(Pressure))) +
          geom_jitter() +
          ggtitle( "Temperature by Pressure") +
          theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
plot(Pressure~Temp,data=df, main="Pressure vs Temperature")
```

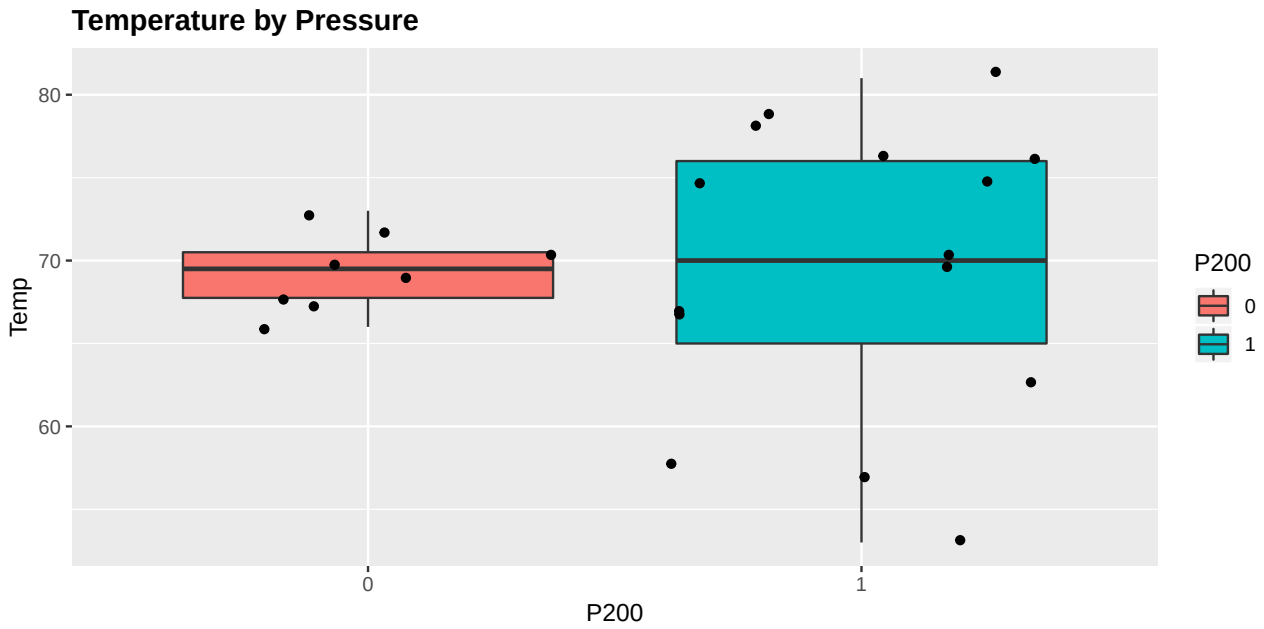


Comparing temperature and pressure directly, we can see that there is a vaguely positive relationship between temperature and pressure. Under basic gas laws, temperature is proportional to pressure, however, this relationship doesn't appear to hold in all cases of o-ring failure.

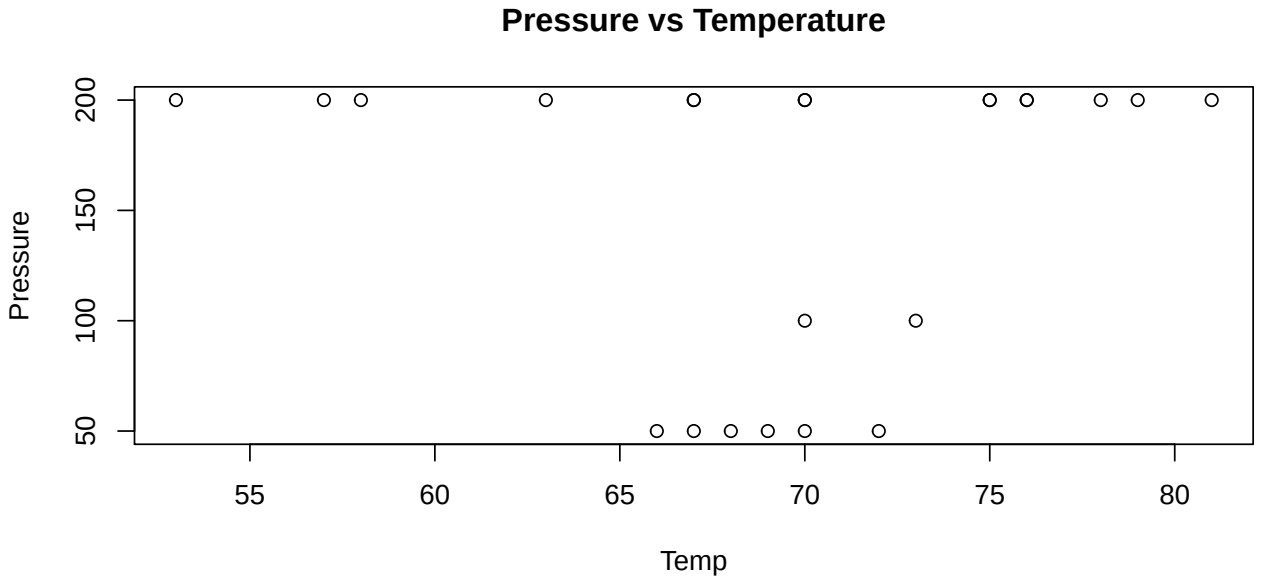
```
#bin pressure to 200 or not
df$P200=df$Pressure
df$P200[df$P200!=200]=0
df$P200[df$P200==200]=1
df$P200=factor(df$P200)

#how much does temperature vary by each pressure stage
ggplot(df, aes(P200, Temp)) +
          geom_boxplot(aes(fill = P200)) +
          geom_jitter() +
```

```
        ggtitle( "Temperature by Pressure") +
        theme(plot.title = element_text(lineheight=1, face="bold"))
```

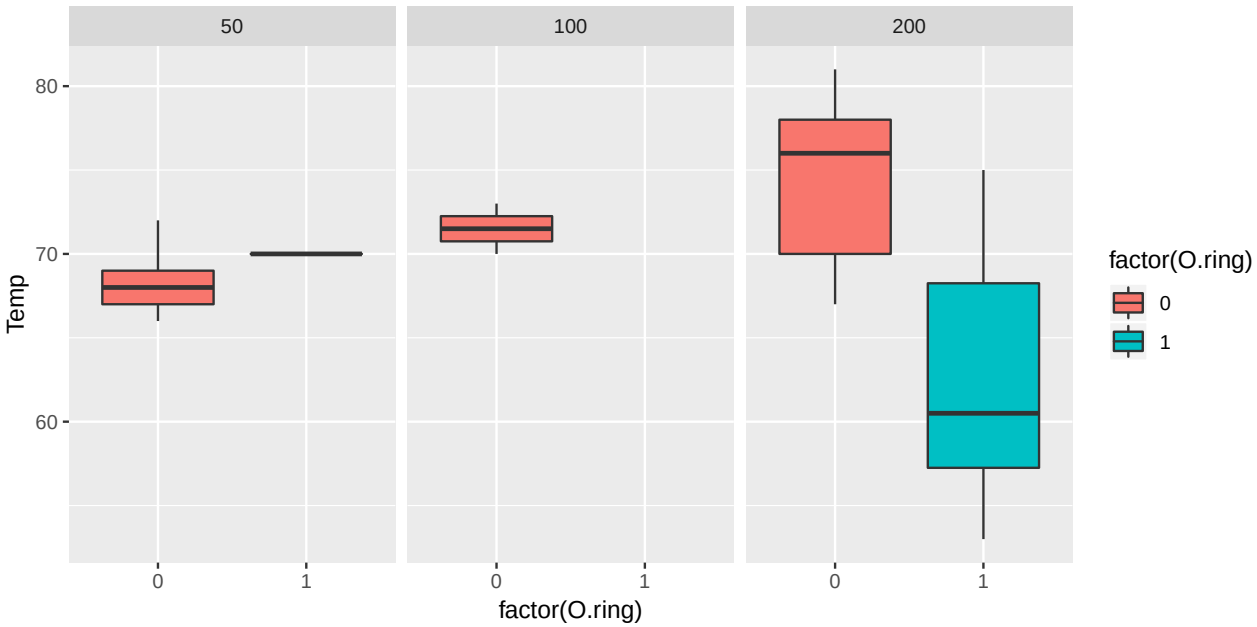



```
plot(Pressure~Temp,data=df, main="Pressure vs Temperature")
```



Above we isolate the data points where pressure is 200 psi. The right boxplot (when pressure is 200 psi) has a much wider temperature range then when pressure is either 50 psi or 100 psi. This could be due to a lack of data points when the pressure is not 200 psi.

```
#slicing by pressure, any distinct relations between failure and temperature?
ggplot(df, aes(x = factor(O.ring), y = Temp, fill = factor(O.ring))) + geom_boxplot() +
facet_wrap(~ Pressure, ncol = 3)
```



Combining all three variables into the above boxplot, we can see that almost all o-ring failures occured at 200 psi with much lower temps than those that did not fail. Without an overlay of data points, we aren't able to determine how many points make up each of these ranges.

III. Book Questions

Question 4

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

This independence assumption is necessary for deriving the likelihood-based solution as we can take products of the probabilities. 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. Moreover damage in one O-ring could affect the probability of damage in subsequent O-rings (might be easier for the system to collapse as a whole).

4 (b) Estimate logistic regression model

```
# Initial model
mod.fit1<-glm(formula=O.ring~Pressure+Temp,data=df,family=binomial(link = logit))
summary(mod.fit1)
```

```
##
## Call:
## glm(formula = O.ring ~ Pressure + Temp, 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 .
```

```
## Pressure     0.010400   0.008979   1.158   0.2468
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## ---
## 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
```

```
#mod.fit2<-glm(formula=O.ring~Pressure+Temp+Pressure:Temp,data=df,family=binomial(link = logit))
```

Our first estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = 13.292360 + 0.0104\text{Pressure} - 0.228671\text{Temp}$$

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

```
Anova(mod.fit1,Test='LRT')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring
##          LR Chisq Df Pr(>Chisq)
## Pressure   1.5331  1   0.215648
## Temp       7.7542  1   0.005359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test evaluates the importance of each explanatory variable and interaction 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 is $-2log(\Lambda) = 1.5331$ and the p-value is 0.215648 so we fail to reject the null hypothesis that pressure has no effect on o-ring failure. For the second explanatory variable, temperature, we test the hypothesis: $H_0 : \beta_2 = 0$ vs $H_a : \beta_2 \neq 0$. The test statistic is $-2log(\Lambda) = 7.7542$ and the p-value is 0.005359 so we reject the null hypothesis and conclude that there is evidence of an interaction between temperature and o-ring failure.

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. Potential problems could be losing precision on temp as well as in probability, especially for edge cases.

Question 5

Next we estimate the model with only one regressor: Temperature.

5 (a) Estimate the model

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

```r
mod.fit2<-glm(formula=O.ring~Temp,data=df,family=binomial(link = logit))
summary(mod.fit2)
```

```
##
## 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 estimated logistic regression model is

$$\text{logit}(\hat{\pi}) = 15.0429 - 0.2322\text{Temp}$$

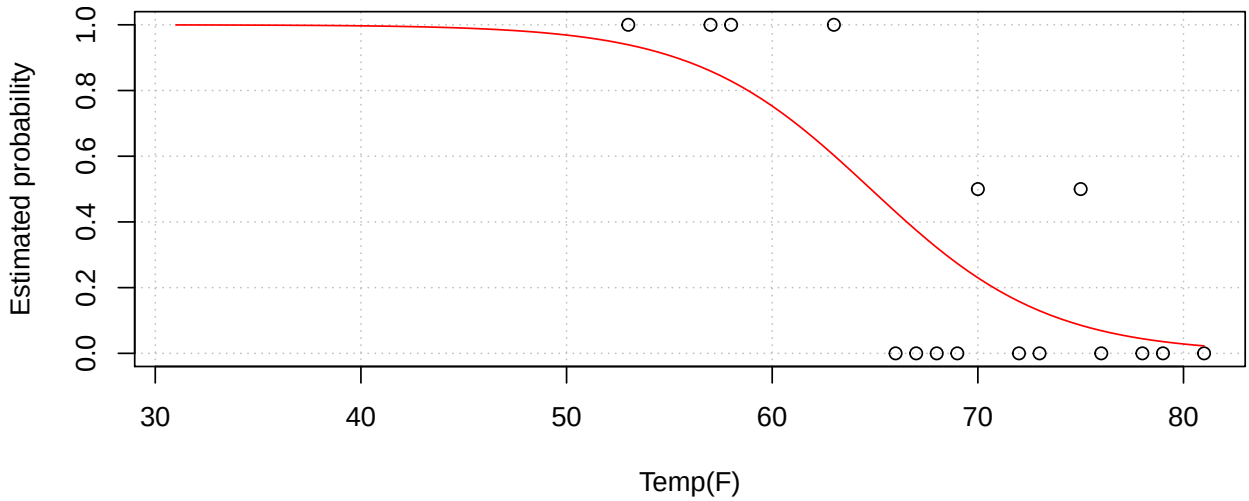
5 (b)

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.

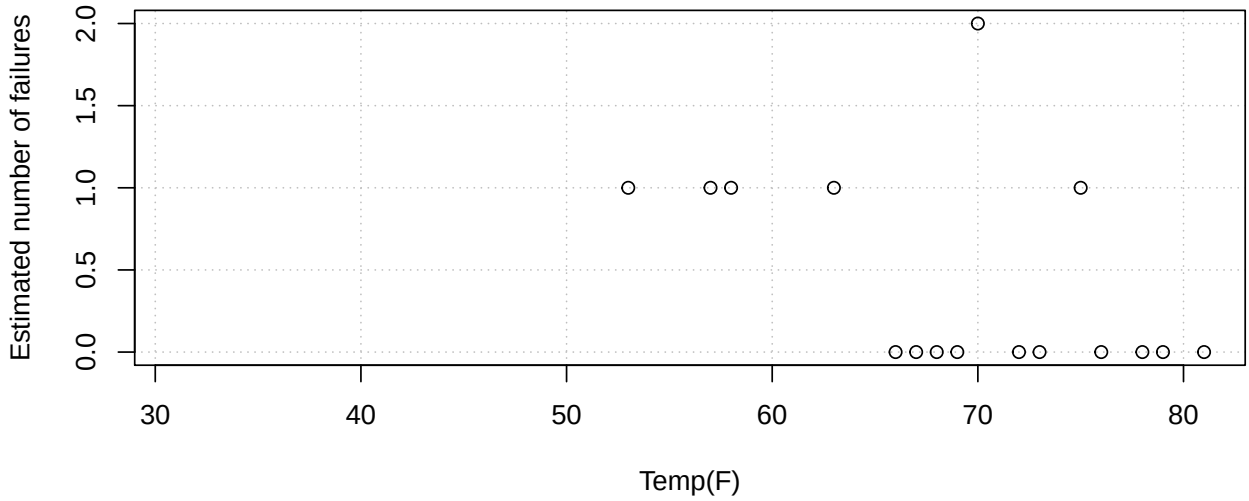
```r
#
w<-aggregate(formula=O.ring~Temp, data=df, FUN=sum)
n<-aggregate(formula=O.ring~Temp, data=df, FUN=length)
w.n<-data.frame(Temperature=w$Temp, Failure=w$O.ring, trials=n$O.ring, proportion=round(w$O.ring/n$O.ri
head(w.n)
```

```
##   Temperature Failure trials proportion
## 1          53       1      1          1
## 2          57       1      1          1
## 3          58       1      1          1
## 4          63       1      1          1
## 5          66       0      1          0
## 6          67       0      3          0
```

```r
# Plot pi vs Temp
plot(x=w$Temp, y=w$O.ring/n$O.ring, xlab="Temp(F)", ylab="Estimated probability", panel.first=grid(col=
curve(expr=predict(object=mod.fit2, newdata=data.frame(Temp=x), type="response"), col="red", add=TRUE)
```




```
# Plot Expected number of failures vs.Temp.
plot(x=w$Temp, y=w$O.ring, xlab="Temp(F)", ylab="Estimated number of failures", panel.first=grid(col="g
```




```
#curve(expr=predict(object=mod.fit2, newdata=data.frame(Temp=x), type="response"), col="red", add=TRUE)
```

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?

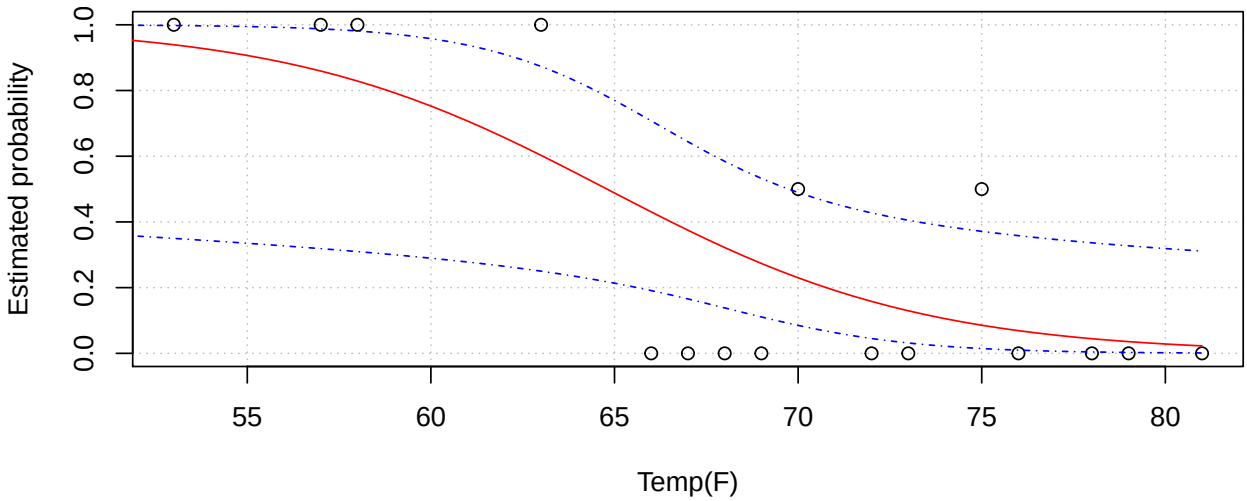
```
plot(x=w$Temp,y=w$O.ring/n$O.ring,xlab="Temp(F)", ylab = "Estimated probability", panel.first =grid(col
curve(expr=predict(object=mod.fit2,newdata=data.frame(Temp = x), type = "response"), col = "red", add=T

ci.pi<-function(newdata,mod.fit.obj,alpha){linear.pred <- predict(object = mod.fit.obj, newdata =newdata
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.fit2, alpha = 0.05)$lower, col = "blue",

curve(expr=ci.pi(newdata=data.frame(Temp=x),mod.fit.obj = mod.fit2, alpha = 0.05)$upper, col = "blue",
```



Bands are wider due to change in probability across temperature gradient. There is a much steeper drop in temperature below and above 65 (similar to complete separation problem). Less of a drastic change in higher temperatures due to two "middle" values between 70 and 75. We also have fewer observations for the low temperature range.

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.

```
alpha=0.05
# Set data to predict on
predict.data <- data.frame(Temp=31)
# Linear part of model
linear.pred=predict(object = mod.fit2, newdata = predict.data,
                    type = "link", se = TRUE)
# Estimate probability
pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit))
# Confidence interval before exponentiation
CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se
# Actual interval
CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred))
```

The estimated probability of o-ring failure at 31 F is 0.9996 and the corresponding confidence interval is 0.4816, 1.

5 (e)

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

```
##
## 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
```

```r
# 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$O.star <- rbinom(n, 1, samp_df$pred)
    # Fit new model on these generated outcomes
    out.star <- glm(O.star ~Temp, family = binomial(link = logit),data=samp_df)
    test=data.frame(Temp<-72.27) #test temperature
    # Predict probability of at least one O-ring failure 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$O.star <- rbinom(n, 1, samp_df$pred)
    # Fit new model on these generated outcomes
    out.star <- glm(O.star ~Temp, family = binomial(link = logit),data=samp_df)
    test=data.frame(Temp<-31) #test temperature
    # Predict probability of at least one O-ring failure 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 bootstrapped 90% confidence interval for 31°F is 0.9558, 1 and for 72.27°F is 0, 0.

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

```
mod.fit.Ha<-glm(formula=O.ring~Temp,data=df,family=binomial(link = logit))
anova(mod.fit2,mod.fit.Ha,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Temp
## Model 2: O.ring ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        21     20.315  0        0
```

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.

IV. Estimate Linear Regression Model

3 (a). Interpret the main result of your final model in terms of both odds and probability of failure

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 ~ Temp+P200, data = df)
summary(mod.lm)
```

```
##
## Call:
## lm(formula = O.ring ~ Temp + P200, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50074 -0.18773 -0.08518  0.07481  0.89861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74586    0.81119   3.385  0.00294 **
## Temp        -0.03778    0.01153  -3.277  0.00377 **
## P2001        0.28602    0.16708   1.712  0.10239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3816 on 20 degrees of freedom
## Multiple R-squared:  0.402,  Adjusted R-squared:  0.3422
## F-statistic: 6.723 on 2 and 20 DF,  p-value: 0.005847
```

The estimated linear regression model is

$$\hat{\pi} = 2.658765 + 0.001962\text{Pressure} - 0.038136\text{Temp}$$

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

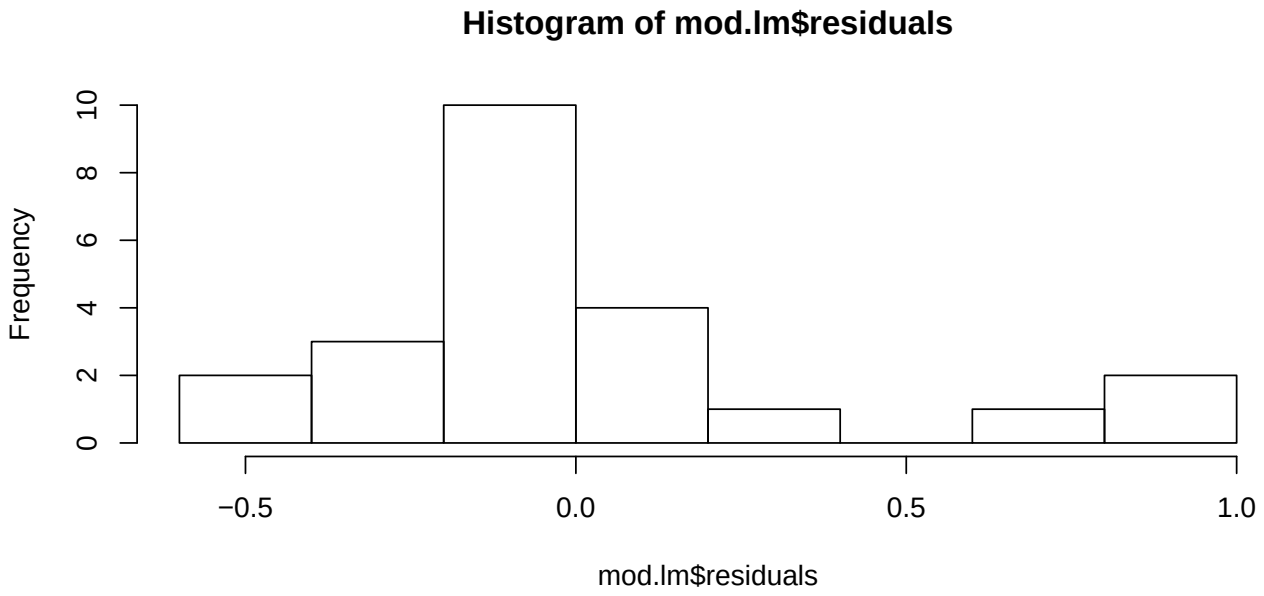
1. The model is linear in it's parameters.
2. The conditional mean of the errors is 0.
3. There is a random sampling of observations. *don't have choice here*
4. There is no multi-collinearity/perfect collinearity amongst explanatory variables. *passed in EDA*
5. The errors have common constant variance (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.).

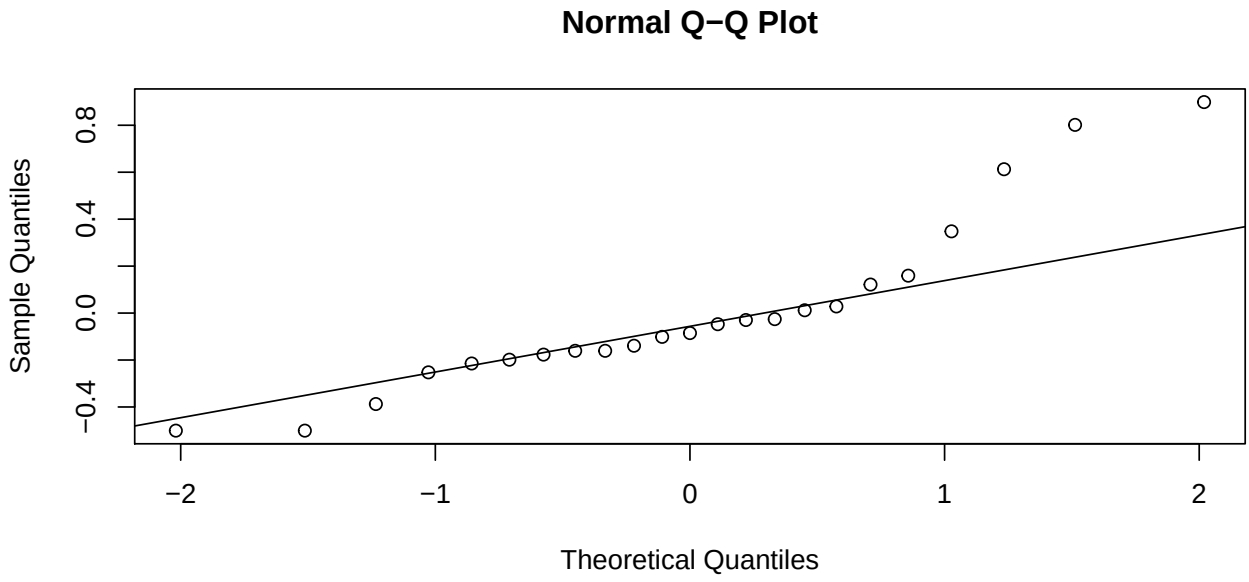
7. The errors are normally distributed.

```
# Justification for these packages: used commonly in 203 to asses LR assumptions
suppressPackageStartupMessages(library(lmtest))
suppressPackageStartupMessages(library(plm))
suppressPackageStartupMessages(library(gvlma))

hist(mod.lm$residuals)
```




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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  mod.lm$residuals
## W = 0.87805, p-value = 0.00917
```

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

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.7458627  0.4653069  5.9012 9.003e-06 ***
## Temp        -0.0377782  0.0064335 -5.8721 9.600e-06 ***
## P2001        0.2860186  0.1739501  1.6443    0.1158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

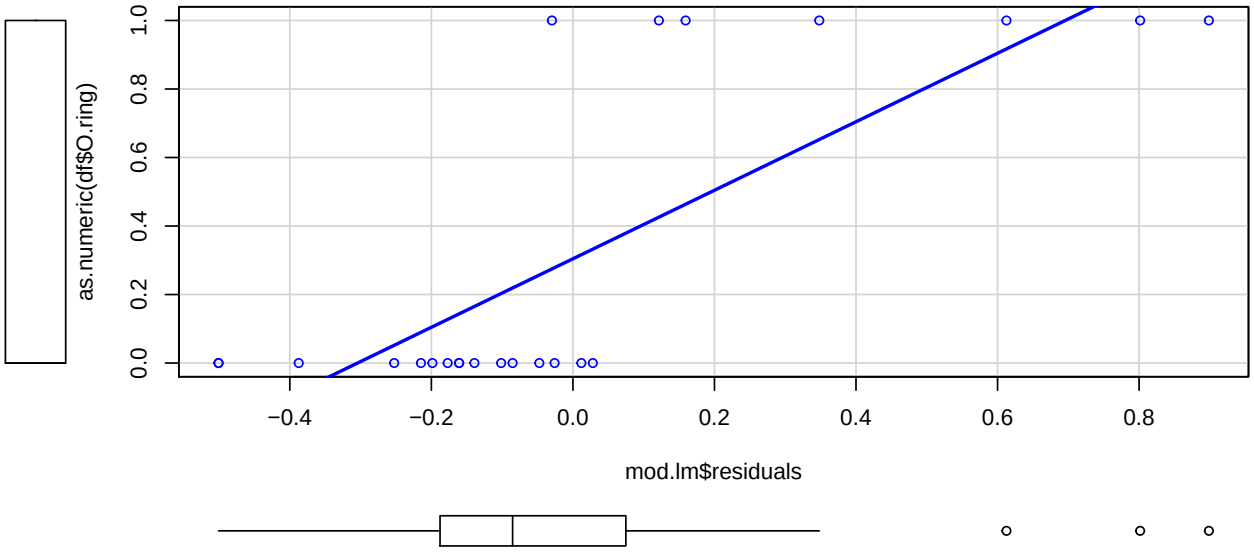
We see both from the histogram and the qq-plot that the distribution deviates from normality at the tail ends. We particularly see a positive skew present in our distribution. With a p-value 0.009 for the Shapiro-Wilk test we can safely reject the null hypothesis that the residuals follow a normal distribution. Hence, we can't gaurantee precision on our standard errors. We can switch to robust standard errors as done in the last line.

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.02815  0.10139  0.19852  0.30435  0.44407  1.02964
```

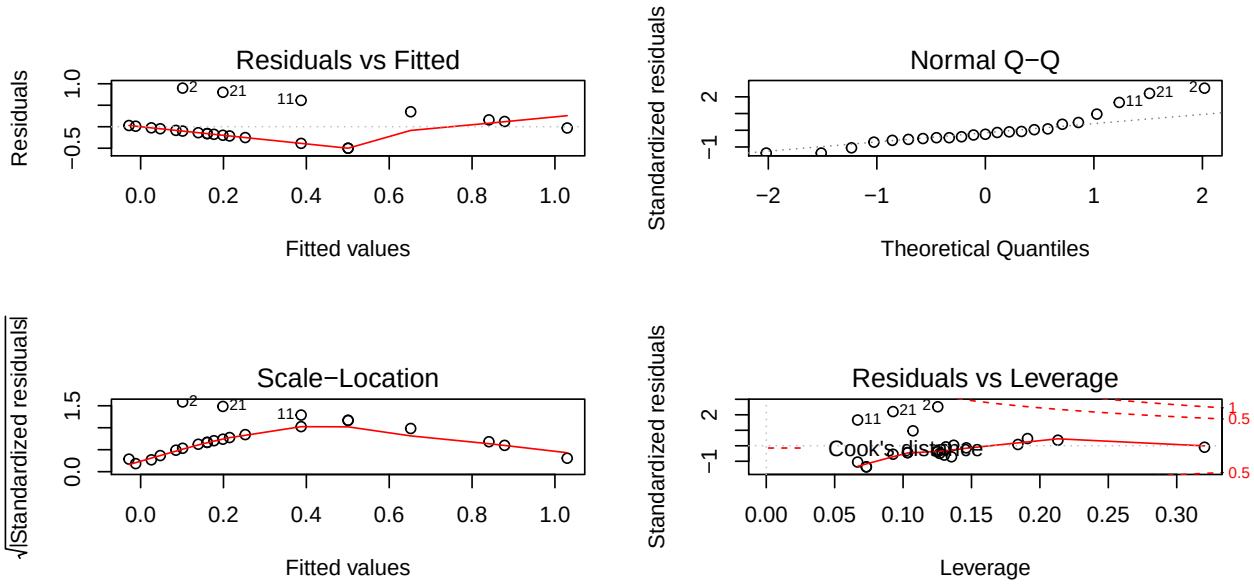
We see that the predictions are outside of the range for a probability. This is concerning as we want to assess risk of O-ring failure and probability would be well-suited for that task. Having a range that doesn't correspond gives us little sense of what's going on as the scale becomes relatively arbitrary.

```
scatterplot(mod.lm$residuals,as.numeric(df$O.ring))
```



We already knew from the binary nature of the outcome that a continuous seemed innapropriate, however the line of residuals vs the O.ring failure outcome does manage to somewhat separate distinct cases (we see 4 of the observations are on the "wrong" side of the decision boundary). A linear relation seems roughly justified.

```
par(mfrow=c(2,2))
plot(mod.lm)
```



We

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

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

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.04975908, Df = 1, p = 0.82348
```

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

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.lm
## BP = 0.066718, df = 2, p-value = 0.9672
```

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.

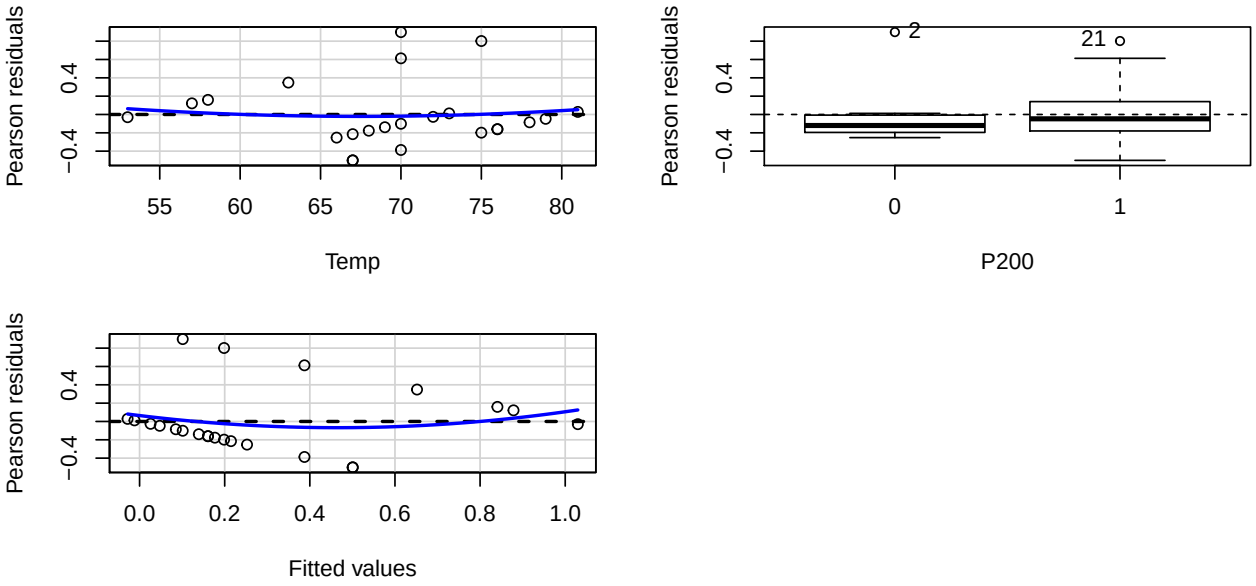
2. The conditional mean of the errors is 0.

6. The errors are independent of one another.

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01573965      2.000875   0.748
##  Alternative hypothesis: rho != 0
```

```
par(mfrow=c(1,1))
residualPlots(mod.lm)
```




```
##            Test stat Pr(>|Test stat|)
## Temp          0.3503           0.7300
## P200
## Tukey test    0.6227           0.5335
```

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.

```
gv.mod.lm <- gvlma(mod.lm)
summary(gv.mod.lm)
```

```
##
## Call:
## lm(formula = O.ring ~ Temp + P200, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50074 -0.18773 -0.08518  0.07481  0.89861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74586    0.81119   3.385  0.00294 **
## Temp        -0.03778    0.01153  -3.277  0.00377 **
## P2001        0.28602    0.16708   1.712  0.10239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3816 on 20 degrees of freedom
## Multiple R-squared:  0.402,  Adjusted R-squared:  0.3422
## F-statistic: 6.723 on 2 and 20 DF,  p-value: 0.005847
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mod.lm)
##
##                      Value p-value                   Decision
## Global Stat        5.86862 0.20918    Assumptions acceptable.
## Skewness           4.77997 0.02879 Assumptions NOT satisfied!
## Kurtosis           0.52919 0.46695    Assumptions acceptable.
## Link Function      0.45997 0.49764    Assumptions acceptable.
## Heteroscedasticity 0.09949 0.75245    Assumptions acceptable.
```

From the gvlma we are added some assurance in that Global Stat- fail to reject relationship between X and Y are roughly linear, 4.failed to reject the Link function was appropriate , 5. failed to reject variance of the residuals seem constant. From Skewness, we see a potential need for variable transformation but we need to keep the model comparable to the logistic version.

After review of both models, we would opt for a logistic regression for several reasons. 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. Conclusion

In conclusion, we found that