W271 Group Lab 1

Due 4:00pm Pacific Time Monday June 1 2020

Instructions (Please Read Carefully):

- 20 page limit (strict)

- Do not modify fontsize, margin or line_spacing settings

- One student from each group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded.

- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation

- Submit two files:

  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file

  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted

- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:

  – `StanCartman_KennyKyle_Lab1.Rmd`
  – `StanCartman_KennyKyle_Lab1.pdf`

- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files

- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modelling must be clearly shown and explained

- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc

- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file

- Incorrectly following submission instructions results in deduction of grades

- Students are expected to act with regard to UC Berkeley Academic Integrity.

Investigation of the 1989 Space Shuttle Challenger Accident

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

Load libraries

```r
library(ggplot2) # for plotting
library(GGally) # for scatterplot matrices
library(gridExtra) # for arranging multiple plots together
library(car) # For Likelihood Ratio Tests on the fly
library(data.table) # to enable creation and coercion of data tables
library(stargazer) # to tabulate regression tables
library(skimr) # for basic EDA
#library(Hmisc) # for basic EDA
library(mcprofile) # for profile likelihoods
library(dplyr) # for coercing dataframes and summary statistic
library(boot) # framework for bootstrapping
#library(plotly) # For 3d scatterplot
```

Load data

```r
df = read.table("challenger.csv",
                header=T, sep=",")
```

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.

EDA

The dataset used in the examination of probability of O-ring failure in Challenger's previous space shuttle launches by Dalel et al. (1989) included 23 rows and 5 variables. The Flight column is made up of unique identifiers representing the flight number of a launch. The Temp column contains information about the temperature (F) at launch, and ranges between 53 and 81 degrees for the 23 data points. The Pressure column contains information about the combustion pressure (psi) at launch, and takes on only 3 values among the 23 data points: 50 100, or 200. The O.ring column represents the number of primary field O-ring failures in a launch, which is our outcome of interest. For this dataset, the O.ring column contains only the values 0, 1, or 2. The Number column contains

the total number of primary field O-rings, which is 6 for all launches (three each for the two booster rockets). Each of the 5 variables were treated as `int` classes.

Because the O.ring column contains more than 2 outcomes, we also created a new factor variable to binarize the dependent variable. This new column, O.ringBinary, would have an outcome of 1 if the number of primary O-ring failures in a launch was 1 or greater.

```r
#If we don't use a binary model, then we should just delete this

# Create binary column for O.ring failture
df$O.ringBinary = ifelse(test = df$O.ring > 0, yes = 1, no = 0)
# Convert O.ringBinary to factor
df$O.ringBinary = as.factor(df$O.ringBinary)
# Sanity check
str(df)
```

```
## 'data.frame':    23 obs. of  6 variables:
##  $ Flight      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Temp        : int  66 70 69 68 67 72 73 70 57 63 ...
##  $ Pressure    : int  50 50 50 50 50 50 100 100 200 200 ...
##  $ O.ring      : int  0 1 0 0 0 0 0 0 1 1 ...
##  $ Number      : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ O.ringBinary: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 ...
```

```r
# Summary statistics
#skim(df) this is not knitting

summary(df)
```

```
##      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  O.ringBinary
##  Min.   :6   0:16
##  1st Qu.:6   1: 7
##  Median :6
##  Mean   :6
##  3rd Qu.:6
##  Max.   :6
```

```r
# What are the unique values?
lapply(df[c('Temp','Pressure','O.ring')], unique)
```

```
## $Temp
##  [1] 66 70 69 68 67 72 73 57 63 78 53 75 81 76 79 58
##
## $Pressure
```

```
## [1]  50 100 200
##
## $O.ring
## [1] 0 1 2
```

To ensure that the data do not contain anomolous observations, we next inspected the unique values for each variables. The values for each variable seem reasonable according to their description and we do not observe any top coded or bottom coded values. We also noted that there are no missing values in any of the variables.

Univariate Analysis

The three variables of interest for modeling purposes in the paper by Dalel et al. were temperature, pressure, and number of O-ring failures.

Distribution of Temperature
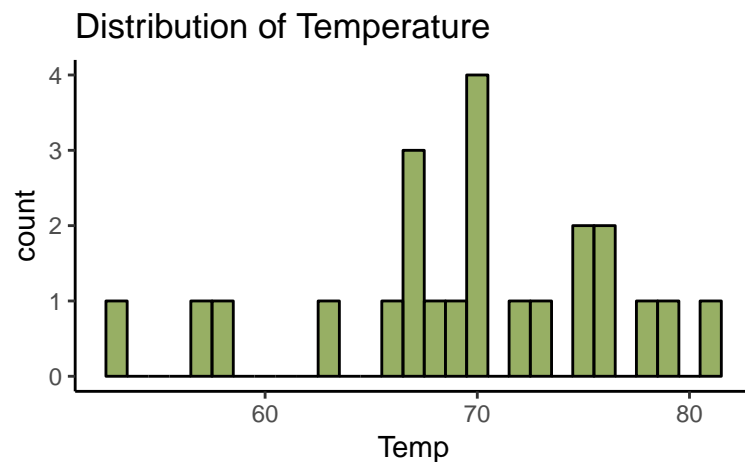
From the histogram, the temperature is seen to peak at 70 degrees with 4 observations and tapered off on either side. Based on the plot, the distribution seems to most closely resemble a normal distribution.

Distribution of Pressure
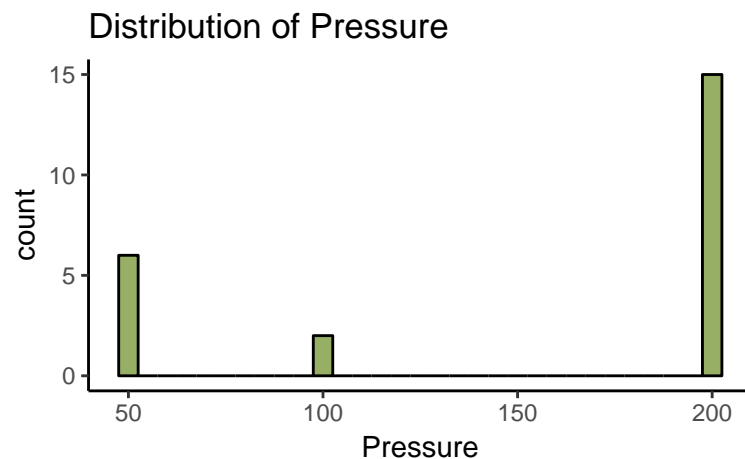
Pressure is seen to only take on 3 unique values, with the most observations having a pressure of 200. Though it is difficult to say with such few unique values, the distribution seems to most closely resemble a bimodal distribution.

Figure 1: Distribution of O−ring failures
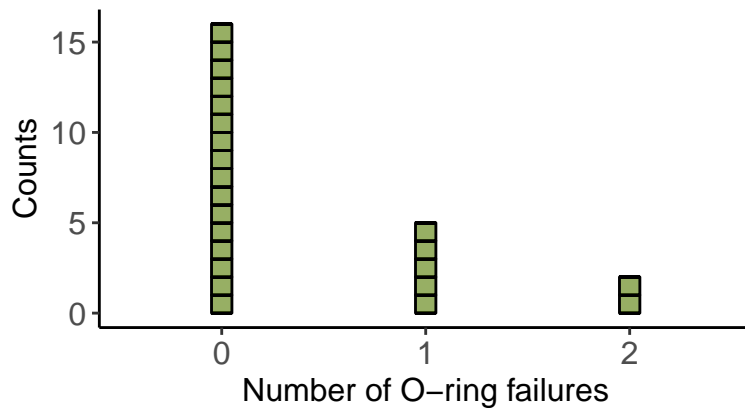
The response variable `O.ring` is positively skewed with the majority of O-ring failures at 0. We observe that there are no O-ring failures in 16 flights, one O-ring failure in 5 flights, and two O-ring failures in 2 flights.

Bivariate Analysis

- How are O.Rings related to putative predictors, like temperature?

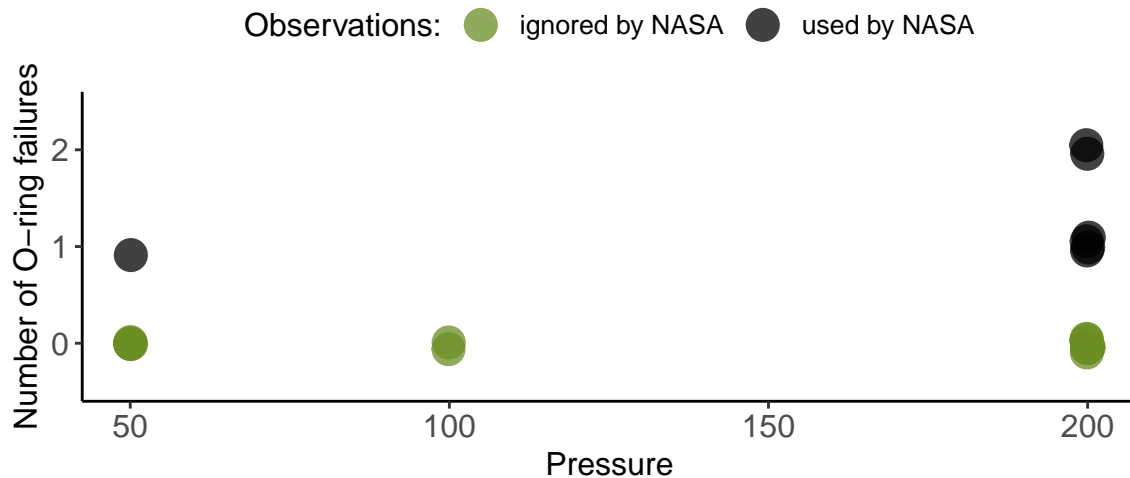**Figure 2:** O−ring failure as a function of temperature (°F)
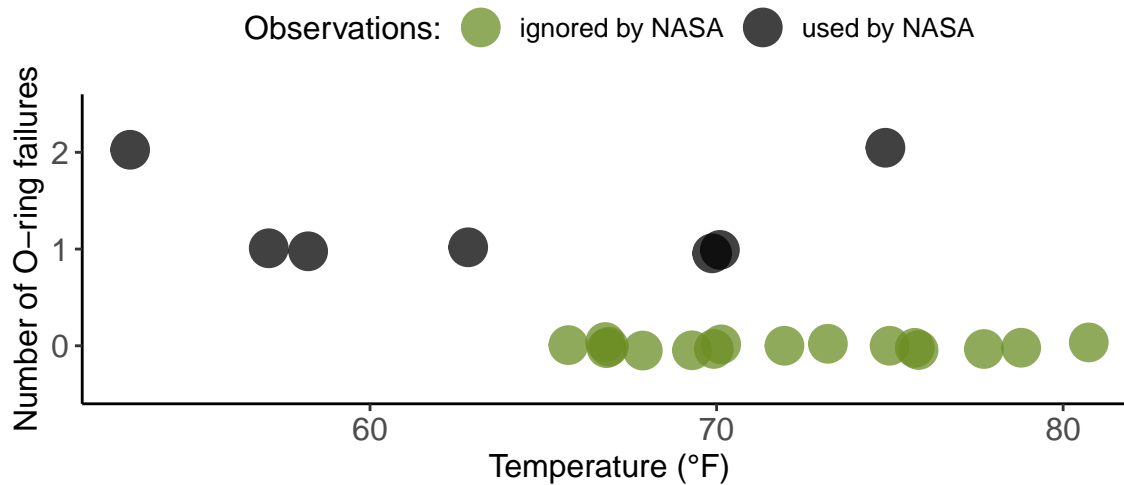
- At what temperature do 2 O.rings fail?

```r
# Cross tabulation
table(df$O.ring, df$Temp)
## At what temperatures are there 0,1 and 2 failures
# 0
sort(df$Temp[df$O.ring == 0])
# 1
sort(df$Temp[df$O.ring == 1])
# 2
sort(df$Temp[df$O.ring == 2])
```

- And pressure?

Figure 3: O−ring failure as a function of pressure

Observations: ● ignored by NASA ● used by NASA

- Are O-ring failures related to time/flight number

Figure 4: O−ring failure as a function of flight number/time

Observations: ● ignored by NASA ● used by NASA
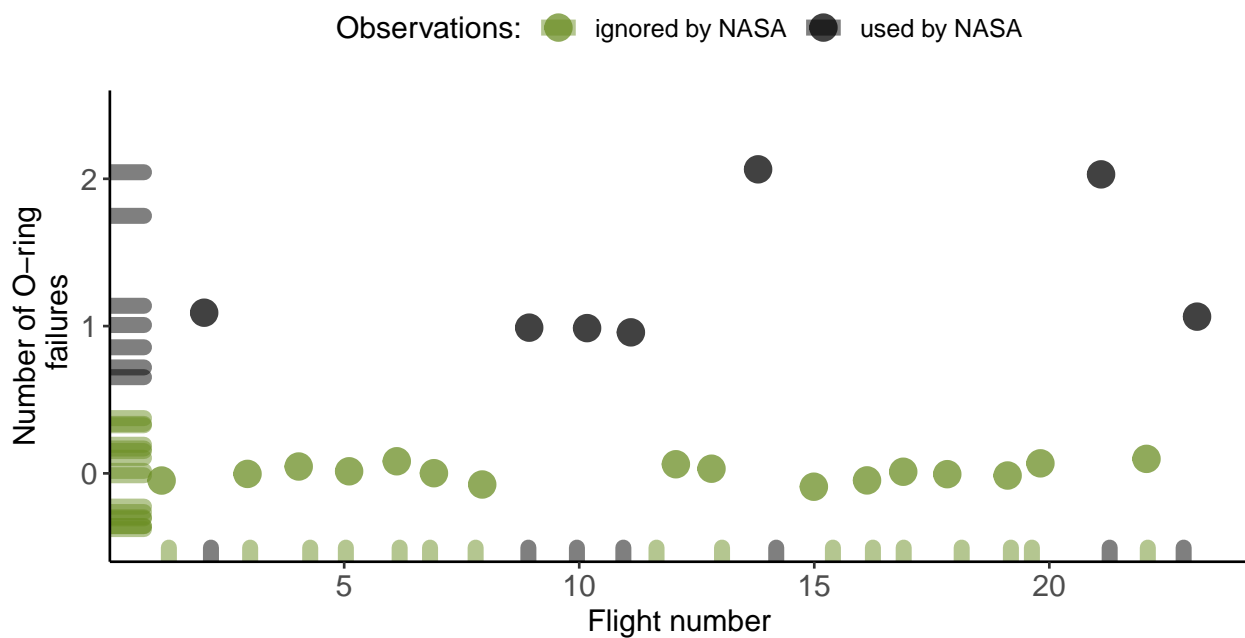
Multivariate Analysis

- How do both `Temp` and `Pressure` affect `O.ring` failure?

```
# Specify levels of O.ring failure
df$O.ring[which(df$O.ring == 0)] <- 0
df$O.ring[which(df$O.ring == 1)] <- 1
```

```
df$O.ring[which(df$O.ring == 2)] <- 2
# Define 3 level factor of O.ring failures
df$O.ringFactor <- as.factor(df$O.ring)
# Sanity check
levels(df$O.ringFactor)
```

```
## [1] "0" "1" "2"
```

```
# Build 3d scatter plot
#fig <- plot_ly(df, x = ~ Temp, y = ~ Pressure, z = ~ O.ringFactor,
#               color = ~ O.ringFactor,
#               colors = c('olivedrab', 'orange2',"orangered2"))
#fig <- fig %>% add_markers()
#fig <- fig %>% layout(scene = list(xaxis = list(title = 'Temperature (°F)'),
#                     yaxis = list(title = 'Pressure'),
#                     zaxis = list(title = 'Number of O-ring failures')))


#fig
```

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 state that the binomial but not the binary model suffer from the assumption that each O-ring failure is independent. More specifically, the binomial model assumes that each of the six 0-rings fail independently at each temperature, t, and pressure, s.

- The independence assumption is necessary for the binomial model becuase it allows us to use the binomial probability mass function.

- Moreover, the assumption is likely violated since the failure of 1 0-ring might expose other 0-rings to damaging conditions and or might exert stress on the remaining 0-rings.

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

- Logistic regression

```
# Estimate logistic regression model
model_a <- glm(O.ring/Number ~ Temp + Pressure,
               family = binomial(link = logit),
               weights = Number,
               data = df)

# Summary of coefficients, standard errors and p-values
stargazer(model_a, type="text")
```

```
##
```

```
## =============================================
##                  Dependent variable:
##                  ---------------------------
##                       O.ring/Number
## ---------------------------------------------
## 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
```

$$logit\left(\hat{\pi}\right) = 2.520 - 0.098\text{Temp} + 0.008\text{Pressure}$$

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

Below we test: $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$, for both `Temp` and `Pressure`:

```
Anova(model_a, 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
```

As discussed below `Temp` but not `Pressure` is important for 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?

- `Pressure` is not an important explanatory variable.
- Above, we tested $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$, for `Pressure`. The $-2log(\Lambda) = 1.5406572$ and the p-value of $P\left(\chi_1^2 > 1.5406572\right) = 0.21452 > \alpha$ 0.05.
- Thus, `Pressure` was probably removed in response to it lack of importance detected by the likelihood ratio test & because it does not contribute much systematic variation to the model (i.e `Pressure` can only take values 50, 100, 200).

- Although it was removed, it *could* be an important covariate to include in the model. For example, recall from **Figure 3** that (i) both instances where 2 O.rings failed occured when Pressure was high = 200 and (ii) 4 of 5 instances where 1 O.ring failed occured when Pressure was high = 200. Similarly, the Dalal *et al* 1989 report (i) very weak evidence of a pressure effect and (ii) overlapping confidence intervals for the expected number of incidents when pressure was 50 and 200. Together these bits of evidence support the hypothesis that Pressure could contribute to the mechanism by which O.rings fail.
- Several problems could arise by the omission of Pressure from the model. These problems are both mechanistic and statistical. The mechanistic problem is outline in Dalal *et al* 1989. O.ring function and failure are highly influenced by Pressure. Thus, while a simple model without Pressure might be *sufficient* for ones purposes, a model with Pressure and, if more data existed, an interaction term between Temp and Pressure might help guide interpretation. The statistical problem(s) of removing Pressure could be numerous. For one, if Pressure is *correlated* with both the explanatory variable Temp and the outcome O.ring, removal of Pressure could introduce omitted variable bias.

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.

```
# Estimate logistic regression model
model_b <- glm(O.ring/Number ~ Temp,
               family = binomial(link = logit),
               weights = Number,
               data = df)
# Summary of coefficients, standard errors and p-values
stargazer(model_b, type="text")
```

```
##
## =============================================
##                     Dependent variable:
##                 ----------------------------
##                          O.ring/Number
## ---------------------------------------------
## Temp                       -0.116**
##                            (0.047)
##
## Constant                    5.085*
##                            (3.052)
##
## ---------------------------------------------
## Observations                  23
## Log Likelihood             -15.823
## Akaike Inf. Crit.           35.647
## =============================================
```

$$logit\left(\hat{\pi}\right) = 5.085 - 0.116Temp$$

(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 $53°$.

- (1) π vs. Temp:

```r
plot(df$Temp, df$O.ring/df$Number,
     xlab='Temperature (°F)',
     ylab=expression(pi),
     pch=20, cex=1.5,
     col=ifelse(test=df$O.ring > 0,
                yes="black",
                no="olivedrab"),
     xlim=c(31,81),
     ylim=c(0,1))
# Figure legend
text(x=75, y=0.9,labels = "Figure 4", cex=0.9)
# Betas
b0 <- model_b$coefficients[1]
b1 <- model_b$coefficients[2]
# Curve
curve(expr = exp(b0+b1*x)/(1+exp(b0+b1*x)), add=T)
```
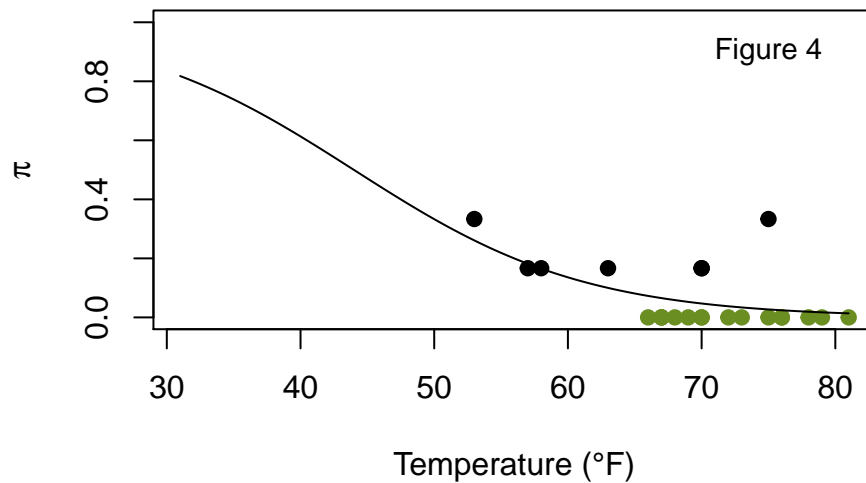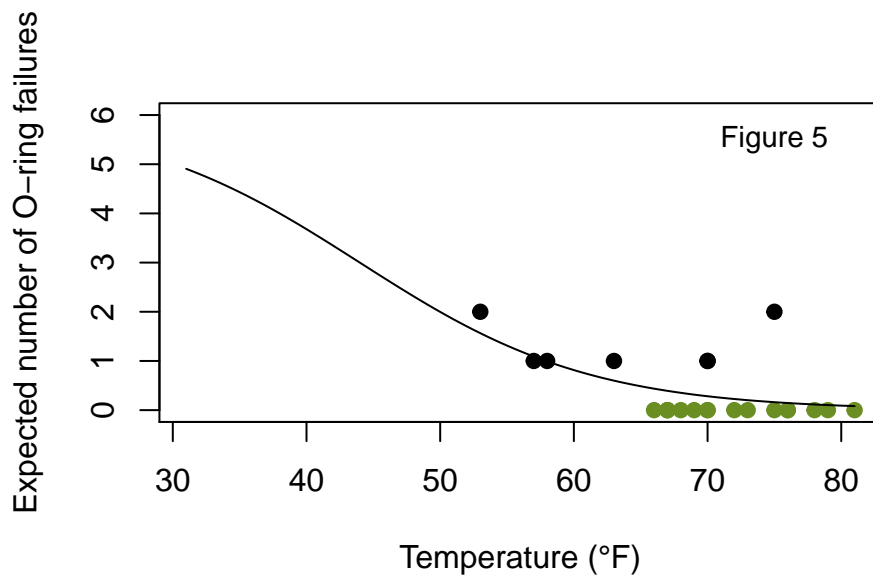

- (2) Expected number of failures vs. Temp.

```r
plot(df$Temp, df$O.ring,
     xlab='Temperature (°F)',
     ylab="Expected number of O-ring failures",
     pch=20, cex=1.5,
     col=ifelse(test=df$O.ring > 0,
                yes="black",
```

```
                 no="olivedrab"),
      xlim=c(31,81),
      ylim=c(0,6))
# Figure legend
text(x=75, y=5.5,labels = "Figure 5", cex=0.9)
# Betas
b0 <- model_b$coefficients[1]
b1 <- model_b$coefficients[2]
# Curve
curve(expr = exp(b0+b1*x)/(1+exp(b0+b1*x))*6, add=T)
```


(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?

```
# Create an array for temperature between 31 and 81 degrees
t <- seq(31,81,1)
alpha=0.05
model_b_predict <- predict(object=model_b, newdata=data.frame(Temp=t), type='link', se=T)
CI_lower_linear <- model_b_predict$fit + qnorm(p=alpha/2)*model_b_predict$se.fit
CI_lower_pi <- exp(CI_lower_linear)/(1+exp(CI_lower_linear))
CI_higher_linear <- model_b_predict$fit + qnorm(p=1-alpha/2)*model_b_predict$se.fit
CI_higher_pi <- exp(CI_higher_linear)/(1+exp(CI_higher_linear))
```
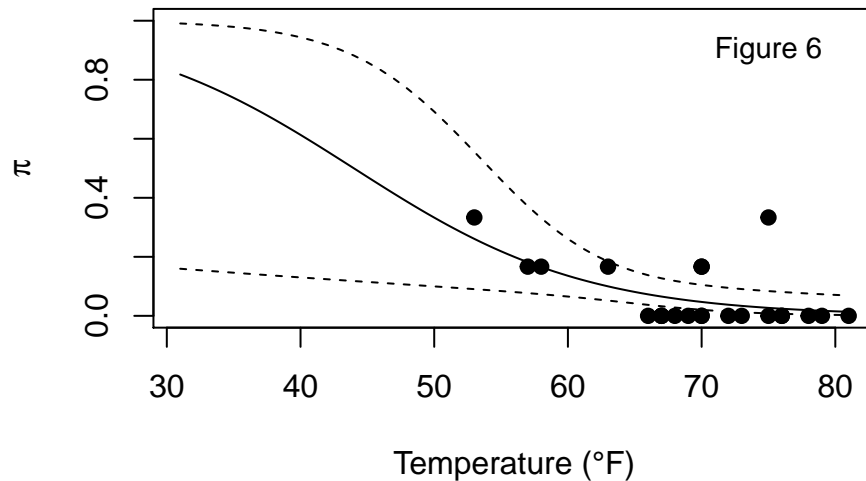
```
plot(df$Temp, df$O.ring/df$Number,
     xlab='Temperature (°F)',
     ylab=expression(pi),
     pch=20, cex=1.5,
     xlim=c(31,81),
     ylim=c(0,1))
# Figure legend
text(x=75, y=0.9,labels = "Figure 6", cex=0.9)
# Betas
```

```
b0 <- model_b$coefficients[1]
b1 <- model_b$coefficients[2]
curve(expr = exp(b0+b1*x)/(1+exp(b0+b1*x)), add=T)
lines(t, CI_lower_pi, lty = 'dashed')
lines(t, CI_higher_pi, lty = 'dashed')
```


Figure 6

- The confidence band is wider when temperatures are lower than 53 because there are no observations below 53.

(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.

- The probability of O-ring failure when temperature is 31°:

```
# Data to extrapolate
predict.data <- data.frame(Temp = 31)

# Predict surface/link
predict.linear <- predict(object = model_b,
                          newdata = predict.data,
                          type = "link")

# Predict response
predict.pi <- predict(object = model_b,
                      newdata = predict.data,
                      type = "response")

# Confidence interval for 31deg was computed above in part (c)
data.frame(estimate=predict.pi, lower=CI_lower_pi[[1]], upper=CI_higher_pi[[1]])

##     estimate     lower     upper
## 1 0.8177744 0.1596025 0.9906582

#THIS CAN BE DELETED
# Data to extrapolate
```

12

```
predict.data <- data.frame(Temp = 72)

# Predict surface/link
predict.linear <- predict(object = model_b,
                          newdata = predict.data,
                          type = "link",
                          se=T)

# Predict response
predict.pi <- predict(object = model_b,
                      newdata = predict.data,
                      type = "response")
predict.pi
```

```
##          1
## 0.03774935
```

```
#pi_hat <- exp(predict.linear$fit)/(1+exp(predict.linear$fit))
alpha=0.1
CI_linear <- predict.linear$fit + qnorm(p=c(alpha/2, 1-alpha/2))*predict.linear$se.fit
exp(CI_linear)/(1+exp(CI_linear))
```

```
## [1] 0.01690869 0.08213102
```

- Thus, the probability of O-ring failure at 31° is 0.0377493 with a very wide confidence interval of 0.1596025, 0.9906582.
- The assumptions needed to apply the inference procedures are:
- We are assuming a linear relationship between the log odds of failure of an O-ring and Temperature.
- We are assuming that the observations are independent. i.e. The failure of one O-ring does not increase the odds of the other to fail

(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

For parametric bootstrap, we do the following steps. Create new column o.ring2 and initialize to o.ring. This column will be updated in each iteration. Do the following steps in a loop (1..100): * sample 23 points from dataset with replacement * fit the model using the above points * Calculate Z using intercept + coefTemp * Temp This results in a vector of 23 values * Calculate pi using exp(z) / (1+exp(z)) (vector of size 23) * Generate vector of size 23 of binomials 0-6 using the built in rbinom function by passing in the pi's generated in prior step. * replace o.ring2 with above generated binom values, updating only the row numbers corresponding to sample. * run predictions for 31 and 72 degrees and save results

At the end, the CI will be the 0.05 and 0.95 quantile of the accumulated results.

```r
# Do not change original # of failures
df[, "O.ring2"] = df$O.ring

# Dataframe to populate with results
results <- data.frame(pred.31 = numeric(), pred.72 = numeric())

# For each sample in 1:n
for (s in 1:100){
  # Collect sample of size 23 from the whole data frame with replacement
  I.sample <- sample(x = 1:nrow(df),
                     size = 23,
                     replace = TRUE)
  # Populate d with samples
  d <- df[I.sample,]

  # Estimate model with bootstrap sample data
  mod <- glm(O.ring2/Number ~ Temp,
             family = binomial(link = logit),
             weights = Number,
             data = d)

  if (s==1){
    # Z is the the beta0 + beta1X
    z = mod$coefficients["(Intercept)"] + mod$coefficients["Temp"] * d$Temp

    # Pi is the binary Logistic Regression
    pi = exp(z)/(1+exp(z))
  }
  # Generate new pi values from pi values generated with bootstrap samples
  pi.O.ring <- rbinom(n=23, # sample size
                      size=6, # number of trials
                      prob=pi) # probability

  # update the failures generated from model
  df[I.sample,"O.ring2"] <- pi.O.ring
  # Estimate confidence interval for temp = 31
  temp.31.data <- data.frame(Temp=31)
  temp.31 <- predict(object = mod, newdata = temp.31.data,
                     type = "response")
  # Estimate confidence interval for temp = 72
  temp.72.data <- data.frame(Temp=72)
  temp.72 <- predict(object = mod, newdata = temp.72.data,
                     type = "response")
  results <- results %>% add_row(pred.31 = temp.31, pred.72 = temp.72)
}
```

```r
paste("For 31 degrees CI =")
```

```
## [1] "For 31 degrees CI ="
```

```r
quantile(results[,1], probs=c(0.05, 0.95))
```

```
##              5%           95%
## 0.0001558126 0.9114315255
```

```r
paste("For 72 degrees CI =")
```

```
## [1] "For 72 degrees CI ="
```

```r
quantile(results[,2], probs=c(0.05, 0.95))
```

```
##           5%          95%
## 0.009172457 0.109221024
```

```r
# Plot histograms of confidence intervals
par(mfrow=c(2,1))
# Histogram for temp = 31
hist(results[,1],
     breaks=50,
     col="lightskyblue1",
     main="Bootstrapped confidence intervals for temperature (°F)= 31")

# Histogram for temp = 72
hist(results[,2],
     breaks=50,
     col="orange",
     main="Bootstrapped confidence intervals for temperature (°F)= 72")
```
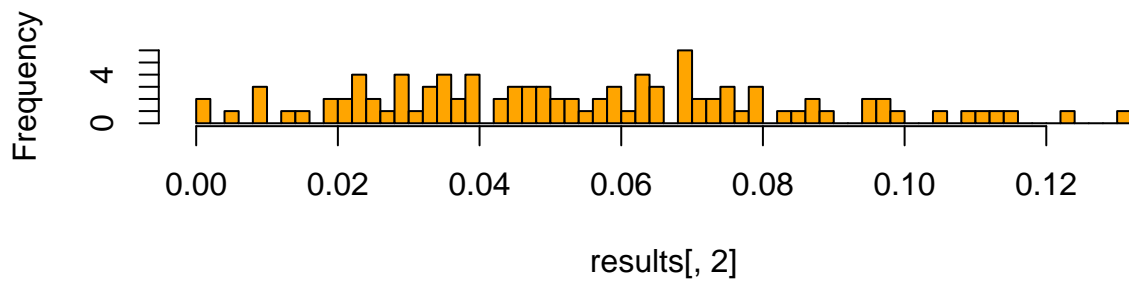
Bootstrapped confidence intervals for temperature (°F)= 31
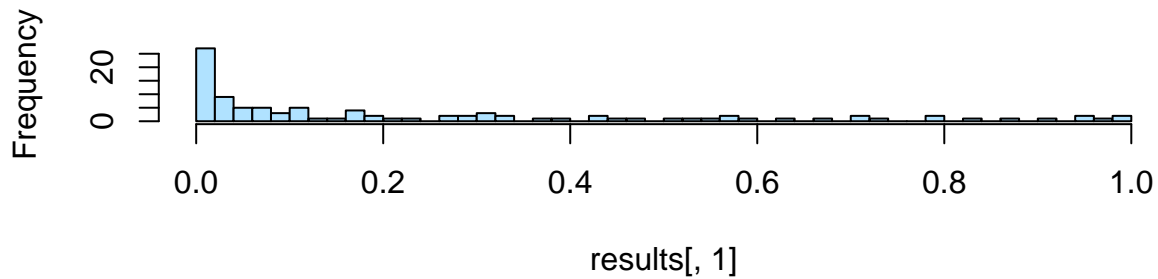
Bootstrapped confidence intervals for temperature (°F)= 72


```r
# KATIE'S ATTEMPT
set.seed(1)

# Use pi estimated from the model
z = model_b$coefficients["(Intercept)"] + model_b$coefficients["Temp"] * df$Temp
pi = exp(z)/(1+exp(z))

# Save the pi array to the dataframe
df$O.ring.pi <- pi

# Dataframe to populate with results
results <- data.frame(pred.31 = numeric(), pred.72 = numeric())

for (s in 1:1000){

  I.sample <- sample(x = 1:nrow(df),
                     size = 23,
                     replace = TRUE)
  # Populate d with samples
  d <- df[I.sample,]

  # Simulate outcomes using rbinom
  O.ring2 <- rbinom(n=23, # sample size
                size=6, # number of trials
                prob=d$O.ring.pi) # probability
  d <- data.frame(d, O.ring2)
```

```r
  # Estimate model with rbinom bootstrap outcomes
  mod <- glm(O.ring2/Number ~ Temp,
             family = binomial(link = logit),
             weights = Number,
             data = d)

  # Estimate confidence interval for temp = 31
  temp.31.data <- data.frame(Temp=31)
  temp.31 <- predict(object = mod, newdata = temp.31.data,
                     type = "response")
  # Estimate confidence interval for temp = 72
  temp.72.data <- data.frame(Temp=72)
  temp.72 <- predict(object = mod, newdata = temp.72.data,
                     type = "response")
  results <- results %>% add_row(pred.31 = temp.31, pred.72 = temp.72)
}

paste("For 31 degrees CI =")
```

```
## [1] "For 31 degrees CI ="
```

```r
quantile(results[,1], probs=c(0.05, 0.95))
```

```
##           5%           95%
## 7.152891e-05 9.852102e-01
```

```r
paste("For 72 degrees CI =")
```

```
## [1] "For 72 degrees CI ="
```

```r
quantile(results[,2], probs=c(0.05, 0.95))
```

```
##           5%           95%
## 0.002080109 0.079570875
```

```r
# Plot histograms of confidence intervals
par(mfrow=c(2,1))
# Histogram for temp = 31
hist(results[,1], breaks=50, col="lightskyblue1",
     main="Bootstrapped confidence intervals for temperature (°F)= 31")

# Histogram for temp = 72
hist(results[,2], breaks=50, col="orange",
     main="Bootstrapped confidence intervals for temperature (°F)= 72")
```
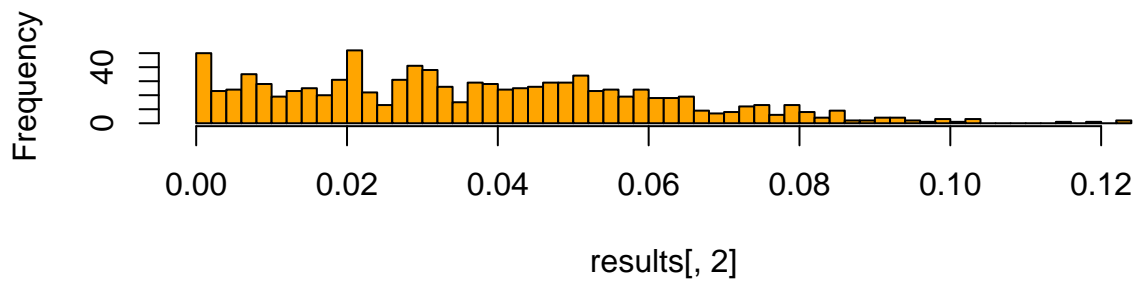
Bootstrapped confidence intervals for temperature (°F)= 31
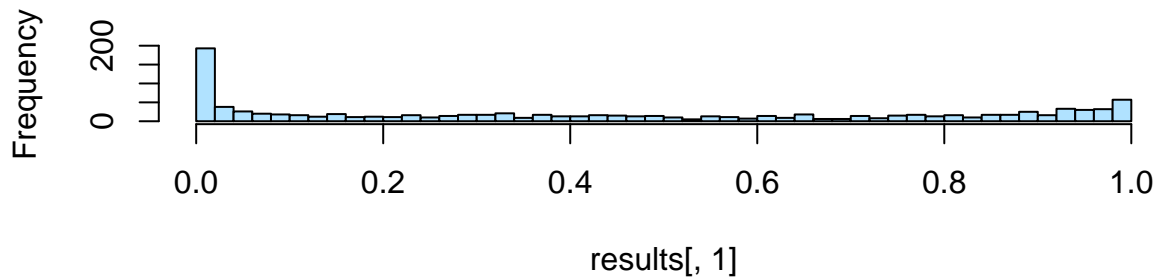
Bootstrapped confidence intervals for temperature (°F)= 72

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

- Add quadratic temperature term

```
# Estimate logistic regression model
model_3f <- glm(O.ring/Number ~ Temp + I(Temp^2),
                family = binomial(link = logit),
                weights = Number, data = df)
# Summary of coefficients, standard errors and p-values
stargazer(model_3f, type="text")
```

```
##
## =========================================
##                  Dependent variable:
##              ----------------------------
##                       O.ring/Number
## ---------------------------------------------
## Temp                     -0.651
##                          (0.741)
##
## I(Temp2)                  0.004
##                          (0.006)
##
## Constant                 22.126
##                          (23.794)
##
## ---------------------------------------------
```

```
## Observations                        23
## Log Likelihood                  -15.576
## Akaike Inf. Crit.                37.152
## =============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

$$logit\left(\hat{\pi}\right) = 22.126 - 0.651Temp + 0.004Temp^2$$

- Conduct LRT:

```
# Compare models with LRT
anova(model_b, model_3f, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Temp
## Model 2: O.ring/Number ~ 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
```

- No, there is no need for a quadratic term.
- We tested $H_0 : \beta_2\text{temp}^2 = 0$ vs $H_A : \beta_2\text{temp}^2 \neq 0$.
- For `temp^2`, $-2log(\Lambda) = 0.4947004$ and the p-value of $P\left(\chi_1^2 > 0.4947004\right) = 0.481838 > \alpha = 0.05$. Thus, `temp^2` is not important, when we hold `temp` constant.

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.

- Estimation of linear regression model:

```
# Estimate model
model_4 = lm(formula = O.ring/Number ~ Temp, data = df, weights = Number)
# Summary
stargazer(model_4, type="text")
```
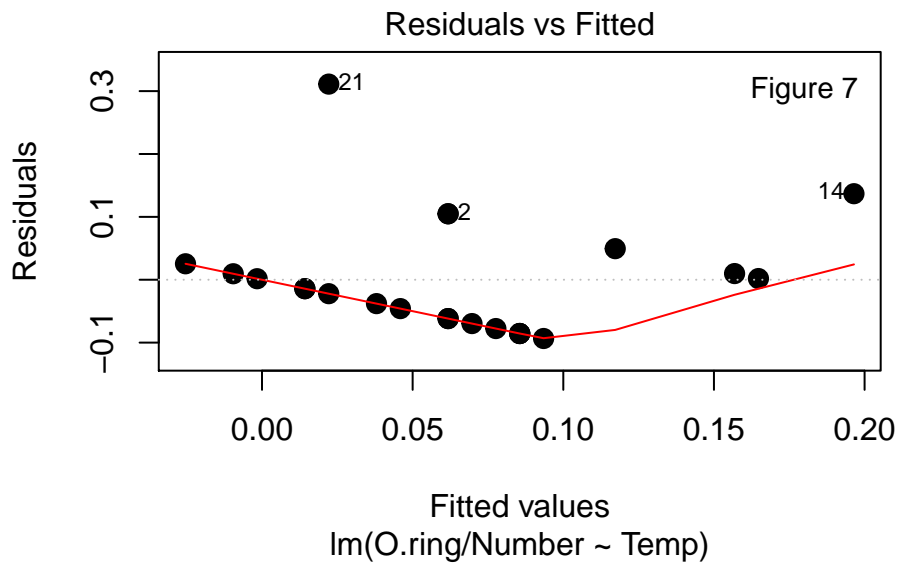
```
##
## =============================================
##                        Dependent variable:
##                    -------------------------
##                            O.ring/Number
## -------------------------------------------
## Temp                         -0.008**
##                               (0.003)
##
## Constant                     0.616***
##                               (0.203)
##
```

```
## -------------------------------------------------
## Observations                      23
## R2                             0.261
## Adjusted R2                    0.226
## Residual Std. Error      0.236 (df = 21)
## F Statistic           7.426** (df = 1; 21)
## =================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```
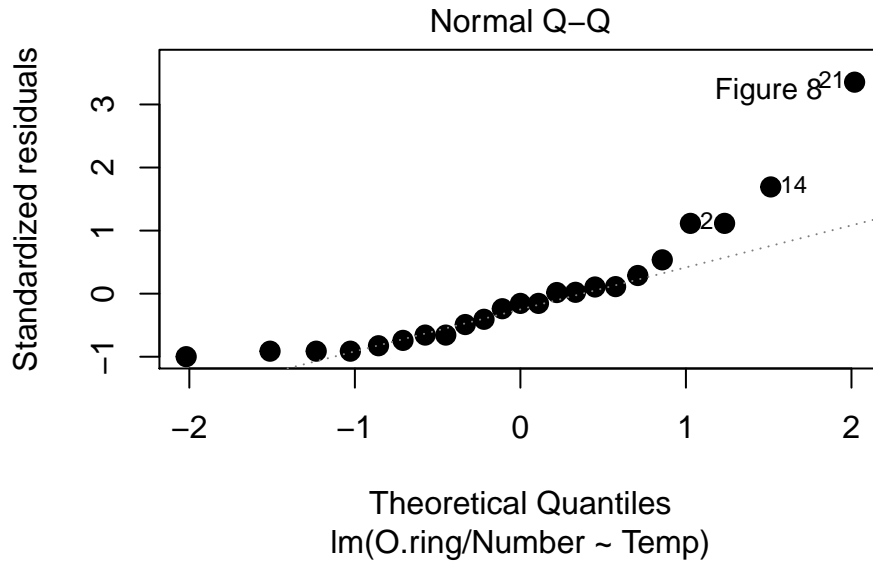
$$\frac{O.ring}{Number} = 0.616 - 0.008Temp$$

- Explanation of the model results:

- Model diagnostics:

```
# diagnotics
plot(model_4, which=1, pch=20, cex=2)
# Figure legend
text(x=0.18, y=0.3,labels = "Figure 7", cex=0.9)
```



```
# diagnotics
plot(model_4, which=2, pch=20, cex=2)
# Figure legend
text(x=1.5, y=3.2,labels = "Figure 8", cex=0.9)
```

Normal Q–Q

Figure 8

Theoretical Quantiles
lm(O.ring/Number ~ Temp)

- Assess the validity of the model assumptions:

- **Assumption MLR.1 (Linear in Parameters):**

  - The model is linear in its parameters.The $\hat{\beta}s$ are expressed as additive/linear functions of the response: O-ring failure.

- **Assumption MLR.2 (Random Sampling):**

  - It seems doubtful that these O-rings were sampled randomly, or even *sampled* at all- **all** O-rings were likely collected from 23 flights. Further survival bias could be introduced by the inability of NASA to collect all O-rings from flights that malfunctioned.

- **Assumption MLR.3 (No Perfect Collinearity):**

  - ince there is only one explanatory variable, it is not perfectly collinear with any other explanatory variable.

- **Assumption MLR.4 (Zero Conditional Mean):**

  - The zero-conditional mean assumption,$E(\mu|x_1, x_2, x_3, ..., x_k)$, is not satisfied, sensu stricto. The residual versus fitted values plot: documents the extent to which this model violates this assumption. If the zero mean assumption were satisfied, then we should expect to see a horizontal red line, centered at zero. In contrast we see that our estimates of $\hat{\mu}$, the residuals, deviate from this expectation dramatically.

- **Assumption MLR.5 (Homoskedasticity):**

  - Homoskedasticity does not appear to be satisfied. Evidence for heteroskedasticity is again provided by the residual versus fitted values plot: . If variances were homoskedastic we should see uniform horizontal bands/scatters of points across these plots. However, note that we see non-constant variability in the residuals across the range of fitted values.

- **Assumption MLR.6 (Normality):**

  - The assumption of normality is not satisfied. Evidence of this assumption is presented in the qq-plot: & a histogram of the response variable: .

- **Lastly, this model formulation possesses another shortcomings**:
    - The probabilities are linearly related to the predictors for all of their possible values.
- We select the binary logistic regression model because:
- the response, O-ring failure is either binary (e.g. 0 or 1), binomial (e.g. a proportion from 0 to 1), or counts (e.g. 0,1,2,3,...,n) which can ultimately be expressed as a binary or binomial variable.
-
- 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.

Our final model model is

$$logit\left(\hat{\pi}\right) = 5.085 - 0.116Temp$$

This represents the log odds of the fraction of O-rings that fail. To calculate the odds, we exponentiate the coefficients as below.

```
1/exp(model_b$coefficients)
```

```
## (Intercept)       Temp
## 0.006189028 1.122548073
```

Here we see that an increase in temperature changes the odds of failure by 0.89 times for each degree increase. In other words, a unit decrease in temperature increases the odds of failure by 1.12 times for each unit temperature decrease.

In terms of probability of failure, as is seen in the chart of probability of failure vs temperature in part 3c, we clearly see an increase in the probability of failure with lower temperature. For temperature going from 50F, 40F to 30F, we see corresponding probability of failure π values of 0.33, 0.61 to 0.83, which clearly shows that lower temperatures lead to a higher probability of failure.