W271 Group Lab 1

Due 11:59pm Pacific Time Sunday February 9 2020

Investigation of the 1989 Space Shuttle Challenger Accident

Carefully read the Dalal et al (1989) paper (Skip Section 5).

Part 1 (25 points)

Conduct a thorough EDA of the data set. 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.

```
path = "/Users/jeff/Documents/MIDS/W271/w271_lab1/challenger.csv"

challenger<-read.table(file =  path, header = TRUE, sep = ",") #Import table

knitr::kable(
  challenger[1:5,1:5 ], caption = 'Top 6 Rows of Admissions Dataset')
```

Table 1: Top 6 Rows of Admissions Dataset

Flight	Temp	Pressure	O.ring	Number
1	66	50	0	6
2	70	50	1	6
3	69	50	0	6
4	68	50	0	6
5	67	50	0	6

First, we read in the data set to an object called 'challenger'. From here, we see that we have 5 variables in a dataframe with 23 observations. We review each of the 5 below:

- Flight: Simply akin to counting the row of the dataframe. Will not be integral to the analysis
- Temp: An integer variable that records the takeoff temperature in degrees Fahrenheit
- Pressure: An integer variable measuring the amount of pressure on the O.rings, measured in pounds per square inch
- O.ring: An integer variable that counts the number of O.ring failures on the flight in question
- Number: An integer variable recording the number of O.rings on each flight

```
#output in text format for view in latex
stargazer(challenger, header= F, title = "Summary Table of O.Ring Failure")
```

```
#output in text format for view in R
#stargazer(challenger, header= F, title = "Summary Table of O.Ring Failure", type='text')
```

Table 2: Summary Table of O.Ring Failure

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Flight	23	12.000	6.782	1	6.5	17.5	23
Temp	23	69.565	7.057	53	67	75	81
Pressure	23	152.174	68.221	50	75	200	200
O.ring	23	0.391	0.656	0	0	1	2
Number	23	6.000	0.000	6	6	6	6

This summary tallies each variable by range, so we get the see quartiles, minimums, maximums, and medians/means. A few conclusions can be drawn:

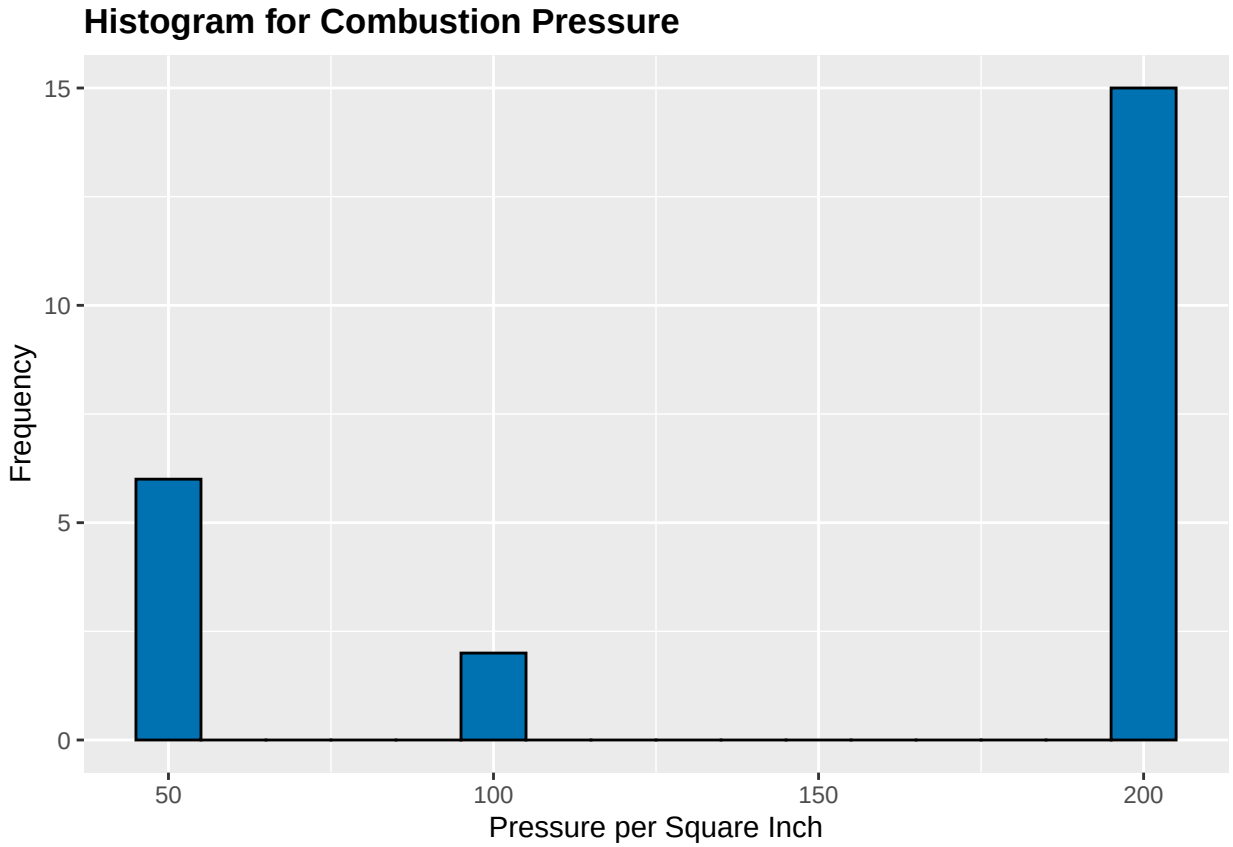
- Temp: Temperature in the dataset ranges between 53 and 81 degrees Fahrenheit
- Pressure: Pressure in the dataset ranges between 50 and 200 pounds per square inch, and appears to occur in increments of 50
- O.ring: It appears that in the dataset, most O.ring failures are 0 (meaning no failure), but there are some flights that did fail, and at least one flight that had as many as two O.ring failures.
- Number: This variable does not appear to be informative to the analysis, because it is 6 for all measurements.

```
sum(is.na(challenger))
```

```
## [1] 0
```

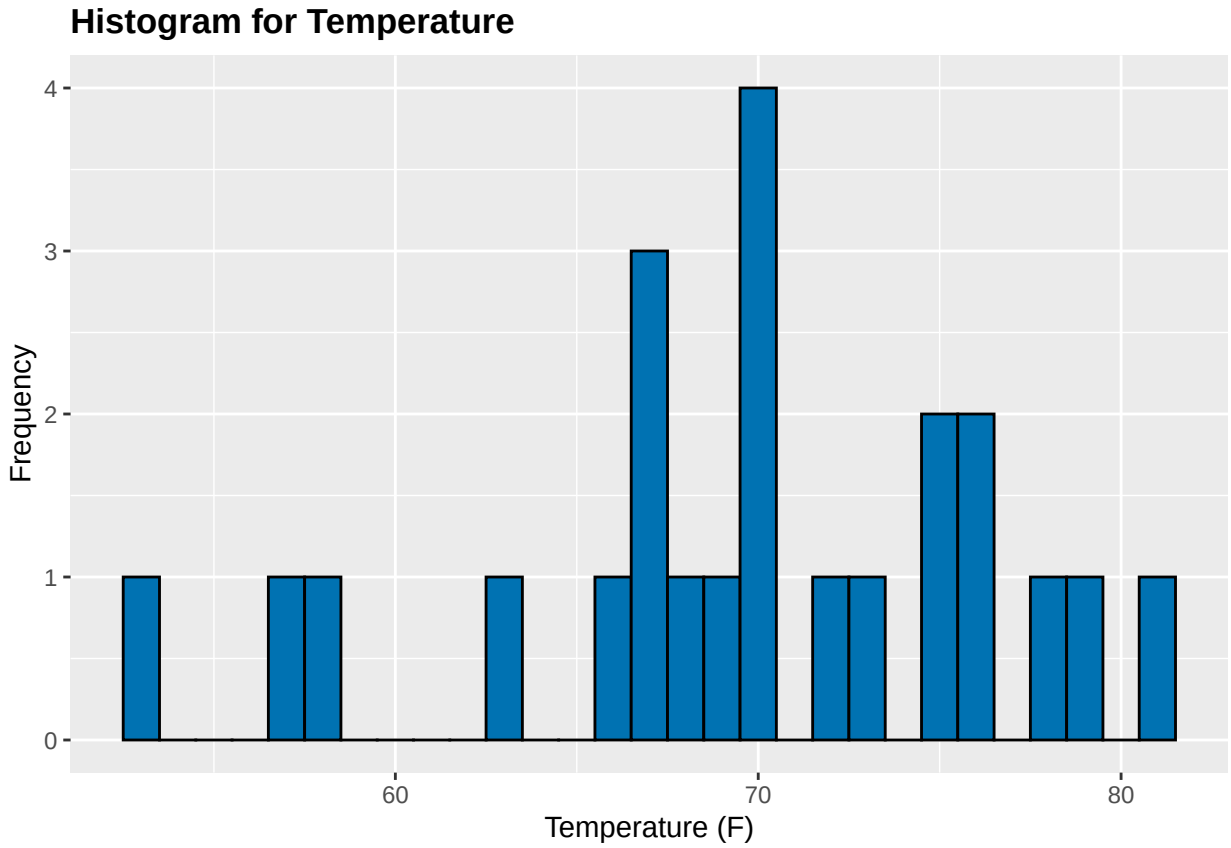
In the cell above, we examine the dataset for missing data, and find none. The EDA we have performed so far indicates that 'Flight' and 'Number' are not informative as explanatory variables. Temperature and Pressure are informative as explanatory variables, while O.ring will be our response variable. It may be worthwhile to transform the O.ring variable into a binary categorical variable such that any values over 0 all register as failures, and any zeros register as non-failures.

```
# Distribution of Pressure
ggplot(challenger, aes(x = Pressure)) +
  geom_histogram(aes(x = Pressure), binwidth = 10, fill="#0072B2", colour="black") +
  ggtitle("Histogram for Combustion Pressure") +
  xlab("Pressure per Square Inch") +
  ylab("Frequency") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



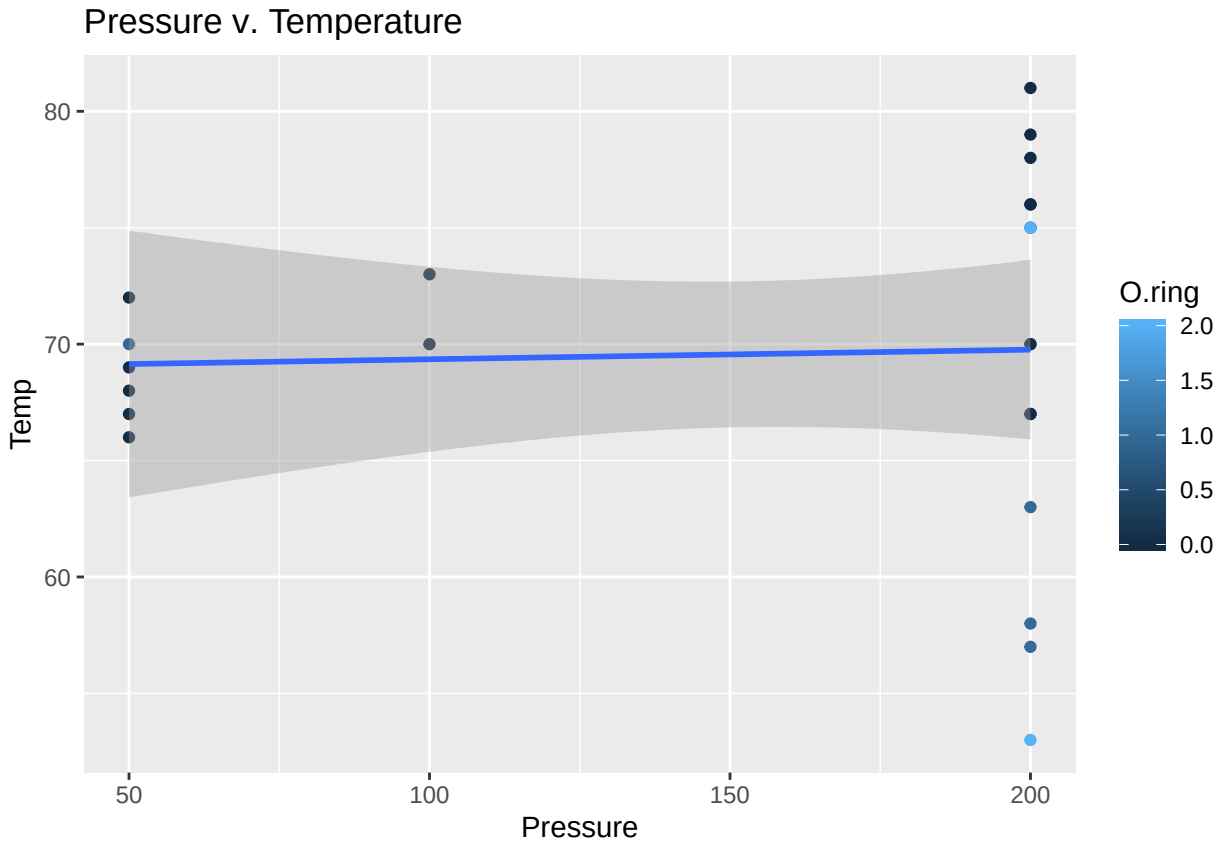
The histogram above confirms our notion that PSI measurements in the dataset occur in regular increments, with most measurements taking place at 200psi, but other measurements also taking place at 50 and 100 psi as well. We had considered converting Pressure to a categorical variable, since there are only 3 distinct values for Pressure (50, 100, 200) in the given dataset. However, we were also concerned that in converting Pressure to categorical, we would be losing some of the scaling factor of Pressure, thereby losing some of the predictive power for the variable. We choose to leave as a discrete integer variable.

```
# Distribution of Temp
ggplot(challenger, aes(x = Temp)) +
  geom_histogram(aes(x = Temp), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Histogram for Temperature") +
  xlab("Temperature (F)") +
  ylab("Frequency") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



We extend our histogram analysis further, this time examining temperature. The distribution has two major peaks, right around upper 60s, and mid 70s, with only a few observations warmer or cooler than those areas. This is a continuous variable.

```
ggplot(challenger, aes(x=Pressure, y=Temp, color=O.ring)) + geom_point() +
  geom_smooth(method='lm') +
  ggtitle('Pressure v. Temperature')
```



Taking into consideration the two explanatory variables that matter for our analysis, we create a scatterplot involving Temperature and Pressure. In addition to being a simple scatterplot, we include two other features. The first is 'best-fit' line with confidence bands provided by R. This is not terribly informative, because of the fact that Pressure is a discrete variable, so a linear model is of limited use. However, the second feature, in which we color the points according to the value of the response variable, is highly useful.

From this chart, we can draw several conclusions. First, most O-ring failures seem to occur at higher pressures. Second, most O-ring failures seem to occur at lower temperatures, with the only datapoint having two failures occuring at the lowest recorded temperatures.

```
chart.a <- ggplot(challenger, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  ggtitle("Temperature by O.ring Failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

chart.b <- ggplot(challenger, aes(factor(O.ring), Pressure)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  ggtitle("Pressure by O.ring Failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

# cast side-by-side
```

The two boxplots we examine here further illustrate the distribution of our two key explanatory variables, but using color to bring out O.ring failure. The first boxplot serves to reaffirm our first

hypothesis that O.ring failures tend to occur at lower temperatures, while the second boxplot indicates that apart from one outlier at 50psi, O.ring failures all occur at 200psi (higher pressure). Thus, without having constructed any formal models or analytics, an initial view of the data may seem to indicate that lower temperatures and higher pressures lead to O.ring failure.

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.

The authors use two types of regression to estimate the probability that an O-ring will fail. The first model that they use is a binomial logistic regression model, which $p(t, s)$ denotes the probability per joint of some thermal distress, t being the temperature and s being the pressure.

We can replicate the first model of the paper using the following R code (see below):

```
challenger_binomial <- read.table(path, header=TRUE, sep=",")
O.ring_binomial <- glm(formula=cbind(O.ring,6-O.ring) ~
                         Temp + Pressure, family=binomial(link="logit"), data=challenger_binom

#summary(O.ring_binomial)
stargazer(O.ring_binomial, header=F, title = "Regression Output for Binomial Model")
```

Table 3: Regression Output for Binomial Model

	Dependent variable:
	cbind(O.ring, 6 - O.ring)
Temp	−0.098**
	(0.045)
Pressure	0.008
	(0.008)
Constant	2.520
	(3.487)
Observations	23
Log Likelihood	−15.053
Akaike Inf. Crit.	36.106
Note:	*p<0.1; **p<0.05; ***p<0.01

A key feature of the binomial distribution is that each trial has to be independent. This may be problematic in trying to model out the likelihood of O.ring failure when using the first model. As noted in the paper (p. 947), the failure of the secondary O-ring may be conditional on the performance of the primary O-ring.

The second model for occurence of O-ring incidents, creates a binary response variable for where there is 1 if there is an incident in the flight and 0 otherwise. In this scenario, there is some "information loss" associated with reducing our dependent variable to a binary response, but we do not make the assumption that the independence of each joint is required. We agree with the paper's initial authors that a binary response variable is the optimal means of modeling because while there may be information loss, it does not make unrealistic assumptions on which the analysis will rest.

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

```
challenger2 <- read.table(path, header=TRUE, sep=",")
challenger2$O.ring = ifelse(challenger2$O.ring > 0, 1, 0)

O.ring <- glm(formula=O.ring ~ Temp + Pressure, family=binomial(link="logit"),
              data=challenger2)
#summary(O.ring)
stargazer(O.ring, header = F, type = 'latex', omit.table.layout= 'n',
          title='Binary Logistic Regression Results for Model 1')
```

Table 4: Binary Logistic Regression Results for Model 1

	Dependent variable:
	O.ring
Temp	-0.229^{**}
	(0.110)
Pressure	0.010
	(0.009)
Constant	13.292^{*}
	(7.664)
Observations	23
Log Likelihood	−9.391
Akaike Inf. Crit.	24.782

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

```
Anova(O.ring)
```

Analysis of Deviance Table (Type II tests)

Response: O.ring LR Chisq Df Pr(>Chisq)
Temp 7.7542 1 0.005359 ** Pressure 1.5331 1 0.215648
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

(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 most likely removed Pressure due to its p-value being above 0.21 (with 0.05 being the generic cutoff for statistical significance). The problem with this is that there could be interactions between temperature and pressure that their models did not consider, and we also have a relatively small sample size to begin with. EDA did seem to show that there might be some sort of relationship.

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.

```
O.ring.temp <- glm(formula= O.ring ~ Temp, family=binomial(link=logit),
                   data=challenger2)
stargazer(O.ring.temp, header = F, type = 'latex', omit.table.layout= 'n',
          title='Regression Results for Model 2')
```

Table 5: Regression Results for Model 2

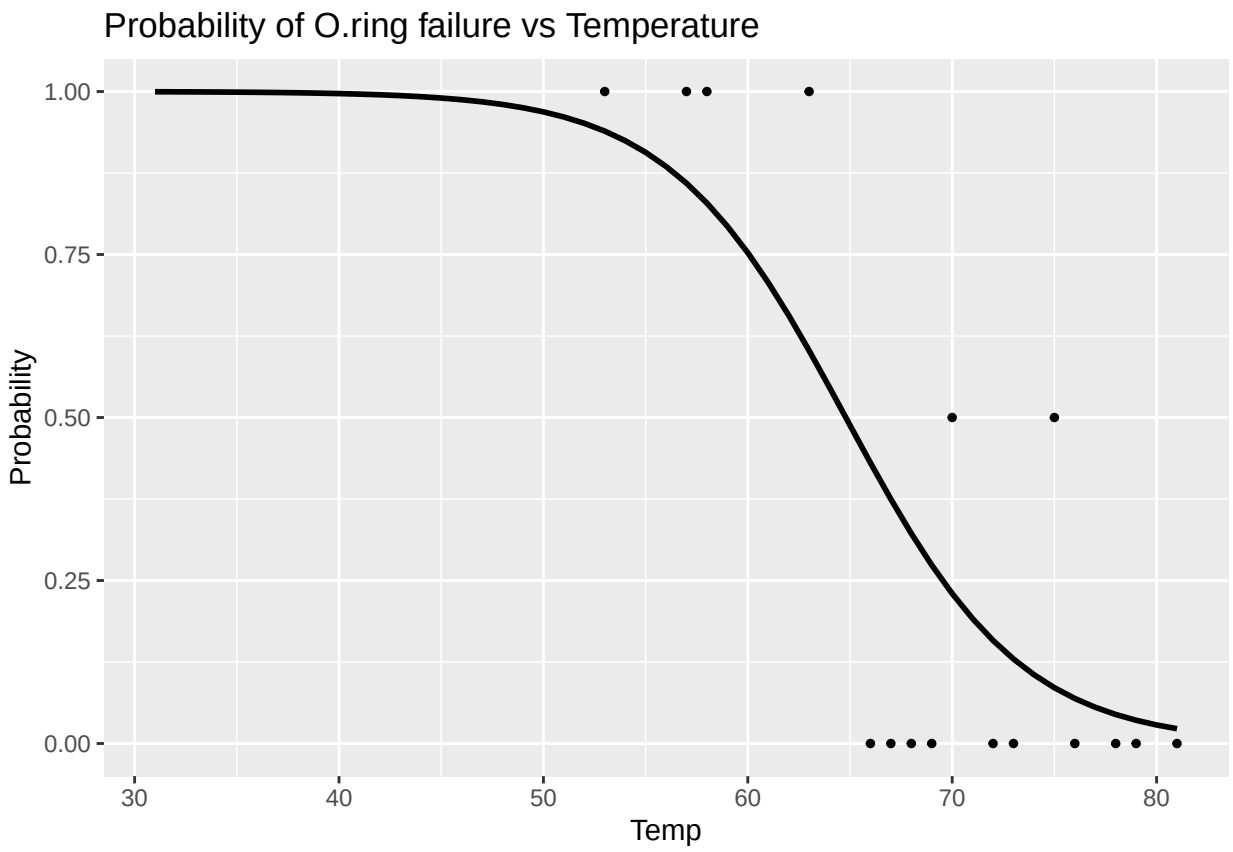
	Dependent variable:
	O.ring
Temp	-0.232^{**}
	(0.108)
Constant	15.043^{**}
	(7.379)
Observations	23
Log Likelihood	-10.158
Akaike Inf. Crit.	24.315

(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°.

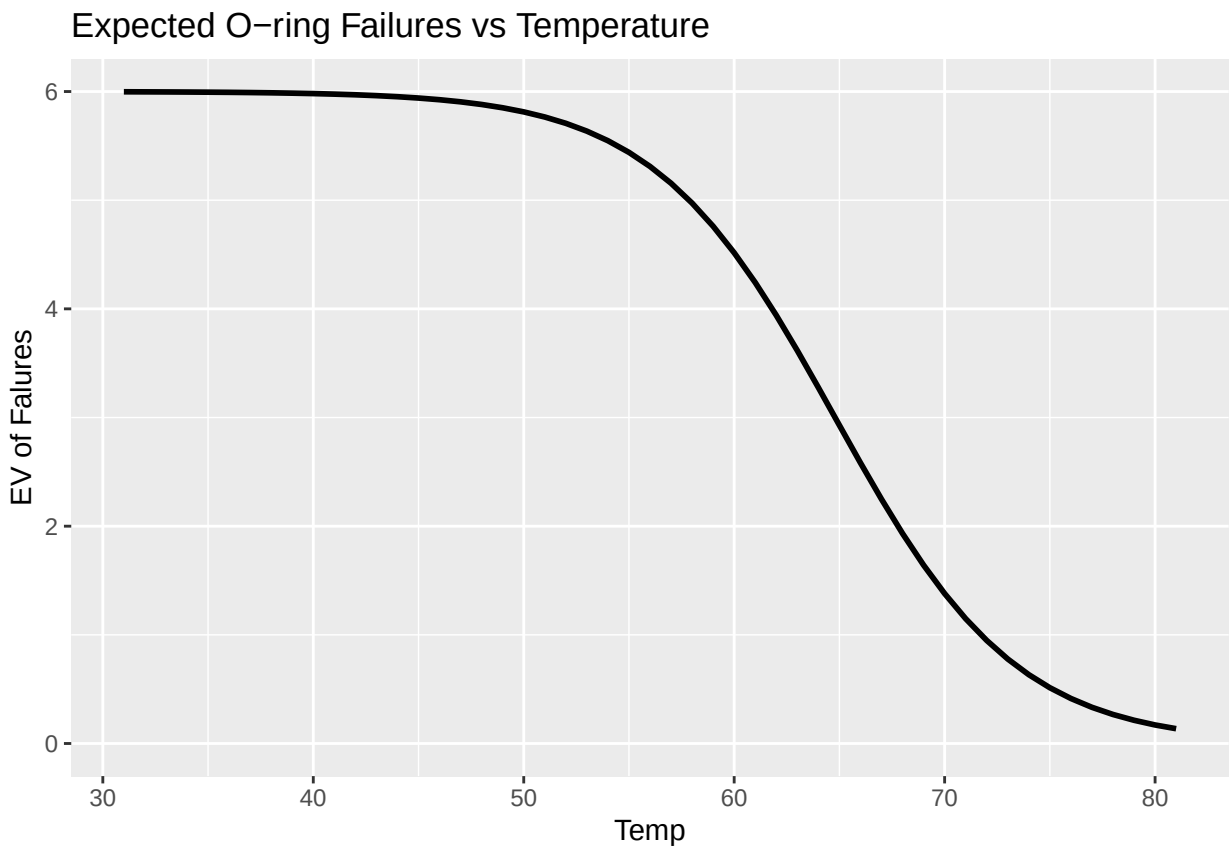
(p.91 and 92 of textbook)

```
###Jeff's code
predict.data <- as.data.frame.table(array(31:81))
predict.data$Temp <- predict.data$Freq

linear.pred<-predict(object = O.ring.temp, newdata = predict.data, type = "link",
                     se = TRUE)
#linear.pred
alpha=.05
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
```

```r
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))
CI.pi<-exp(linear.pred$fit) / (1 + exp(linear.pred$fit))

df3b <- bind_cols(temp = array(31:81), pi = CI.pi, lowerpiCI = CI.pi.lower,
                  upperpiCI = CI.pi.upper)

w <- aggregate(formula = O.ring ~ Temp, data= challenger2, FUN=sum)
n <- aggregate(formula = O.ring ~ Temp, data = challenger2, FUN=length)
w_n <- data.frame(Temperature=w$Temp, success=w$O.ring, trials=n$O.ring,
                  proportion = round(w$O.ring / n$O.ring, 4))
w_n$combine <- w$O.ring / n$O.ring
#w_n

p = ggplot() +
  geom_point(data = w_n, aes(x = Temperature, y = w$O.ring / n$O.ring), size=1) +
  geom_line(data = df3b, aes(x = temp, y = pi), size=1) +
  xlab('Temp') +
  ylab('Probability') +
  ggtitle("Probability of O.ring failure vs Temperature") +
  scale_x_continuous(limits = c(31,81)) +
  scale_y_continuous(limits = c(0, 1))
print(p)
```



```
q = ggplot() +
  geom_line(data = df3b, aes(x = temp, y = pi*6), size=1) +
  xlab('Temp') +
  ylab('EV of Falures') +
  ggtitle("Expected O-ring Failures vs Temperature") +
  scale_x_continuous(limits = c(31,81)) +
  scale_y_continuous(limits = c(0, 6))
print(q)
```

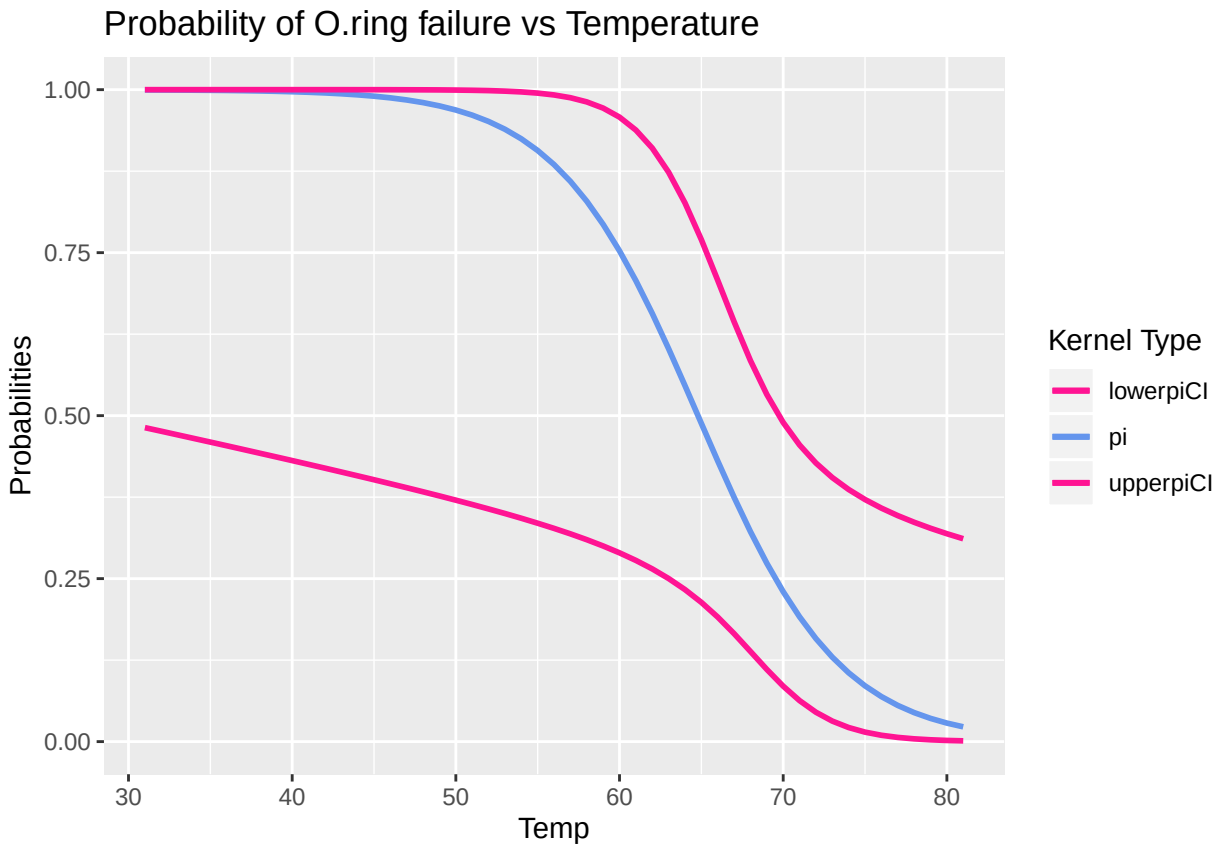


We are making an assumption of independence when calculating the Expected Value for Failures when given the tempreature. We are multiplying the probability of an incident on a flight 6 times (for 6 O-rings). From the paper, it was state that independence may be a wrong assumption to make when trying to estimate the number of total failures.

(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?

```
p = ggplot() +
  geom_line(data = df3b, aes(x = temp, y = pi, color = "pi"), size=1) +
  geom_line(data = df3b, aes(x = temp, y = lowerpiCI, color = "lowerpiCI"), size=1) +
  geom_line(data = df3b, aes(x = temp, y = upperpiCI, color = "upperpiCI"), size=1) +
  xlab('Temp') +
  ylab('Probabilities') +
  ggtitle("Probability of O.ring failure vs Temperature") +
```

```
  scale_x_continuous(limits = c(31,81)) +
  scale_y_continuous(limits = c(0, 1))  +
  scale_color_manual(values = c(
    'pi' = 'cornflowerblue',
    'lowerpiCI' = 'deeppink',
    'upperpiCI' = 'deeppink')
    ) +
  labs(color = 'Kernel Type')
print(p)
```



The bands appear much wider for the lower temperatures because there are much fewer observations for rocket launches at 31 degrees. As we saw from our EDA, the min of the temperature from the given data was 53 and the max was 81.

(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.

```
predict.data <- data.frame(Temp = 31)
predict(object=O.ring.temp, newdata=predict.data, type="response")
```

```
##         1
## 0.9996088
```

```
K <- matrix(data = c(1,31), nrow=1, ncol=2)
linear.combo <- mcprofile(object=O.ring.temp, CM=K)
ci.logit.profile <- confint(object = linear.combo, level=0.95)
#ci.logit.profile
exp(ci.logit.profile$confint) / (1+exp(ci.logit.profile$confint))
```

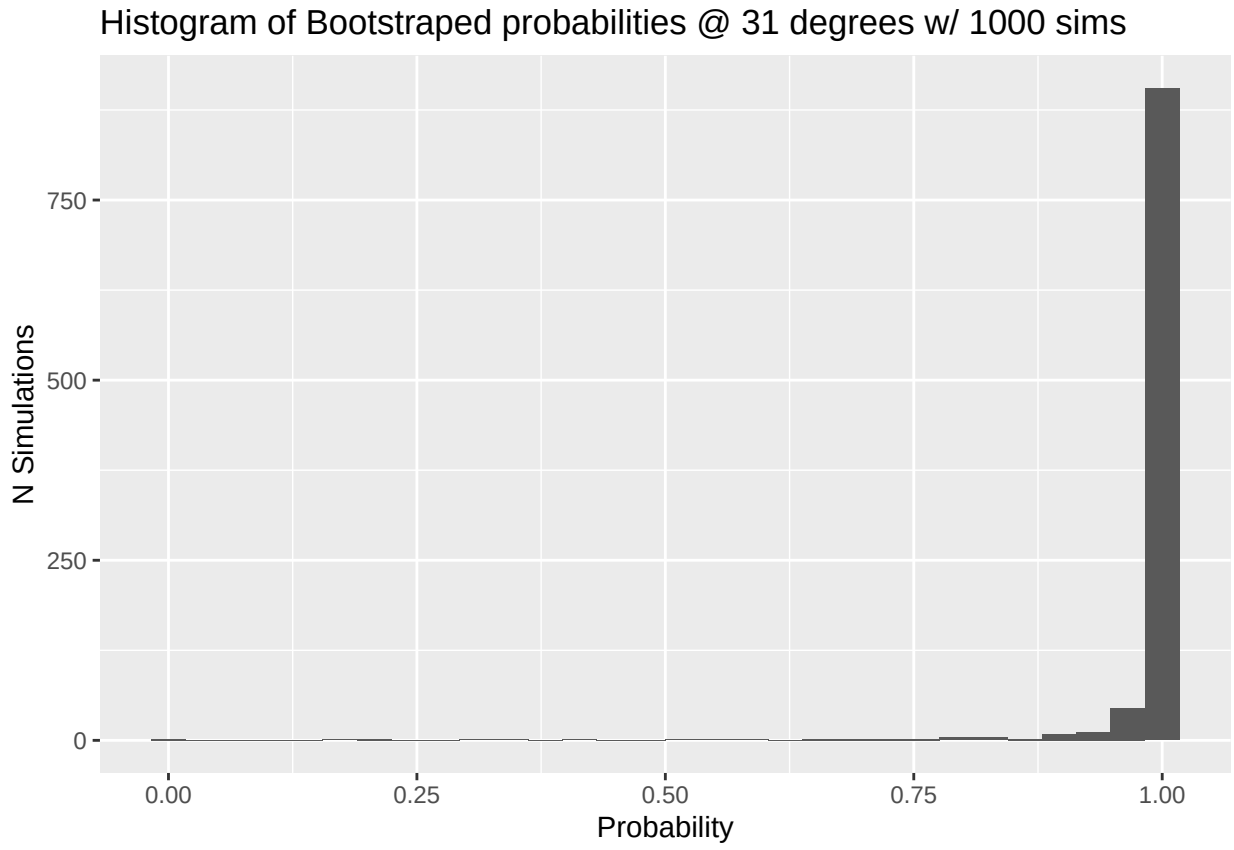
```
##       lower upper
## 1 0.8036982     1
```

In order to apply the inference procedures, you need to make assumptions about the probability distribution of pi as a statistic. According to the textbook, statistics from maximum likelihood estimation have an approximately normal probability distribution when the sample size is sufficiently large. However, at n=23, our sample size is somewhat small, and the assumption may not hold in reality, thus leading to a situation where the number of calculated intervals containing the parameter pi might actually be below our stated confidence level.

(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 31° and 72°.27

```
#start with the parameter estimates from our model and our Temp

beta0 <- O.ring.temp$coefficients[1]
beta1 <- O.ring.temp$coefficients[2]

x <- challenger$Temp
set.seed(808)

simulate <- function(){
  #Sample temp data with replacement (bootstrap)
  bootstrap_x <- sample(x, 23, replace = TRUE)
  bootstrap_x

  #Calculate pi
  pi <- 1 / (1 + exp(-(beta0 + beta1*(bootstrap_x))))
  pi
  #Estimate y with probability pi
  bootstrap_y <- rbinom(n = 23, size = 1, prob = pi)
  bootstrap_y

  #estimate with bootstrapped data
  #https://stackoverflow.com/questions/8596160/why-am-i-getting-algorithm-did-not-converge-and
  mod.fit.bs <- suppressWarnings(glm(bootstrap_y ~ bootstrap_x, family = binomial(link=logit),
  beta0.bs = mod.fit.bs$coefficients[1]
  beta1.bs = mod.fit.bs$coefficients[2]
```

```
  pi.31 <- 1 / (1 + exp(-(beta0.bs + beta1.bs*31)))
  pi.72 <- 1 / (1 + exp(-(beta0.bs + beta1.bs*72)))
  return(c(pi.31,pi.72))
}

n=1000
sim_matrix <- replicate(n,simulate())

#plot histogram of 31 degrees
p31 <- ggplot() +
        geom_histogram(aes(sim_matrix[1,]), bins=30) +
        xlab('Probability') +
        ylab('N Simulations') +
        ggtitle("Histogram of Bootstraped probabilities @ 31 degrees w/ 1000 sims")
print(p31)
```

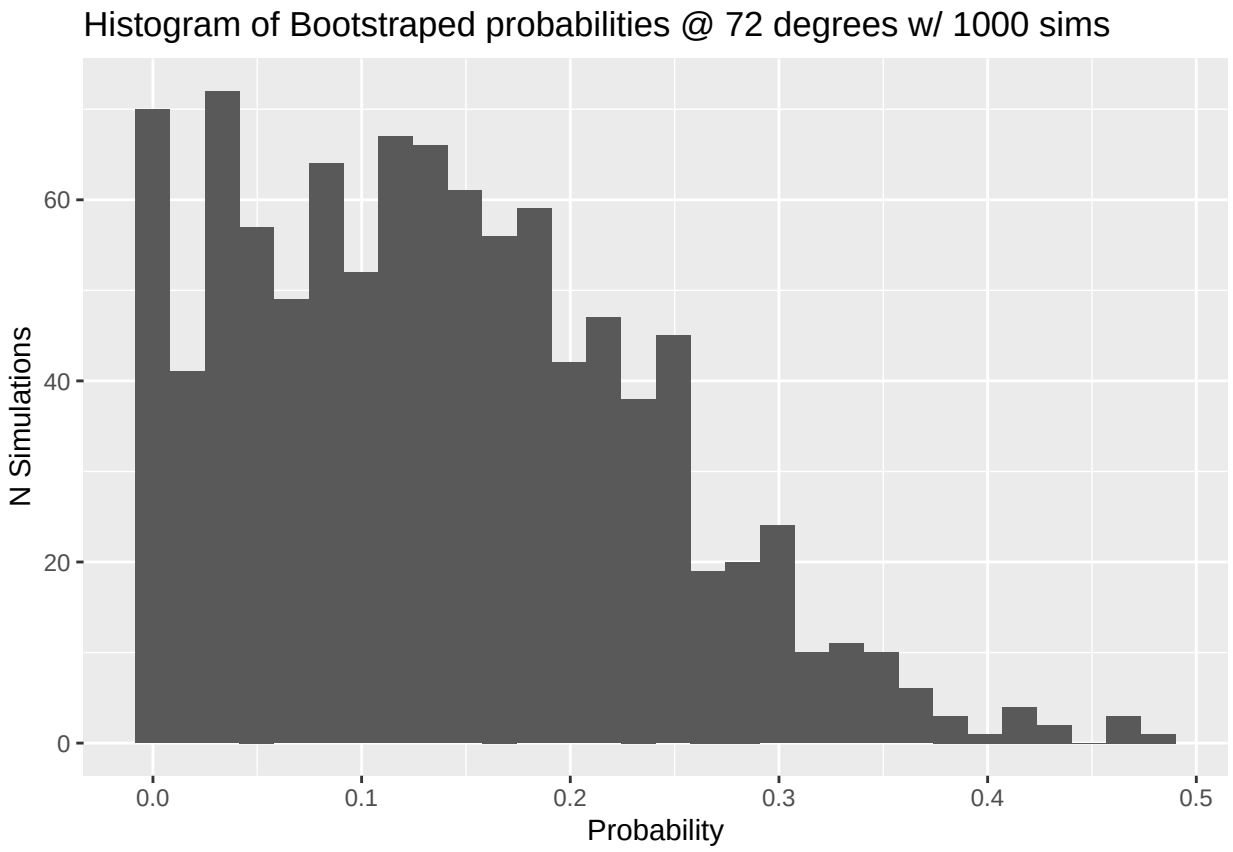



```
#return 90th conf interval
quantile(sim_matrix[1,],c(0.05,0.95))
```

```
##        5%       95%
## 0.9535073 1.0000000
```

```
#plot histogram of 72 degrees
p72 <- ggplot() +
```

```
        geom_histogram(aes(sim_matrix[2,]), bins=30) +
        xlab('Probability') +
        ylab('N Simulations') +
        ggtitle("Histogram of Bootstraped probabilities @ 72 degrees w/ 1000 sims")
print(p72)
```




```
#return 90th conf interval
quantile(sim_matrix[2,],c(0.05,0.95))
```

```
##           5%          95%
## 0.0008097175 0.3076519624
```

Because our dataset is so small here (n=23), a relatively small number of the bootstrapped models we generate do not converge. Given that we are running n=1000 simulations, we include them in our bootstrapped distributions for 31 and 72 degrees regardless.

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

```
O.ring.quadratic <- glm(formula=O.ring ~ Temp + Pressure + I(Temp^2), family=binomial(link="log
O.ring.quadratic.temp <- glm(formula=O.ring ~ Temp + I(Temp^2), family=binomial(link="logit"),

Anova(O.ring.quadratic, test="LR")
```

Analysis of Deviance Table (Type II tests)

Response: O.ring LR Chisq Df Pr(>Chisq) Temp 0.35659 1 0.5504 Pressure 0.83948 1 0.3595

I(Temp^2) 0.23287 1 0.6294

```
Anova(O.ring.quadratic.temp, test="LR")
```

Analysis of Deviance Table (Type II tests)

Response: O.ring LR Chisq Df Pr(>Chisq) Temp 1.19322 1 0.2747 I(Temp^2) 0.92649 1 0.3358

Based on the p-values for both a model with Temp and Pressure, as well as a model with only Temp, the Anova LR test shows no indication of statistically significant relationships in the presence of a quadratic term.

Part 4 (10 points)

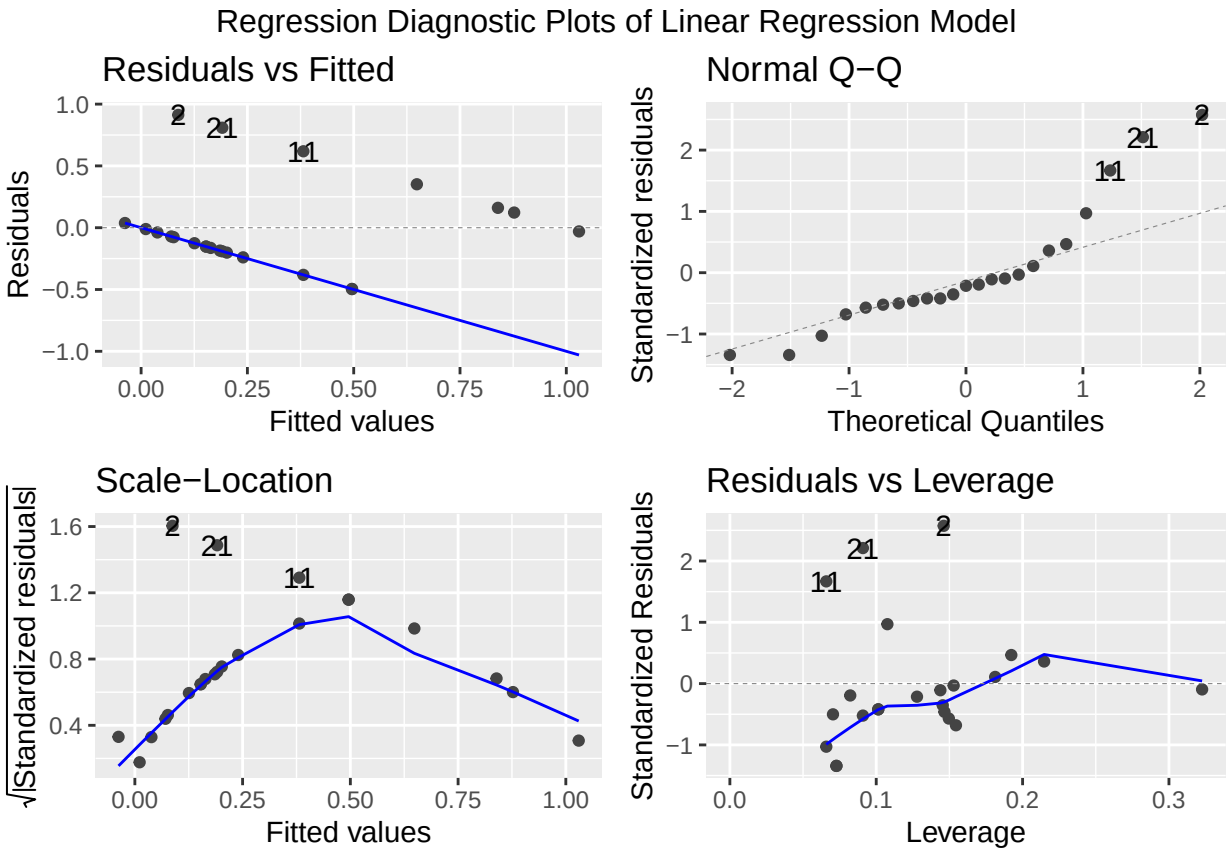
With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case? Explain why.

```
O.ring.linear <- lm(formula = O.ring ~ Temp + Pressure, data=challenger2)
#summary(O.ring.linear)
stargazer(O.ring.linear, header=F, title = "Linear Regression Model Output")
```

Table 6: Linear Regression Model Output

	Dependent variable:
	O.ring
Temp	-0.038^{***}
	(0.012)
Pressure	0.002
	(0.001)
Constant	2.659^{***}
	(0.824)
Observations	23
R^2	0.395
Adjusted R^2	0.335
Residual Std. Error	0.384 (df = 20)
F Statistic	6.534^{***} (df = 2; 20)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# Residual plots
p<-autoplot(O.ring.linear, top="Regression Diagnostic Plots of Linear Regression Model")

gridExtra::grid.arrange(grobs = p@plots,
                        top="Regression Diagnostic Plots of Linear Regression Model")
```



Linearity in parameters holds for this model. Random sampling, the next assumption behind classical linear models, is something that is dependent on the methodology behind constructing the data set. Because we are dealing with space shuttle launches, which occur infrequently, there are likely to not be an over-abundance of shuttle launches, and we may actually be dealing with a dataset that is close to or exactly matches the population of total observations available.

Below, the variance inflation factor shows that none of the values by variable are low, meaning the assumption of no perfect multicollinearity holds.

```
stargazer(vif(O.ring.linear), header=F, title = "VIF for Model 2")
```

Table 7: VIF for Model 2

Temp	Pressure
1.002	1.002

```
#vif(O.ring.linear)
```

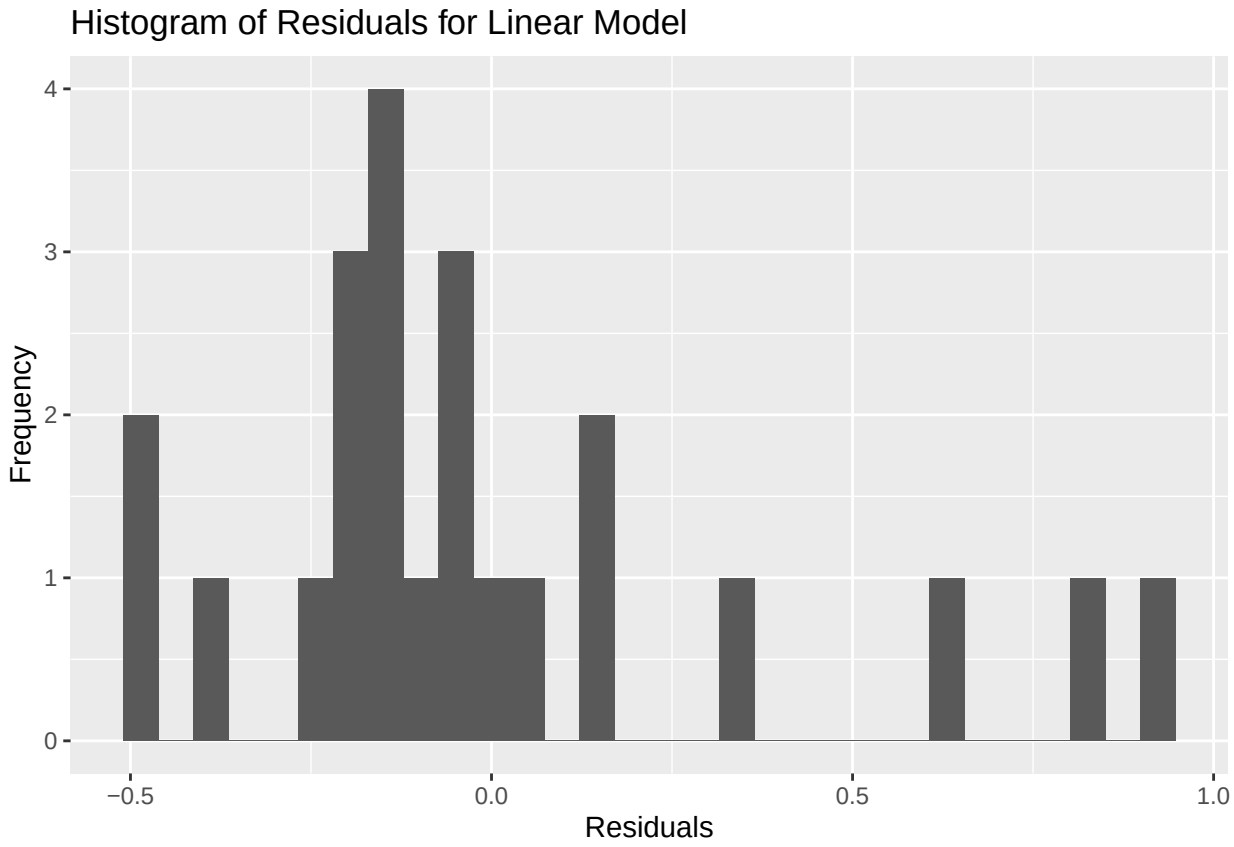
Our Residuals vs fitted chart shows that the assumption of zero conditional mean is violated, because the line strays quite far from zero.

The Scale-Location chart is not flat which shows this regression violates the assumption of homoskedasticity. Heteroskedasticity does not cause biased estimates in itself, but it does the formulas for standard errors to be inaccurate, which means we would need to use robust standard errors to

avoid problems here.

The charts below show that the normality of error assumption is not met because the residuals deviate from the straight line in our Normal-Q-Q plot (they look like a logistic function), and they do not follow a normal distribution in the second.

```
p_residuals <- ggplot() +
        geom_histogram(aes(O.ring.linear$residuals), bins=30) +
        xlab('Residuals') +
        ylab('Frequency') +
        ggtitle("Histogram of Residuals for Linear Model")
print(p_residuals)
```



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.

```
O.ring.final <- glm(formula = O.ring ~ Temp, family=binomial(link="logit"), data=challenger2)
#summary(O.ring.final)

stargazer(O.ring.final, header=F, title = "Final Model Output")
```

Table 8: Final Model Output

	Dependent variable:
	O.ring
Temp	-0.232^{**}
	(0.108)
Constant	15.043^{**}
	(7.379)
Observations	23
Log Likelihood	-10.158
Akaike Inf. Crit.	24.315
Note:	*p<0.1; **p<0.05; ***p<0.01

Ahana take because of stargazer

Don't need interaction