W271 Group 8, Lab 1: Katie Mo, Rajiv Nair, David Linnard Wheeler

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(mcprofile) # for profile likelihoods
library(dplyr) # for coercing dataframes and summary statistic
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

In January of 1986, the launch of the space shuttle Challenger ended up in a tragic accident. After subsequent investigation, the root cause of the accident was identified to be caused from a gas leak through a booster rocket joint, typically sealed by an O-ring. The Dalal *et al.* (1989) paper explores the use of statistical modeling to predict the probability of failure of the O-rings. In this report, we aim to perform a similar analysis.

The dataset used to examine the probability of O-ring failure in Challenger's previous space shuttle launches 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 °F 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
```

```
# Unique values
lapply(df[c('Temp','Pressure','O.ring')], unique)
```

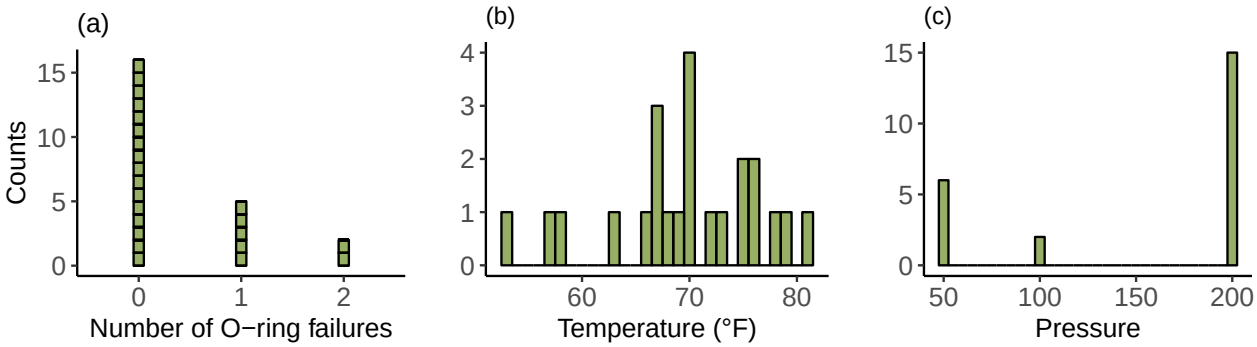
```
## $Temp
##  [1] 66 70 69 68 67 72 73 57 63 78 53 75 81 76 79 58
##
## $Pressure
## [1]  50 100 200
##
## $O.ring
## [1] 0 1 2
```

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 < 1 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.

Figure 1



In **Figure 1a**, 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.

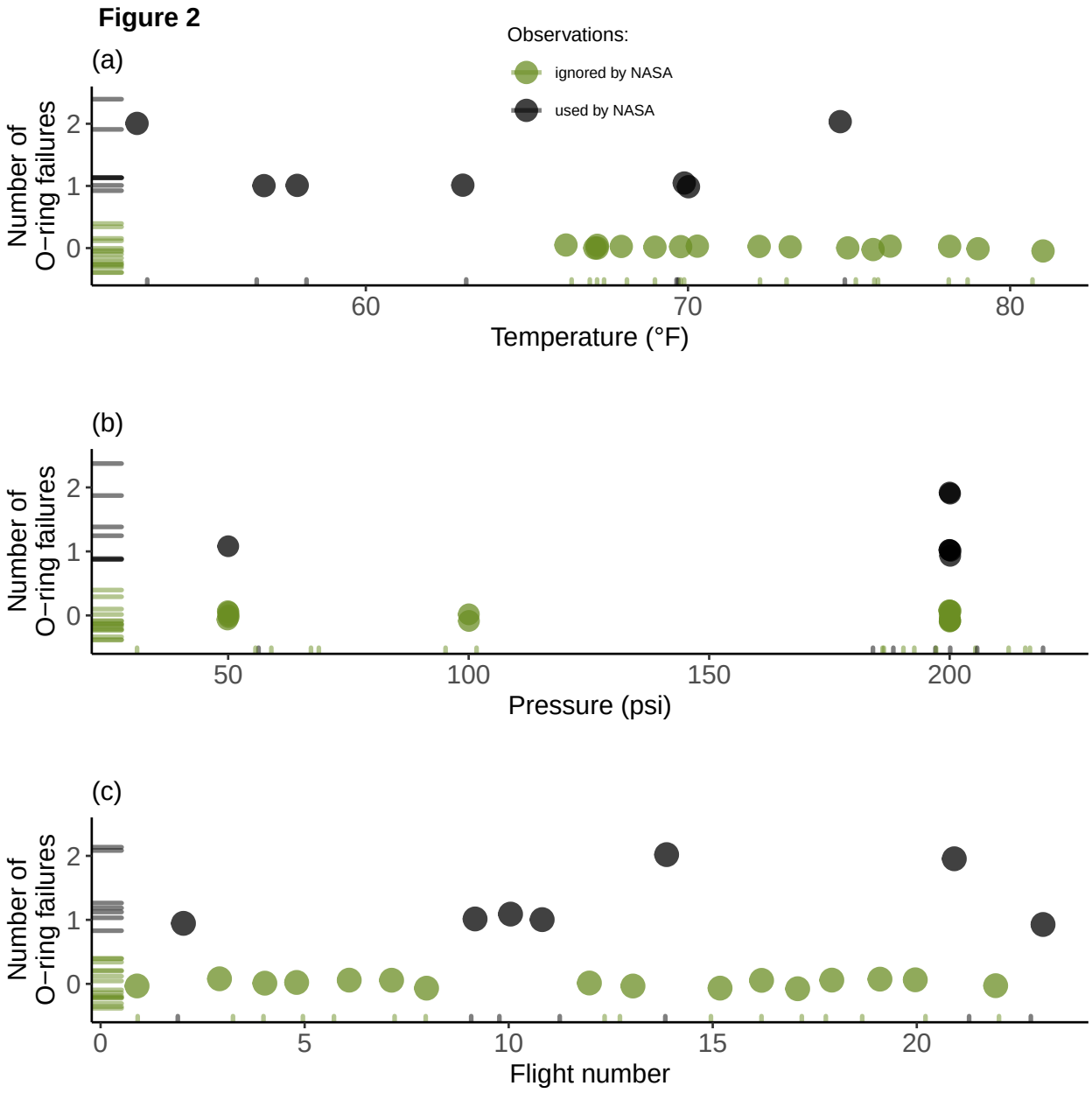
From the histogram in **Figure 1b**, the temperature variable, `Temp`, is seen to peak at 70 °F with 4 observations and tapered off on either side. Based on the plot, the distribution seems to most closely resemble a normal distribution.

`Pressure` is seen to only take on 3 unique values, with the most observations having a pressure of 200, as seen in **Figure 1c**. 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 there are only 3 levels, the consequences of this coercion would be estimates of each factor level **with very few observations per level**. Thus we will treat it as an `int`.

Bivariate Analysis

When plotting the temperature against the number of O-ring failures, as shown in **Figure 2a** 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.

In **Figure 2b**, 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.

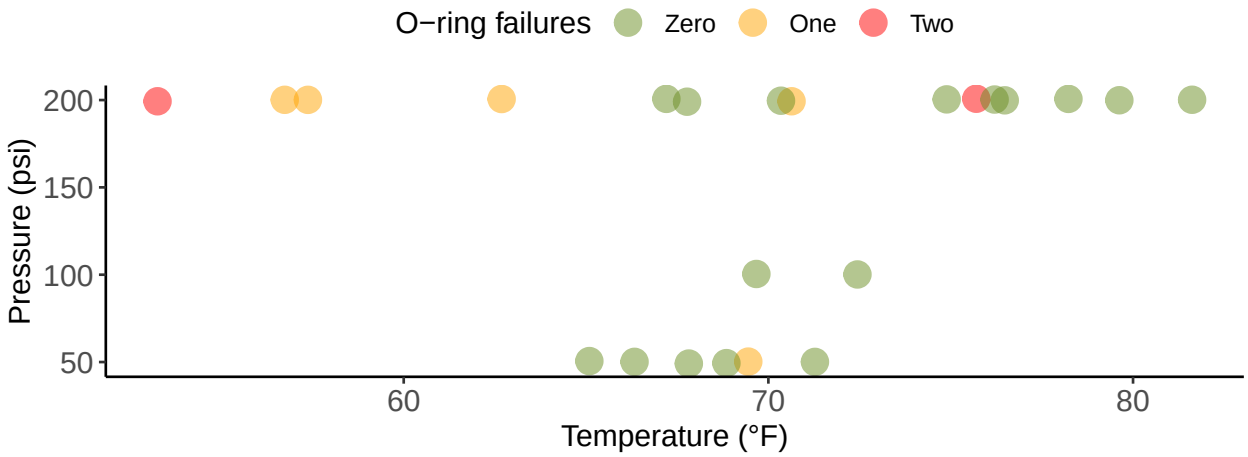


Based on the description in the paper, we do not anticipate there to be any meaningful relationship between temperature and pressure as they are both arbitrarily determined. Pressure is a condition set by the test procedure, which was first at 50 psi, then progressively updated to 100 psi and 200 psi. Temperature is merely a condition at the time of launch. The low correlation between the two variables, 0.04, supports our understanding.

Lastly, **Figure 2c** documents the relationship between the number of O-ring failures and flight number. If flight numbers were not assigned chronologically, then there is little to learn from this plot. However, if flight numbers are ordered chronologically then (i) note that the occurrence of failures *over time* appear nearly random- that is, there is no overt pattern- and (ii) the history of flight attempts is punctuated by episodes of failure.

Multivariate Analysis

Figure 3



In order to understand if there is a multivariate relationship between the number of O-ring failures, temperature, and pressure, we created a scatterplot in **Figure 3**. The number of O-ring failures is expressed as a function of both temperature and pressure. Green points represent those temperature and pressure conditions where 0 O-rings failed- these were the data ignored by NASA. Similarly, orange and red points represent those conditions under which 1 or 2 O-rings failed, respectively. As we showed earlier, there was little correlation between pressure and temperature. There does seem to be a larger range in temperatures as well as more O-ring failures for 200 psi, but this could just be due to the a greater number of data points at that level.

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 explore the use of the binomial and the binary models for the given launch data. The independence assumption is necessary for both models because it allows the use of the binomial probability mass function. However, the binomial model likely violates the assumption that each of the six 0-rings would fail independently at each temperature, t, and pressure, s. The failure of 1 or more of the O-rings could potentially influence the failure of the remaining O-rings, possibly due to exposure to damaging conditions or additional stress. On the other hand, the binary logistic regression model satisfies this assumption since each trial combines the success or failure of all 6 O-rings as one outcome, which is more likely to be independent from launch to launch. To enable the binary logistic regression model, the authors introduced a binary flag for failure, where one or more O-ring failures were simply marked as a failure.

The authors found that the fits between the the binomial and the binary models were comparable. Furthermore, the outcomes for the binary and binomial models are not substantially different, as

only 2 launches had 2 primary O-ring failures and the majority of the launches corresponded to 0 or 1 O-rings failures. Because of the analysis, we were comfortable with using the binomial model, despite the likely violation of the independence assumption. Because the dataset itself is quite limited with only 23 flight launches, a binomial model is able to make use of all 6 data points from each launch, while the binary model compresses the information for the 6 O-rings into a single data point.

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

```
# Create new variable Fail for binary model
df["Fail"] <- ifelse(df$O.ring>=1, 1, 0)

# Estimate a binomial logistic regression model
model_binomial <- glm(O.ring/Number ~ Temp + Pressure,
               family = binomial(link = logit),
               weights = Number, data = df)
# Estimate binary logistic regression model
model_binary <- glm(Fail ~ Temp + Pressure,
               family = binomial(link = logit),
               data = df)
# Summary of coefficients, standard errors and p-values
stargazer(model_binomial, model_binary, 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
```

Here we estimate the bimonial (1) and binary (2) regression models separately. For the binary model, we introduced a new flag variable "Fail" that is set to 1 if at least one failure, and 0 otherwise, which helps satisfy the independence assumption. From the above model summary, we see that

Temperature is the only significant variable at the 0.05 level for both models. By exponentiating the coefficients and taking inverse, we can interpret the effect of temperature on odds of failure.

Binomial Model

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

```
# Inverse odds ratio for binomial model
1/exp(summary(model_binomial)$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

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

```
# Inverse odds ratio for binary model
1/exp(summary(model_binary)$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`, for the binomial model:

```
Anova(model_binomial, 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
```

(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?

Based on the likelihood ratio test above, `Pressure` is not an important explanatory variable and could be removed from the model. 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 and 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 still 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 authors 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.

Potential problems could arise by the omission of `Pressure` from the model, however we are unable to draw conclusions on the value of `Pressure` with such limited variation in the data. While a simple model without `Pressure` might be sufficient, a model with `Pressure`, if more data existed, and an interaction term between `Temp` and `Pressure` might help guide interpretation. 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
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

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

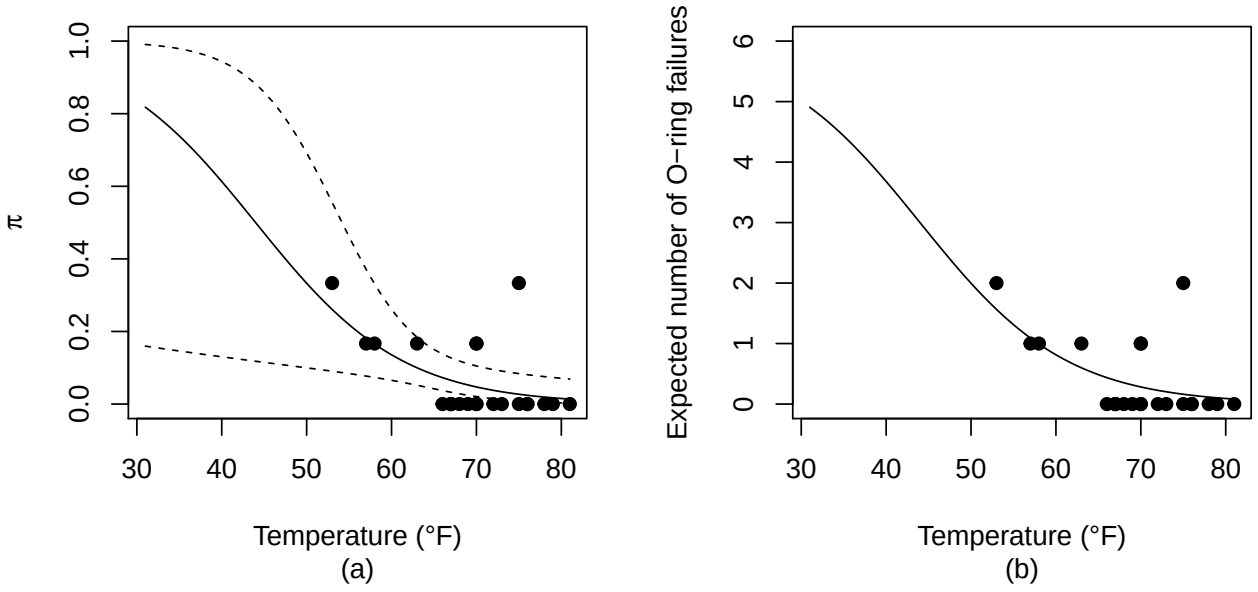
(b) Construct two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

Please see **Figure 4** below for the plots π vs `Temp` with confidence intervals (**Figure 4a**) and the expected number of failures vs `Temp`(**Figure 4b**).

(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_predict <- predict(object=model_b, newdata=data.frame(Temp=t), type='link', se=T)
CI_lower_linear <- model_predict$fit + qnorm(p=alpha/2)*model_predict$se.fit
CI_lower_pi <- exp(CI_lower_linear)/(1+exp(CI_lower_linear))
CI_higher_linear <- model_predict$fit + qnorm(p=1-alpha/2)*model_predict$se.fit
CI_higher_pi <- exp(CI_higher_linear)/(1+exp(CI_higher_linear))
```

Figure 4



The confidence band is wider when temperatures are lower than 53(°F) because there are no observations below 53(°F).

(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.

```r
# 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
```

The probability of O-ring failure at 31(°F) is 0.8177744 with a very wide confidence interval of 0.1596025, 0.9906582.

In order to infer the probability of O-ring failure at 31(°F), we need to make the assumption that there is a linear relationship between the log odds of failure of an O-ring and Temperature. Because we use the binomial model, we do not meet the independence assumption. However, the fits between the binomial and binary models are comparable, as discussed above.

(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 our parametric bootstrap procedure, we performed the following steps:

1. Using our binomial model with temperature as the explanatory model, we calculate Z, the linear outcome, using $Z = \beta_0 + \beta_1 \cdot Temp$, which results in a vector of size 23.
2. Next, we calculate $\hat{\pi} = \dfrac{e^Z}{1 + e^Z}$, which remains a vector of size 23.
3. The calculated $\hat{\pi}$ is saved to the dataframe as `O.ring.pi`.
4. The original dataset is resampled with replacement to create a new dataset d of size 23.
5. A vector of size 23 of binomials 0-6 is generated as $O.ring2$, represent new outcomes, using the built in rbinom function by passing in the associated calculated $\hat{\pi}$ values.
6. A new binomial logistic regression model is fitted with the resampled dataset d and outcomes $O.ring2$.
7. The predictions for 31 and 72 °F are found and the estimated probabilities are saved to *results*.

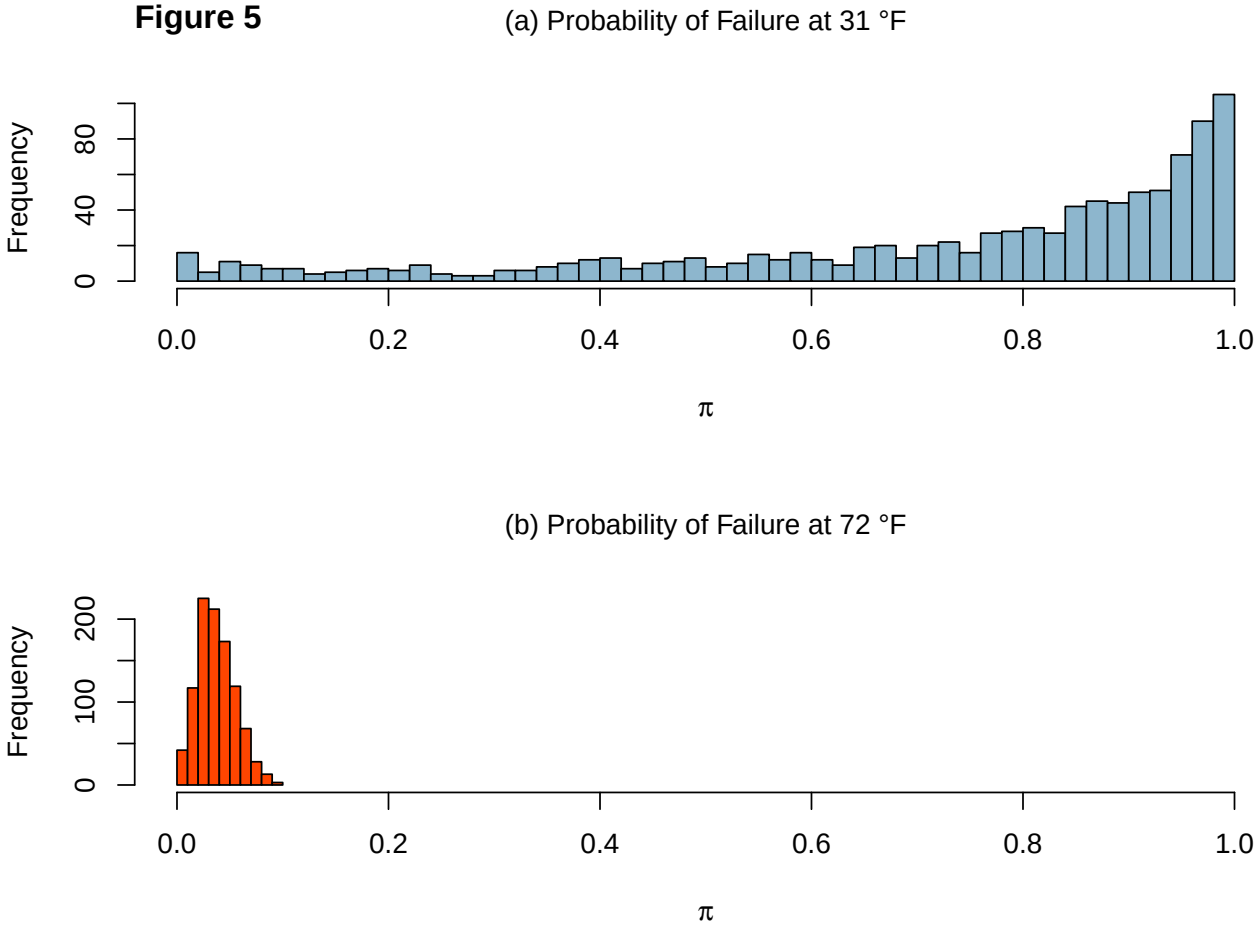
8. Steps 4-7 are repeated and performed for a total of 1000 times.
9. Finally, the 5% and 95% quantiles from the predictions for 31 and 72 °F accumulated in *results* are reported as the 90% confidence intervals.

```
# Set a seed
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 = T)
  # 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)
  # 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)
}

CI.31 <- quantile(results[,1], probs=c(0.05, 0.95))
CI.72 <- quantile(results[,2], probs=c(0.05, 0.95))
data.frame(temp=c(31,72), lower.CI=c(CI.31[1],CI.72[1]), upper.CI=c(CI.31[2],CI.72[2]))
```

```
##   temp   lower.CI   upper.CI
## 1   31 0.10960271 0.99287144
## 2   72 0.01137686 0.06824987
```

Using the parametric bootstrap method to estimate confidence interval, we again found that the confidence interval for 31 (°F) is quite wide at 0.1096 and 0.9929, due to the lack of data points at low temperatures. The confidence interval estimated for 72 (°F) is much tighter, at 0.0114 and 0.0682.

Figure 5



Out of the 1000 simulated datasets, we plotted a histogram of their probability of failures. The confidence intervals were taken as the 5th and 95th percentile of the distributions. In **Figure 5a**, the probability of failures for 31(°F) took on the full range of 0 and 1, and had a left skew towards 1. In contrast, the probability of failures for 72(°F) in **Figure 5b** were more concentrated at the left end towards 0.

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

```
# Estimate logistic regression model
model_c <- 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_c, 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
```

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

```
# Compare models with LRT
anova(model_b, model_c, 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
```

To test whether a quadratic term is needed, we specified a new model for `Temp` using both its linear and quadratic form. 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:

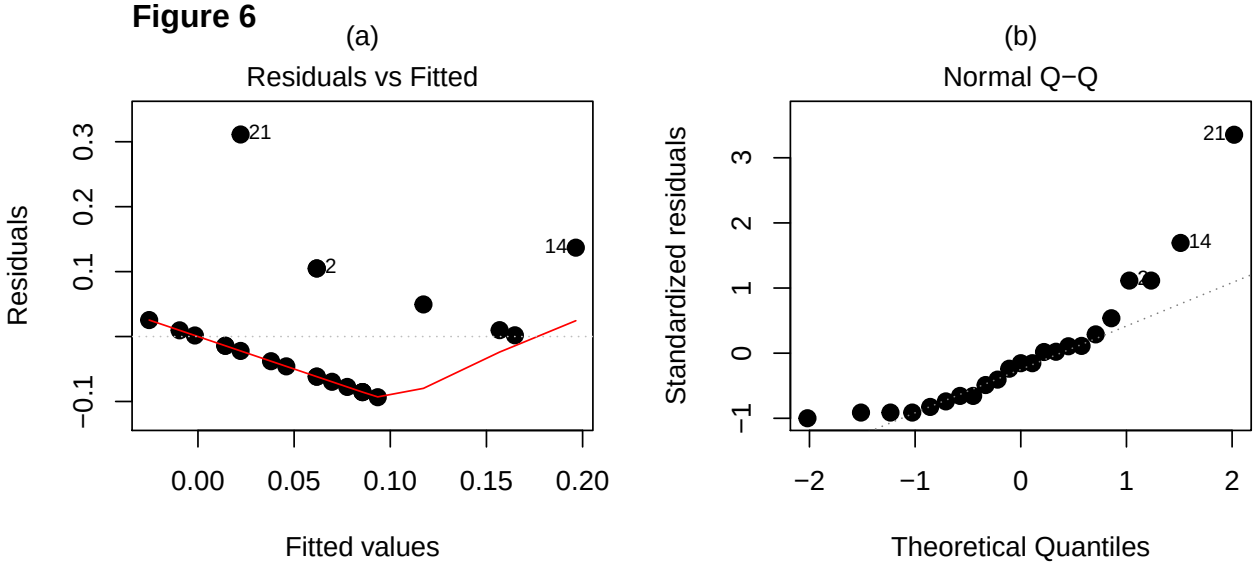
```
# Estimate model
model_d = lm(formula = O.ring/Number ~ Temp, data = df, weights = Number)
# Summary
stargazer(model_d, type="text")
```

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

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

To explain the results of the model, we will first forget that we are violating the assumptions discussed below. O-ring failure is negatively affected by temperature. That is, as temperature increases, the risk of O-ring failure decreases. For every incremental increase in temperature, the expected proportion of O-rings that fail drops by -0.0079233 with a standard error of 0.0029075, and p-value of 0.0126818. Thus, there is a modest effect of temperature on O-ring failure. Note that, even though we are violating several assumptions, we still reach a similar conclusion- launching rockets in cold weather is a bad idea!

Figure 6



Below, we assess the validity of the linear probability 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-ring outcomes were sampled randomly, or even sampled at all, as all of the O-rings data were likely collected from the 23 flights. It is not clear how the 23 flights were selected, however, it appeared that we only had observations for higher temperatures. If there is random sampling of the launch day throughout all seasons of the year, we would expect to have data that represents temperatures from the full range of possible temperatures. 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):

Since 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, strictly speaking. The residual versus fitted values plot (**Figure 6a**) 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 (**Figure 6a**). 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 (**Figure 6b**), as the points deviate from the dotted line.

Lastly, this model formulation possesses another shortcoming, in that the probabilities are linearly related to the predictors for all of their possible values.

We would select the logistic regression model over the linear regresion 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 discussed above do not hold and therefore the linear regression model is not suitable for this problem.

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 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.

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

```
## (Intercept)        Temp
## 161.5762605   0.8908304
```

To calculate the inverse odds, we inverse the exponentiated coefficients, as shown below.

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

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

Here we see that a unit decrease in temperature changes the odds of failure 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 (**Figure 4a-b**), we clearly see an increase in the probability of failure with lower temperature. For temperature going from 50°F, 40°F to 30°F, 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. Although the probability of failure rises at lower temperatures, the confidence intervals widen in response to the absence of data. Even with using parametric bootstrapping to estimate confidence intervals, we found the same result, that the confidence interval at the lower temperature of 31°F remains wide. The fact that there were no previous data point below temperature of 53°F at all, and the confidence interval at the temperature of 31°F was wide, should have been enough justification to postpone the launch on a cold day since it was tempered by uncertainty.

We evaluated both the binary logistic regression models, where O-ring failure is a dichotomous function of the explanatory variables as well as the binomial logistic regression model that considers each O-ring as a separate data point. Though the binary model may be more robust, since it does not have the assumption of independence across O-ring failures, the Dalal *et al.* (1989) paper utilizes both models throughout the analysis. We settled on the binomial model as our final model considering that the observations are limited.

In summary, we explored the relationship to pressure vs O-ring failure and temperature vs O-ring failure. Conclusions on these relationships should be, as noted above, conditioned on the paucity of data. We note a higher incidence of failures at lower temperatures. Moreover, all but one of the failures occurred at higher pressures. Equipped with more data, a test for independence could be used to determine if levels of pressure are independent of levels of temperature (say 30-50, 51-70, 71-90 °F). Similarly, logistic regression models that captured the relationship between O-ring failure, temperature, pressure and the interaction of the two explanatory variables could reveal important information about physical processes that influence O-ring failure.

Regardless of the limitations and challenges, we learned that (i) launching rockets in cold temperatures is a bad idea and (ii) ignoring data, as NASA ignored the "outlying" values where 0 failures occurred, can be deadly. Both groups of models, the binary and binomial models, converged on these conclusions.