w271 Lab 1 Solutions

Aditya Dhara, Arunima Kayath, Peter Grabowski

June 3, 2018

Introduction

In this Lab, we use data from challenger shuttle flights to predict the odds of failure of an O-ring. This is said to have been a root cause of the challenger disaster, wherein statistical techniques were incorrectly used to predict the likelihood of an O-ring failure on the shuttle.

Questions addressed

In this lab we address two broad problems of fitting a logistic regression model of likelihood of O-ring failure on the Temperature and Pressure columns. Then we explore a more focused model using Temperature alone. Finally, we interpret the results in terms of odds and probabilities, and provide an alternative linear model and assess its validity.

Highlight of Results

Using a logistic regression model, and bootstrap confidence intervals, we estimate that the challenger shuttle would have experienced one O-ring failure with 85% chance, and a 90% confidence interval of between 27% and 99% chance.

EDA

```
challenger = read.csv("challenger.csv", header=TRUE)
library(Hmisc)
library(ggplot2)
library(dplyr)
library(car)
library(mcprofile)
```

The data set contains flight of the shuttle program, and included the following columns:

1. Flight (int): flight number, serves as an ID for this data set
2. Temp (int): sea level temperature at time of flight
3. Pressure (int): this was the pressure used to test O-Rings before the flight
4. O. Ring (int): the number of O-rings that failed in the launch
5. Number (int): the number of O-rings that were installed in the aircraft

Please note, the question states the Pressure variable is "Combustion pressure", while Dalal et al define the Pressure variable as the level at which the O-Ring was tested. They suggested that tests with stronger pressures could even cause degradation in the O-rings, causing blowback. We are using Dalal et al's definition here.

```
str(challenger)
```

```
## '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 ...
```

```
head(challenger)
```

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

A few things to note: - There are only 23 total observations. So we will be making observations from a small data-set. Within those 23, there are no missing values. - The lowest temperature for which data is available is 53, which is much higher than the flight conditions temperature of 31. So we will be extrapolating beyond observed data using the model we build to get an estimate for 31 degrees, the temperature at the time of the flight. - Pressure is only available at 3 distinct levels, and of these, there are only 2 data points when the pressure is at 100. - O.ring failure is seen in 7 out of the 23 flights. Of the 7 failures, two instances show 2 failures, and 5 instances shows 1 failure.

For the sake of the analysis, the column `ratio` was calculated to denote the proportion of failed O-Rings per flight

```
challenger$ratio = challenger$O.ring / challenger$Number
plot(
  x=challenger$Temp, y=challenger$ratio,
  main="Proportion of O-ring failures for various launch temperatures",
  xlab="Temperature", ylab="Proportion of failures"
)
```

Proportion of O–ring failures for various launch temperatures
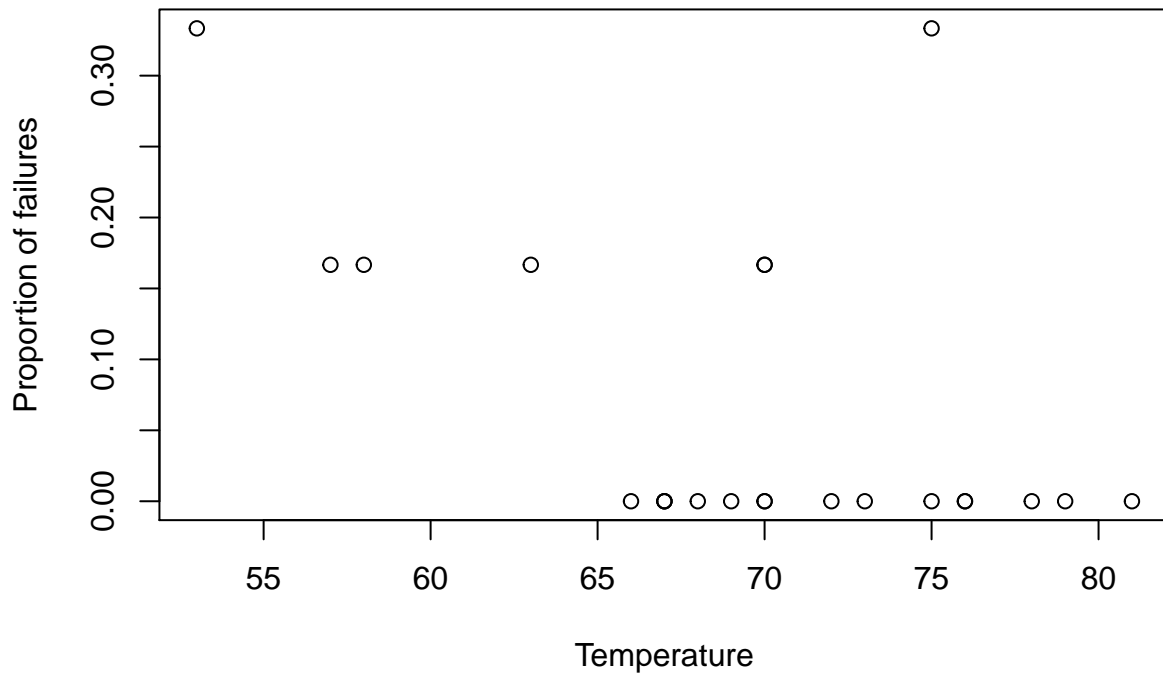

```
plot( x = challenger$Pressure, y = challenger$ratio,
  main="Proportion of O-ring failures for various pressures",
  xlab="Pressure", ylab="Proportion of failures")
```

Proportion of O–ring failures for various pressures
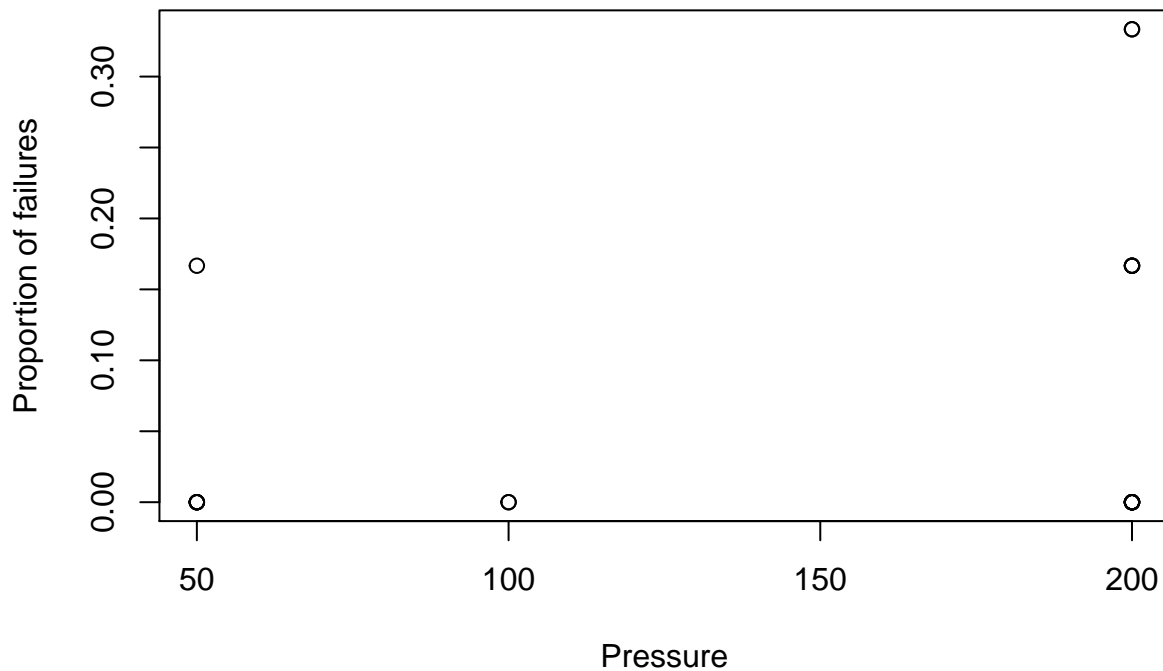
We have also assessed the assumptions of a binomial distribution, which are necessary for our logistic regression to describe the data properly: 1 - *Fixed number of trials*: This is clearly the case as each row in the data set

has a `Number` of 6 for the number of o-rings in a shuttle 2 - *Each observation is independent*: Independence is slightly questionable since the failure of one o-ring has a twisting effect putting pressure on other o-rings in each craft. This means the likelihood of failure for other o-rings in the craft could change after one o-ring fails. This is a strong assumption we will make. 3 - *Two and only two outcomes*: This is also clearly the case as an o-ring failure is unambiguous. We are ignoring the causes and symptoms such as blowback or the melting of the putty. We are simply measuring if the o-ring failed or not 4 - *Probability of success is the same for each trial*: The six o-rings in a craft are placed in different locations. However, the functionality and structures around the o-rings are the same, and so this is an assumption we can make. 5 - *Random variable of interest W is the number of successes*: This is also clearly the case as the `O.ring` variable contains the number of failed O-rings

Q4. Full model: Temperature and Pressure

4A: Independence assumption and problems

The necessity of this assumption stems from the requisite assumptions when using a binomial regression model. The use of the model requires that all observations be independent of one another and have the same probability of failure. If not for this assumption, the regression model would be much more complicated. As a result, the log likelihood function based on the model that is used to perform Maximum Likelihood Estimation will also become increasingly complicated with joint probability terms.

For a space shuttle the independence of failure of O-rings is a strong assumption. For example, failure of one O-ring may cause stress throughout the rest of the system, influencing the likelihood that adjacent O-rings would fail.

4B: Estimating a logistic regression

Here we estimate the logistic regression using Temperature and Pressure as independent variables. We use the `binomial` family with a `logit` link in order to achieve a logistic regression. We use the calculated `ratio` of failed O-rings as the dependent variable. We also set `Number` to be the column of weights for this regression since the table is already in an aggregated form where each row represents an observation of a binomial variable.

```
mod.fit = glm(
  formula = ratio ~ Temp + Pressure,
  family=binomial(link=logit),
  weights=Number,
  data=challenger
)

summary(mod.fit)
```

```
##
## Call:
## glm(formula = ratio ~ Temp + Pressure, family = binomial(link = logit),
##     data = challenger, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

4

```
## (Intercept)   2.520195    3.486784    0.723    0.4698
## Temp          -0.098297    0.044890   -2.190    0.0285 *
## Pressure       0.008484    0.007677    1.105    0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

From this model, we estimate for each 1 degree Fahrenheit of temperature drop, the odds of an O-ring failure increase by $e^{0.098297} \approx 1.1033$. The Wald confidence interval also suggests there is a 97% significance of this effect, suggesting that 97% of similarly sized samples will contain the true co-efficient of temperature within their bounds. Additionally, the Wald hypothesis testing suggests that `Pressure` does not significantly affect the log-likelihood of O-ring failure (More on that in section 4D)

4C: Log Likelihood Ratio Tests (LRT)

In order to assess the effectiveness of this model, we started with the Wald confidence interval which can be seen in the summary of this model. However, since it is very liberal and assumes normality of variance in estimated parameters, we will perform the more effective Likelihood Ratio tests here.

```
library(package = car)
Anova(mod.fit, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ratio
##         LR Chisq Df Pr(>Chisq)
## Temp      5.1838  1     0.0228 *
## Pressure  1.5407  1     0.2145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.fit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: ratio
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                       22     24.230
## Temp      1   6.1440        21     18.086  0.01319 *
## Pressure  1   1.5407        20     16.546  0.21452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Likelihood Ratio test confirms that Temperature has a significant effect on likelihood of failure, while there isn't sufficient evidence to show that Pressure has a significant effect on likelihood of failure.
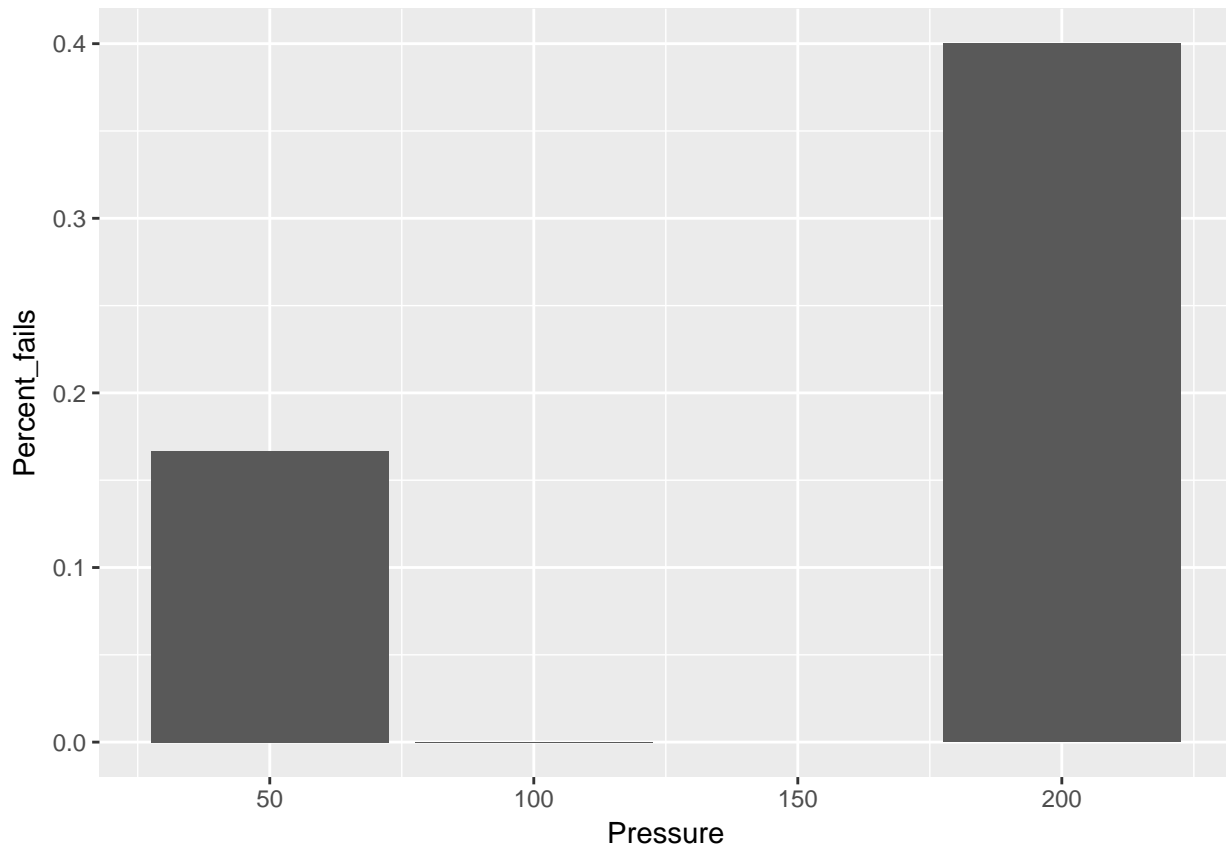
4D: Removing the `Pressure` variable

The effect of pressure was not significant in any of the analyses we ran, which supports the authors' decision to remove it from the model. However, there are concerns that remain while removing it.

```r
cnt_total <- count(challenger, group_by = Pressure)
challenger$fail <- ifelse(challenger$O.ring>0,1,0) #create a fail/ no fail binary variable

cnt_fails <- challenger %>%
  group_by(Pressure) %>%
  summarise(Frequency = sum(fail))

temp = data.frame(Pressure = cnt_fails$Pressure, Percent_fails = cnt_fails$Frequency/cnt_total$n)
ggplot(temp, aes(x=Pressure, y=Percent_fails)) + geom_bar(stat = "identity")
```



```r
ggplot(data = challenger) + geom_point(mapping = aes(x = Temp,y = Pressure))
```
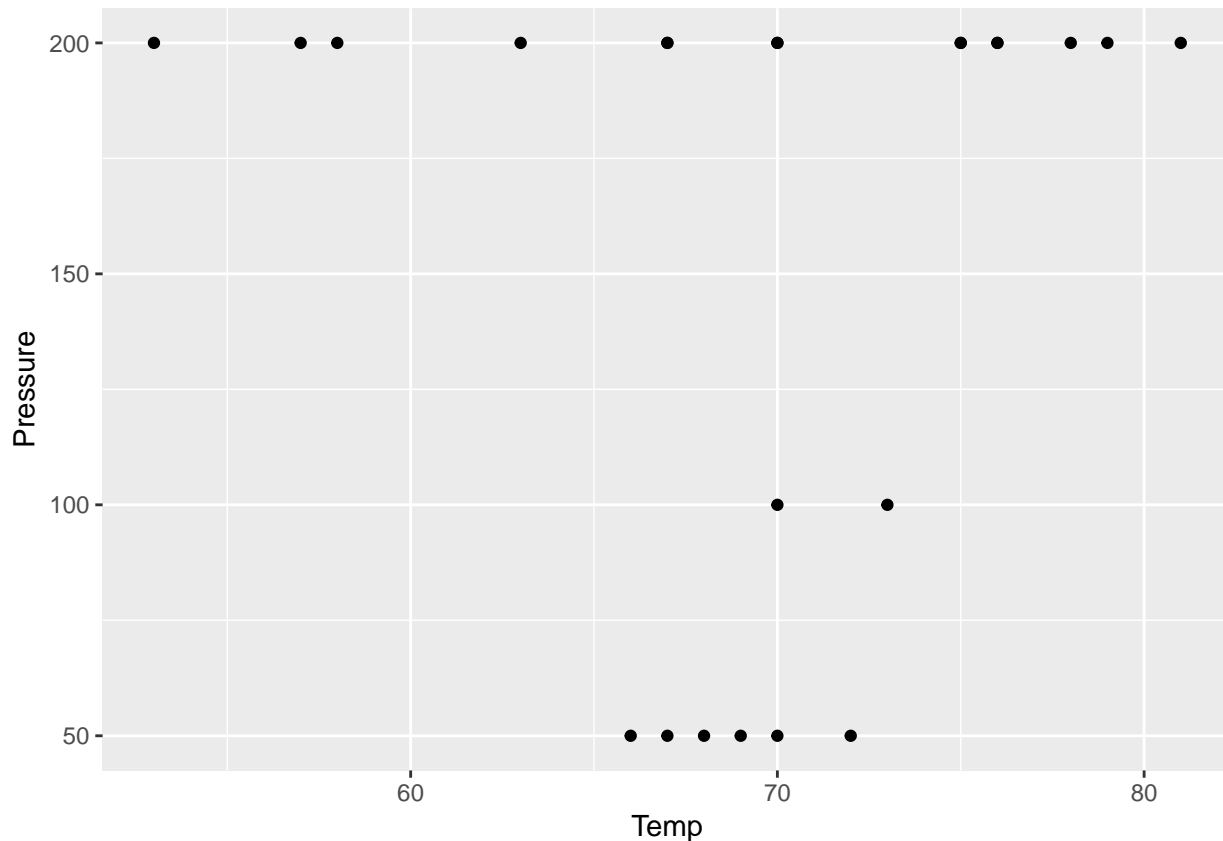
```
cor(challenger$Temp, challenger$Pressure)
```

[1] 0.03981769

The effect of pressure was not significant in any of the analyses we ran, which supports the authors' decision to remove it from the model. However, there are concerns that remain while removing Pressure. As our graphic analysis above shows, there does seem to be some weak correlation with pressure. The sample size is very small, so it's possible this is masking an otherwise weak effect size. We also haven't checked for any interaction effects with pressure, which could influence the model.

We also checked the correlation of Temperature and Pressure to see if that could be masking the effect of Pressure. However, the correlation is small. It is interesting though that there is very little temperature variation at the lower pressure, and most of the temperature variation is seen at the higher pressure.

Q5. Focused model: Temperature only

5A: Creating the model

We perform a similar analysis as 4B, except we exclude `Pressure` from this model. We use the `binomial` family with a `logit` link in order to achieve a logistic regression. We use the calculated `ratio` of failed O-rings as the dependent variable. We also set `Number` to be the column of weights for this regression since the table is already in an aggregated form where each row represents an observation of a binomial variable.

```
mod.fit2 = glm(
  formula = ratio ~ Temp,
  family=binomial(link=logit),
  weights=Number,
```

```
  data=challenger
)

summary(mod.fit2)
```

```
##
## Call:
## glm(formula = ratio ~ Temp, family = binomial(link = logit),
##     data = challenger, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

From this model, we estimate for each 1 degree Fahrenheit of temperature drop, the odds of an O-ring failure increase by $e^{0.11560} \approx 1.1225$. We can also observe the Wald hypothesis testing yields that the intercept as well as the Temperature parameters have achieved at least 95% significance.

5B: Plots of the logistic regression model

We use the estimated parameters to create the logistic regression model formula and plot it using the `curve` built-in function. This formula would plot the estimated expected probabilities of any O-ring's failure given a temperature (note, F is the random variable representing a single O-ring failure).

$$E(\hat{F}) = \hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 temp}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 temp}}$$

We can also utilize this formula to estimate the binomial estimated number of failed O-rings. Since $E(W) = n\pi$, and the number of O-rings per shuttle is $n = 6$, $E(W) = 6\pi$. Thus, using our model, we can simply use the same formula to estimate the expected number of O-rings that will fail at different temperatures:

$$E(\hat{W}) = 6\hat{\pi} = 6\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 temp}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 temp}}$$

```
b0 = mod.fit2$coefficients[["(Intercept)"]]
b1 = mod.fit2$coefficients[["Temp"]]
pi.vs.temp = function(x) { exp(b0 + b1*x) / (1+ exp( b0 + b1*x)) }
exp.failures.vs.temp = function(x) { 6 * pi.vs.temp(x) }
```

```
par(mfrow=c(1, 2))
curve(pi.vs.temp, from=10, to=90, col='blue',
      main="Estimated Pi vs Temperature", xlab="Temperature", ylab="Estimated pi")
curve(exp.failures.vs.temp, from=10, to=90, col='red',
      main="E(O-Ring Failures) vs Temperature", xlab="Temperature", ylab="E(O-Ring Failures)")
```

Estimated Pi vs Temperature   E(O−Ring Failures) vs Temperatu
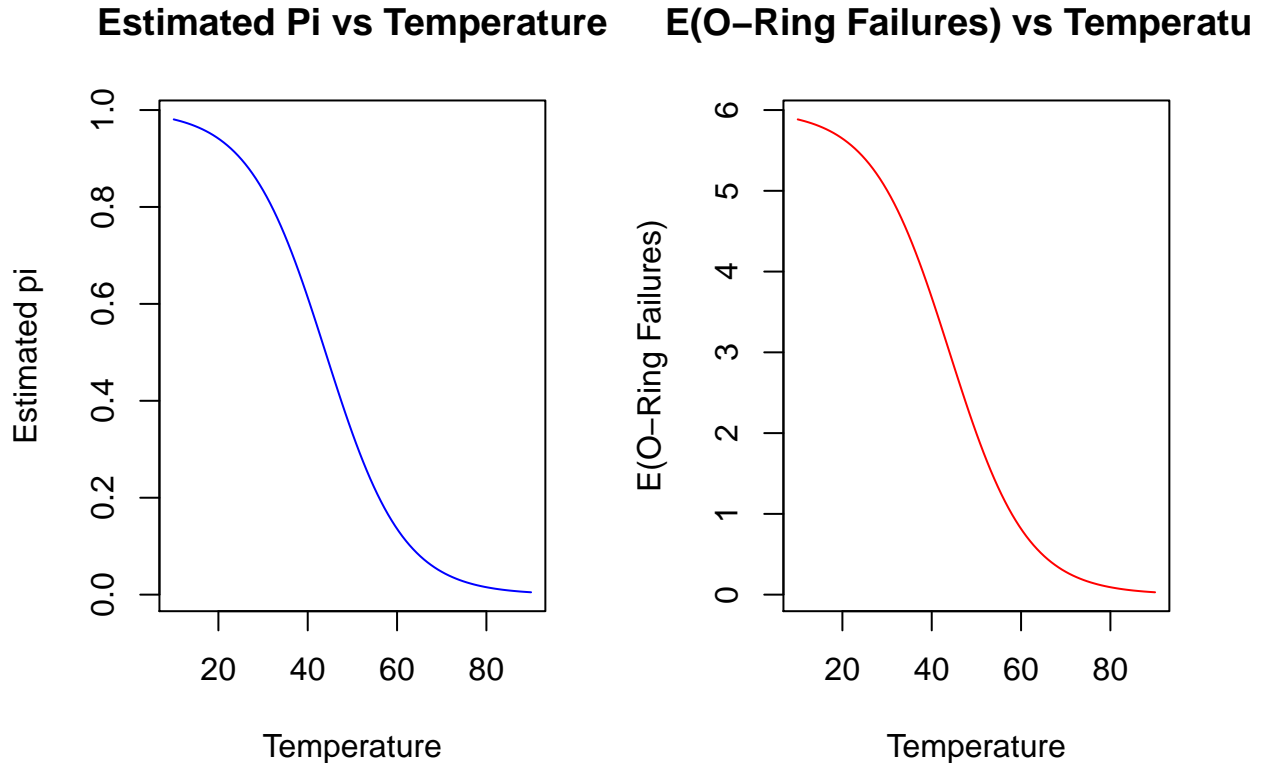
5C: Wald confidence interval bands

We can utilize the model's fitted values to determine the Wald confidence interval bands for π at 95% confidence. We utilize the qnorm function to calculate the value z-statistic for $\alpha = 0.05$, and multiply it with the $\hat{Var}(\hat{W})$ (which can also be found in the model object).

```
predict.data = data.frame(Temp = seq(from = 31, to = 81))

ci.pi = function(newdata, mod.fit.obj, alpha) {
  linear.pred = predict(object = mod.fit.obj, 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)
}

curve(
  ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = mod.fit2, alpha=0.05)$upper,
  lty = "dotdash", xlim=c(0,81), ylim=c(0,1),
  main="Wald confidence intervals for Estimated Probability of O-Ring Failures over Temperature",
  xlab="Temperature", ylab="Estimated Probability"
```
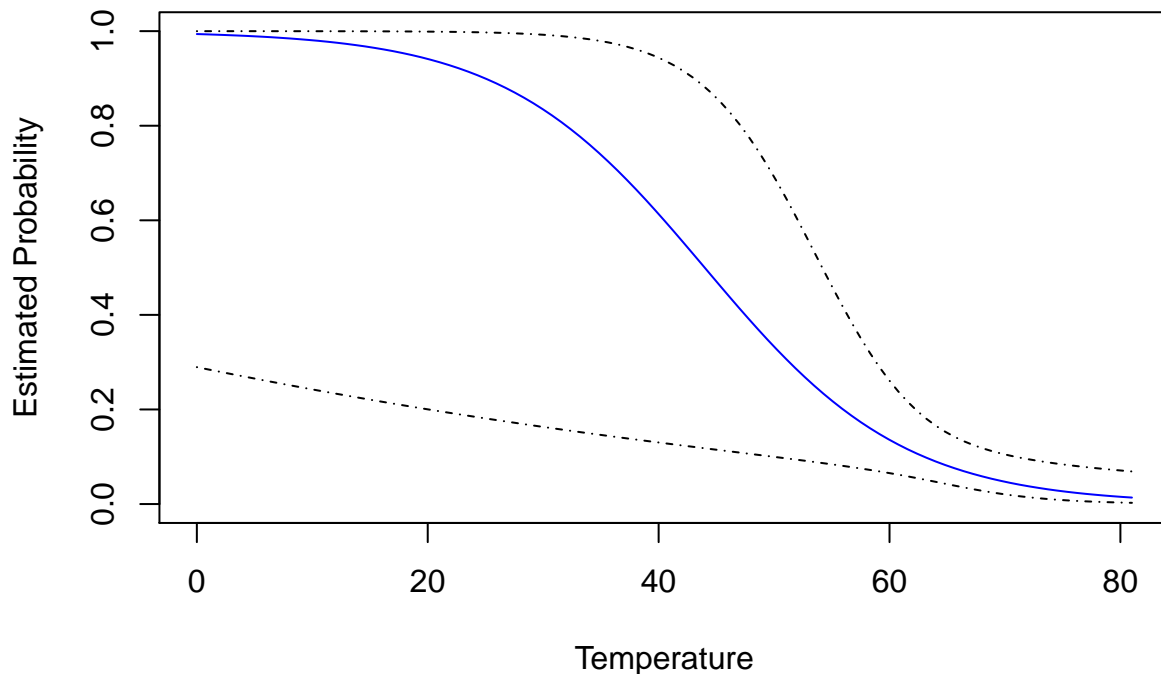
```
)
curve(
  pi.vs.temp,
  xlim=c(0, 81), ylim=c(0,1),
  col='blue', add = TRUE
)
curve(ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = mod.fit2, alpha=0.05)$lower, lty = "dotdash",
```

onfidence intervals for Estimated Probability of O–Ring Failures over T

The bounds are much wider at the lower range of the interval than the upper range because we have many fewer data points in this range. In the extreme lower range (near 31), we don't have any data points from our sample, so the confidence intervals are much larger.

5D: Prediction at 31 degrees and confidence interval

```
alpha = 0.05
predicted_obj = predict(object = mod.fit2, newdata = data.frame(Temp=31), type = "link", se= TRUE)
predicted_value = exp(predicted_obj$fit)/(1 + exp(predicted_obj$fit))

# Wald confidence for sanity checking
CI.lin.pred.lower = predicted_obj$fit - qnorm(p = 1-alpha/2) * predicted_obj$se
CI.lin.pred.upper = predicted_obj$fit + qnorm(p = 1-alpha/2) * predicted_obj$se
CI.lin.pred.pi.lower = exp(CI.lin.pred.lower) / (1+ exp(CI.lin.pred.lower))
CI.lin.pred.pi.upper = exp(CI.lin.pred.upper) / (1+ exp(CI.lin.pred.upper))

# Profile LR interval
contrast.matrix = matrix(data=c(1, 31), nrow=1, ncol=2)
CI.profile = confint(
  object = mcprofile(mod.fit2, CM=contrast.matrix),
```

```
    level = 0.95
)
CI.profile.pi.bounds = exp(CI.profile$confint)/(1 + exp(CI.profile$confint))

data.frame(
  Lower = c(CI.lin.pred.pi.lower, CI.profile.pi.bounds$lower),
  Prediction = c(predicted_value, predicted_value),
  Upper = c(CI.lin.pred.pi.upper, CI.profile.pi.bounds$upper),
  row.names = c('Wald', 'Profile LT')
)
```

```
##               Lower Prediction     Upper
## Wald       0.1596025  0.8177744 0.9906582
## Profile LT 0.1418508  0.8177744 0.9905217
```

From the above, it appears the predicted probability of an O-ring failure is roughly 81.7% with a 95% profile LR confidence interval of roughly 14.1% and 99.1%. We can compare the profile LR results above with the Wald results and conclude the profile LR intervals are valid since they are similar in value. The Wald test assumes the variance in the fitted values follows a normal distribution, which by way of the central limit theorem could be true for large enough samples. However, since this sample is very small with $n = 23$, normality of variance is a strong assumption to make, and so the profile LR confidence intervals are better estimates.

5E: Parametric bootstrap

The paper used a parametric bootstrap method to calculate 90% confidence intervals for various temperatures, so we are replicating it here. We are following the prescribed steps: 1 - simulate a large number of data sets, each $n = 23$ to match our sample in size. Here we created 500 such samples 2 - estimate new models for each data set 3 - compute $\hat{\pi}$ using each of those models at the temperatures of interest - 31 degrees and 72 degrees

```
####### parametric bootstrap ##########

#Original estimated coefficients for model with temperature only from actual flight data.
b0 = mod.fit2$coefficients[["(Intercept)"]]
b1 = mod.fit2$coefficients[["Temp"]]
niter = 1000 #number of simulations

#dataframe to store estimated probabilities from simulated runs
results_sim <- data.frame(iteration = c(1:niter), '31' = 0, '72' = 0)

set.seed(8529)
for (i in c(1:niter)) {
  #Generate random temperatures uniformly with min and max same as in data here. Number of simulated sa
  x1 <- runif(n = 23, min = 53, max = 81)

  #Estimate probability of failure from model based on actual data for each x1.
  pi <- exp(b0 + b1*x1) / (1+ exp( b0 + b1*x1))

  #Estimate corresponding y with some random noise.
  y <- rbinom(n = length(x1), size = 6, prob = pi) / 6

  #fit new model on the simulated data.
  tempdf <- data.frame(y, x1, pi, Number = 6)
  newmod.fit <- glm(formula = y ~ x1, family = binomial(link = logit), weights = Number, data = tempdf)
```

```
  beta.hat0 <- newmod.fit$coefficients[1]
  beta.hat1 <- newmod.fit$coefficients[2]

  #probability of failure with new model at temperature = 31
  pi.hat31 = exp(beta.hat0 + beta.hat1*31)/(1+exp(beta.hat0 + beta.hat1*31))
  results_sim[i, 2]<-pi.hat31

  #probability of failure with new model at temperature = 72
  pi.hat72 = exp(beta.hat0 + beta.hat1*72)/(1+exp(beta.hat0 + beta.hat1*72))
  results_sim[i, 3]<-pi.hat72
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
cat('\n90% confidence interval at temperate = 31 degree\n')
```

```
##
## 90% confidence interval at temperate = 31 degree
```

```
quant31<-quantile(results_sim[,2], probs = c(0.05, 0.5, 0.95), na.rm = TRUE)
formatC(quant31, digits=2, format="f")
```

```
##     5%    50%    95%
## "0.30" "0.83" "0.99"
```

```
cat('\n90% confidence interval at temperate = 72 degree\n')
```

```
##
## 90% confidence interval at temperate = 72 degree
```

```
quant72<-quantile(results_sim[,3], probs = c(0.05, 0.5, 0.95), na.rm = TRUE)
formatC(quant72, digits=2, format="f")
```

```
##     5%    50%    95%
## "0.01" "0.04" "0.07"
```

5F: Exploring a quadratic term

Here we determine whether a quadratic term offers a better fit with our sample. For this, we perform a likelihood ratio test between two models, one with the quadratic term, and the original one.

```
mod.fit.Ho = mod.fit2
mod.fit.Ha = glm(formula = ratio ~ Temp + I(Temp^2) , family=binomial(link=logit), weights=Number, data=
summary(mod.fit.Ha)
```

```
##
## Call:
## glm(formula = ratio ~ Temp + I(Temp^2), family = binomial(link = logit),
##     data = challenger, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84320  -0.72385  -0.61980  -0.01335   2.52101
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) 22.126148  23.794426   0.930    0.352
## Temp         -0.650885   0.740756  -0.879    0.380
## I(Temp^2)     0.004141   0.005692   0.727    0.467
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 17.592  on 20  degrees of freedom
## AIC: 37.152
##
## Number of Fisher Scoring iterations: 5
```

```
anova(mod.fit.Ho, mod.fit.Ha, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: ratio ~ Temp
## Model 2: ratio ~ 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
```

There doesn't seem to be incremental value in adding Temp^2 to the model equation. The Nova test gives a pretty high p value, implying that there isn't enough evidence to support the rejection of the null hypothesis that the co-efficient of the $Temp^2$ term is 0.

Interpretation and Model choice

Questiond 3a. Interpretation:

From our final model, we estimate that for each 1 degree Fahrenheit decrease in temperature, the odds of an O-ring failure increase by $e^{0.11560} \approx 1.1225$. The odds of an O-ring failure at any one temperature can be estimated as:
$$\hat{\beta}_0 + \hat{\beta}_1 Temp = 5.08498 - 0.11560(Temp)$$

As a result, the probability of an O-ring failure can be estimated as:

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 Temp}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 Temp}} = \frac{e^{5.08498 - 0.11560(Temp)}}{1 + e^{5.08498 - 0.11560(Temp)}}$$

We estimate at temperature 31 degrees Fahrenheit, there is a 0.85 probability of O-ring failure with a 90% confidence interval between 0.30 and 0.99. In terms of odds, that's a 5.667 to 1 odds of O-ring failure with a 90% confidence interval between 0.43 to 1 and 99 to 1 odds.

Similarly, we estimate at temperature 72 degrees Fahrenheit, there is a 0.03 probability of O-ring failure with a 90% confidence interval between 0.01 and 0.07 of an O-ring failure. In terms of odds, that's a 32 to 1 odds of O-ring staying intact(0.031 to 1 odds of O-ring failure) with a 90% confidence interval between 99 to 1(0.01 to 1 odds of O-ring failure) and 13 to 1 odds of the O-rings staying intact(0.075 to 1 odds of O-ring failure)

### Questions 3b. Comparison with linear model

For the sake of exploration, we look at a linear regression model that fits `ratio` on `Temp` (since it is our Final model). First we fit a model, and assess it's validity.

```
mod.linear = lm(ratio ~ Temp, weights=Number, data=challenger)
summary(mod.linear)
```

```
##
## Call:
## lm(formula = ratio ~ Temp, data = challenger, weights = Number)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22894 -0.16102 -0.03486  0.04311  0.76223
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.616402   0.203252   3.033  0.00633 **
## Temp        -0.007923   0.002907  -2.725  0.01268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2357 on 21 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.2261
## F-statistic: 7.426 on 1 and 21 DF,  p-value: 0.01268
```

Once again we notice there is a 95% significance in the linear relationship between `Temp` and `ratio`. We can diagnose the linear model using `plot`:

```
par(mfrow=c(2, 2))
plot(mod.linear)
```
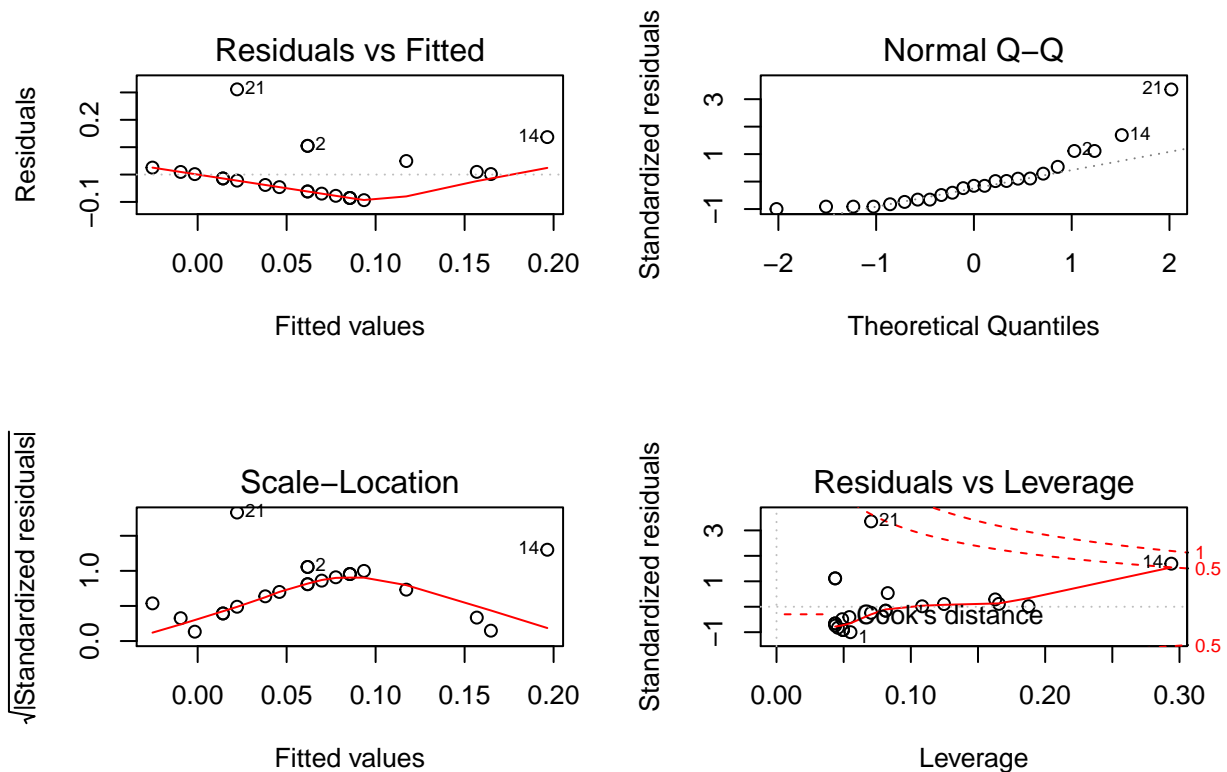

From this we can see the assumptions of a BLUE linear model: 1 - linearity: the QQ plot does not show a linear fit, thus linearity is not evidenced in the sample 2 - variation in independent variables: from the

EDA above, we know the temperature variable has variance (albeit a small one) 3 - no perfect co-linearity: N/A since we are only assessing one independent variable 4 - zero conditional mean: the Residuals vs Fitted line shows this assumption is violated severely 5 - homoskedasticity: this is clearly violated, as can be seen in the residuals vs fitted line. the variance in errors seems to decrease sharply then increase again as fitted values increase. 6 - normality of error term: it appears as though the residuals are clustered around 0 in the Residuals vs. Fitted lines plot. There is also no even spread between.

These diagnostics suggest that a linear model is a poor fit for this sample data since we are violating the majority of the assumptions necessary for a linear model to work: all of linearity, zero-conditional mean, homoskedasticity and normality of error term are violated. Additionally, plotting the linear model also shows that the probability of an o-ring failure goes below 0 at roughly 78 degrees of temperature. This is impractical and doesn't represent the data well.

```
plot(
  challenger$Temp, challenger$ratio,
  main="Linear model of Proportion of O-ring failures against Temperature",
  xlab="Temperature", ylab="Proportion of O-ring failures"
)
abline(mod.linear)
```


Linear model of Proportion of O–ring failures against Temperature
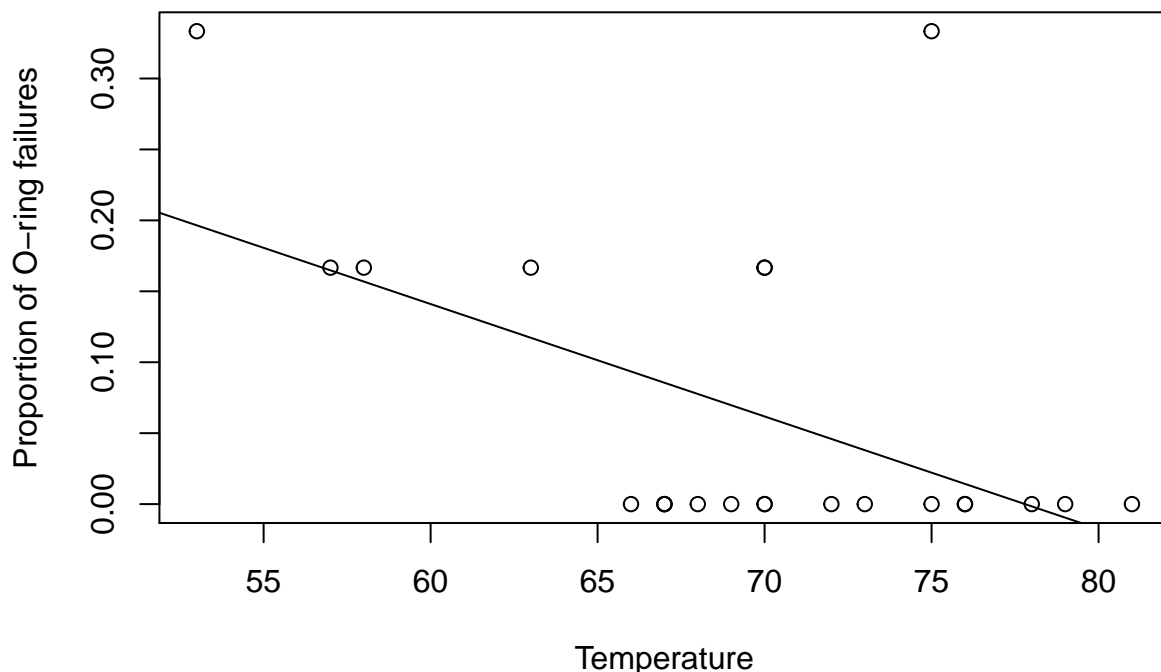
Thus, a logistic regression model is still the better model even though the linear model shows a good fit.

Conclusion

Using a logit model, we tried to fit likelihood of O-ring failure against temperature and pressure. We concluded that pressure does not offer enough predictive value and so we chose our final model to be a regression on temperature alone. We predicted that at 31 degrees and using a bootstrap confidence intervals to estimate that the challenger shuttle would have experienced one O-ring failure with 85% chance, and a 90% confidence interval of between 27% and 99% chance.