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.

```r
# 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

```r
# 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 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
```

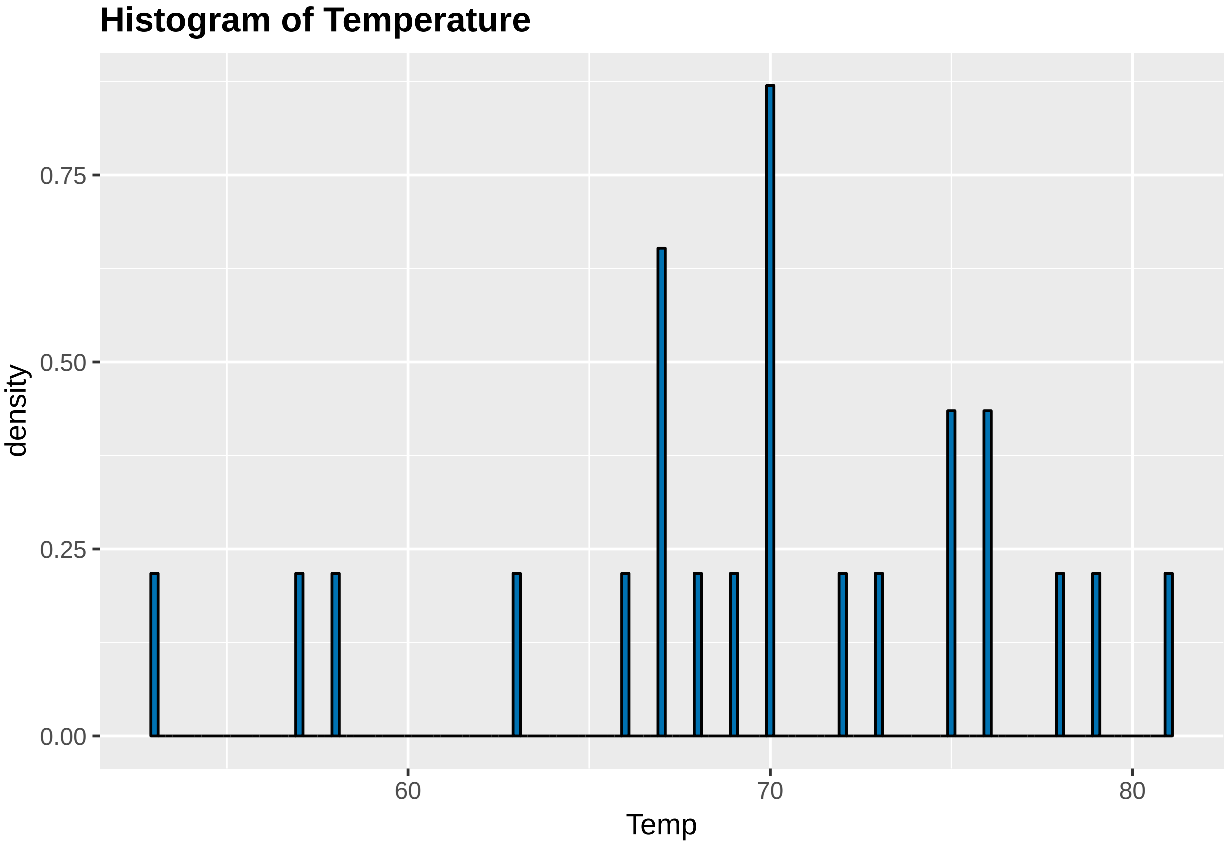
In addition, we see that the dataset contains 23 data points from other shuttle launches and none of the variables are missing any entries. Interestingly, pressure is generally considered to be a continuous variable, however, we see three distinct values of 50, 100, and 200. We could potentially see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

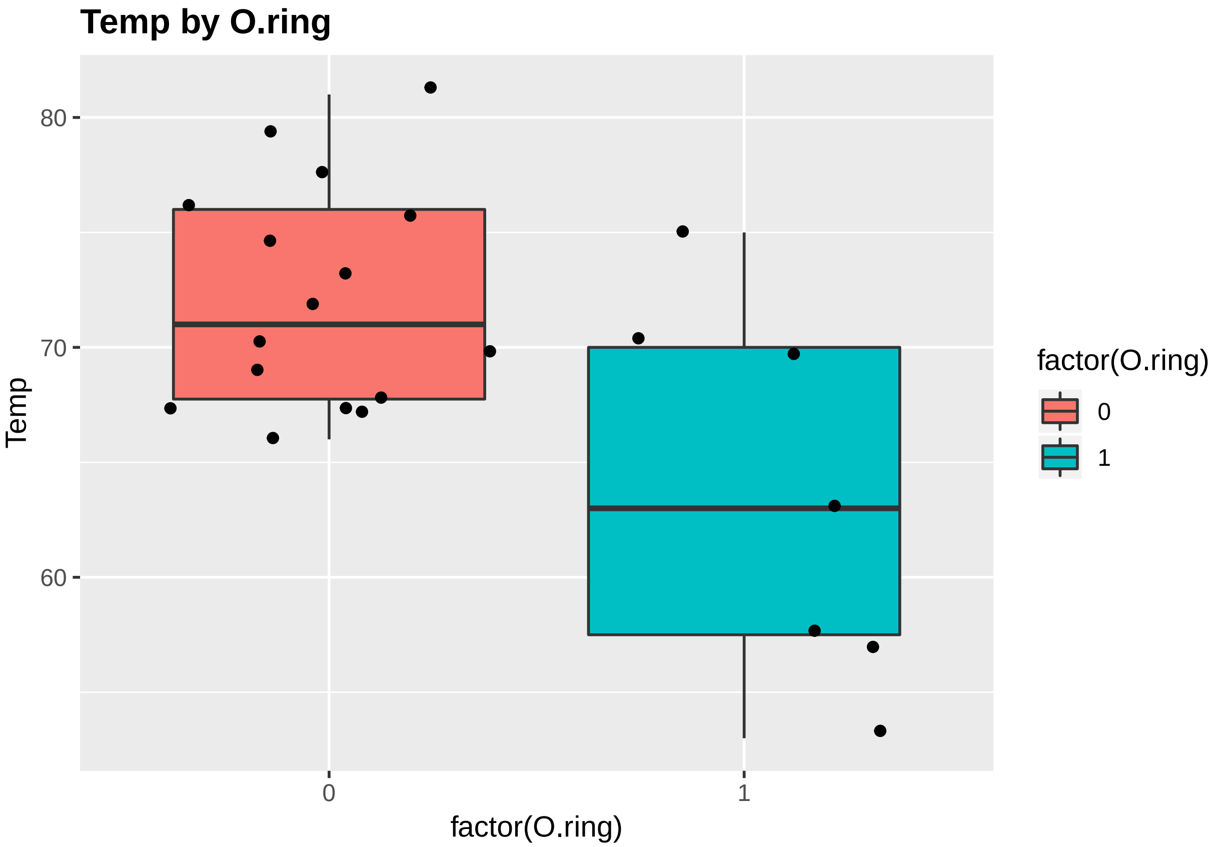
```
#What does the distribution of temperatures look like?
ggplot(df, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill="#0072B2", colour="black") +
  ggtitle("Histogram of Temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```

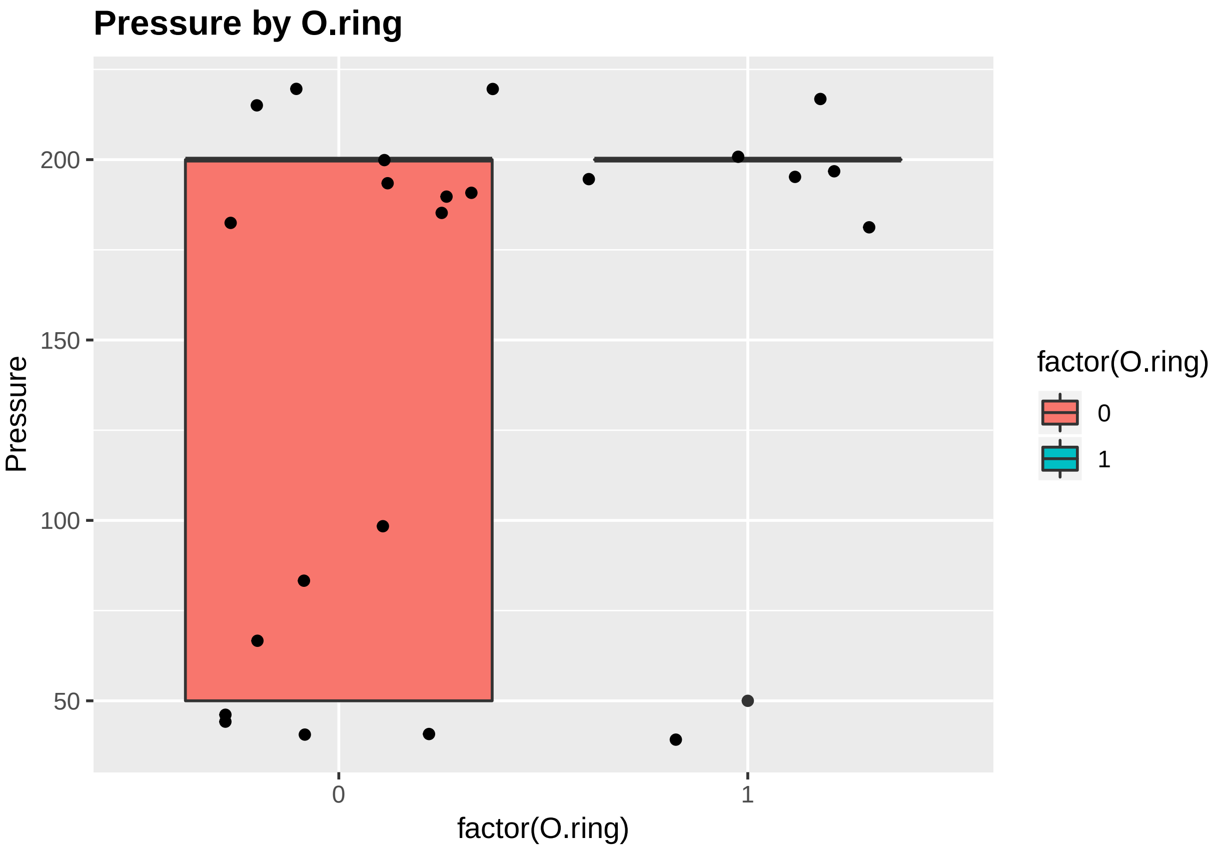


The distribution of the temperature explanatory variable looks to faintly resemble a normal distribution with a very slight negative skew. Due to this, we see no compelling reason to take a log transformation of the variable at this stage. The mode of temperatures is 70 and the range is 53, 81.

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")))
}
```

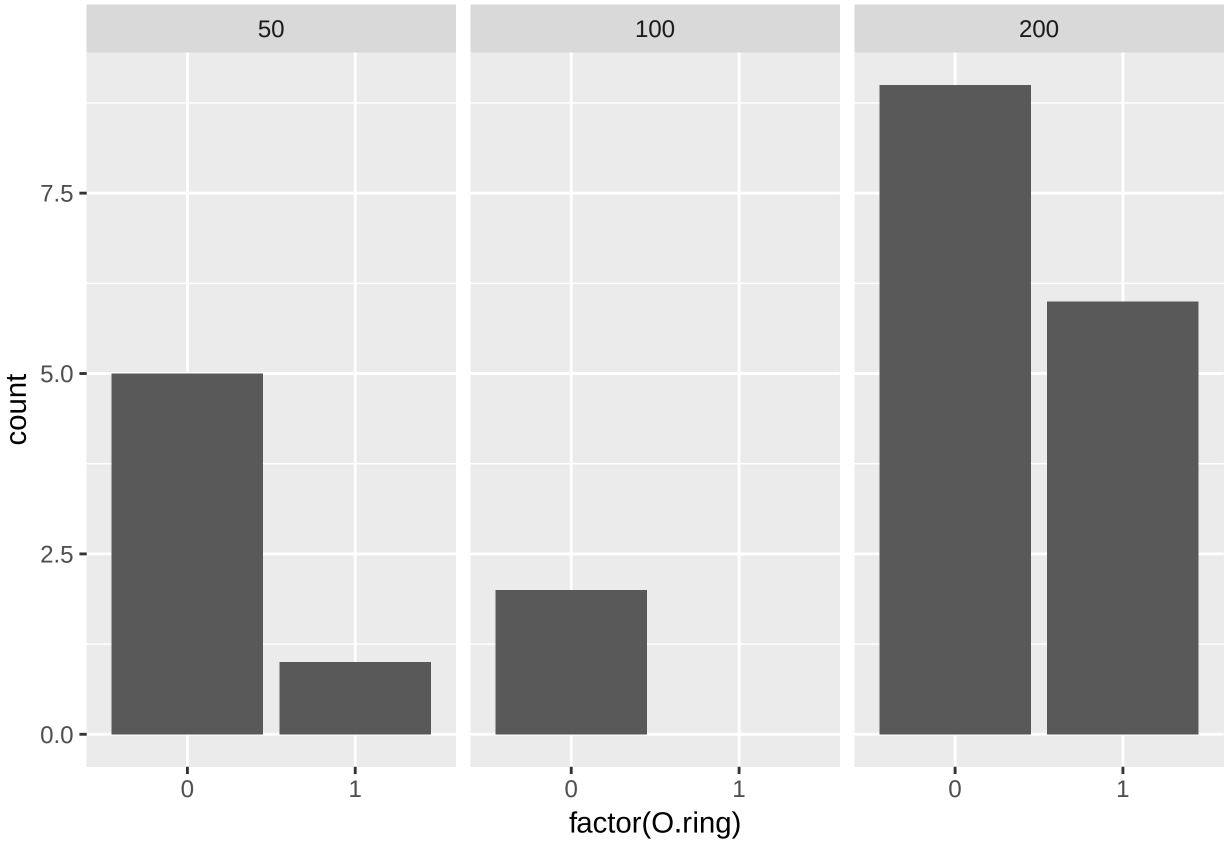




The boxplot above of temperature grouped by o-ring failure shows the average temperature when o-rings failed was lower than when they did not. Also, the overlay scatterplot shows the mismatch of data: there's much more data for non-failures than failures. The second boxplot for pressure shows that all 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.

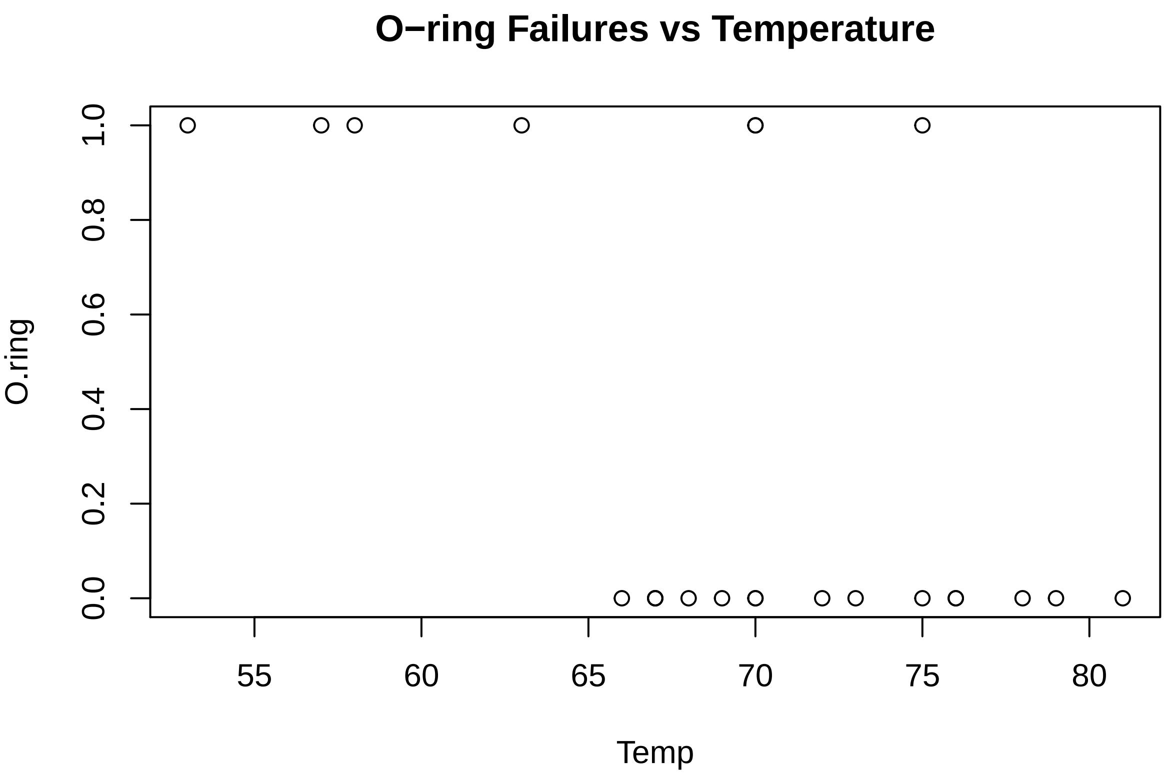
```
#How do failures vary by pressure level?
ggplot(df,aes(x=factor(O.ring))+geom_histogram(stat='count')+facet_grid(~Pressure)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



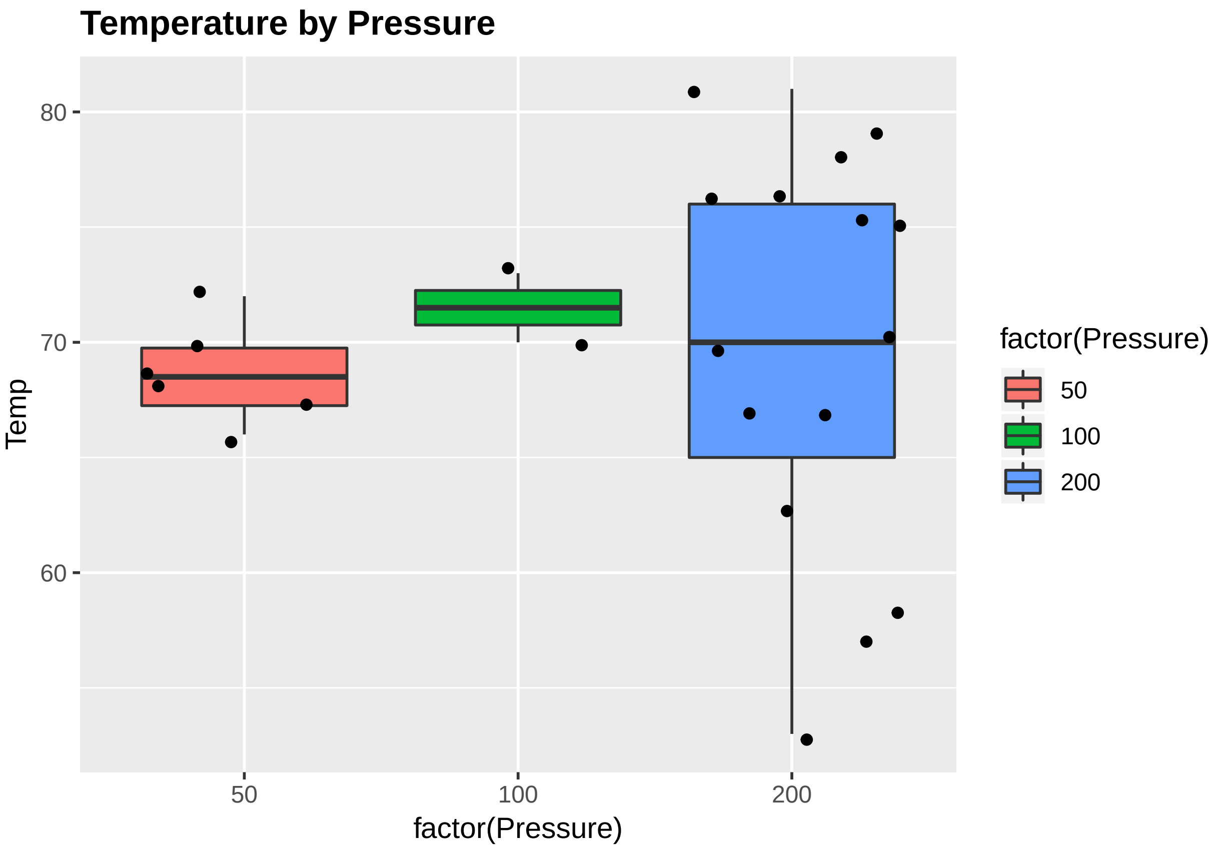
The above plot reinforces the finding from above that the data has a negative skew with most of the data for both failures and non-failures occuring under pressure of 200 psi.

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

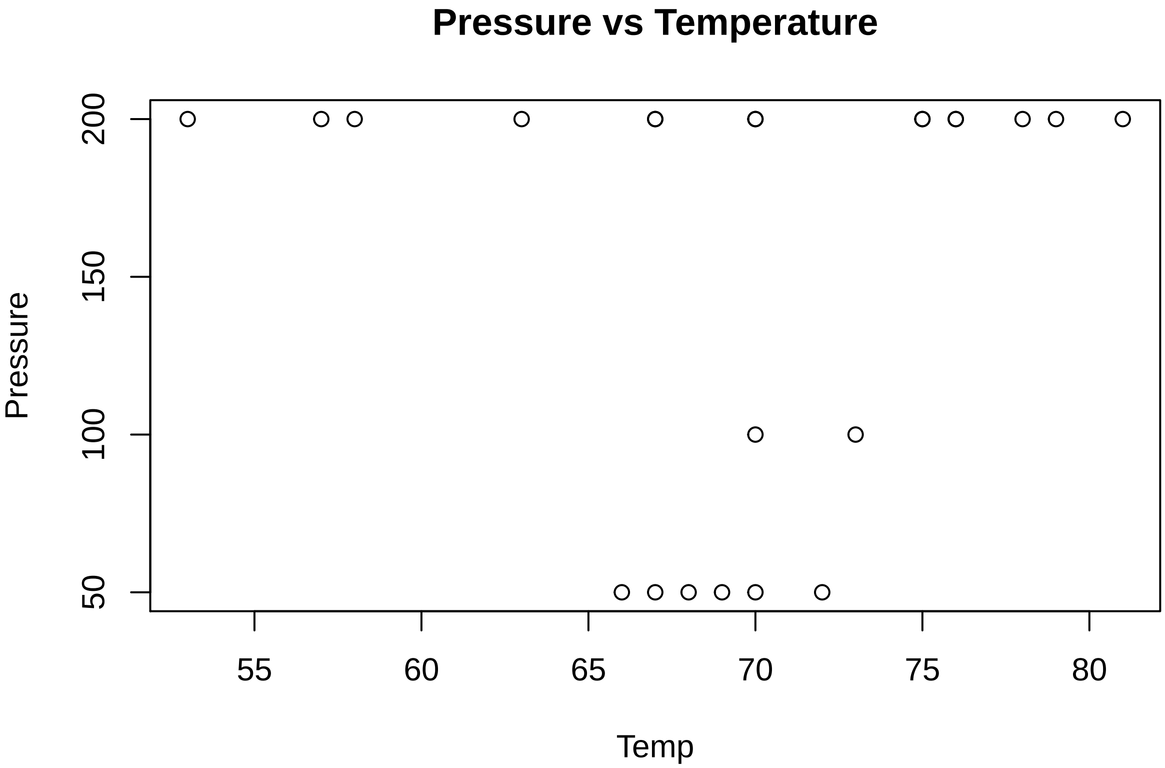


Directly looking at the distribution of o-ring failures to temperature, we can determine our data is not completely separated (hence traditional logistic regression is still a valid possibility for model selection).

```r
#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"))
```



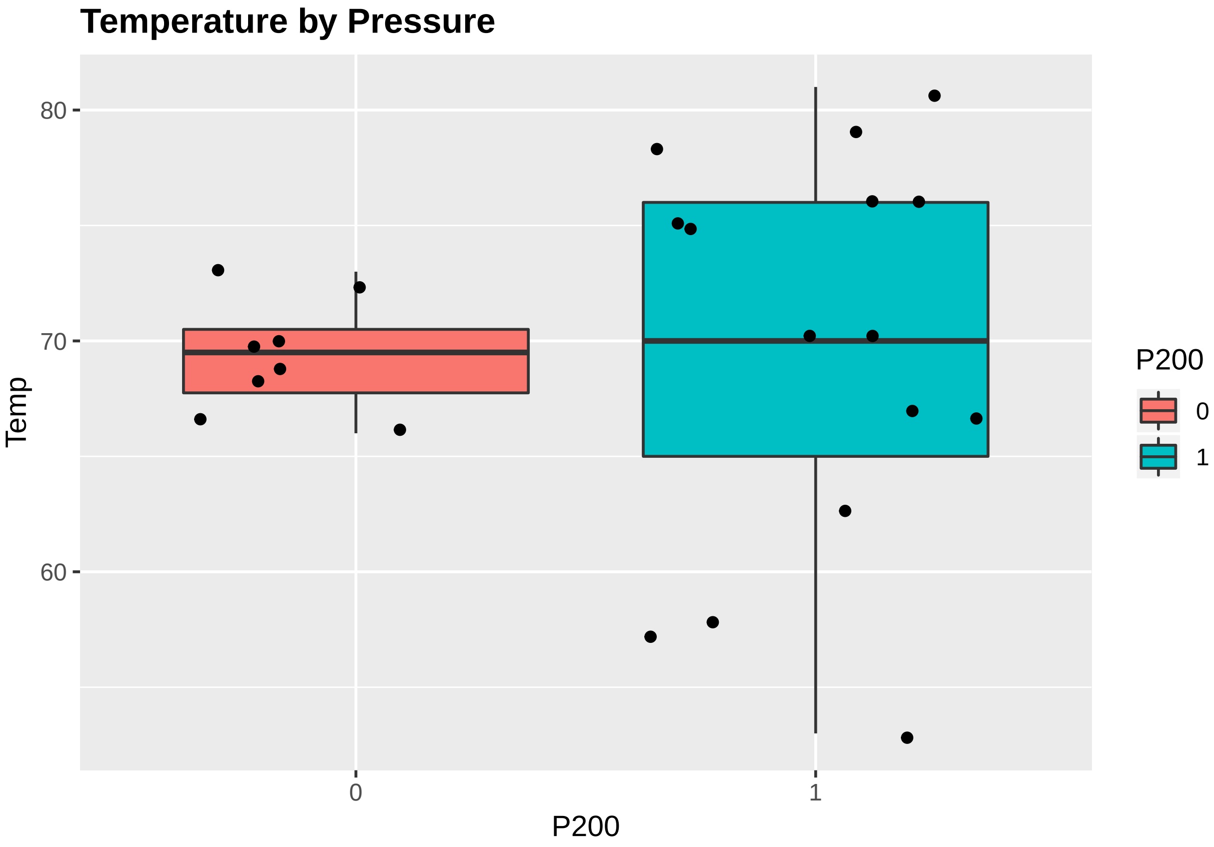
```r
plot(Pressure~Temp,data=df, main="Pressure vs Temperature")
```



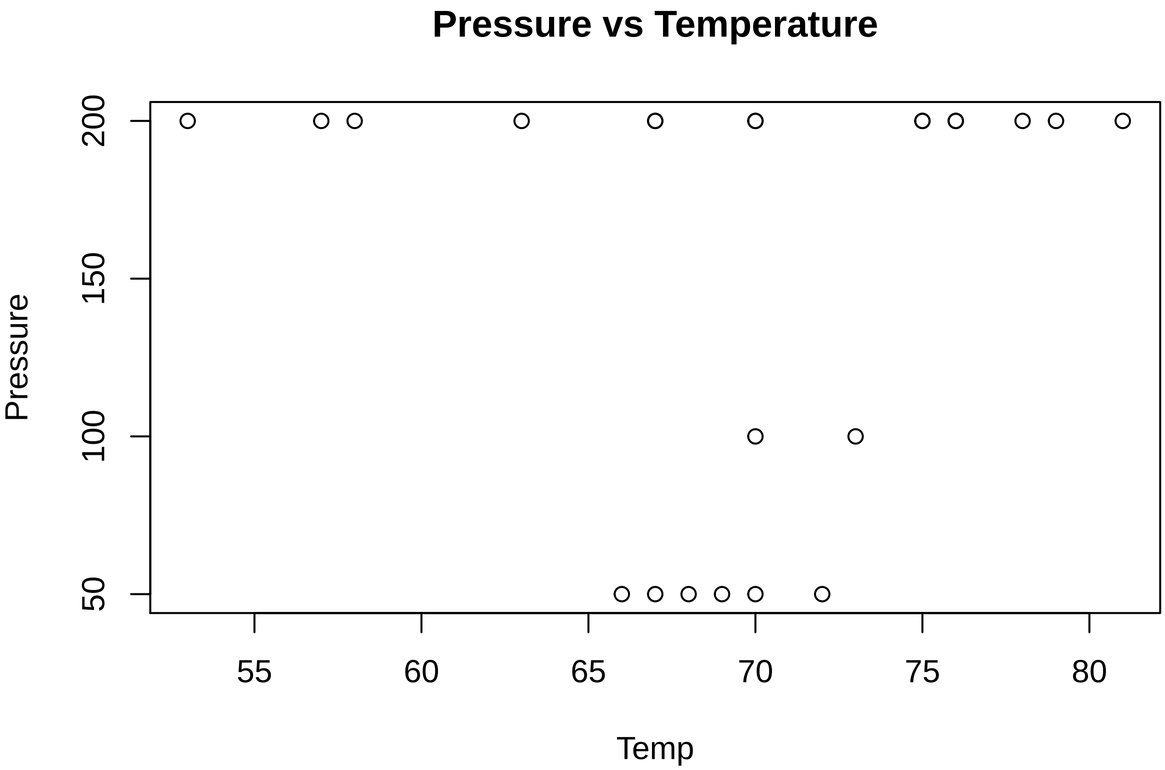
Comparing temperature and pressure directly, we can see that there is a vaguely positive relationship between temperature and pressure. Under basic gas laws, temperature is proportional to pressure, however, this relationship doesn't appear to hold in all cases of o-ring failure.

```
df$P200=df$Pressure
df$P200[df$P200!=200]=0
df$P200[df$P200==200]=1
df$P200=factor(df$P200)
```

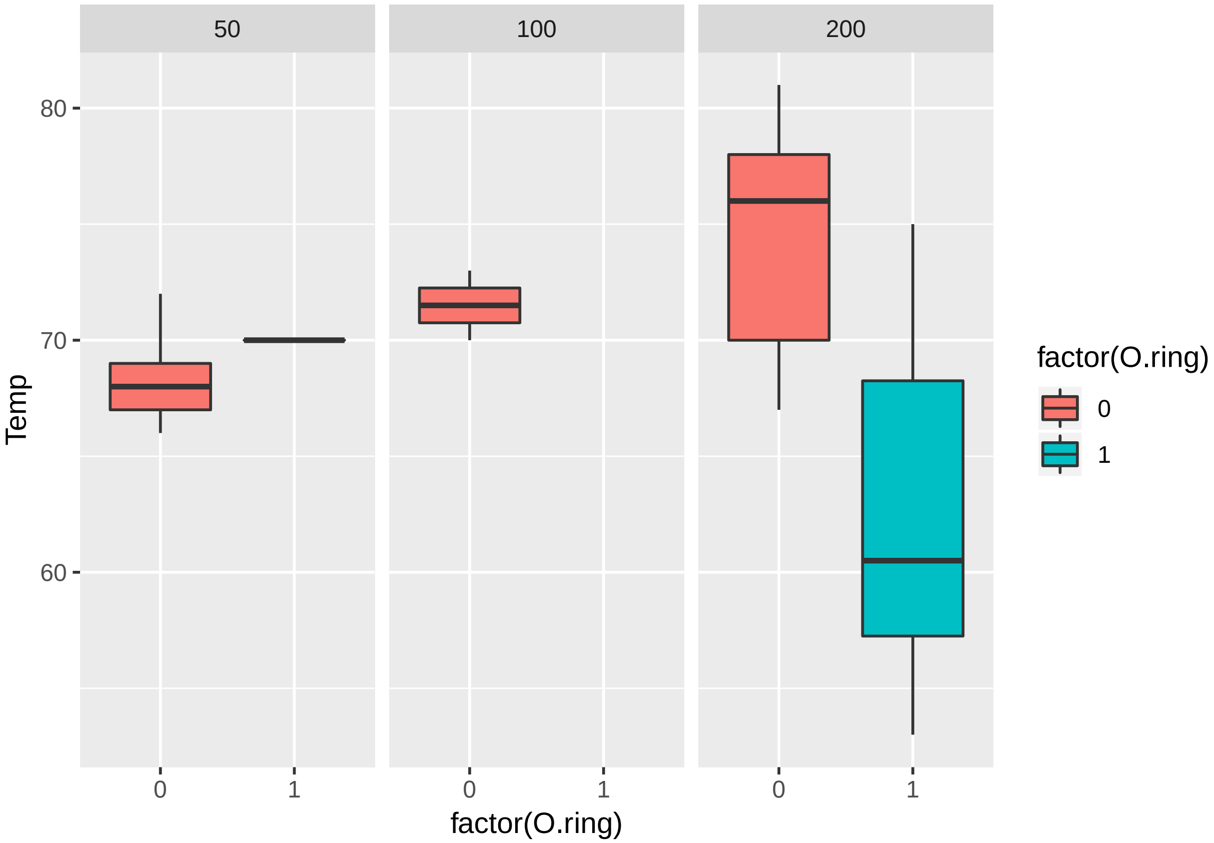
```
#how much does temperature vary by each pressure stage
ggplot(df, aes(P200, Temp)) +
          geom_boxplot(aes(fill = P200)) +
          geom_jitter() +
          ggtitle( "Temperature by Pressure") +
          theme(plot.title = element_text(lineheight=1, face="bold"))
```



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



```
#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)
```



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
```

```
# Include interaction between temperature and pressure to capture positive relationship between those v
mod.fit2<-glm(formula=O.ring~Pressure+Temp+Pressure:Temp,data=df,family=binomial(link = logit))
summary(mod.fit2)
```

```
##
## Call:
## glm(formula = O.ring ~ Pressure + Temp + Pressure:Temp, 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
## Pressure        0.221088   0.258398   0.856    0.392
## Temp            0.358212   0.720644   0.497    0.619
## Pressure:Temp  -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
```

```
# Turn Pressure into a factor
mod.fit3<-glm(formula=O.ring~factor(Pressure)+Temp,data=df,family=binomial(link = logit))
summary(mod.fit3)
```

```
##
## Call:
## glm(formula = O.ring ~ factor(Pressure) + Temp, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2093  -0.6044  -0.4151   0.3635   2.0479
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           13.5106     7.4288   1.819   0.0690 .
## factor(Pressure)100  -15.2969  2761.7586  -0.006   0.9956
## factor(Pressure)200    1.3774     1.3154   1.047   0.2950
## Temp                  -0.2211     0.1078  -2.050   0.0403 *
## ---
## 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: 18.214  on 19  degrees of freedom
## AIC: 26.214
##
## Number of Fisher Scoring iterations: 16
```

Our first estimated logistic regression model is

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

We include an interaction variable in the second

$$\text{logit}(\hat{\pi}) = -27.217458 + 0.221088\text{Pressure} + 0.358212\text{Temp} - 0.003054\text{Pressure x Temp}$$

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

```
#Anova(mod.fit1,Test='LRT')
Anova(mod.fit2,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 **
## Pressure:Temp   0.7069  1   0.400478
## ---
## 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. Finally, we test the interaction variable, Pressure x Temp, with the hypothesis: $H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$. The test statistic is $-2log(\Lambda) = 0.7069$ and the p-value is 0.400478 so we fail to reject the null hypothesis that the effect of temperature and pressure on o-ring failure depend on each other.

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

5 (a) Estimate the model

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

```
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

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

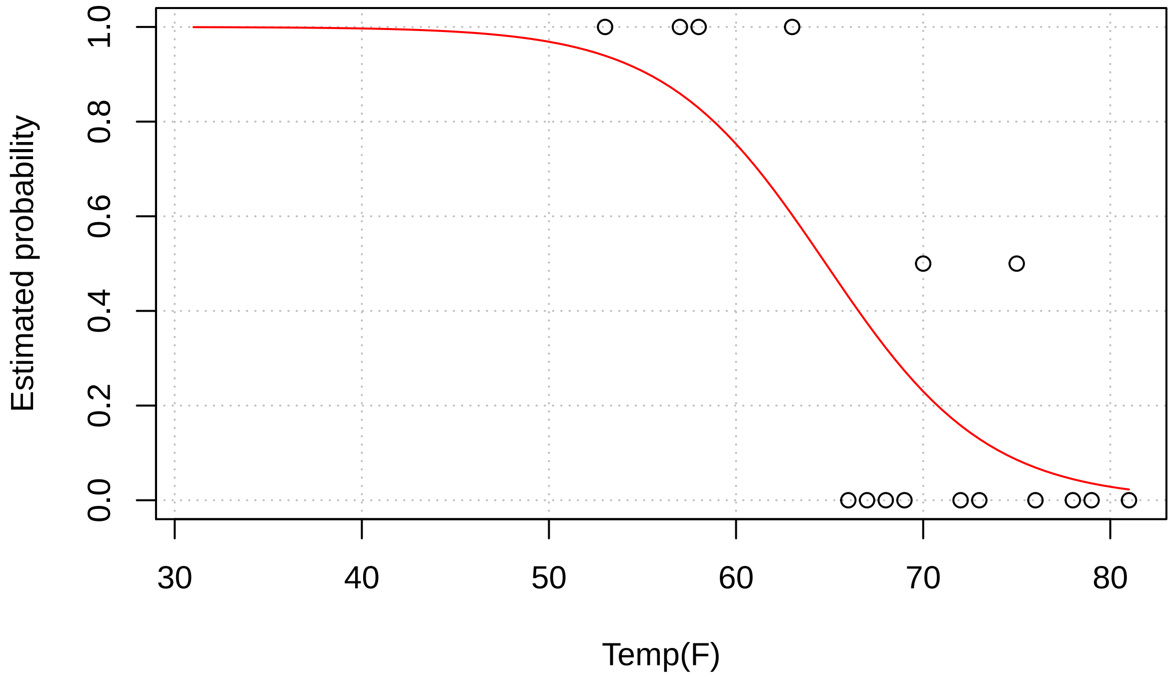
5 (b)

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.

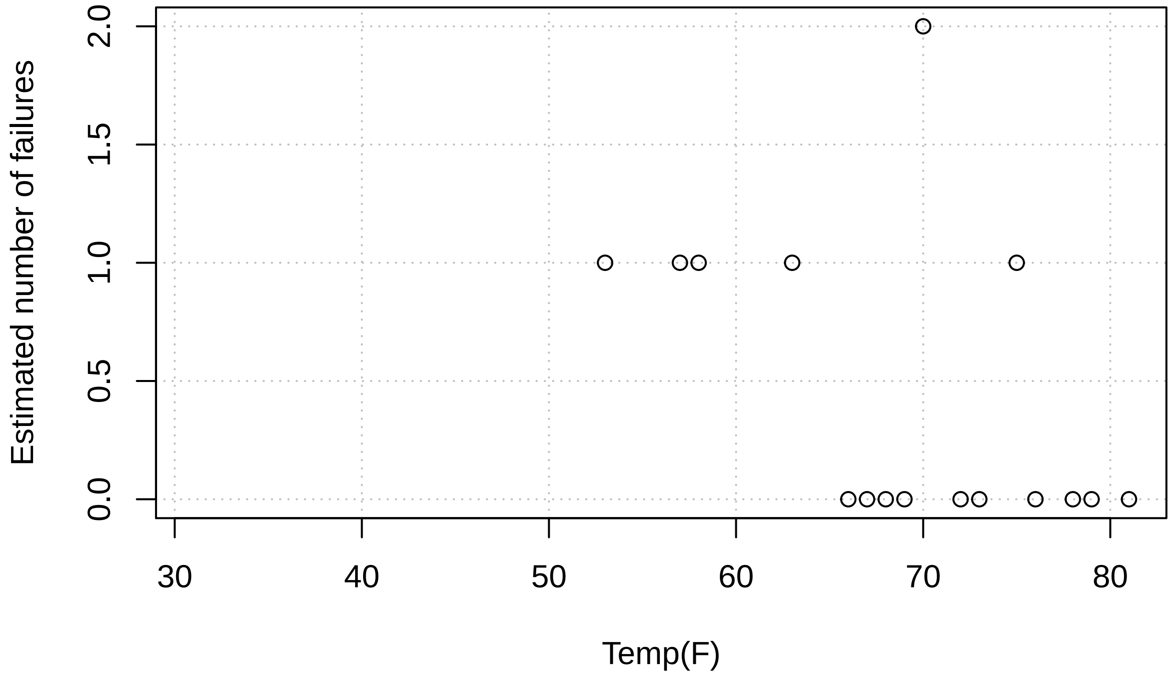
```
#
w<-aggregate(formula=O.ring~Temp, data=df, FUN=sum)
n<-aggregate(formula=O.ring~Temp, data=df, FUN=length)
w.n<-data.frame(Temperature=w$Temp, Failure=w$O.ring, trials=n$O.ring, proportion=round(w$O.ring/n$O.ri
head(w.n)
```

```
##   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
```

```
# 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?

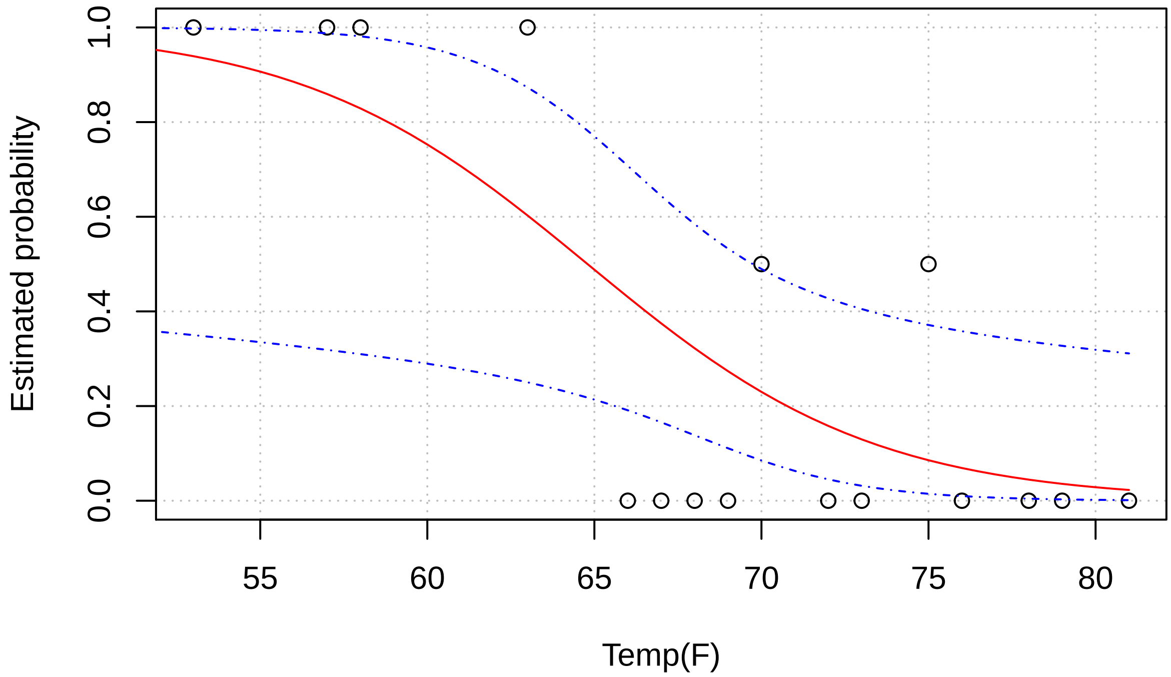
```
plot(x=w$Temp,y=w$O.ring/n$O.ring,xlab="Temp(F)", ylab = "Estimated probability", panel.first =grid(col
curve(expr=predict(object=mod.fit2,newdata=data.frame(Temp = x), type = "response"), col = "red", add=T

ci.pi<-function(newdata,mod.fit.obj,alpha){linear.pred <- predict(object = mod.fit.obj, newdata =newdat
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.

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
predict.data <- data.frame(Temp=31) #data to predict on

linear.pred=predict(object = mod.fit2, newdata = predict.data, #linear part of model
                      type = "link", se = TRUE)

pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit)) #estimated probability

CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se #confidence interval be

CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred)) #actual interval
```

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

5 (e) Rather than using Wald or profile LR intervals for the probability of failure, Dalalet al. (1989) use a parametric bootstrap to compute intervals. Their process was to (1) simulate a large number of data sets (n= 23 for each) from the estimated model of

$$logit(\pi) = \beta_0 + \beta_1 * Temp$$

;(2)estimate new models for each dataset,say

$$logit(\pi) = Beta_0 + Beta_1 * Temp$$

;and (3)compute at a specific temperature of interest. The authors used the 0.05 and 0.95 observed quantiles from the simulated distribution as their 90% confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31 and 72.27

```
library(boot) # Necessary because we're calculating a parametric bootstrap interval and this library is
```

```
##
## Attaching package: 'boot'

## The following object is masked from 'package:car':
##
##     logit

## The following object is masked from 'package:survival':
##
##     aml

## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
# Get probs from mod.fit2
predict.data <- data.frame(Temp=31:81)
linear.pred=predict(object = mod.fit2, newdata = predict.data, #linear part of model
                    type = "link", se = TRUE)
pis = exp(linear.pred$fit)/(1+exp(linear.pred$fit))

# Function to calculate desired statistic
pi.hat.star <- function(data){
  model <- glm(formula=O.ring~Temp,data=df,family=binomial(link = logit))

  predict.data <- data.frame(Temp=c(31,72.27)) #data to predict on

  linear.pred=predict(object = model, newdata = predict.data, #linear part of model
                        type = "link", se = TRUE)
  pi.hat <- exp(linear.pred$fit)/(1+exp(linear.pred$fit)) #estimated probability

  return(pi.hat)
}

# Function to generate random samples
gensamples <- function(data, pis){
  # Randomly determine temps
  x <- data.frame(Temp=round(runif(nrow(data), min=31, max=81),0))
  x$O.ring <- rbinom(n=nrow(data),size=1,prob=pis[x$Temp-30])
  return(x)
}

results <- boot(data=df[,c("Temp", "O.ring")], statistic=pi.hat.star, sim="parametric", R=1000, ran.gen=
boot.ci(results, conf = 0.90, type="bca")
```

```
## [1] "All values of t are equal to  0.999608782884929 \n Cannot calculate confidence intervals"

## NULL
```

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

```
mod.fit.Ha<-glm(formula=O.ring~Temp+I(Temp^2),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 + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     19.389  1  0.92649   0.3358
```

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.

Final Question

3. In addition to the questions in Question 4 and 5, answer the following questions:

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

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.

```
w<-aggregate(formula=O.ring~Temp+P200, data=df, FUN=sum)
n<-aggregate(formula=O.ring~Temp+P200, data=df, FUN=length)
w.n<-data.frame(Temperature=w$Temp, Failure=w$O.ring, trials=n$O.ring,Pressure200=w$P200, proportion=rou
head(w.n)
```

```
##   Temperature Failure trials Pressure200 proportion
## 1          66       0      1           0        0.0
## 2          67       0      1           0        0.0
## 3          68       0      1           0        0.0
## 4          69       0      1           0        0.0
## 5          70       1      2           0        0.5
## 6          72       0      1           0        0.0
```

```
mod.lm <- lm(proportion ~ Temperature+Pressure200, data = w.n)
summary(mod.lm)
```

```
##
## Call:
## lm(formula = proportion ~ Temperature + Pressure200, data = w.n)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52690 -0.11424 -0.00958  0.08846  0.45700
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.828546   0.523503   5.403 7.33e-05 ***
## Temperature  -0.039793   0.007445  -5.345 8.17e-05 ***
## Pressure2001  0.364512   0.114504   3.183  0.00617 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2367 on 15 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.6897
## F-statistic: 19.89 on 2 and 15 DF,  p-value: 6.042e-05
```

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. 4) There is no multi-collinearity/perfect collinearity amongst explanatory variables. 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 for reliable standard errors)

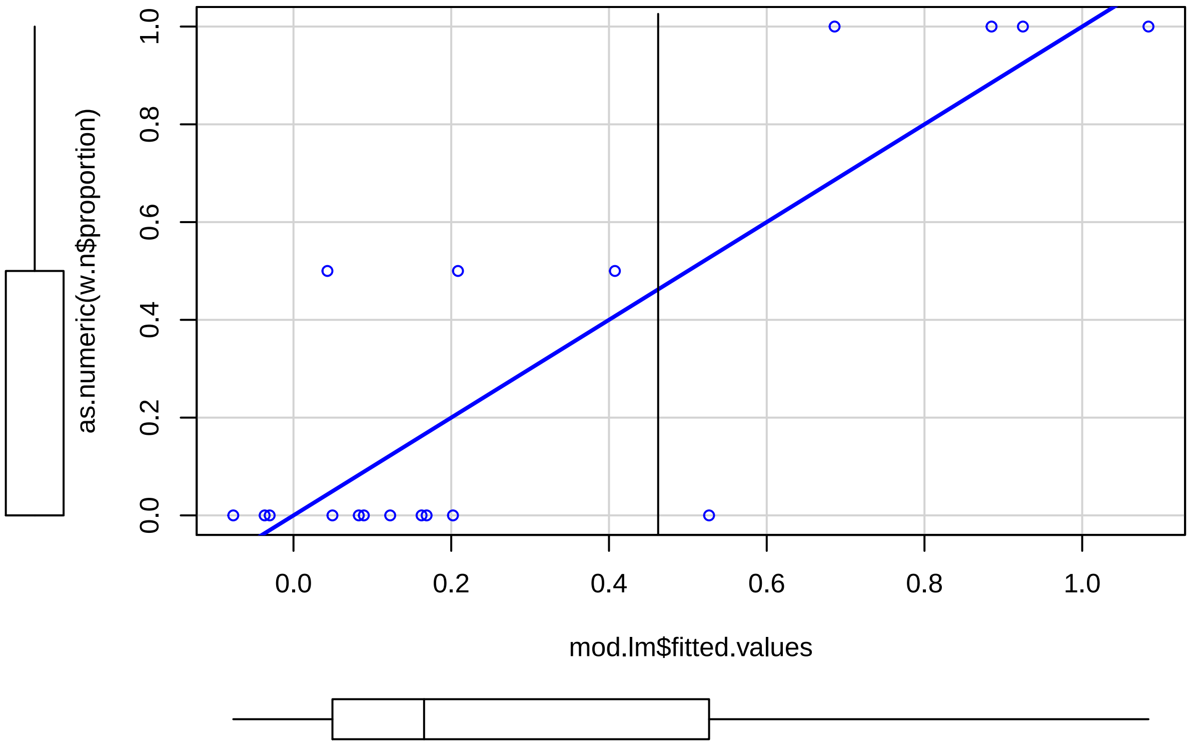
```
summary(mod.lm$fitted.values)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.07638  0.05773  0.16557  0.30556  0.49705  1.08401
```

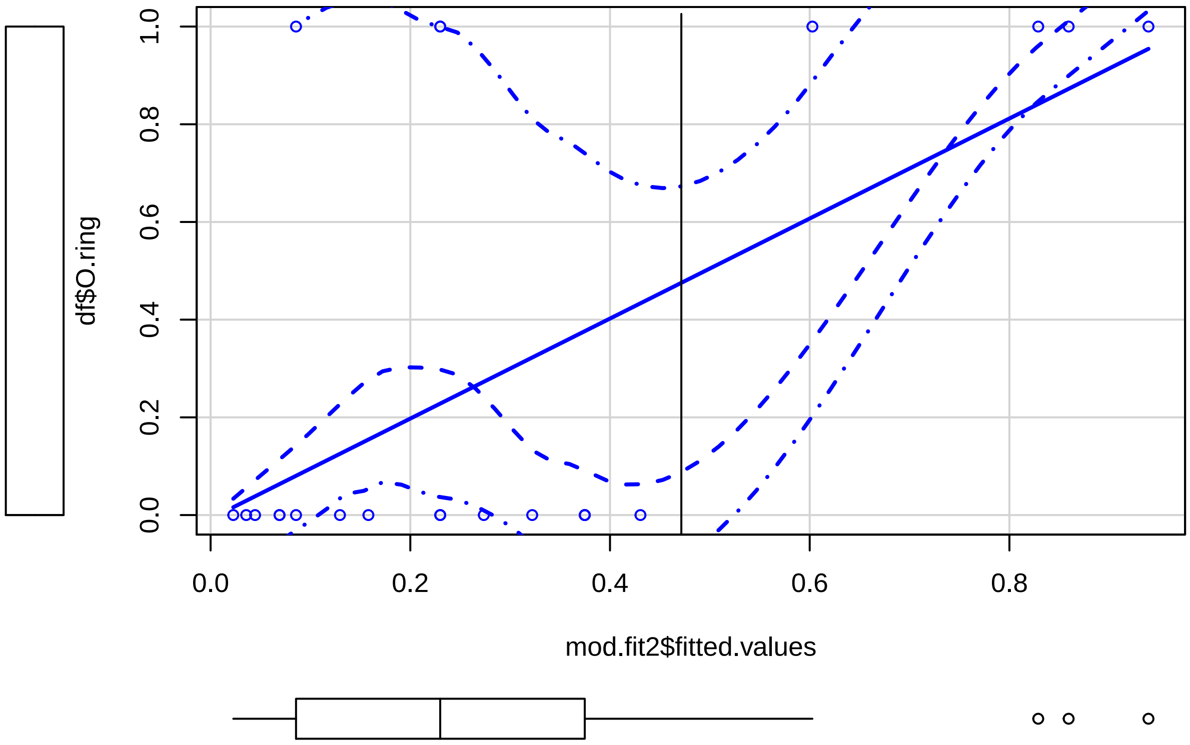
```
scatterplot(mod.lm$fitted.values,as.numeric(w.n$proportion))
```

```
## Warning in smoother(.x, .y, col = col[1], log.x = logged("x"), log.y =
## logged("y"), : could not fit smooth
```

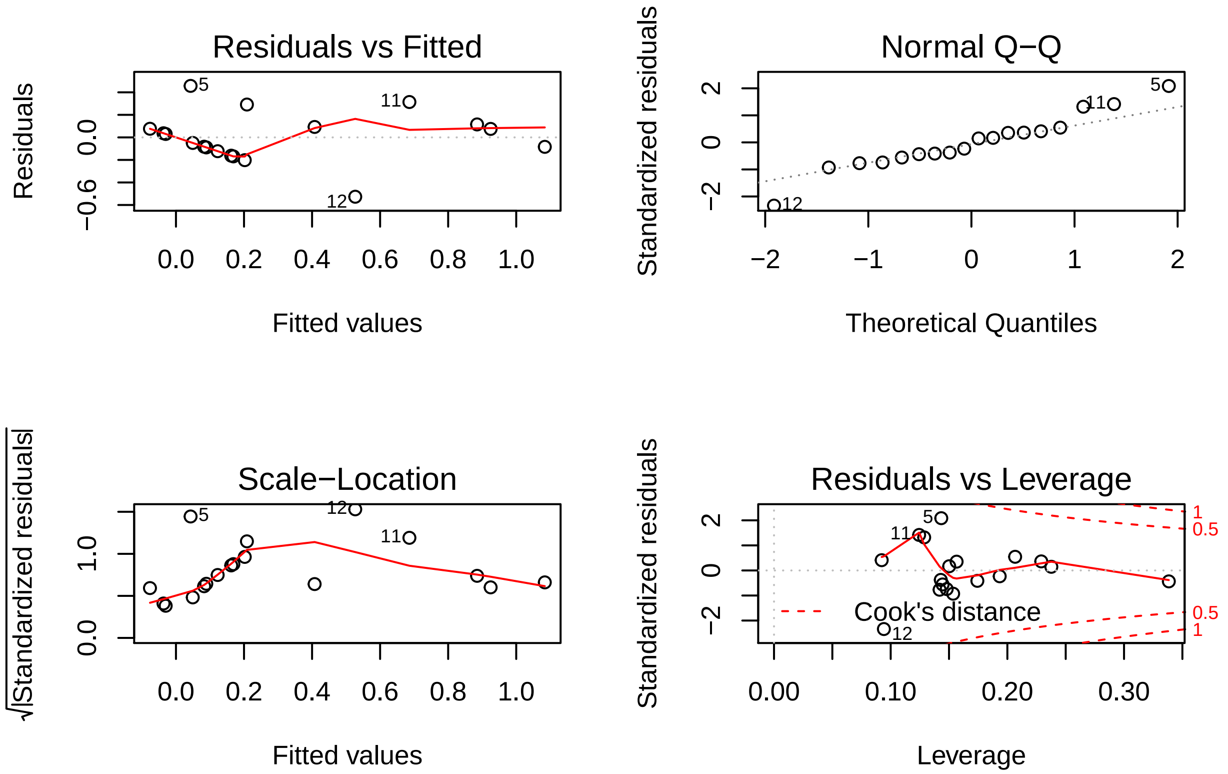
```
abline(v=.5)
```



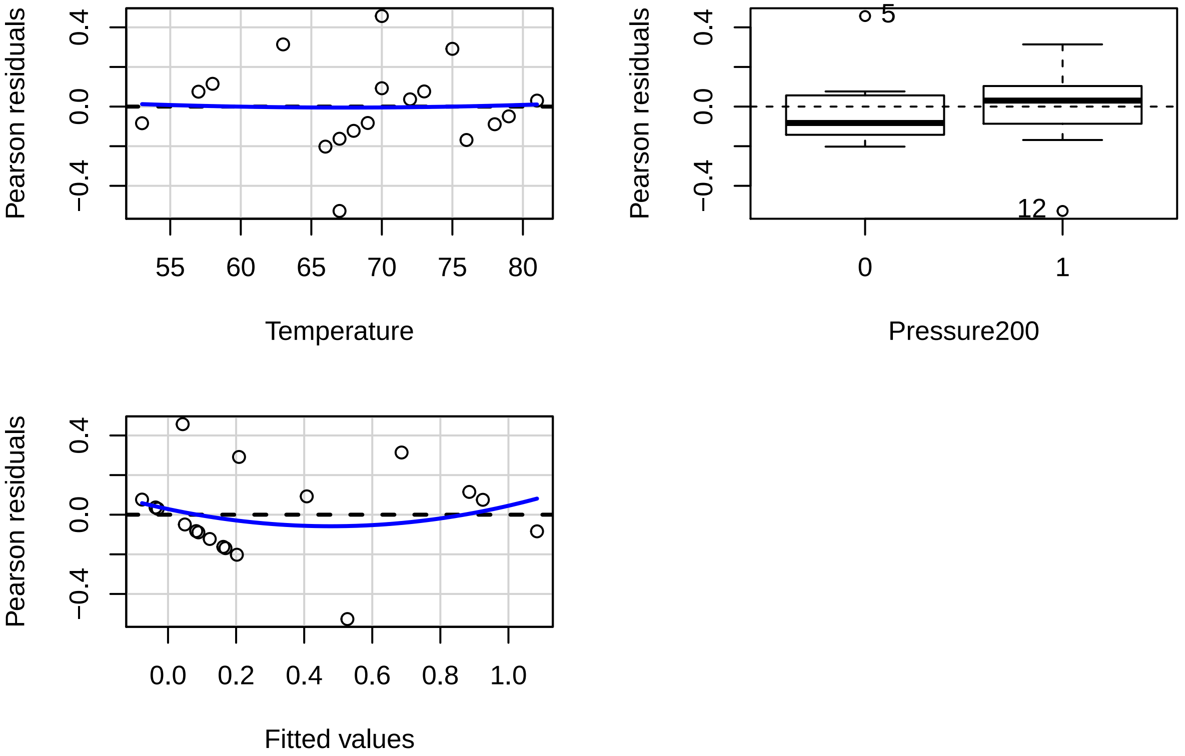
```
scatterplot(mod.fit2$fitted.values,df$O.ring)
abline(v=.5)
```



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

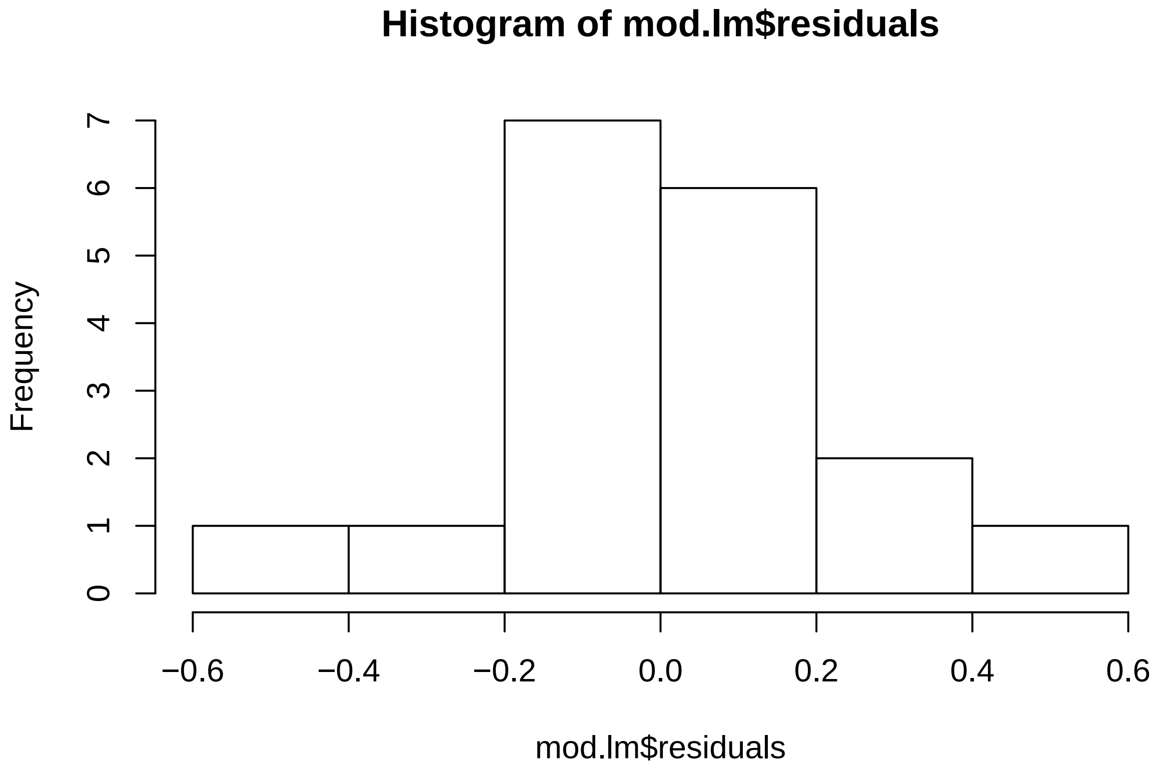


```r
par(mfrow=c(1,1))
residualPlots(mod.lm)
```

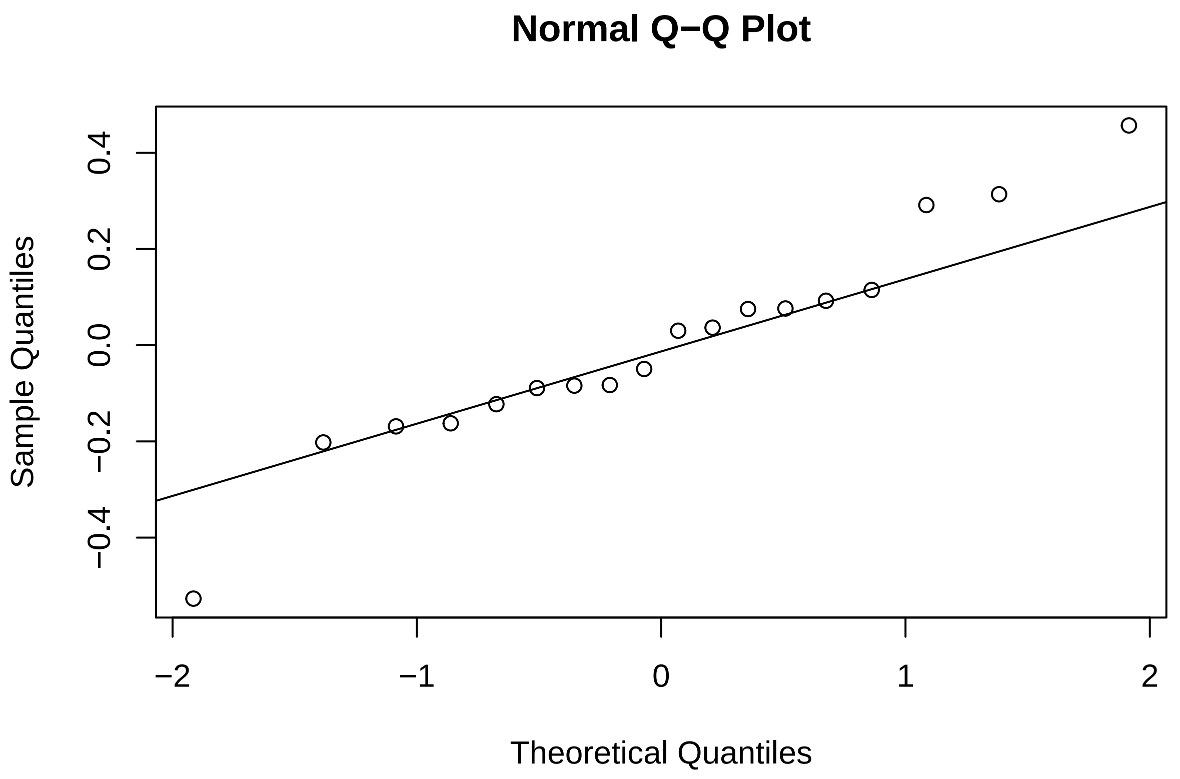


```
##              Test stat Pr(>|Test stat|)
## Temperature    0.1185           0.9073
## Pressure200
## Tukey test     0.6687           0.5037
```

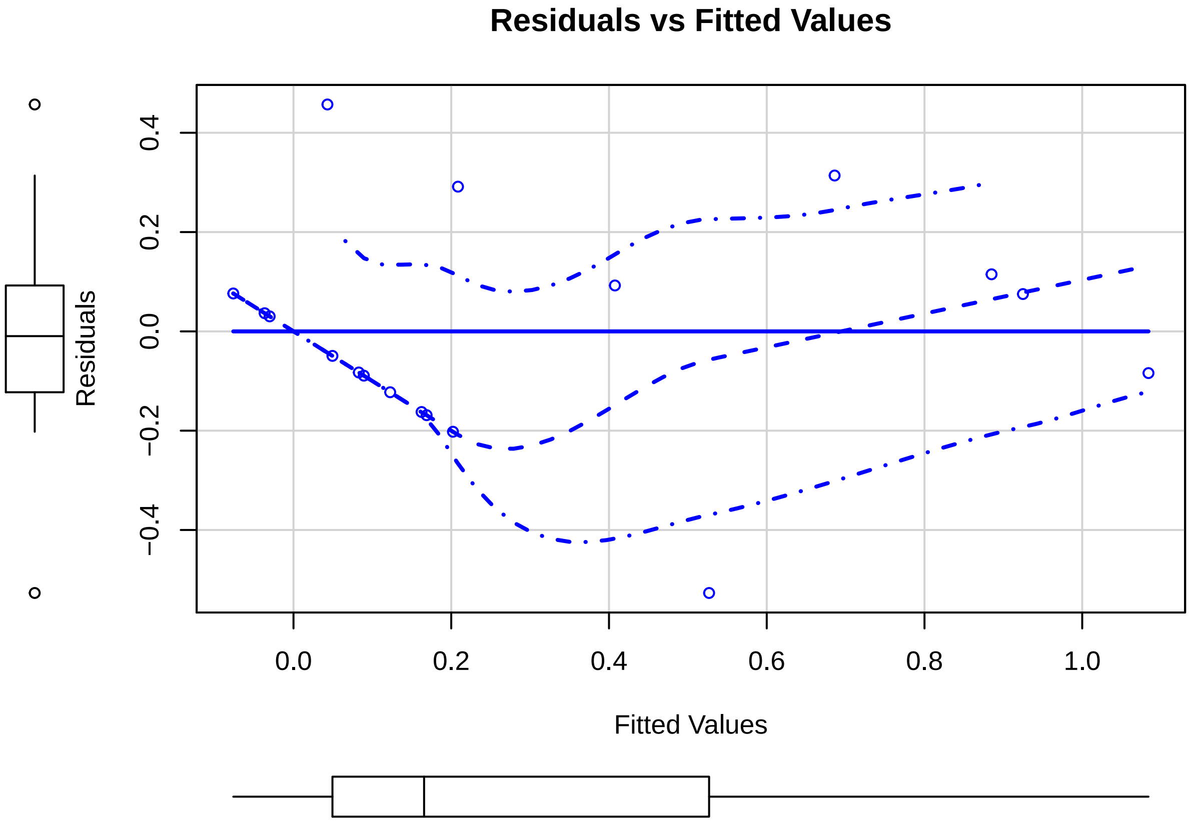
```
hist(mod.lm$residuals)
```



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



```
scatterplot(mod.lm$fitted.values, mod.lm$residuals,
            main = "Residuals vs Fitted Values",
            xlab = "Fitted Values", ylab ="Residuals")
```



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

```
## 
##  Shapiro-Wilk normality test
## 
## data:  mod.lm$residuals
## W = 0.95667, p-value = 0.5389
```

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

```
## Non-constant Variance Score Test 
## Variance formula: ~ fitted.values 
## Chisquare = 0.1046198    Df = 1     p = 0.7463545
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

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

```
## 
##  studentized Breusch-Pagan test
## 
## data:  mod.lm
## BP = 0.080483, df = 2, p-value = 0.9606
```

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1        -0.16876      2.287802   0.924
##  Alternative hypothesis: rho != 0
```

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

```
## 
## Call:
## lm(formula = proportion ~ Temperature + Pressure200, data = w.n)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52690 -0.11424 -0.00958  0.08846  0.45700 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.828546   0.523503   5.403 7.33e-05 ***
## Temperature  -0.039793   0.007445  -5.345 8.17e-05 ***
## Pressure2001  0.364512   0.114504   3.183  0.00617 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 0.2367 on 15 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.6897
## F-statistic: 19.89 on 2 and 15 DF,  p-value: 6.042e-05
##
##
## 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        0.85933  0.9303 Assumptions acceptable.
## Skewness           0.01553  0.9008 Assumptions acceptable.
## Kurtosis           0.27480  0.6001 Assumptions acceptable.
## Link Function      0.55717  0.4554 Assumptions acceptable.
## Heteroscedasticity 0.01182  0.9134 Assumptions acceptable.
```