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

```
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(scatterplot3d) # For 3d scatterplot
```

Load data

```
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 Dalal *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).

```
# Structure
str(df)
```

```
## 'data.frame':    23 obs. of  5 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 ...
```

```
# Summary
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
##  Min.   :6
##  1st Qu.:6
##  Median :6
##  Mean   :6
##  3rd Qu.:6
##  Max.   :6
```

```
# 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
```

Each of the 5 variables were treated as `int` classes. To ensure that the data do not contain anomolous observations, we 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 (aside from the `O.ring` values > 0 that NASA ignored). We also noted that there are no missing values in any of the variables.

Univariate Analysis

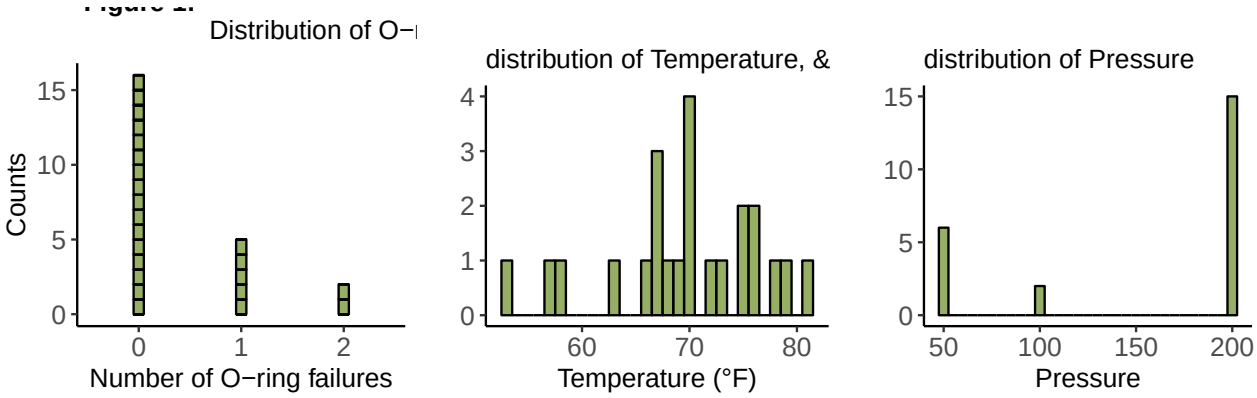
The three variables of interest for modeling purposes in the paper by Dalal *et al.* 1989 were

temperature, pressure, and number of O-ring failures.

From the histogram in **Figure 1**, the temperature variable, `Temp`, 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.

`Pressure` is seen to only take on 3 unique values, with the most observations having a pressure of 200 (**Figure 2**). Though it is difficult to say with such few unique values, the distribution seems to most closely resemble a bimodal distribution. Although one could argue that `Pressure` should be coerced into a factor because of the low ...

```
grid.arrange(three, one, two,
             nrow=1)
```



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

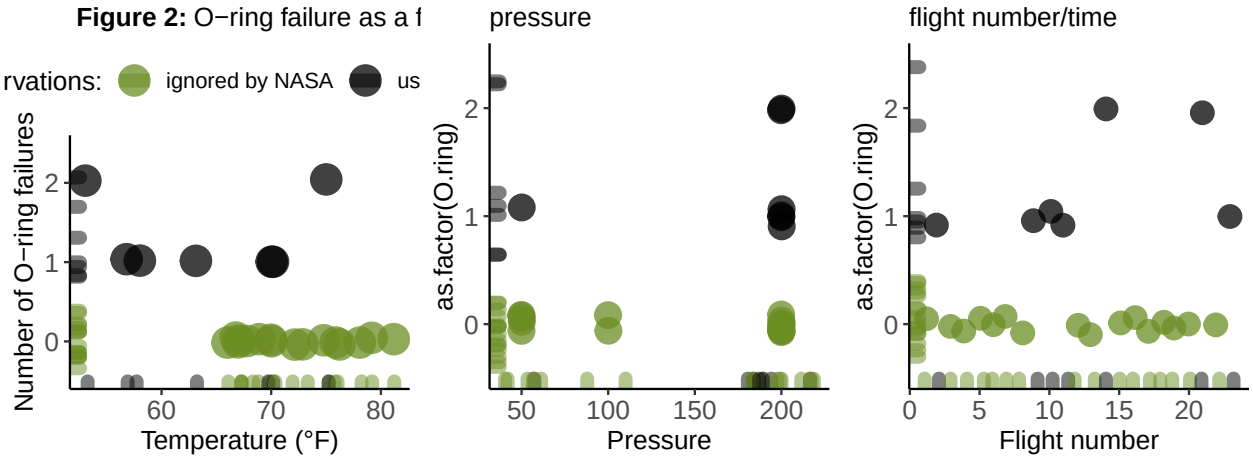
- How are O.Rings related to putative predictors, like temperature?

When plotting the temperature against the number of O-ring failures, the launches with no O-ring failures tend to occur on the right hand side of the graph, where temperature is higher. This is consistent with the negative correlation of -0.511. The lowest temperature at which there were no recorded O-ring failures was observed to be 66 °F. Above this temperature, we have one launch with 1 O-ring failure and one launch with 2 O-ring failures. The launches below 66 °F all had at least 1 O-ring failure. The authors of the paper mentioned that launches with no O-ring failures were not included in the original thermal distress analysis prior to the Challenger accident, which are highlighted in green in the plot above.

There were the most number of recorded launches at 200 psi, and at this pressure, there was the most number of launches with at least 1 O-ring failure. At 200 psi, 4 out of the 15 launches (27%) had at least 1 O-ring failure. At 50 psi, 1 out of the 6 launches (17%) had at least 1 O-ring failure. There were no O-ring failures for launches at 100 psi, though there were only 2 recorded data points. The correlation between the two variables is slightly positive, at 0.285.

Based on the description in the paper, we do not anticipate there to be any meaningful relationship between temperature and pressure as they are both arbitrarily determined. Pressure is a condition set by the test procedure, which was first at 50 psi, then progressively updated to 100 psi and 200

psi. Temperature is merely a condition at the time of launch. The low correlation between the two variables, 0.04, supports our understanding.

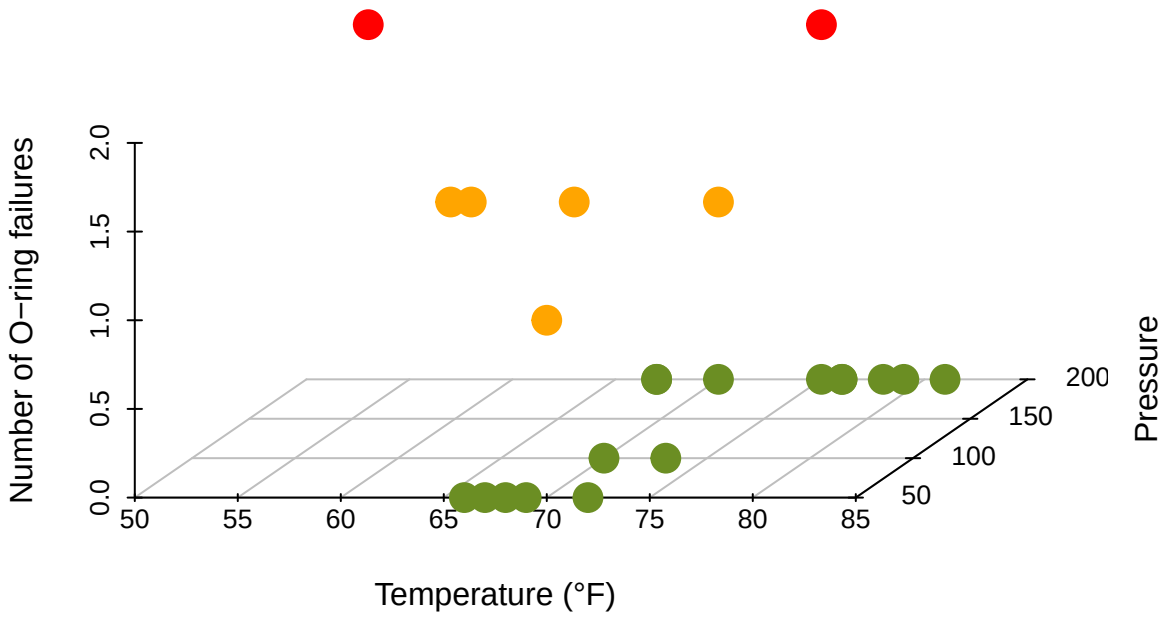


Multivariate Analysis

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

```
#3d scatter plot
#Colors
colors <- c("olivedrab", "orange", "red")
colors <- colors[as.numeric(df$O.ring)+1]
scatterplot3d(x=df$Temp, y=df$Pressure, z=df$O.ring,
              #angle = 80,
              main= expression(~bold('Figure 3:')~' O-ring failure as a function of temperature
              xlab = "Temperature (°F)",
              ylab = "Pressure",
              zlab = "Number of O-ring failures",
              pch = 20, cex.symbols=3,
              color=colors,
              grid=TRUE, box=FALSE)
```

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



Part 2

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. This assumption is likely to be violated since the failure of 1 or more of the O-rings influences the failure of the remaining O-rings. Moreover, the binary logistic regression model does not suffer from this assumption since each *trial* is a flight and we assume the flights are independent.
- The independence assumption is necessary for the binomial model because 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.
- The authors subsequently used a binary logistic regression model by introducing a binary flag for failure. One or more failures were simply marked as a failure, thereby removing the independence assumption.

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

```r
df["Fail"] <- ifelse(df$O.ring>=1, 1, 0)
# Estimate binomial logistic regression model
model_a <- glm(O.ring/Number ~ Temp + Pressure,
               family = binomial(link = logit),
               weights = Number,
               data = df)

# Estimate binary logistic regression model
model_a.1 <- glm(Fail ~ Temp + Pressure,
               family = binomial(link = logit),
               data = df)


# Summary of coefficients, standard errors and p-values

stargazer(model_a, model_a.1,
          type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                      O.ring/Number      Fail
##                           (1)            (2)
## ------------------------------------------------
## Temp                     -0.098**      -0.229**
##                          (0.045)        (0.110)
##
## Pressure                  0.008          0.010
##                          (0.008)        (0.009)
##
## Constant                  2.520         13.292*
##                          (3.487)        (7.664)
##
## ------------------------------------------------
## Observations                23             23
## Log Likelihood           -15.053        -9.391
## Akaike Inf. Crit.         36.106        24.782
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```r
1/exp(summary(model_a.1)$coefficients)
```

```
##                 Estimate   Std. Error   z value  Pr(>|z|)
## (Intercept) 1.687336e-06 0.0004694408 0.1765067 0.9204913
## Temp        1.256928e+00 0.8958446419 7.9968563 0.9630857
## Pressure    9.896537e-01 0.9910609925 0.3140357 0.7813261
```

Here we estimate the bimonial and binary regression models separately. For the binary model,

we introduce a new flag "Fail" that is set to 1 if at least one failure, 0 otherwise. This effectively eliminates the independence assumption. From the above model summary, we see that Temperature is the only significant variable at the 0.05 level. By exponentiating the coefficients and taking inverse, we can interpret the effect of temperature on odds of failure.

Binomial Model

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

```
1/exp(summary(model_a)$coefficients)
```

```
##               Estimate Std. Error   z value  Pr(>|z|)
## (Intercept) 0.08044395 0.03059912 0.4853986 0.6251197
## Temp        1.10329014 0.95610229 8.9325907 0.9718580
## Pressure    0.99155187 0.99235230 0.3311752 0.7640570
```

For the Logistic regression binomial model with `Temperature` and `Pressure` included, we see that a unit decrease in temperature changes the odds of failure by 1.1 times

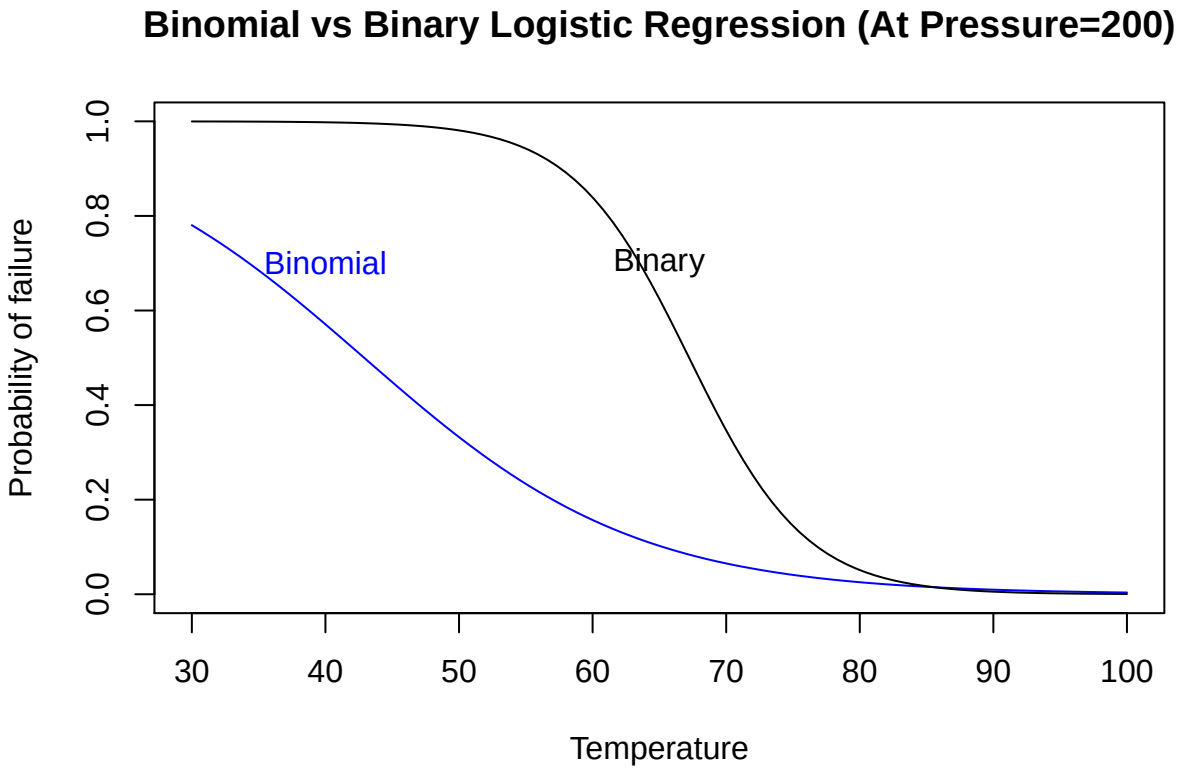
Binary Model

$$logit\left(\hat{\pi}\right) = 13.292 - 0.229\text{Temp} + 0.010\text{Pressure}$$

```
1/exp(summary(model_a.1)$coefficients)
```

```
##                 Estimate   Std. Error   z value  Pr(>|z|)
## (Intercept) 1.687336e-06 0.0004694408 0.1765067 0.9204913
## Temp        1.256928e+00 0.8958446419 7.9968563 0.9630857
## Pressure    9.896537e-01 0.9910609925 0.3140357 0.7813261
```

For the Logistic regression binary model with `Temperature` and `Pressure` included, we see that a unit decrease in temperature changes the odds of failure by 1.25 times.



(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.ring`s 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

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(Fail ~ Temp,
               family = binomial(link = logit),
               data = df)
# Summary of coefficients, standard errors and p-values
stargazer(model_b,  type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Fail
## -----------------------------------------------
## 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
```

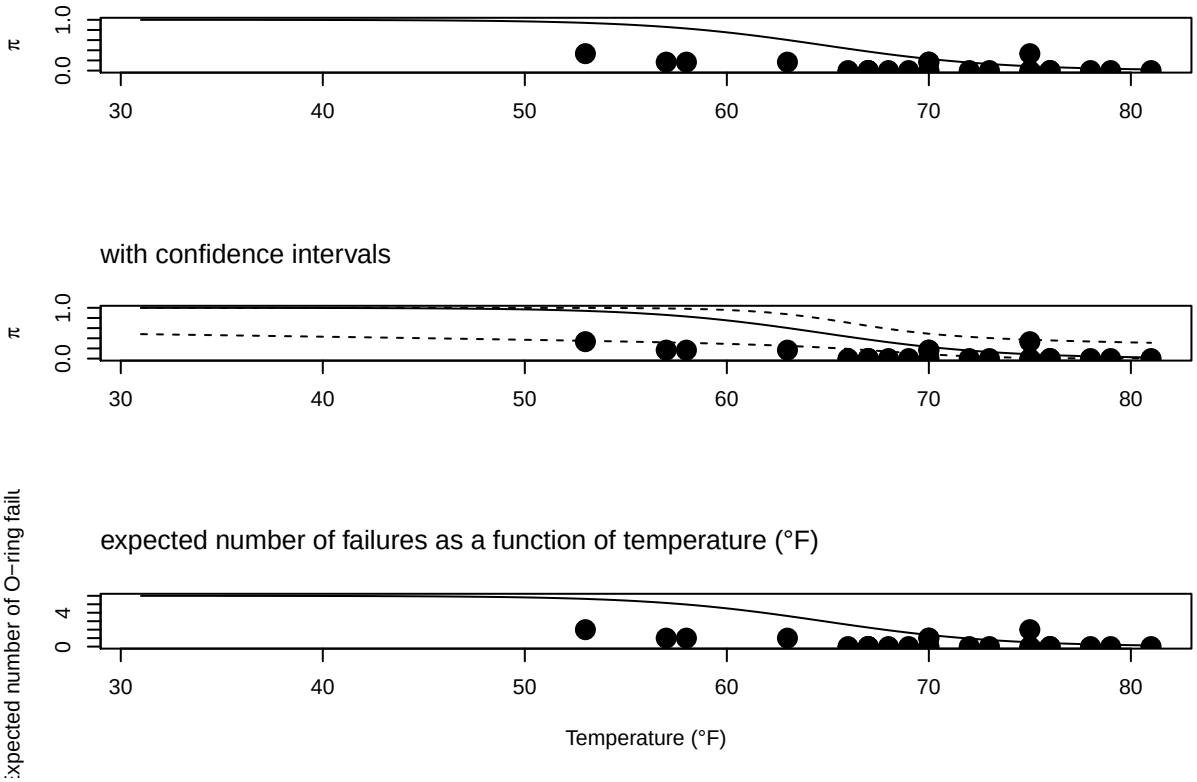
$$logit\,(\hat{\pi}) = 5.085 - 0.116 Temp$$

(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°.

(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))
```

Figure 4: π as a function of temperature (°F)



- The confidence band is wider when temperatures are lower than53 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 0.9996088 0.4816106 0.9999999
```

- Thus, the probability of O-ring failure at 31° is 0.9996088 with a very wide confidence interval of 0.4816106, 0.9999999.
- 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.
- Since we are using the binary logistic model, each flight/*trial* is independent of other flights.

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

For parametric bootstrap, we do the following steps. 1. Create new column o.ring2 and initialize to o.ring. This column will be updated in each iteration. 2. Calculate Z the linear outcome using $Z = \beta_0 + \beta_{Temp} * Temp$ (vector of size 23) 3. Calculate $\hat{\pi} = \frac{e^Z}{1+e^Z}$ (vector of size 23) 4. Do the following steps in a loop (1..1000): * sample 23 points d from original dataset with replacement * subset π values corresponding to the random subset generated above * Generate vector of size 23 of binomials 0-6 $O.ring2$ that represent the outcomes, using the built in rbinom function by passing in the pi's generated in prior step. * Fit the binomial logistic regression model with new datapoints d and outcomes $O.ring2$ * run predictions for 31 and 72 degrees and save results 5. select the 5% and 95% quantiles from the predictions accumulated in the loop above and report

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

```r
# 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,]
  # subset the probabilities tied to the sample
  pi.b <- df$O.ring.pi[I.sample]

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

  # Estimate model with rbinom bootstrap outcomes
  mod <- glm(Fail ~ Temp,
              family = binomial(link = logit),
              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%
## 0.9768686 1.0000000
```

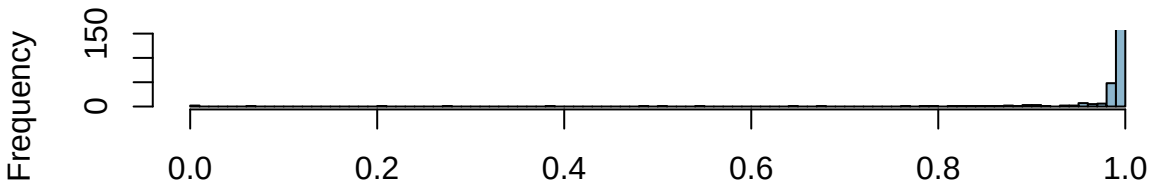
```
paste("For 72 degrees CI =")
```

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

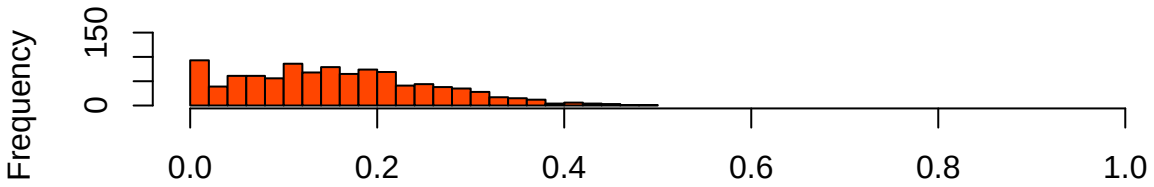
```
quantile(results[,2], probs=c(0.05, 0.95))
```

```
##          5%         95%
## 0.01070611 0.33372384
```

Figure 5: Probability of failure when temperature: **31 (°F)**



72 (°F)



(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(Fail ~ Temp + I(Temp^2),
                family = binomial(link = logit),
                data = df)
# Summary of coefficients, standard errors and p-values
stargazer(model_3f, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Fail
## -----------------------------------------------
## Temp                          -2.094
##                               (2.185)
##
## I(Temp2)                       0.014
```

```
##                               (0.016)
##
## Constant                      78.483
##                              (76.674)
##
## ------------------------------------------
## Observations                    23
## Log Likelihood                -9.694
## Akaike Inf. Crit.             25.389
## ==========================================
## 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: Fail ~ Temp
## Model 2: Fail ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     19.389  1  0.92649   0.3358
```

- 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`, $-2log(\Lambda) = 0.9264902$ and the p-value of $P\left(\chi_1^2 > 0.9264902\right) = 0.3357766 > \alpha = 0.05$. Thus, `temp^2` is not important, when we hold `temp` constant.

Part 4

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 = Fail ~ Temp, data = df)
# Summary
stargazer(model_4, type="text")
```

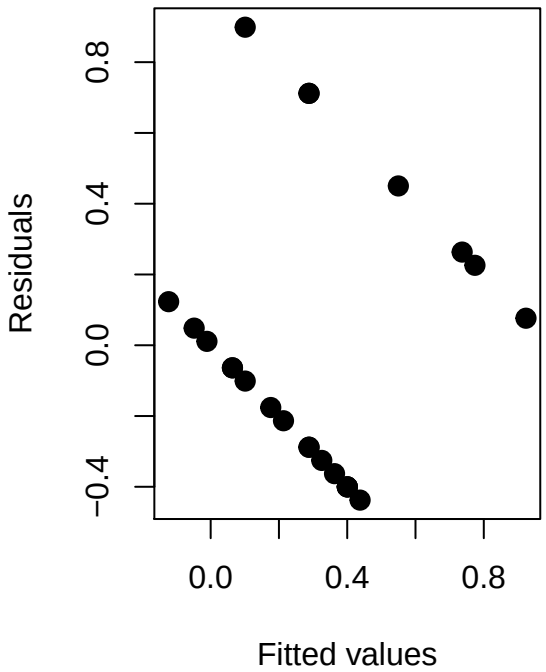
```
##
## ==============================================
```

```
##                         Dependent variable:
##                     ---------------------------
##                                 Fail
## ------------------------------------------------
## Temp                          -0.037***
##                                (0.012)
##
## Constant                      2.905***
##                                (0.842)
##
## ------------------------------------------------
## Observations                     23
## R2                              0.314
## Adjusted R2                     0.282
## Residual Std. Error        0.399 (df = 21)
## F Statistic             9.630*** (df = 1; 21)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

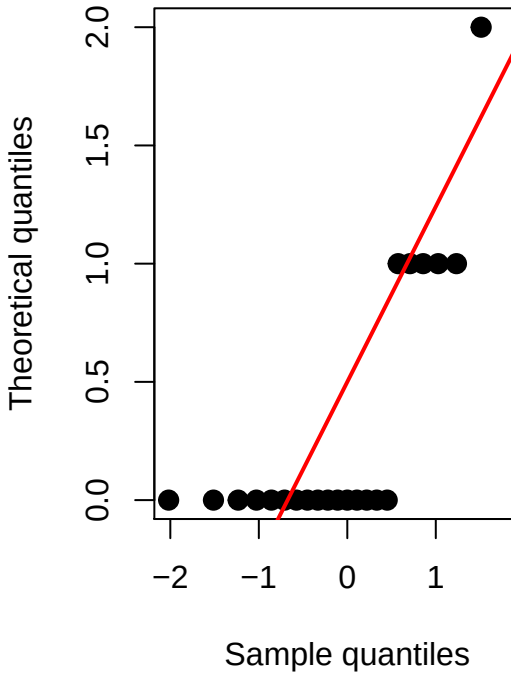
$$\frac{O.ring}{Number} = 0.616 - 0.008Temp$$

- Explanation of the model results:

Figure 6: Residual vs. fitted



Normal q−q plot



- Model diagnostics:
- 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.

- Explain why:
- We select the binary logistic regression model because:
- The problem is clearly asking a success/failure question. 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.
- The MLR assumptions as seen above does not hold and therefore the linear regression model is not suitable for this problem.

Part 5

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 the binomial logistic regression model

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

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

```
exp(model_b$coefficients)
```

```
##  (Intercept)         Temp
## 3.412315e+06 7.928171e-01
```

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

```
##  (Intercept)         Temp
## 2.930561e-07 1.261325e+00
```

Here we see that a unit decrease in temperature increases the odds of failure increases by 1.12 times.

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.

In summary, we explored the relationship to pressure vs failure and temperature vs failure. We note a higher incidence of failures at lower temperatures, wheras no strong evidence of a relationship to pressure was identified.

- EDA
- Potential importance of pressure
- Reasons for binary vs binomial
- Interpretation of odds ratios & pi at extreme temperaturs
- Challenges introduced by small data set (at lower temperatures)