W271 Group Lab 1

Due 4:00pm Pacific Time Monday June 1 2020

Maria Auslander, Megha Bhardwaj, Atit Wongnophadol

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.

```
# data inspection
challenger<-read.csv("challenger.csv")
attach(challenger)
summary(challenger)
```

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

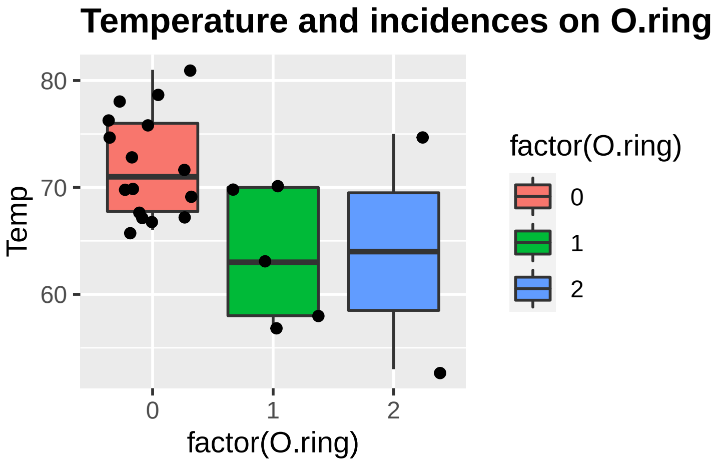
In the Challenger dataset, there are 23 observations with 5 columns. The column *Flight* appears to be just for the index of an observation, so it shouldn't have any meaningful information for our study. *Temp* and *Pressure* seem to be in a reasonable range. *O.ring* also contains a reasonable range of integers from 0 and not greater than 6 (the total number of O rings in each flight, which is basically the variable *Number*). In short, by looking at these individual variables separately without prior knowledge of the dataset, they all seem reasonable. The variables that contain the needed information in this study are *Temp*, *Pressure* as potential explanatory variables, and *O.ring* as the dependent variable of a model.

Next, pairwise relationships among *Temp*, *Pressure* and *O.ring* are plotted to identify any meaningful pattern.

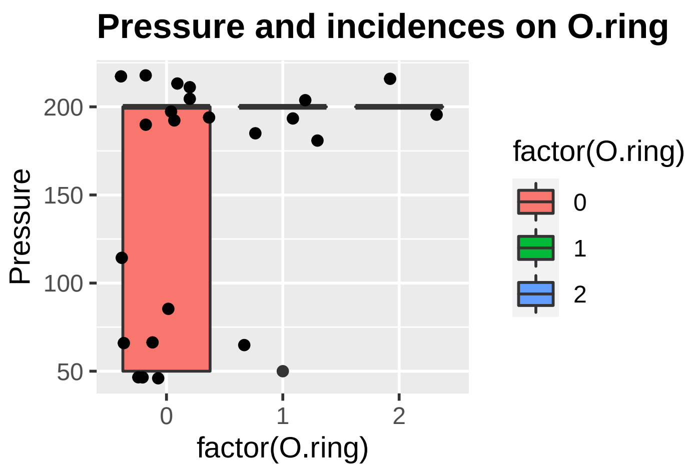
```
# pairwise relationships between variables of interest: $Temp$, $Pressure$ and $O.ring$
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

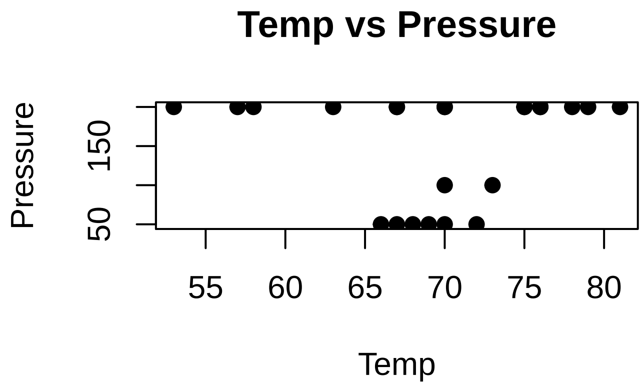
```
library(ggplot2)

# Distribution of temperature (Temp) by O.ring incidences (O.ring)
ggplot(challenger, aes(factor(O.ring), Temp)) +
geom_boxplot(aes(fill = factor(O.ring))) +
geom_jitter() +
ggtitle("Temperature and incidences on O.ring") +
theme(plot.title = element_text(lineheight=1, face="bold"))
```



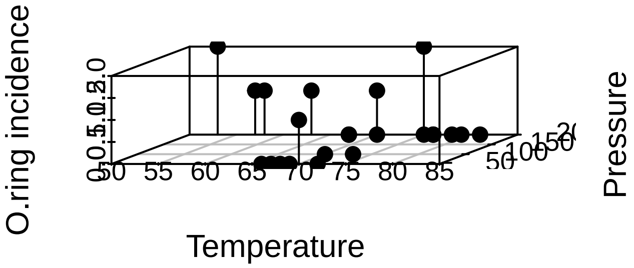
```
# Distribution of pressure (Pressure) by O.ring incidences (O.ring)
ggplot(challenger, aes(factor(O.ring), Pressure)) +
geom_boxplot(aes(fill = factor(O.ring))) +
geom_jitter() +
ggtitle("Pressure and incidences on O.ring") +
theme(plot.title = element_text(lineheight=1, face="bold"))
```



```
# Relationship between Temp and Pressure
plot(challenger$Temp, challenger$Pressure,
     main="Temp vs Pressure",
     xlab="Temp", ylab="Pressure", pch=19)
```



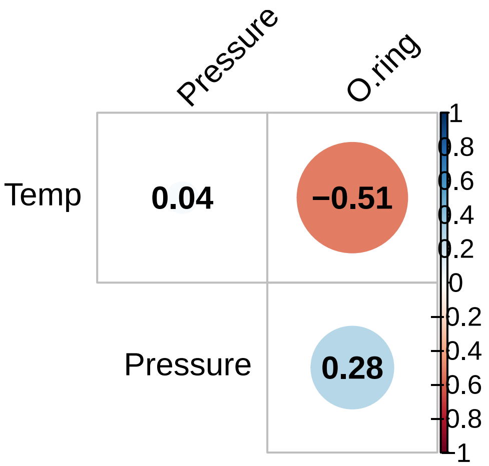
```
# 3d scatter plot for the relatinoship between all 3 variables
library("scatterplot3d")
scatterplot3d(challenger$Temp, challenger$Pressure, challenger$O.ring, pch = 19, type="h",
              xlab = "Temperature", ylab = "Pressure", zlab = "O.ring incidence")
```



```
# correlation matrix
res <- challenger[ ,c("Temp","Pressure","O.ring")]
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(res), type = "upper", order = "hclust",
         addCoef.col = "black", diag = FALSE,
         tl.col = "black", tl.srt = 45)
```



```
# reference: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram
```

From the graphs above, there is an obvious relationship between temperature and O.ring incidence, whereas temperature lower than 60 some degree tends to associate with at least one O.ring incidence. From the correlation matrix, this relationship is rather strong at -0.51.

In terms of pressure, O.ring seems to be able to withstand a wide range of pressure, as can be seen by the high density of zero-incidence plots at both end of the pressure spectrum, from the low of 50s to the high of 200s. However, there is concerning evidence that 6 flights with at least one occurred O.ring incidence were observed with the high pressure level; there were two exception flights in which O.ring incidence occurred with the pressure below 50. So the relationship between pressure and O.ring incidence might exist; from the correlation matrix, there is a slight to moderate relationship between pressure and O.ring failure at the correlation of 0.28.

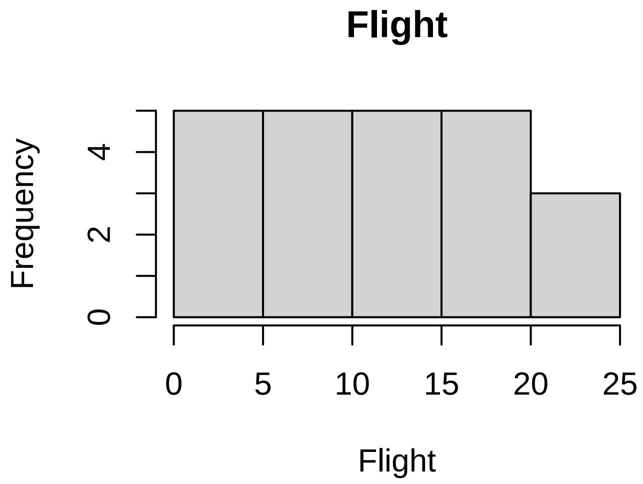
The subsequent scatter plot between temperature and pressure gives no apparent relationship that can be drawn from the data. This is confirmed by the correlation matrix in which the correlation between the two variables is close to zero at 0.04.

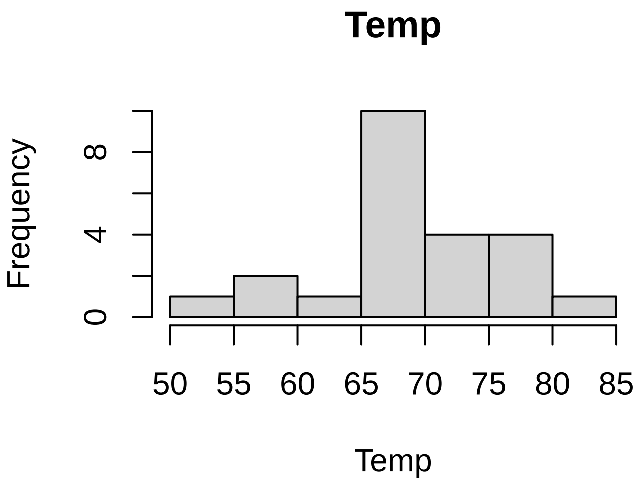
The 3d plot graph attempts to visualize a relationship between an interaction between temperature and pressure and the resulting O.ring incidence. There is no apparent evidence that an interaction effect between temperature and pressure on O.ring incidence exists.

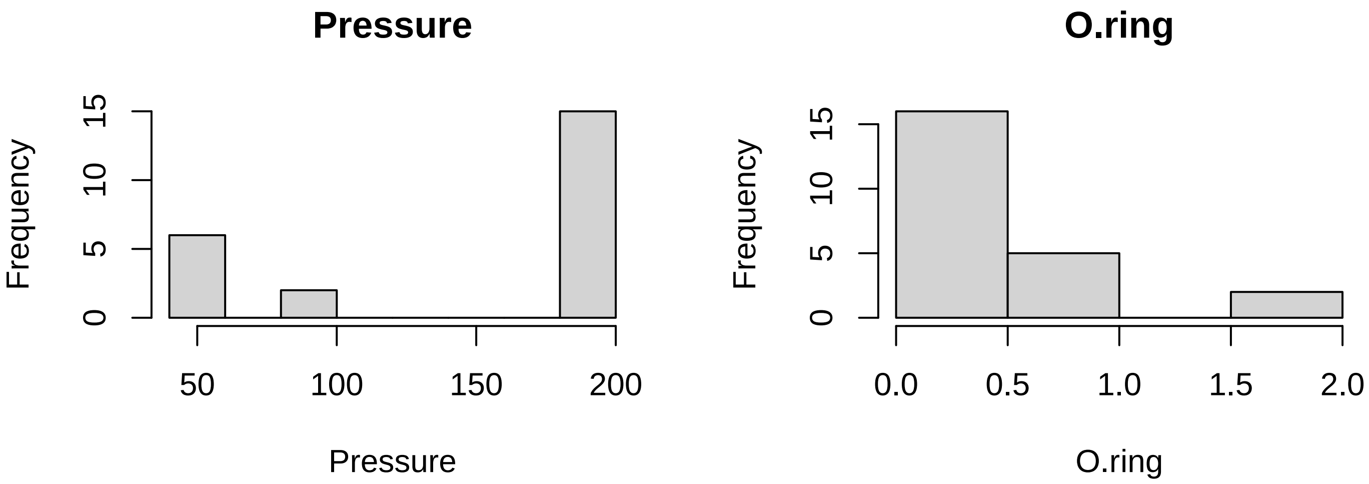
There is a relationship between temperature and O.ring incidence that justifies further exploration and model formulation. Light evidence exists for a relationship between pressure and O.ring incidence; a hypothesis on pressure factor may be formed and tested to verify if it plays a significant role in explaining O.ring incidence.

Below, each histogram of values for columns are shown:

```
columns<-names(challenger)
challenger.eda.matrix <- matrix(data=NA,nrow=length(challenger),ncol=2)

counter=1
for (col in columns) {
    if (col!="Number") {
      hist(challenger[[col]],main=col,xlab=col)
      num_missing<-sum(is.na(challenger[[col]]))
      challenger.eda.matrix[counter,] = c(col,num_missing)
      counter = counter+1}
}
```







Looking at the histograms and summary statistics (shown earlier in the EDA), the low *Pressure* and *Temp* data points may be anomalies and a cause for concern when creating a model. With a low concentration of values at the lower end, the models created will be less reliable for cases where low *Pressure* and *Temperature* values occur. The large O.ring values seem to be anomalies and potentially a cause for concern as a solid predictive model would benefit from more evenly spread distributions of independent variables. Number maintains a value of 6 throughout the dataset.

The data below shows the count of missing values for each column within the dataset, there is no missing data.

```
challenger.eda.matrix[,2]
```

```
## [1] "0" "0" "0" "0" NA
```

Part 2 (20 points)

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

The response variable is O.ring, and the explanatory variables are Temp and Pressure. Complete the following:

(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.

Independence assumption for logistic regression - Characterizing the distributions of sums of random variables usually rest on two key results (1) law of large numbers (L.L.N) and (2) central limit theorems (C.L.T) Independence of random variables (standard i.i.d assumptions) are usually essential to obtaining such results. In a few cases we still have LLN and CLT theorems that relax either the independence assumption or the identical assumption. However these are generally much weaker, are more obscure and are not as general and widely applicable as the theorems that use i.i.d assumptions. Typically, in the absence of i.i.d, central limit theorems that show asymptotic normality or the Gaussianness of sums of independent identical random variables will often not hold. We want limit theorems to hold.The epistemological value of probability theory is revealed only by limit theorems and limit theorems are much easier to obtain when we have independent observations. In simple terms, the independence assumption when true, helps cancel out variations, which helps you guarantee consistency of a method and lets you converge to your true result faster or with fewer samples. The authors suggest two models, one which uses binomial distribution estimating probabilty of O-ring failure at each joint and treating the 6 O-rings as 6 independent trials of the binomial distribution. The assumption of independence is important in this model, since the problem in this case is modelled as a binomial logistic regression with each O.ring failure treated as an independent trial. The trial size in this case, n = 6 and the number of failures, w = number of O.ring failed. The issue is that the O-rings are on every rocket (6/rocket), so there may be some dependencies on each other given they are located on the same entity.

A subsequent analysis done by the author alleviates this concern, as the author performed a binary logistic regression model. In this model, the assumption was that a failure was counted if at least one O.ring on the rocket failed. In this case the outcome y =0 when number of O.ring failure equal to 0, otherwise 1. This model is not based on the same assumption of independence as the binomial distribution since the probability of success/failure is only depending on whether there was any O.ring failure or not.

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

There are two possible models–binomial and binary. We predict and analyze both models throughout the analysis. The binary model has binary outcome for O.ring failures, 1 if any of the 6 O.rings failed and 0 otherwise. The binomial model uses the proportion of O.rings that failed of 6 as the dependent variable.

```r
challenger$percent.O.ring.fail<- with(challenger, O.ring/Number)
challenger_glm<-glm(percent.O.ring.fail~Temp+Pressure,
                  family=binomial(link="logit")
                  ,data=challenger,weights=Number)
summary(challenger_glm)
```

Binomial Regression

```
##
## Call:
## glm(formula = percent.O.ring.fail ~ Temp + Pressure, family = binomial(link = "logit"),
##     data = challenger, 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
```

$$logit(\hat{\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$logit(\hat{\pi}) = 2.520195 - 0.098297 x_1 + 0.008484 x_2$$

Where $x_1 = Temp$ and $x_2 = Pressure$

Binary Regression In the case of binary regression, we assume at least one O.ring failed is countered as a failed flight (1) while no failures is counted as a successful flight (0).

```
bin.o.ring<-ifelse(O.ring>0,1,0)
challenger_glm_binary<-glm(bin.o.ring >0 ~Temp+Pressure,
                  family=binomial(link="logit")
                  ,data=challenger)
summary(challenger_glm_binary)
```

```
##
## Call:
## glm(formula = bin.o.ring > 0 ~ Temp + Pressure, family = binomial(link = "logit"),
##     data = challenger)
##
## 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 .
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## Pressure     0.010400   0.008979   1.158   0.2468
## ---
## 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
```

$$logit(\hat{\pi}) = 13.292360 - 0.228671x_1 + 0.010400x_2$$

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

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: ’car’
```

```
## The following object is masked from ’package:dplyr’:
##
##     recode
```

```
Anova(challenger_glm, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: percent.O.ring.fail
##          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
```

```
Anova(challenger_glm_binary, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: bin.o.ring > 0
##          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
```

```
anova(challenger_glm_binary, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bin.o.ring > 0
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        22     28.267
## Temp      1   7.9520        21     20.315 0.004804 **
## Pressure  1   1.5331        20     18.782 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?

In both models, *Pressure* was likely taken from the model because the according p-value was large (p>alpha=0.05) while the p-value for *Temp* was small (p<alpha=0.05). There is not enough evidence to assume that *Pressure* is important in the explanatory model, however, removing *Pressure* from the model may be an issue if *Pressure* could be the part of an interaction term or a transformation in a future model.

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.

Binomial Model

```
challenger_glm_2<-glm(percent.O.ring.fail~Temp,family=binomial(link="logit"),data=challenger,w
summary(challenger_glm_2)
```

```
##
## Call:
## glm(formula = percent.O.ring.fail ~ Temp, family = binomial(link = "logit"),
##     data = challenger, 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
```

$$logit(\hat{\pi}) = \beta_0 + \beta_1 x_1$$
$$logit(\hat{\pi}) = 5.0850 - 0.1156 x_1$$

Where $x_1 = Temp$

Binary Model

```
challenger_glm_binary<-glm(bin.o.ring ~Temp,family=binomial(link="logit"),data=challenger)
summary(challenger_glm_binary)
```

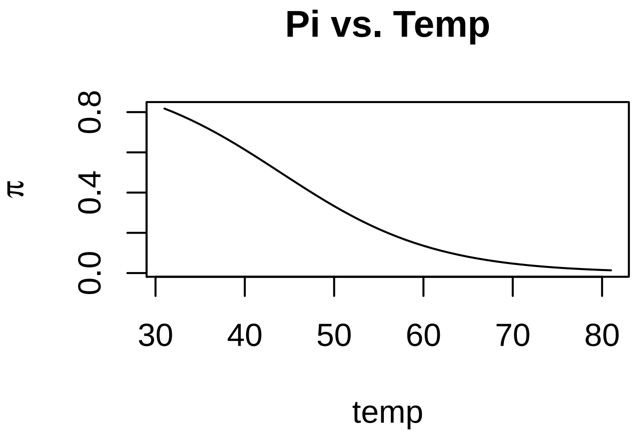
```
##
## Call:
## glm(formula = bin.o.ring ~ Temp, family = binomial(link = "logit"),
##     data = challenger)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0611  -0.7613  -0.3783   0.4524   2.2175
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (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
```

$$logit(\hat{\pi}) = 15.0429 - 0.2322x_1$$

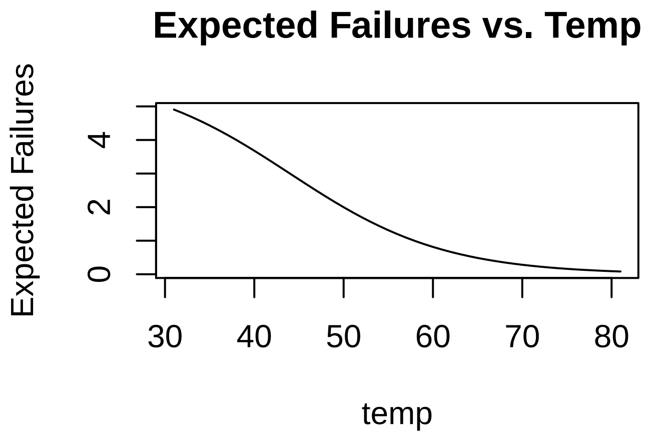
(b) Construct two plots: (1) π vs. Temp and (2) Expected Failures vs. Temp. Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

Plots are constructed for the binomial model.

```
temp <-31:81 # range of temperature
press<-rep(c(200), times = 51)
pi.hat.binomial <- predict(challenger_glm_2, list(Temp = temp), type="response")

plot(x = temp, y = pi.hat.binomial, type = "l", ylab = expression(pi), main = "Pi vs. Temp")
```



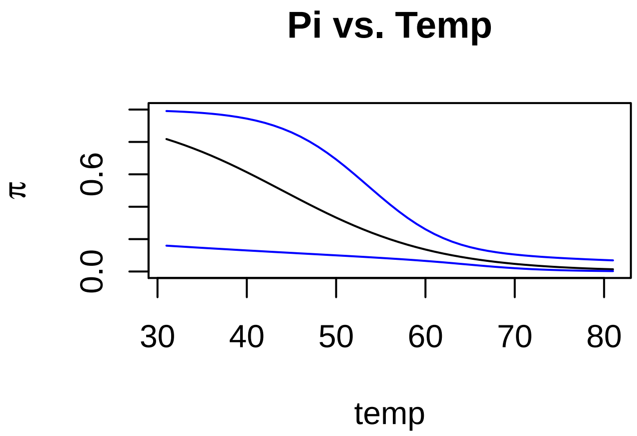
```
plot(x = temp, y= pi.hat.binomial*6, type="l", ylab = "Expected Failures", main="Expected Failu
```



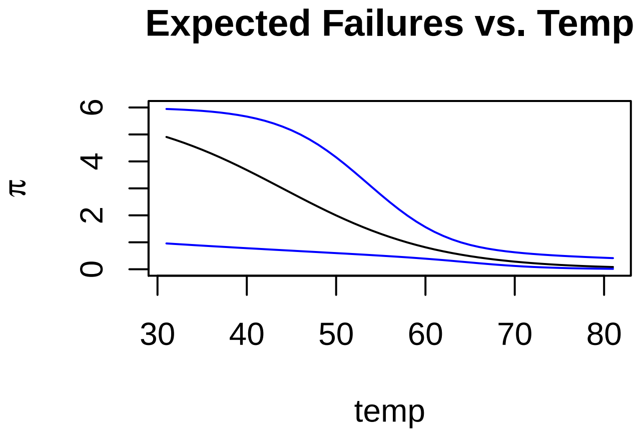
(c) Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

Binomial Model

```
temp <-31:81
pi_val <- predict(challenger_glm_2, list(Temp = temp),type="response")
plot(x=temp, y=pi_val,type="l",ylab=expression(pi),main="Pi vs. Temp",ylim=c(0,1))

inverse_logit <- function(x){
  exp(x)/(1+exp(x))
}

#Wald interval
predicted <- predict(challenger_glm_2, list(Temp = temp), se.fit = TRUE)
pred0 <- predicted$fit
pred <- inverse_logit(predicted$fit)
alpha <- 0.05
sc <- abs(qnorm(alpha/2))  ## Normal approx. to likelihood
lwr<-inverse_logit(pred0-sc*predicted$se.fit)
upr<-inverse_logit(pred0+sc*predicted$se.fit)
lines(temp, upr, col="blue")
lines(temp, lwr, col="blue")
```



```
plot(x=temp, y=pi_val*6,type="l",ylab=expression(pi),main="Expected Failures vs. Temp",ylim=c(
lines(temp, upr*6, col="blue")
lines(temp, lwr*6, col="blue")
```



There are fewer estimations for the lower temperatures in the dataset, so there is higher uncertainty at lower temperatures, this is reflected in the wide confidence interval range.

(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.

```
single_pred<-predict(challenger_glm_2, list(Temp = c(31)),type="link",se.fit=TRUE)

pred0 <- single_pred$fit
pred <- inverse_logit(single_pred$fit)
ci <- 0.95
```

```
sc <- abs(qnorm((1-ci)/2))  ## Normal approx. to likelihood
lwr=inverse_logit(pred0-(sc*single_pred$se.fit))
upr=inverse_logit(pred0+(sc*single_pred$se.fit))

lwr
```

```
##         1
## 0.1596025
```

```
upr
```

```
##         1
## 0.9906582
```

The 95% confidence interval is (0.1596025,0.9906582). We need to assume the same trend that occurs at the temperature range available in the dataset (53,81) also occurs at the lower temperature, that the model is still applicable.

(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

```
# bootstrap function for simulating pi.hat
bootstrap <- function(temp){
  # bootstrap resampling of the dataset
  challenger.rep <- sample_frac(challenger, replace = TRUE)
  # fit the model
  challenger.rep$percent.O.ring.fail <- with(challenger.rep, O.ring/Number)
  mod.rep.fit <- suppressWarnings(glm(percent.O.ring.fail~Temp,
                    family=binomial(link="logit"),
                    data=challenger.rep))
  # predict pi.hat
  pi.hat <- predict(mod.rep.fit, list(Temp = temp), type="response")

  return(pi.hat)
}
# number of trials
set.seed(2020)
N <- 10000
# save the list of simulated pi.hat
rep.pi.hat.31 <- replicate(N, bootstrap(31))
```

```r
rep.pi.hat.72 <- replicate(N, bootstrap(72))
# 90% confidence interval of pi.hat
quantile(rep.pi.hat.31, c(.05, .50, .95)) # CI+median at 31 degrees
```

```
##        5%       50%       95% 
## 0.1913729 0.8308555 0.9919977
```

```r
quantile(rep.pi.hat.72, c(.05, .50, .95)) # CI+median at 72 degrees
```

```
##          5%         50%         95% 
## 0.005952112 0.035032178 0.078022870
```

```r
# average pi.hat
mean(rep.pi.hat.31)
```

```
## [1] 0.7503743
```

```r
mean(rep.pi.hat.72)
```

```
## [1] 0.03782724
```

The parametric bootstrap method yields the following confidence intervals [0.20435, 0.99203] and [0.00616, 0.07922] for temperature 31 degree and 72 degree respectively.

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

Binomial Model

```r
quad_model<-glm(percent.O.ring.fail~Temp+I(Temp**2),family=binomial(link=logit),data=challenge
anova(challenger_glm_2,quad_model,test="LR")
```

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

Using the binomial model, the p-value of the likelihood ratio test comparing the model with a quadratic term of temperature and without a quadratic term of temperature is large (0.4818 >0.05), indicating that there is not enough evidence suggesting a quadratic term is needed in the model.

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.

In the final model, we use only *Temp* as a independent variable as *Pressure* was not statistically significant and there is no evidence to suggest a quadratic term of *Temp* is needed. We use the dependent variable from the binomial model (`percent.O.ring.fail`) in the linear model, as the continuous nature of the variable is more conducive to the linear model.

```r
linear_model_temp_only<-lm(percent.O.ring.fail~Temp,data=challenger)
summary(linear_model_temp_only)
```

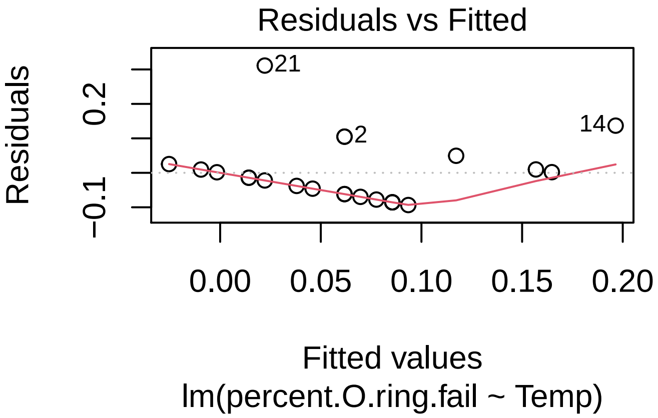
```
##
## Call:
## lm(formula = percent.O.ring.fail ~ Temp, data = challenger)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09347 -0.06573 -0.01423  0.01760  0.31118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.616402   0.203252   3.033  0.00633 **
## Temp        -0.007923   0.002907  -2.725  0.01268 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.09624 on 21 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.2261
## F-statistic: 7.426 on 1 and 21 DF,  p-value: 0.01268
```

CLM Assumptions and Model Diagnostics

Linear Population Model Under the linear population model assumption, the relationship between dependent and independent variables is meant to be linear. Looking at the global stat p-value (1.670532e-05) it appears this is not the case. This p-value indicates a non-linear relationship between variables.

Random Sampling The random sampling assumption is best tested by reviewing the methods of data collection for the dataset. If the data was collected in a random, unbiased way the assumption would be fulfilled, if not, the assumption would be rejected.

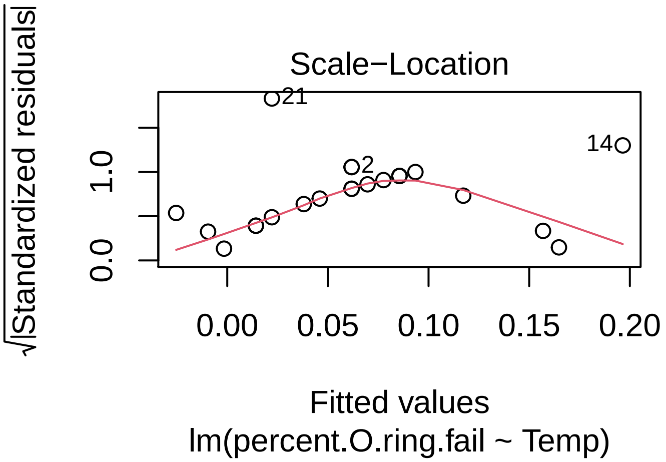
```r
plot(linear_model_temp_only, which = 1)
```



Zero Conditional Mean

Looking at the residual vs. fitted plot above, there is a violation of the zero conditional mean assumption. In the case of the zero conditional mean assumption being fulfilled, we would expect residuals to be spaced evenly around the 0 line, this is not the case. Through the graph it appears there may be some sort of curved relationship present.

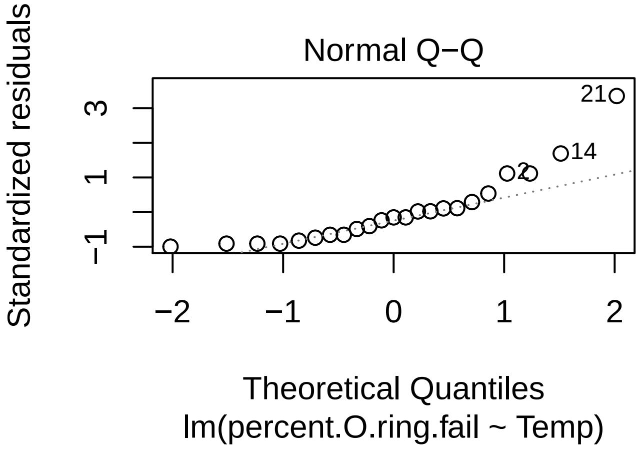
```r
plot(linear_model_temp_only, which = 3)
```



Homoscedasticity

The homoscedasticity assumption states that the variance of a residual should be about the same for any value of x. The scale-location plot shows a violation of this assumption. If the assumption were fulfilled we would expect the points to be spaced evenly and randomly above and below the line, in this case there appears to be evidence of a curved relationship.

```
plot(linear_model_temp_only, which = 2)
```



Normality of Errors

Looking at the normal Q-Q plot, the normality of errors assumption is not satisfied. The normal Q-Q plot shows many residuals straying far from the line, indicating a distribution that is not normal.

Based on the analyses above, linear regression model would not be our choice considering it violates several key assumptions. Moreover, a prediction from the linear model may not be in the range [0,1], required in estimating the probability of failure in our study. Given these reasons and assuming that any individual O-ring failure is a cause for concern our choice of model would be the binary logit model.

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.

We will evaluate the odds ration of O-ring failure for a 10 degree increase in temperature. For a 10 degree increate in temperature, the odds ratio of O.ring failure changes by 0.098 Next, we do a profile LRT test for probability of failure at temperature = 30 degrees and temperature = 80 degrees. With 95% confidence the probability of atleast 1 O.ring failure is between 0.8 to 1 with estimated probabilty = .99 At temperature = 80 degrees - With 95% confidence the probability of atleast 1 O.ring failure is between 0.0 to .22 and the estimated probabilty = .03

```
library (package = mcprofile)
c=10
OR = exp(coefficients(challenger_glm_binary)[2]*c)
OR
```

```
##       Temp
## 0.09811378
```

```
K <- matrix (data = c (1,30), nrow = 1, ncol = 2)
linear.combo <- mcprofile (object = challenger_glm_binary, CM = K)
ci.logit.profile <- confint (object = linear.combo, level = 0.95)
inverse_logit(ci.logit.profile$estimate)
```

```
##     Estimate
## C1 0.9996898
```

```
inverse_logit(ci.logit.profile$confint)
```

```
##       lower upper
## 1 0.8134177     1
```

```
K <- matrix (data = c (1,80), nrow = 1, ncol = 2)
linear.combo <- mcprofile (object = challenger_glm_binary, CM = K)
ci.logit.profile <- confint (object = linear.combo, level = 0.95)
inverse_logit(ci.logit.profile$estimate)
```

```
##      Estimate
## C1 0.02846733
```

```
inverse_logit(ci.logit.profile$confint)
```

```
##           lower    upper
## 1 0.0007961551 0.219677
```