Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 1

Robert Deng, Tiffany Jaya, Shan He, Joanna Huang

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Introduction

- An introduction section that summarize the question(s) being asked, the methodology employed (including the final model specification), and a highlight of the results.

The failure of an O-ring on the space shuttle Challenger's booster rockets led to its destruction in 1986. Using data on previous space shuttle launches, Dalal et al. (1989) examine the probability of an O-ring failure as a function of temperature at launch and combustion pressure.

EDA

- A comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Please remember that your report will have to "walk me through" your analysis.
- A thorough analysis of the given dataset, which includ examiniation of anomalies, missing values, potential of top and/or bottom code, etc, in each of the variables.

Variables in the dataset: • Flight: Flight number • Temp: Temperature (F) at launch • Pressure: Combustion pressure (psi) • O.ring: Number of primary field O-ring failures • Number: Total number of primary field O-rings (six total, three each for the two booster rockets)

```
challenger <- read.table(file = "./dataset/challenger.csv", header = TRUE,
    sep = ",")
summary(challenger)
```

```
##      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
```

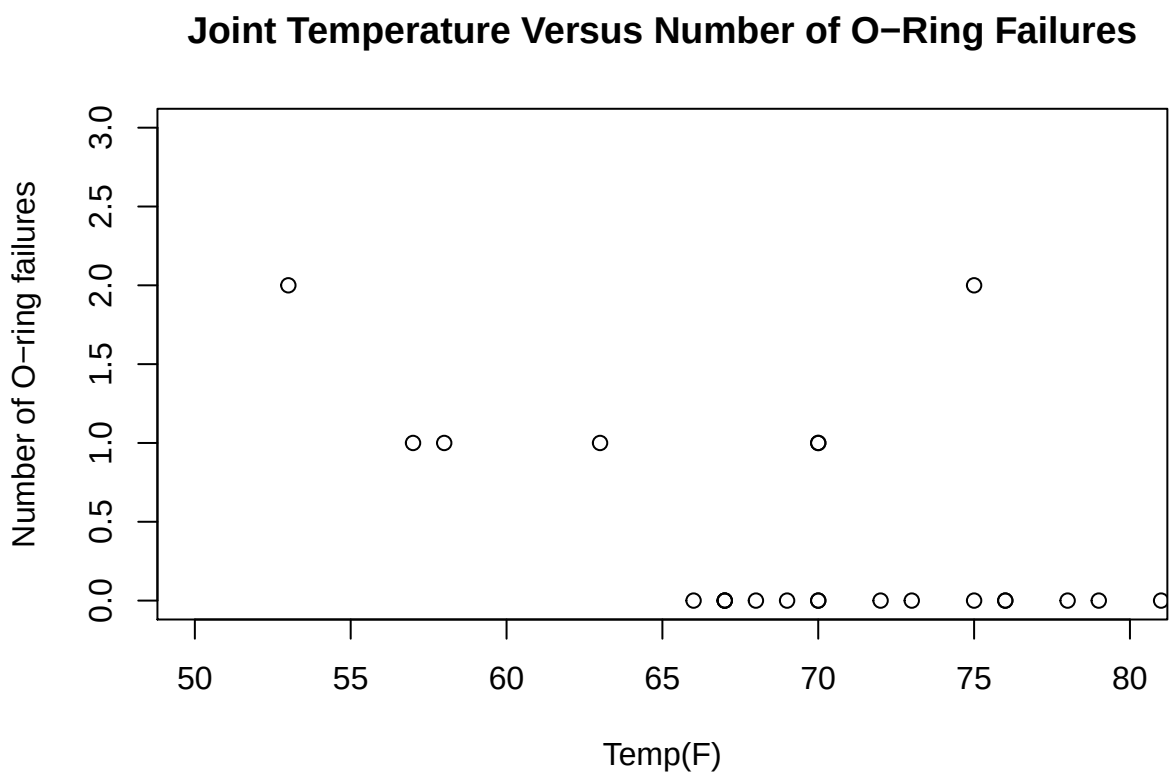
```
##  3rd Qu.:6
##  Max.   :6
```

```
# list rows of data that have missing values
challenger[!complete.cases(challenger), ]  # no missing values
```

```
## [1] Flight   Temp     Pressure O.ring   Number
## <0 rows> (or 0-length row.names)
```

There is a total of 23 data points in this dataset. As mentioned in the article, there were 24 launches prior to Challenger but one flight lost its motors at sea, thus we end up with 23 flights with motor data recorded here. All space shuttles have a total numebr of 6 primary field O-rings, so the Number column can be removed. No columns have any abnormal values.

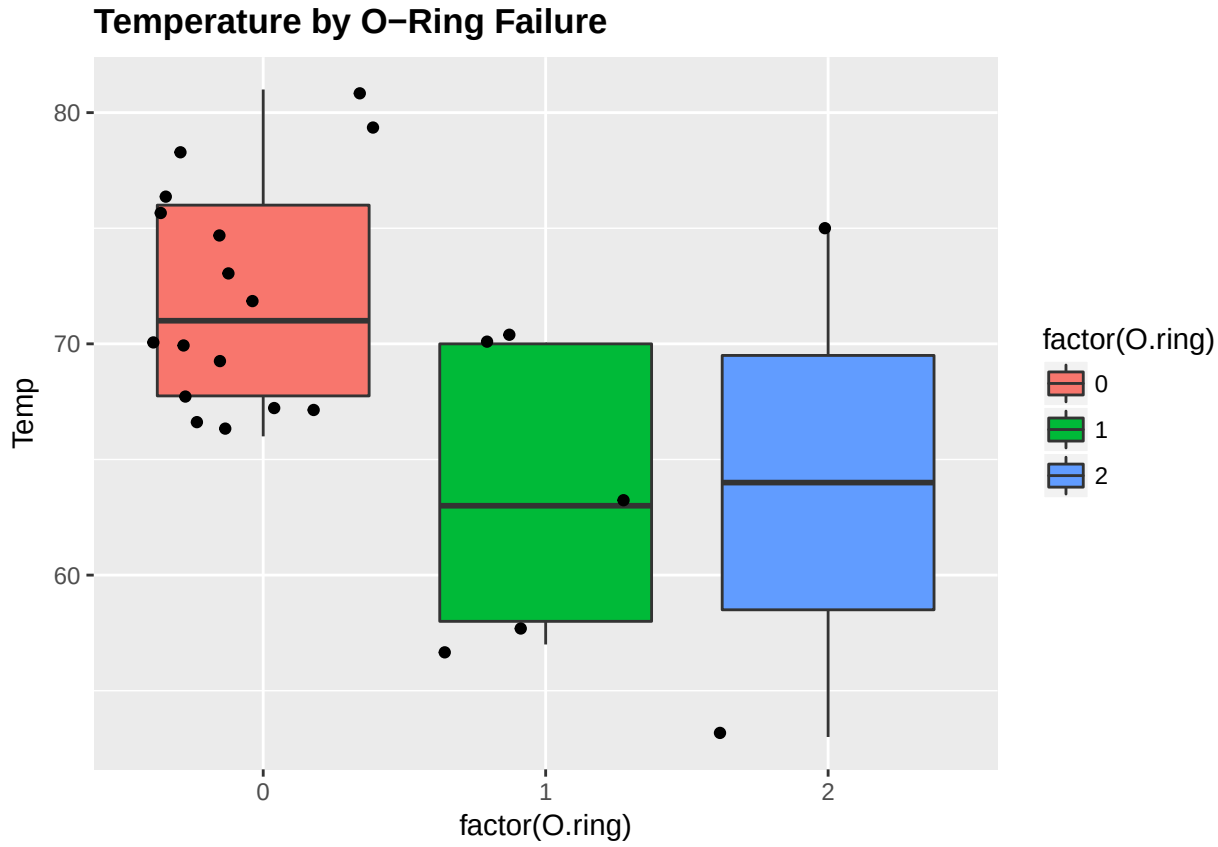
```
plot(x = challenger$Temp, y = challenger$O.ring, ylab = "Number of O-ring failures",
    xlab = "Temp(F)", main = "Joint Temperature Versus Number of O-Ring Failures",
    xlim = c(50, 80), ylim = c(0, 3))
```



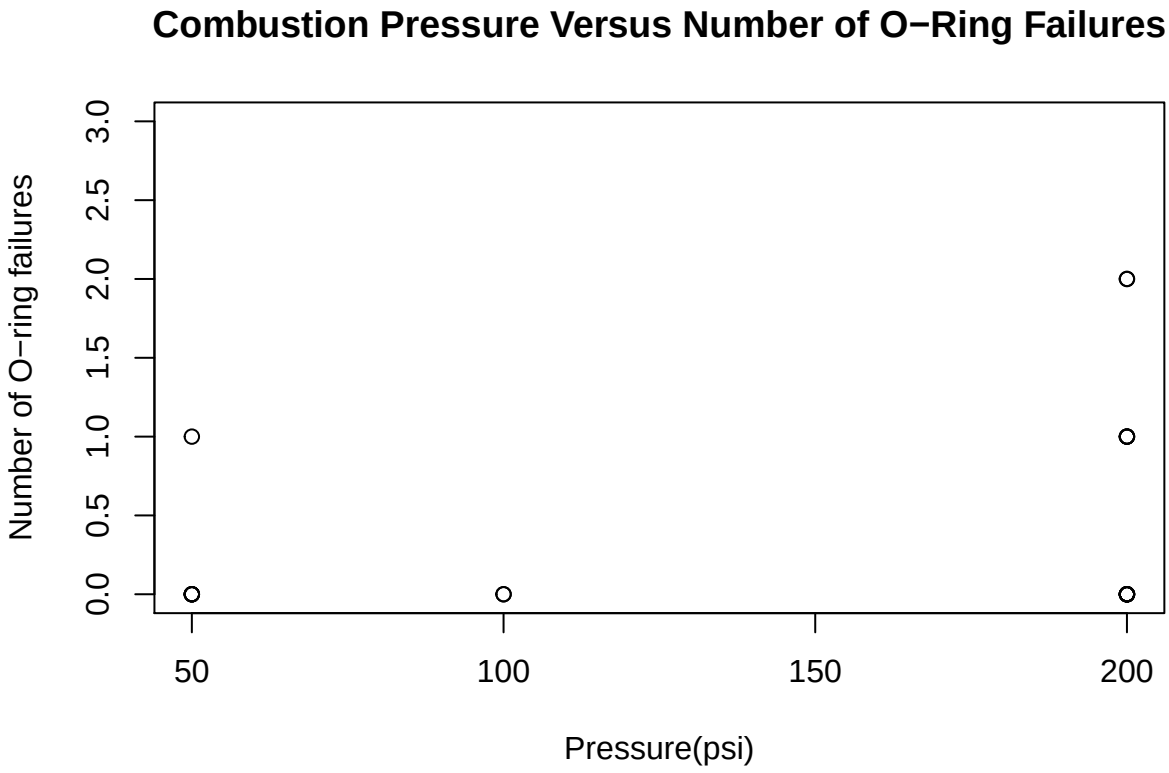
Looking at this plot, there appears to be a correlation of O-ring failures with lower temperatures, with the exception of two points.

```
library(ggplot2)
# temperature by o-ring failure
```

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

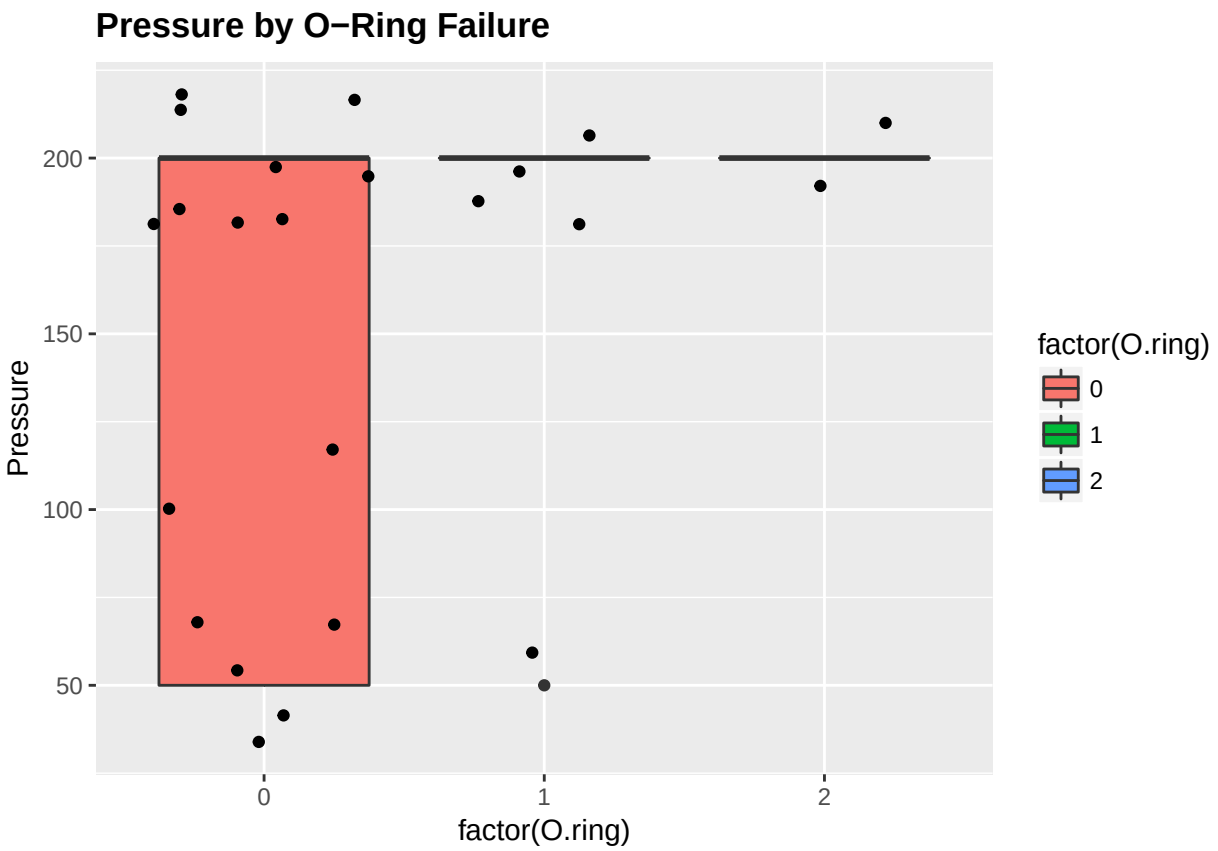



```
plot(x = challenger$Pressure, y = challenger$O.ring, ylab = "Number of O-ring failures",
    xlab = "Pressure(psi)", main = "Combustion Pressure Versus Number of O-Ring Failures",
    xlim = c(50, 200), ylim = c(0, 3))
```



Looking at this plot, there doesn't appear to be any particular relationship between O-ring failures and pressure.

```
# pressure by o-ring failure
ggplot(challenger, aes(factor(O.ring), Pressure)) + geom_boxplot(aes(fill = factor(O.ring))) +
    geom_jitter() + ggtitle("Pressure by O-Ring Failure") + theme(plot.title = element_text(li
    face = "bold"))
```



Modeling

- A modeling section that include a detailed narrative.
- The rationale of decisions made in your modeling, supported by sufficient empirical evidence. Use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
- All the steps used to arrive at your final model; these steps must be clearly shown and explained.

The response variable is O.ring, and the explanatory variables are Temp and Pressure. Complete the following:
(a) The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.
This assumption is necessary because in order to sum up the Bernoulli random variables to a binomial model, the random variables need to be independent and have the same probability of success. The potential issue is that such independence cannot always be assumed.

(b) Estimate the logistic regression model using the explanatory variables in a linear form. **Since we are interested in examining the probability of an O-ring failure, regardless of how many O-rings failed, we can include an additional column in the data with a value of 1 if O-ring failure !=0 and 0 if it does.**

```
challenger$Fail <- ifelse(challenger$O.ring >= 1, 1, 0)
head(challenger)
```

```
##   Flight Temp Pressure O.ring Number Fail
## 1      1   66       50      0      6    0
## 2      2   70       50      1      6    1
## 3      3   69       50      0      6    0
## 4      4   68       50      0      6    0
## 5      5   67       50      0      6    0
## 6      6   72       50      0      6    0
```

```
challenger.glm <- glm(Fail ~ Temp + Pressure, family = binomial,
    data = challenger)
summary(challenger.glm)
```

```
##
## Call:
## glm(formula = Fail ~ Temp + Pressure, family = binomial, data = challenger)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1993  -0.5778  -0.4247   0.3523   2.1449
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.292360   7.663968   1.734   0.0828 .
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## Pressure     0.010400   0.008979   1.158   0.2468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 18.782  on 20  degrees of freedom
## AIC: 24.782
##
## Number of Fisher Scoring iterations: 5
```

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

```
challenger.glm.H0 <- glm(Fail ~ Temp, family = binomial, data = challenger)

anova(challenger.glm.H0, challenger.glm, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Fail ~ Temp
## Model 2: Fail ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     18.782  1   1.5331   0.2156
```

```
library(car)
Anova(mod = challenger.glm, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Fail
##          LR Chisq Df Pr(>Chisq)
## Temp       7.7542  1   0.005359 **
## Pressure   1.5331  1   0.215648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

Both LRTs output a p-value of 0.215648 for the variable Pressure. Thus we cannot reject the null hypothesis model. There is insufficient evidence to indicate that pressure has an effect on the probability of O-ring failure given temperature is in the model.

5. Continuing Exercise 4, consider the simplified model logit( ) = 0 + 1Temp, where is the probability of an O-ring failure. Complete the following:

(a) Estimate the model.

```
challenger.glm.temp <- glm(formula = Fail ~ Temp, family = binomial(link = logit),
    data = challenger)
summary(challenger.glm.temp)
```

```
##
## Call:
## glm(formula = Fail ~ Temp, family = binomial(link = logit), data = challenger)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0611  -0.7613  -0.3783   0.4524   2.2175
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## Temp         -0.2322     0.1082  -2.145   0.0320 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

(b) Construct two plots: (1) 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.

```
# Find the observed proportion of failure at each Temp
w <- aggregate(formula = Fail ~ Temp, data = challenger, FUN = sum)
n <- aggregate(formula = Fail ~ Temp, data = challenger, FUN = length)
w.n <- data.frame(Temp = w$Temp, success = w$Fail, trials = n$Fail,
    proportion = round(w$Fail/n$Fail, 4))
head(w.n)
```

```
##   Temp success trials proportion
## 1   53       1      1          1
## 2   57       1      1          1
## 3   58       1      1          1
## 4   63       1      1          1
## 5   66       0      1          0
## 6   67       0      3          0
```

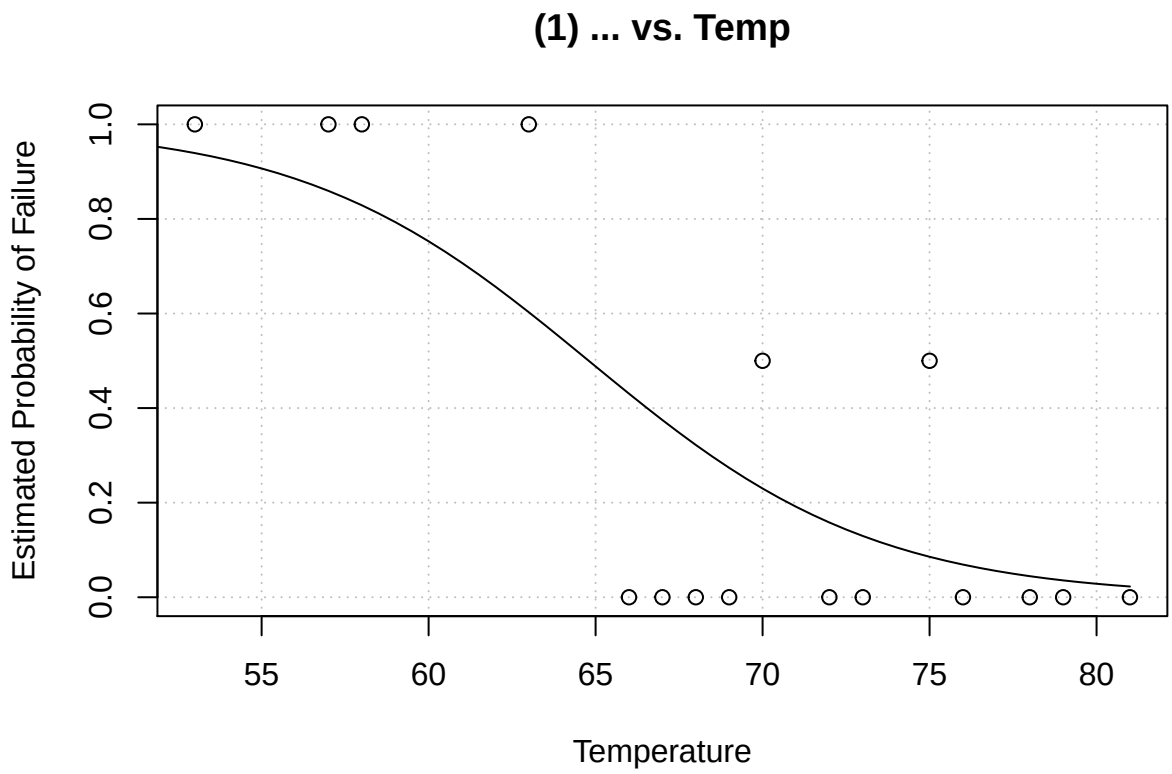
```
# Plot of the observed proportions with logistic regression
# model

plot(x = w$Temp, y = w$Fail/n$Fail, xlab = "Temperature", ylab = "Estimated Probability of Fai
    panel.first = grid(col = "gray", lty = "dotted"), main = "(1)  vs. Temp")
```

```
## Warning in title(...): conversion failure on '(1)  vs. Temp' in
## 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in title(...): conversion failure on '(1)  vs. Temp' in
## 'mbcsToSbcs': dot substituted for <87>
```

```
## Warning in title(...): conversion failure on '(1)  vs. Temp' in
## 'mbcsToSbcs': dot substituted for <a1>
```

```
# Put estimated logistic regression model on the plot
curve(expr = predict(object = challenger.glm.temp, newdata = data.frame(Temp = x),
    type = "response"), add = TRUE, xlim = c(31, 81))
```



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

```
plot(x = w$Temp, y = w$Fail/n$Fail, xlab = "Temperature", ylab = "Estimated Probability of Fail
    panel.first = grid(col = "gray", lty = "dotted"), main = "  vs. Temp with C.I. Bands")
```

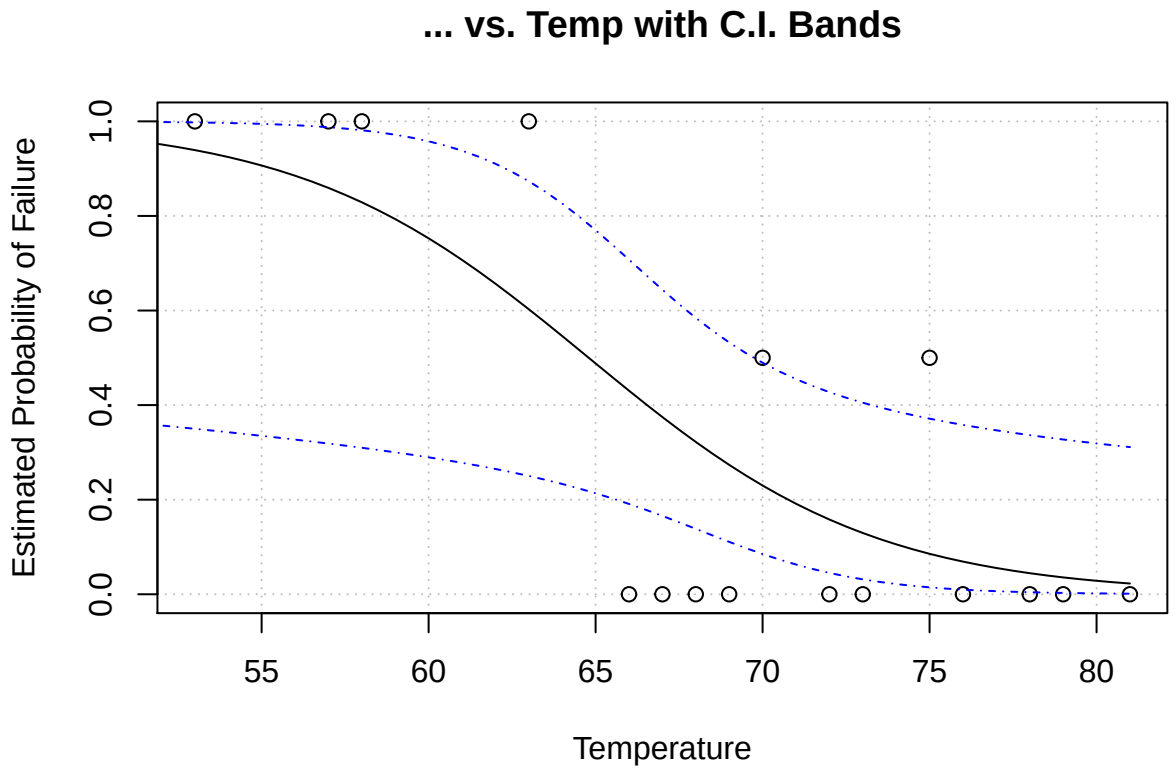
```
## Warning in title(...): conversion failure on '  vs. Temp with C.I. Bands'
## in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in title(...): conversion failure on '  vs. Temp with C.I. Bands'
## in 'mbcsToSbcs': dot substituted for <87>
```

```
## Warning in title(...): conversion failure on '  vs. Temp with C.I. Bands'
## in 'mbcsToSbcs': dot substituted for <a1>
```

```
# Put estimated logistic regression model on the plot
curve(expr = predict(object = challenger.glm.temp, newdata = data.frame(Temp = x),
    type = "response"), add = TRUE, xlim = c(31, 81))

# Create C.I. function
conf.int <- function(newdata, mod, alpha) {
    linear.pred <- predict(object = mod, newdata = newdata, type = "link",
        se = TRUE)
    CI.lin.pred.lower <- linear.pred$fit - qnorm(p = 1 - alpha/2) *
```

```
        linear.pred$se
    CI.lin.pred.upper <- linear.pred$fit + qnorm(p = 1 - alpha/2) *
        linear.pred$se
    CI.pi.lower <- exp(CI.lin.pred.lower)/(1 + exp(CI.lin.pred.lower))
    CI.pi.upper <- exp(CI.lin.pred.upper)/(1 + exp(CI.lin.pred.upper))
    list(lower = CI.pi.lower, upper = CI.pi.upper)
}

# Add 95% Wald C.I. bands
curve(expr = conf.int(newdata = data.frame(Temp = x), mod = challenger.glm.temp,
    alpha = 0.05)$lower, col = "blue", lty = "dotdash", add = TRUE,
    xlim = c(31, 81))
curve(expr = conf.int(newdata = data.frame(Temp = x), mod = challenger.glm.temp,
    alpha = 0.05)$upper, col = "blue", lty = "dotdash", add = TRUE,
    xlim = c(31, 81))
```



(d) The temperature was 31 at launch for the Challenger in 1986. Estimate the probability of an O-ring failure using this temperature, and compute a corresponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

(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 logit(ˆ) = ˆ0 +

ˆ1Temp; (2) estimate new models for each data set, say logit(ˆ?) = ˆ0? + ˆ1?Temp; and (3) compute ˆ? at a specific temperature of interest. The authors used the 0.05 and 0.95 observed quantiles from the ˆ? sim- ulated distribution as their 90% confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31 and 72 .27

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

3. In addition to the questions in Question 4 and 5, answer the following questions:
   a. Interpret the main result of your final model in terms of both odds and probability of failure
   b. 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. Why? Or, why not?

Conclusion

- A conclusion that summarize the final result with respect to the question(s) being asked and key takeaways from the analysis.