W271 Lab 1 – Investigation of the 1989 Space Shuttle Challenger Accident

Lingyao Meng and Devin Robison

Part 1 (25 points)

Conduct a thorough EDA of the data set, including univariate, bivariate and trivariate analysis. This should include both graphical and tabular analysis as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Since the report has a page-limit, you will have to be selective when choosing visuals to illustrate your key points, associated with a concise explanation of the visuals. This EDA should begin with an inspection of the given dataset; examination of anomalies, missing values, potential of top and/or bottom code etc.

```
#load the packages
library(car)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(corrplot)
```

```
#load and summarize data
data <- read.csv("challenger.csv")
summary(data)
```

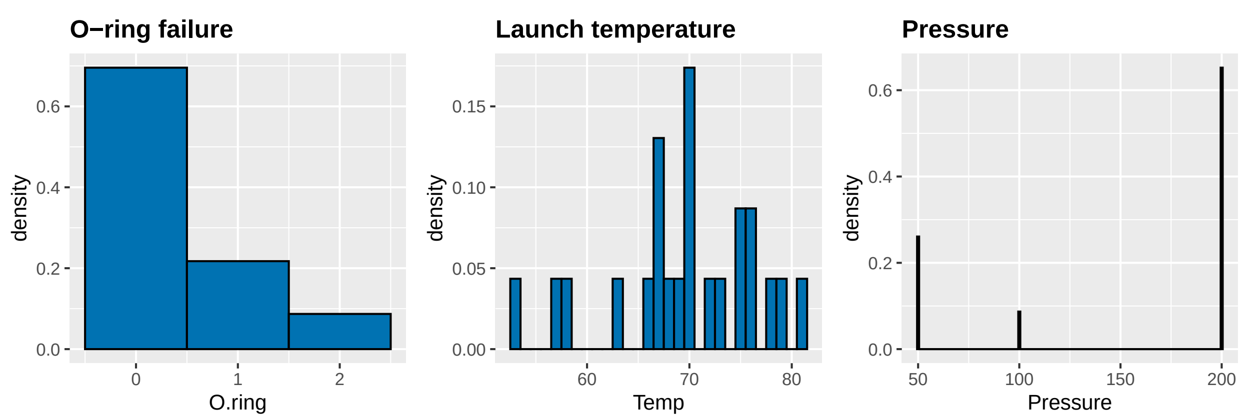
```
##      Flight          Temp          Pressure         O.ring
##  Min.   : 1.0   Min.   :53.00   Min.   : 50.0   Min.   :0.0000
##  1st Qu.: 6.5   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.:0.0000
##  Median :12.0   Median :70.00   Median :200.0   Median :0.0000
##  Mean   :12.0   Mean   :69.57   Mean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

The initial inspection shows that there are five variables in the dataset and there is no missing value in any avariable. *O.ring* denotes the number of O-ring failures in a flight while *Number* denotes the total number of O-rings, which is a constant 6 for all flights. *O.ring* is the response variable of our interest. *Temp* denotes the ambient environment temperature at launch and is a potential explanatory variable. *Pressure* denotes the leak test pressure, which is the internal pressure inside the solid rocket motor. The increase of leak test pressure could possibly cause "blow holes" in the putty which is used to protect the O-ring, and thus lead to erosion of the O-ring. Therefore, *Pressure* is also considered a potential explanatory variable.

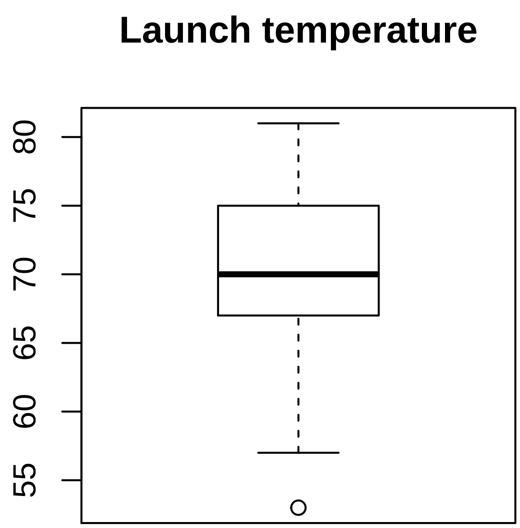
Next we performed the univariate analysis for *O.ring*, *Temp* and *Pressure*.

Univariate analysis of *O-ring*, *Temp* and *Pressure*

```
Oring.plot <- ggplot(data, aes(x = O.ring)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("O-ring failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

temp.plot <- ggplot(data, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Launch temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

pressure.plot <- ggplot(data, aes(x = Pressure)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

grid.arrange(Oring.plot, temp.plot, pressure.plot, ncol=3)
```



```r
boxplot(data$Temp, main = "Launch temperature")
```



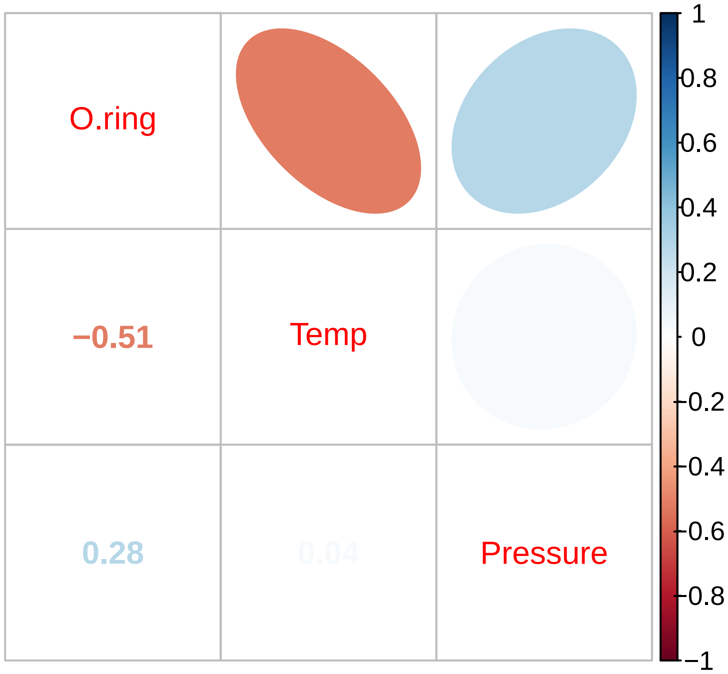

```r
outlier <- data %>% filter(Temp < 55)
outlier
```

```
##   Flight Temp Pressure O.ring Number
## 1     14   53      200      2      6
```

From the histogram of *O.ring*, we can tell that about 70% of the flights had no O-ring failure, 20% had only 1 failure and 10% had 2 failures. Visually, the distribution of *Temp* is slightly right skewed. From the data summary, we can also see that the median (70.00 F) is a bit larger than the mean (69.57 F). From the boxplot, we observed an outlier with the temperature lower than 55 F. After checking, we found that the outlier had 2 O-ring failures. According the Dalal's paper, the leak test pressure was originally 50 psi, then increased to 100 psi and finally 200 psi, in the sequence of flights launching. Based on the histogram of *Pressure*, we found that 200 psi leak test pressure was used in over 60% of the flights, 100 psi used in less than 10% and 50 psi used in less than 30% of the flights.

We initiated the explore on the correlations between three variables by plotting the correlation matrix.

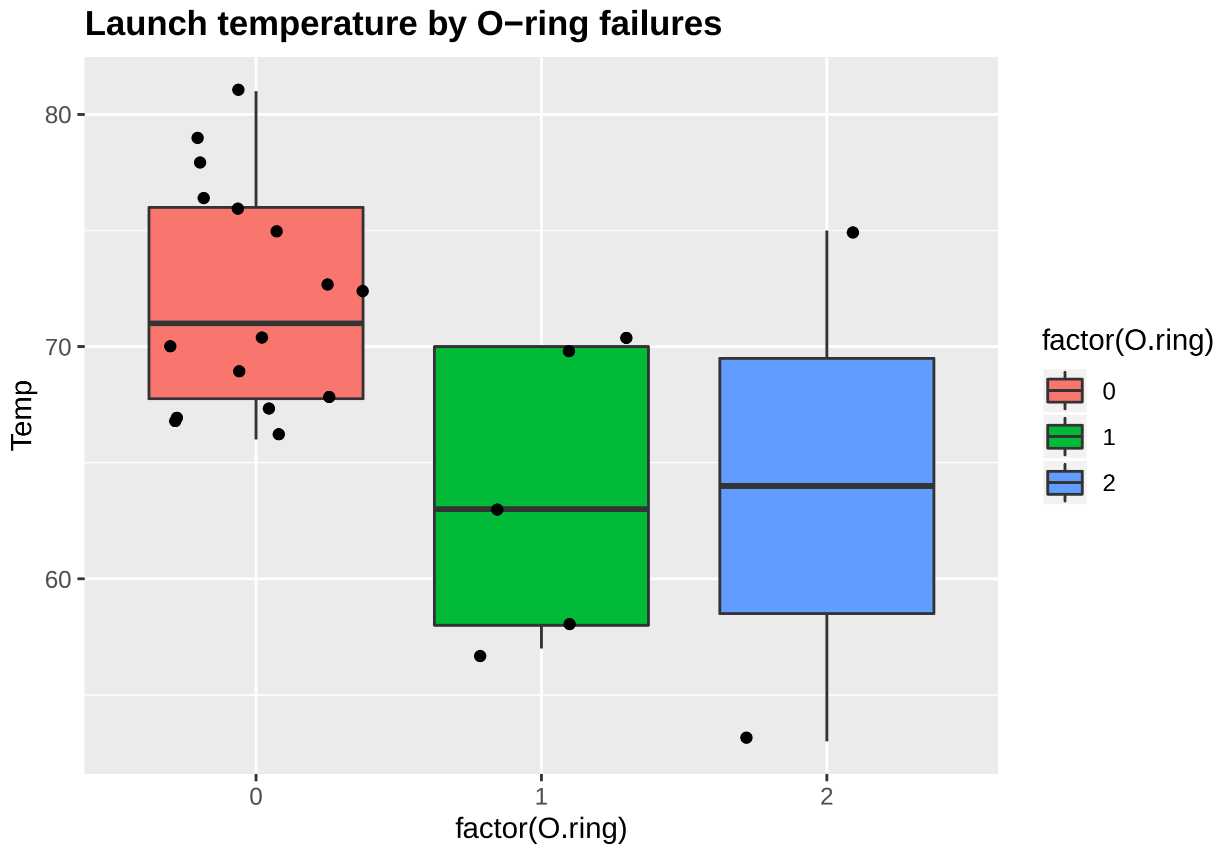
```r
correlations = cor(data[c('O.ring', 'Temp', 'Pressure')])
corrplot.mixed(correlations, lower='number', upper='ellipse')
```



There seems to be some negative correlation between *O.ring* and *Temp* and some positive correlation between *O.ring* and *Pressure*. No correlation was found between *Temp* and *Pressure*. Next, we performed the bivariate analysis between each pairs of variables.

Bivariate analysis of *O.ring* vs. *Temp*

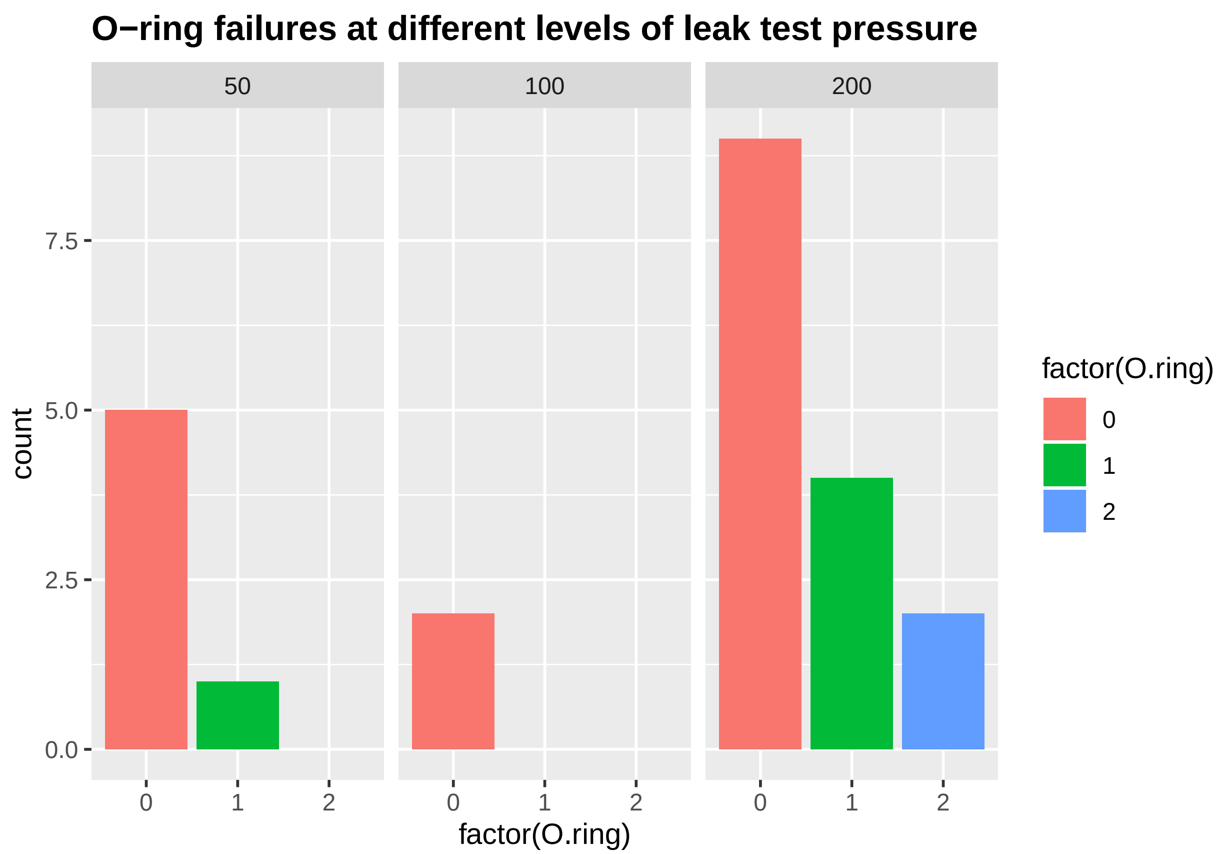
```
ggplot(data, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  geom_jitter() +
  ggtitle("Launch temperature by O-ring failures") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



We firsly grouped launch temperature by the number of O-ring failures and ploted each group using boxplot. Apparently, the flights with 0 O-ring failure were launched under higher temperature, than that for the flights with 1 or 2 failures. However, it worths to notice that the data size of flights with 1 or 2 O-ring failures is smaller than that of 0 failure flights. Especially, there are only 2 flights with 2 O-ring failures.

Bivariate analysis of *O.ring* vs. *Pressure*

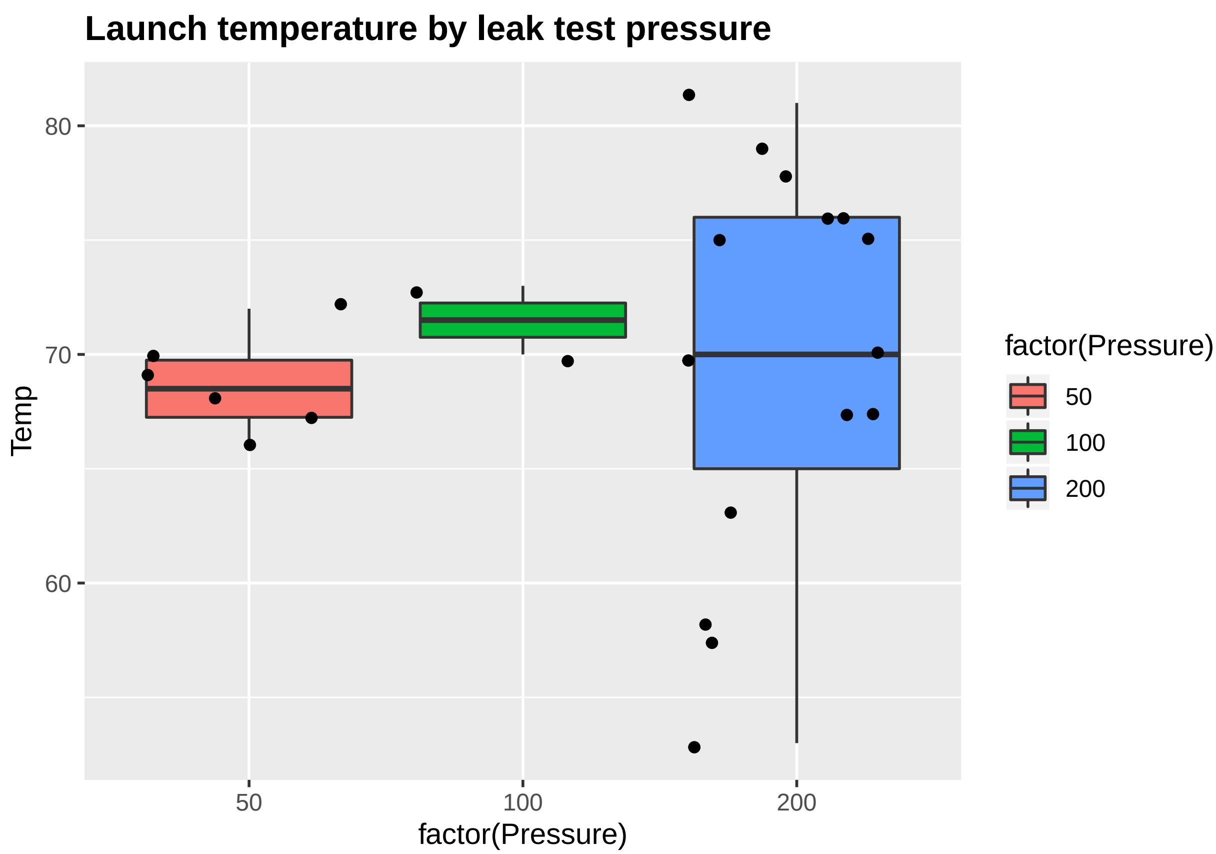
```
ggplot(data, aes(x = factor(O.ring), fill = factor(O.ring))) +
  geom_bar() +
  facet_wrap(~Pressure) +
  ggtitle("O-ring failures at different levels of leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



In the bivariate analysis between O-ring failures and leak test pressure, 0, 1 and 2 O-ring failures were counted, respectively, at different levels of pressure. From the corresponding plot, we cannot tell obvious correlation between *Pressure* and *O.ring*. Further analysis is to be conducted for elucidating potential correlation.

Bivariate analysis of *Temp* vs. *Pressure*

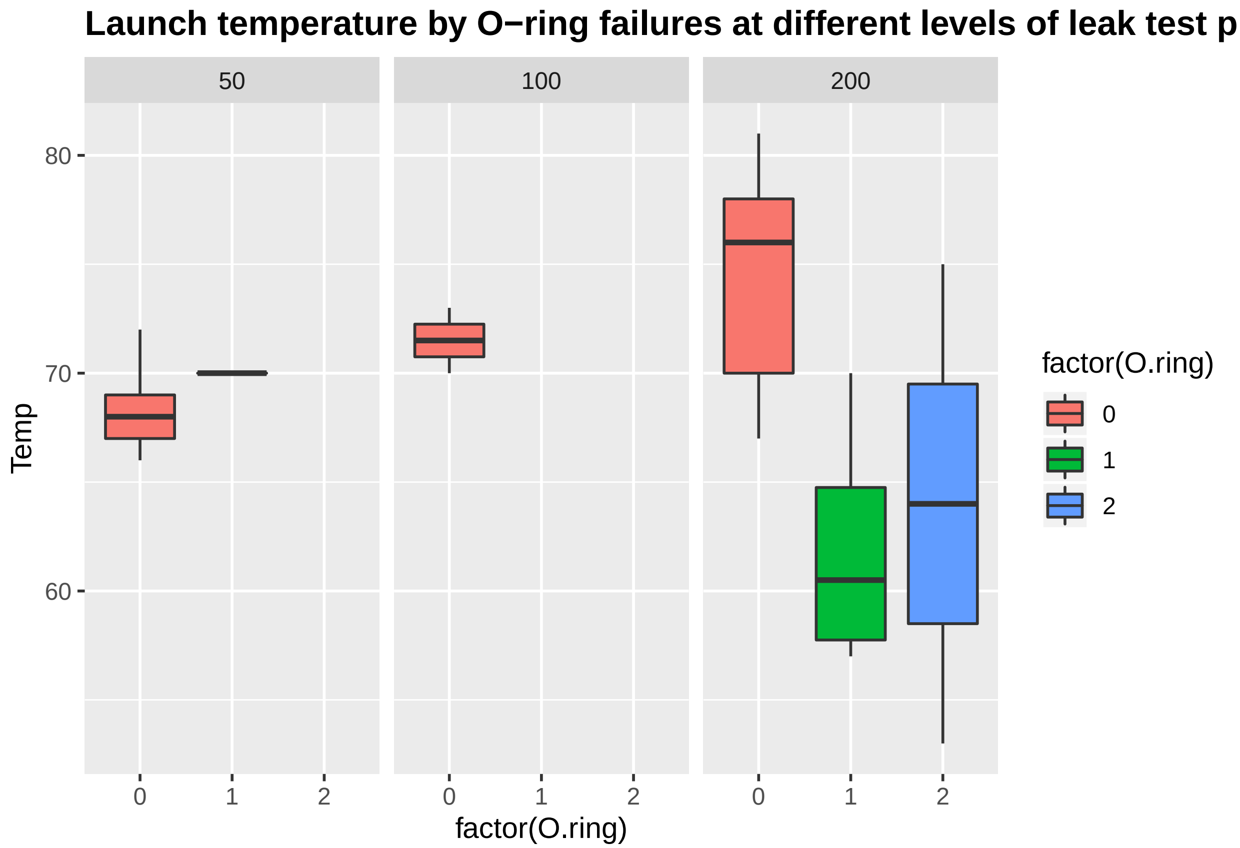
```
ggplot(data, aes(factor(Pressure), Temp)) +
  geom_boxplot(aes(fill = factor(Pressure))) +
  geom_jitter() +
  ggtitle("Launch temperature by leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



Bivariate analysis between launch temperature and pressure for leak test was also performed to check for potential dependence between the explanatory variables. The boxplot of temperature, grouped by pressure, seems to show that when 100 psi pressure was used, the launch temperature is higher than that when 50 or 200 psi was used. However, we cannot ensure that two variables are dependent solely based on this observation because there are only two data points for 100 psi.

Trivariate analysis

```
ggplot(data, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  facet_wrap(~Pressure) +
  ggtitle("Launch temperature by O-ring failures at different levels of leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



Trivariate analysis was performed by examing launch temperature by O-ring failures at three different levels of leak test pressure. Similar correlation between *Temp* and *O-ring* was observed when the pressure of 200 psi was used. However, the plot at pressure levels of 50 or 100 psi didn't provide much useful information due to the small data size.

Part 2 (20 points)

Answer the following from Question 4 of Bilder and Loughin Section 2.4 Exercises (page 129):

(a) The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

By using logistic regression to estimate the probability an O-ring will fail, the authors assumed the number of failed O-rings in a given launch to be a binomial variable. In other words, the response variable of the logistic regression was assumed to have the binomial distribution. One of the assumptions for a process to be modeled by binomial distribution is that the trials are independent of each other. In the O-ring case, each O-ring is a trial, so it's necessary to assume that each O-ring is independent for each launch to ensure the validity of model used. However, this assumption is not necessarily true. For instance, the 6 primary O-rings locate at 2 rocket motors. It's possible for the O-rings locating at the same motor to have more similar probabilities to fail. Also, no information was provided regarding the production of O-rings. The devices produced from the same batch may have more similar probabilities to fail. If so, the assumption of independence doesn't hold any more and the logistic regression model used here is invalid. To check on potential violations of independence, the authors fit another model using a binary response to indicate whether there was an incident in a given launch. The second model doesn't require independence of each O-ring. In fact, the second model was quite close to the original model, which alleviated the authors' concerns about independence.

(b) Estimate the logistic regression model using the explanatory variables in a linear form.

From the exploratory data analysis, we found some correlation between *Temp* and *O-ring* while the correlation between *Pressure* and *O-ring* was not very obvious. However, we still want to include *Pressure* as an explanatory variable for the first logistic regression model. For a given launch i, we denote the probability for an O-ring to fail as π_i, launch temperature as t_i and leak test pressure as p_i. The first model has the following equation:

$$logit\left(\pi_i\right) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 t_i + \beta_2 p_i$$

This model was fit and estimated using the *glm* function:

```
mod1 <- glm(O.ring/Number ~ Temp + Pressure, weights = Number,
               family = binomial (link = logit), data = data)
summary(mod1)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure, family = binomial(link = logit),
##     data = data, 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
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## Pressure     0.008484   0.007677   1.105   0.2691
## ---
## 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
```

```r
cbind(Estimate = coef(mod1), confint(mod1))
```

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

##                 Estimate        2.5 %      97.5 %
## (Intercept)  2.520194641 -4.322926283  9.77264497
## Temp        -0.098296750 -0.194071699 -0.01356289
## Pressure     0.008484021 -0.004346403  0.02885221
```

```r
c.temp <- -5
exp(c.temp*coef(mod1)['Temp'])
```

```
##    Temp
## 1.63474
```

The coefficient of t_i was estimated to be -0.0983 with the 95% Wald confidence interval of -0.1941 to -0.0136, indicating that the decrease on the launch temperature would cause the increase on the odds for an O-ring to fail. Specifically, a decrease of 5 F would increase the odds for failure by around 63%. The coefficient of p_i was estimated to be 0.0085 while 0 was included in the 95% Wald confidence interval, indicating that leak test pressure may not be an important factor for explaining O-ring failure.

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

Because the Wald interval usually has lower true coverage than the cofidence level, we performed likelihood ratio test using the *Anova* function to judge the importance of the explanatory varialbes in the first model.

```r
Anova(mod1, test = "LR")
```

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

For the test of *Temp* with $H_0 : \beta_1 = 0$ vs. $H_\alpha : \beta_1 \neq 0$, we obtained the LRT statistic of 5.184 with a p-value of 0.0228. Using the Type I Error rate $alpha = 0.05$, we would reject the null hypothesis and claim that there is marginal evidence that *Temp* is important to be included in the model, given that *Pressure* is in the model. For the test of *Pressure* with $H_0 : \beta_2 = 0$ vs. $H_\alpha : \beta_2 \neq 0$, we obtained the LRT statistic of 1.541 with a p-value of 0.2145. Using the Type I Error rate $alpha = 0.05$, we could not reject the null hypothesis. Therefore, there is a lack of evidence to claim that *Pressure* is important to be included in the model, given that *Temp* is in the model.

(d) The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

The authors fit a model using both *Temp* and *Pressure* and then fit another model using only *Temp*. By comparing the residual deviances of two models, they found that keeing only *Temp* in the model just increased the residual deviance by 1.54, which was not significant, indicating that *Pressure* may had a very weak effect. This is consistent with our LRT results in the above section. Actually the LRT statistic we computed for *Pressure* was equivalent to the difference in residual deviances computed by the authors. However, the apparently weak effect of *Pressure* might be due to limited data for 50 and 100 psi. Also, the interaction between *Temp* and *Pressure* was not considered by the authors. Given that high pressure can cause "blow holes" in the putty, it can make it easier for the hot gasses to escape. If so, the effect of pressure could be very significant when temperature is very low. Therefore, removing *Pressure* from the model could possibly cause serious information loss.

Part 3 (35 points)

Answer the following from Question 5 of Bilder and Loughin Section 2.4 Exercises (page 129-130):

Continuing Exercise 4, consider the simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$, where π is the probability of an O-ring failure. Complete the following:

(a) Estimate the model.

(b) Construct two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31Â° to 81Â° on the x-axis even though the minimum temperature in the data set was 53Â°.

(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?

(d) The temperature was 31Â° 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.

(e) Rather than using Wald or profile LR intervals for the probability of failure, Dalal et 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 Temp; (2) estimate new models for each data set, say 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 31Â° and 72Â°.27

```r
mod2 <- glm(O.ring/Number ~ Temp, weights = Number,
               family = binomial (link = logit), data = data)
summary(mod2)
```

```
##
```

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

```r
mod2$coefficients
```

```
## (Intercept)        Temp
##   5.0849772  -0.1156012
```

```r
beta.0 <- mod2$coefficients[1]
beta.1 <- mod2$coefficients[2]

sim_fun <- function(){
  temp <- sample(data$Temp, 23, replace=TRUE)
  logit <- exp(beta.0 + beta.1*temp)
  pi <- logit/(1+logit)
  oring.fail <- rbinom(n = 23, size = 6, prob = pi)

  model <- glm(oring.fail/data$Number ~ temp, weights = data$Number,
               family = binomial (link = logit))
  beta0.hat <- model$coefficients[1]
  beta1.hat <- model$coefficients[2]

  logit.31 <- exp(beta0.hat + beta1.hat*31)
  pi.31 <- logit.31/(1+logit.31)

  logit.72 <- exp(beta0.hat + beta1.hat*72)
  pi.72 <- logit.72/(1+logit.72)
```

```r
  return(c(pi.31, pi.72))
}

set.seed(666)
n = 2000
simulations <- replicate(n, sim_fun())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
quantile(simulations[1,],c(0.05,0.95))
```

```
##         5%        95%
## 0.09707027 0.99387349
```

```r
quantile(simulations[2,],c(0.05,0.95))
```

```
##          5%         95%
## 0.009969481 0.069986416
```

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

Part 4 (10 points)

With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case? Explain why.

Part 5 (10 points)

Interpret the main result of your final model in terms of both odds and probability of failure. Summarize the final result with respect to the question(s) being asked and key takeaways from the analysis.