W271 Group Lab 1

Hardy,M., Solanki, H., Wang, P.

Investigation of the 1989 Space Shuttle Challenger Accident

Part 1 (25 points) In this analysis, we are investigating the ability to understand space shuttle O-ring performance, which is the suspected fail-point of the Challenger disaster. We are to do this based on the limited number of trial observations, and compare our results to Dalal et al (1989), who performed a similar analysis. We begin by performing exploratory data analysis to better understand the dataset.

```r
library(car)
library(plyr)
library(ggplot2)
library(dplyr)
library(skimr)
library(knitr)
library(stargazer)
library(psych) ##This library is added for convenient visualization during EDA
library(gridExtra) ## This library is added for paneling graphs to save space
mod_stargazer <- function(...){
    # This function removes the first three lines of stargazer output when converting to Latex
  output <- capture.output(stargazer(...))
  output <- output[4:length(output)]
  cat(paste(output, collapse = "\n"), "\n")}
challenger.data <- read.csv("challenger.csv", header=TRUE, sep=",")
skim_without_charts(challenger.data,)
```

Table 1: Data summary

Name	challenger.data
Number of rows	23
Number of columns	5
Column type frequency:	
numeric	5
Group variables	None

Variable type: numeric

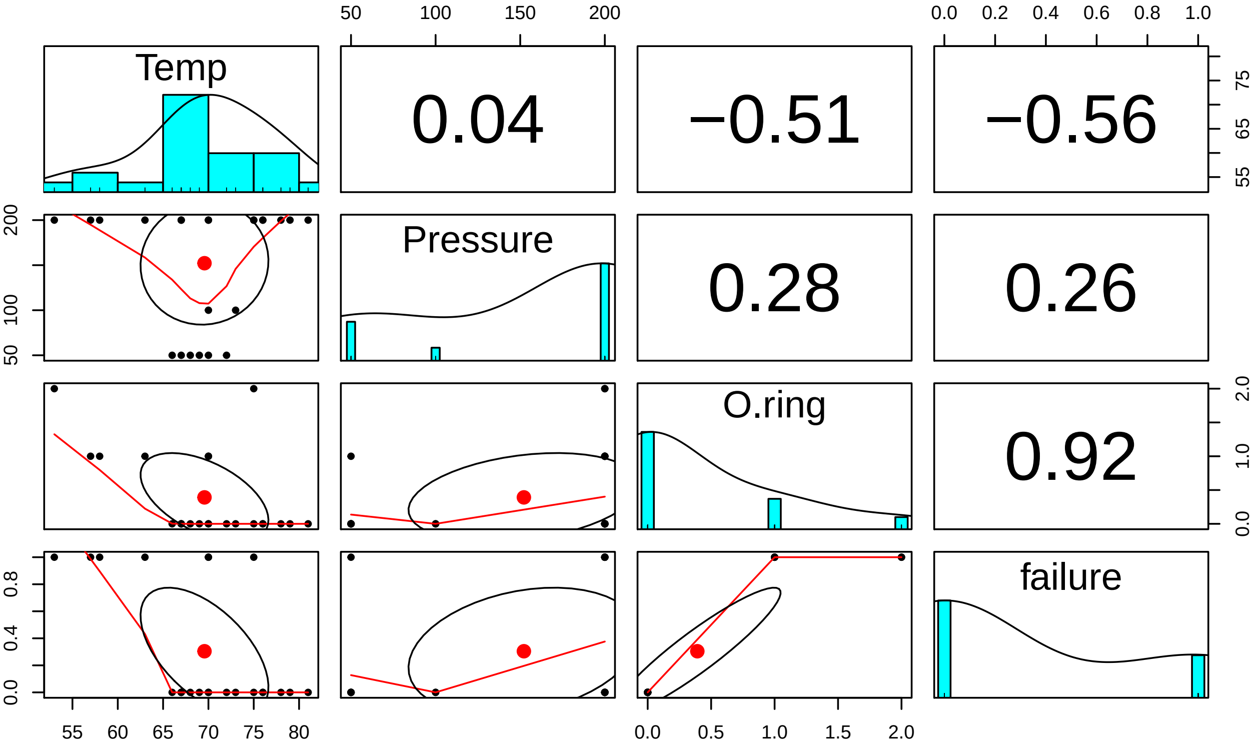
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
Flight	0	1	12.00	6.78	1	6.5	12	17.5
Temp	0	1	69.57	7.06	53	67.0	70	75.0
Pressure	0	1	152.17	68.22	50	75.0	200	200.0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	
O.ring	0	1	0.39	0.66	0	0.0	0	1.0	
Number	0	1	6.00	0.00	6	6.0	6	6.0	
There are 5 dist	inct variabl	es that describe	the cond	itions o	f O-r	ing tes	ts: $f	light$,	*temp*, *pr*

Variable Relationships We use `pairs.panels` to quickly model the relationship between all of the variables. As expected, we see a strong Pearson correlation between $O.ring$ and our new variable, $failure$, since the latter is derivative of the former. Correlations give us indications that $Pressure$ and $Temp$ are not correlated, and it appears that $Temp$ is more correlated with both $O - ring$ and $failure$ than $Pressure$. However, the trimodal distribution of $Pressure$ suggested that we may also need to explore it as a `factor` variable, because it only has 3 unique values.

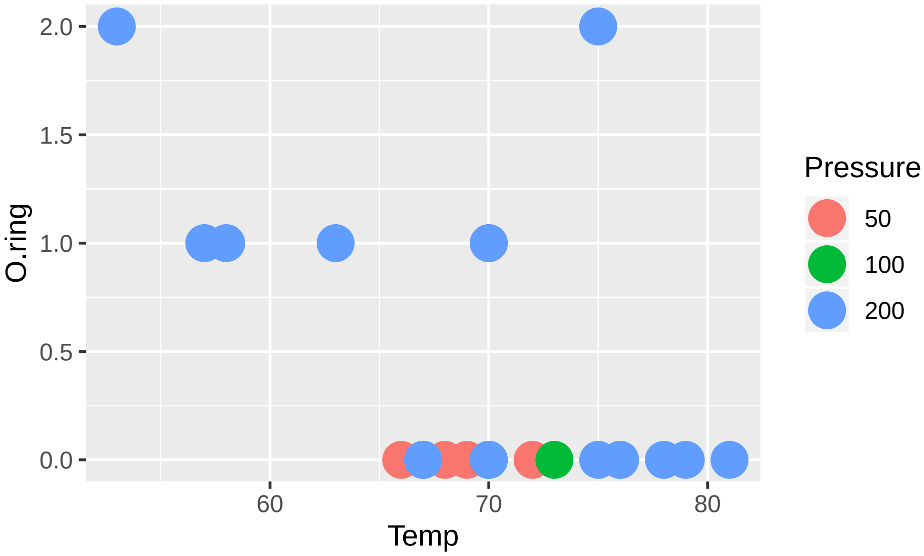
Looking at the `LOESS` lines across the variables, we can see how the average relationship changes, revealing the unusual distributions and lack of data.

```r
# Create a new column if the flight failed
challenger.data$failure <- ifelse(challenger.data$O.ring > 0, 1, 0)
pairs.panels(challenger.data[,c(2,3,4,6)])
```



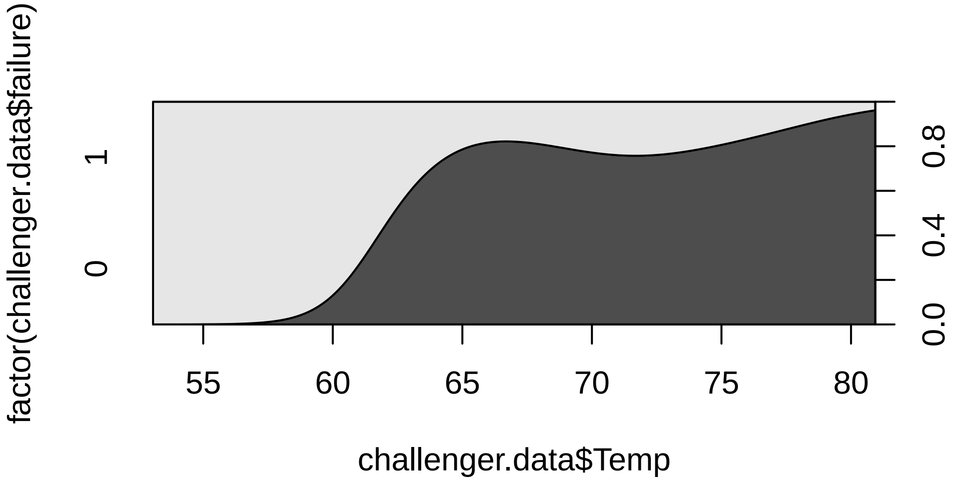
Looking at the Pressure variable as a factor, we see that only pressures of 200 are associated with failure. We may need to look at $Pressure$ as a binary categorical variable of >100 psi or not.

```r
# Looking closer at the relationship between Temperature and Pressure and O-ring failures
data.p.factor <- challenger.data
data.p.factor$Pressure <- as.factor(data.p.factor$Pressure)
data.p.factor$PressBin <- as.factor(ifelse(challenger.data$Pressure > 100, 1, 0))
ggplot(data.p.factor, aes(x = Temp, y = O.ring, color = Pressure)) +geom_point(size=6)
```



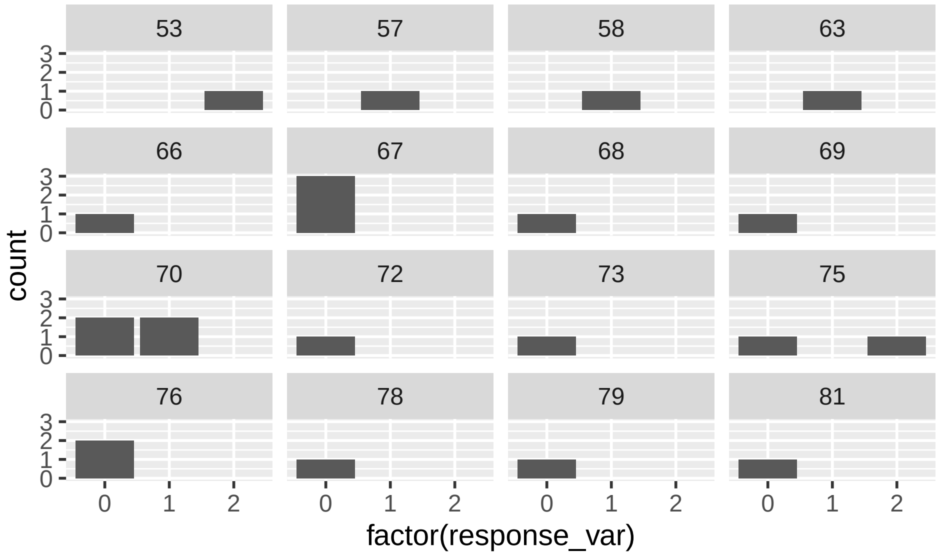
In terms of values, the mean temperature was 69.6(7.06) degrees F, and the mean pressure was 152(68.2) psi. The average amount of O rings that failed were 0.391, but the 0.656 SD along with the lower bound of 0 implies a heavy amount of variation within failure. All the tests had 6 rings total, in the 23 flights that were taken.

```
cdplot(factor(challenger.data$failure) ~ challenger.data$Temp)
```



In this CD plot, it is clear that as temperature increases, probability of success goes up too. However, it is interesting to look at the density function.

```
response_var <- challenger.data$O.ring
ggplot(challenger.data, aes(x=factor(response_var))) + geom_bar() + facet_wrap(~Temp)
```



Here, we can see the failure at every temperature. This allows us to visualize the results.

Part 2 (20 points)

Answer 2(a): In order to fit the bionomial model to the dataset (Y=1 when an incident occur and Y=0 when none occurs), the authors assume that at a specific temperautre and pressure, each of the six O-rings would suffer damage independently with the same probability. However, this is a faulty assumption; if the primary O-ring should fail due to erosion or blowby, the hot gases could escape through the gap between the tang and lead to a failure of a secondary O-ring. Therefore, we can not assumpe independency among each of the O-ring variable. Nevertheless, there are certain desirable properties by maintaining the assumption of independence among the dependent variables: the increas in information gain may be useful as we seek to extrapolate the data far outside of the range of the data set.

Answer 2(b): Below are our two logistic regression models that we will be using to evaluate O-ring failure: 1) a binary regression that does not assume independence of O-ring failure by launch and 2) a binomial count model that makes the faulty assumption of independence, but provides more informatino. Both will be useful for different purposes in our analysis.

In our first models, we include the *Temp* and *Pressure* as explanatory variables. We remove number as an explanatory variable from the binary model, but use it as the weightings within the binomial model, as it represents the number of O-ring installed on a shuttle flight. (Since all shuttle flights have 6 O-rings as the default design, this variable bares no explanatory impact on our model.)

For the binomial logistic regression, we model a probability of an individual O-ring failing (by dividing by 6, and then using the *Number* variable to act as the count).

In both models, we find that *Temp* is significant at $\alpha = 0.05$. We can't compare the deviances of the two models, since they are neither nested nor do they share the same dependent variable.

```
## Using Pressure as continuous
mod.binary.HA <-glm(formula=failure ~ Pressure + Temp, family = binomial (link=logit), data=cha
mod.binom.HA <- glm(formula=O.ring/6 ~ Pressure + Temp, family = binomial (link=logit), data=cl
## Including Pressure as a binary factor
```

```
mod.binary.HA.p <-glm(formula=failure ~ PressBin + Temp, family = binomial (link=logit), data=c
mod.binom.HA.p <- glm(formula=O.ring/6 ~ PressBin + Temp, family = binomial (link=logit), data=
mod_stargazer(mod.binary.HA, mod.binary.HA.p, mod.binom.HA,mod.binom.HA.p, title = 'Checking th
```

Table 3: Checking the Effect of Pressure as Binary

	Dependent variable:			
	failure		O.ring/6	
	(1)	(2)	(3)	(4)
Pressure	0.010		0.008	
	(0.009)		(0.008)	
PressBin1		1.588		1.293
		(1.294)		(1.104)
Temp	-0.229^{**}	-0.227^{**}	-0.098^{**}	-0.097^{**}
	(0.110)	(0.109)	(0.045)	(0.045)
Constant	13.292^{*}	13.678^{*}	2.520	2.842
	(7.664)	(7.583)	(3.487)	(3.241)
Observations	23	23	23	23
Log Likelihood	−9.391	−9.287	−15.053	−14.946
Akaike Inf. Crit.	24.782	24.575	36.106	35.892
Note:		*p<0.1; **p<0.05; ***p<0.01		

Answer 2(c): From the Anova (Type II) tests on both models, *Temp* is statistically signficant at $\alpha = 0.05$, but *Pressure* is not. Confirming through a second `anova()` Chi-squared test on each of our two models, we do not have sufficient evidence to reject the null hypothesis for *Pressure*.

```
mod.binary.full <-glm(formula=failure ~ Pressure*Temp, family = binomial (link=logit), data=cha

mod.binary.Ha <-glm(formula=failure ~ Temp, family = binomial (link=logit), data=challenger.da
Anova(mod.binary.HA)
```

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

```
anova(mod.binary.Ha, mod.binary.HA, test="Chisq")
```

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

```
mod.binom.H0 <-glm(formula=O.ring/6 ~ Temp, family = binomial (link=logit), data=challenger.da
Anova(mod.binom.HA)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/6
##          LR Chisq Df Pr(>Chisq)
## Pressure   1.5407  1     0.2145
## Temp       5.1838  1     0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.binom.H0, mod.binom.HA, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/6 ~ Temp
## Model 2: O.ring/6 ~ Pressure + Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     18.086
## 2        20     16.546  1   1.5407   0.2145
```

Answer 2(d): The authors test the models with and without the pressure variable, removing it based on the LRTs. The incremental G^2 (18.086-16.546=1.54) is not significant but shows that there may be a very weak pressure effect. We test a full model for *Pressure* where we include its interactions with *Temp* and a quadratic term, and then perform both `Anova()` type II tests to verify each separate contributing factor, in addition to testing it against our null hypotheses from above using `anova()`. We also check for parsimony by comparing the deviances via `AIC()`, and find that just *Temp* is the most parsimonious. However, when we use pressure as a binary categorical variable, it has less overall residual deviance than when we use it as continuous. Thus, if we were to include it, it would be categorical.

The authors then create 90% bootstrap CI for the expected number of incidents holding pressue at 50 psi and 200 psi. The intervals overlap greatly which indicate that a pressure effect may be present but it can not be estimated with enough precision to include it in the model. The authors then decide to drop the pressure variable since erosion and blowby (two methods of O-ring failures) were examined separately. Removing pressure as a variable presents the model limitation as the authors later investigate and discover that that primary nozzle O-rings at 100 and 200 psi are far more prone to heat damage than the corresponding field O-rings.

Since we are dealing with such a small sample with a very unusual distribution, removing these variables may also contribute to more intense violations of the assumptions of GLM.

```
## Pressure as continuous and quadratic
mod.binary.full <-glm(formula=failure ~ Pressure*Temp + I(Pressure^2), family = binomial (link
Anova(mod.binary.full)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: failure
##                LR Chisq Df Pr(>Chisq)
## Pressure         0.4740  1   0.491148
## Temp             7.3827  1   0.006585 **
## I(Pressure^2)    1.1993  1   0.273468
## Pressure:Temp    1.3384  1   0.247312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.binary.full, mod.binary.Ha, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: failure ~ Pressure * Temp + I(Pressure^2)
## Model 2: failure ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        18     16.876
## 2        21     20.315 -3  -3.4393   0.3287
```

```
mod.binom.full <- glm(formula=O.ring/6 ~ Pressure*Temp+  I(Pressure^2), family = binomial (lin
Anova(mod.binom.full)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/6
##                LR Chisq Df Pr(>Chisq)
## Pressure         0.5747  1    0.44839
## Temp             4.9465  1    0.02614 *
## I(Pressure^2)    1.1329  1    0.28715
## Pressure:Temp    0.7512  1    0.38611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.binom.full, mod.binom.H0, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/6 ~ Pressure * Temp + I(Pressure^2)
## Model 2: O.ring/6 ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        18     15.124
## 2        21     18.086 -3  -2.9623   0.3975
```

```
data.frame('Continuous Model' = c('Temp only', 'Temp + Press', 'Temp + full press'), 'Binomial
```

```
##    Continuous.Model Binomial.AIC Binary.AIC
## 1         Temp only     35.64654   24.31519
## 2      Temp + Press     36.10589   24.78209
## 3 Temp + full press     38.68422   26.87594
## Pressure as categorical
mod.binary.full.p <-glm(formula=failure ~ PressBin*Temp , family = binomial (link=logit), data
Anova(mod.binary.full.p)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: failure
##                LR Chisq Df Pr(>Chisq)
## PressBin         1.7403  1   0.187097
## Temp             7.6438  1   0.005697 **
## PressBin:Temp    0.6201  1   0.431025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.binary.full.p, mod.binary.Ha, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: failure ~ PressBin * Temp
## Model 2: failure ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        19     17.955
## 2        21     20.315 -2  -2.3604   0.3072
```

```
mod.binom.full.p <- glm(formula=O.ring/6 ~ PressBin*Temp, family = binomial (link=logit), data
Anova(mod.binom.full.p)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/6
##                LR Chisq Df Pr(>Chisq)
## PressBin         1.7546  1    0.18530
## Temp             5.0724  1    0.02431 *
## PressBin:Temp    0.2481  1    0.61839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.binom.full.p, mod.binom.H0, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/6 ~ PressBin * Temp
## Model 2: O.ring/6 ~ Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        19     16.084
## 2        21     18.086 -2  -2.0027   0.3674
```

```
data.frame('Factor Model' = c('Temp only', 'Temp + Press', 'Temp + full press'), 'Binomial AIC
```

```
##         Factor.Model Binomial.AIC Binary.AIC
## 1         Temp only     35.64654    24.31519
## 2      Temp + Press     35.89194    24.57487
## 3 Temp + full press     37.64380    25.95481
```

Part 3

Answer 3(a) We estimated the simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$, where π is the probability of an O-ring failure, both for the binary and binomial response models using the simplified model from above.

```
mod.binary.Ha <-glm(formula=failure ~ Temp, family = binomial (link=logit), data=challenger.da
mod.binom.H0 <-glm(formula=O.ring/6 ~ Temp, family = binomial (link=logit), data=challenger.da
mod_stargazer(mod.binary.Ha, mod.binom.H0,  title = "Binary and Binomial models", no.space = T
```

Table 4: Binary and Binomial models

	Dependent variable:	
	failure	O.ring/6
	(1)	(2)
Temp	-0.232^{**}	-0.116^{**}
	(0.108)	(0.047)
Constant	15.043^{**}	5.085^{*}
	(7.379)	(3.052)
Observations	23	23
Log Likelihood	−10.158	−15.823
Akaike Inf. Crit.	24.315	35.647
Note:	*p<0.1; **p<0.05; ***p<0.01	

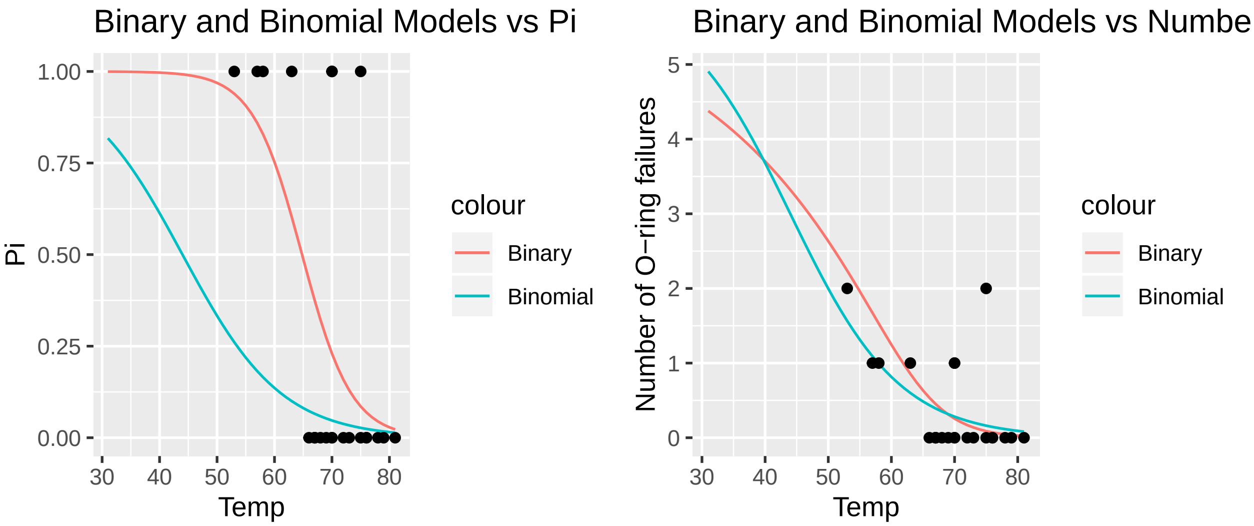
Answer 3(b) We constructed two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp from 31 degrees to 81 degrees, and tested both our binary and binomial models.

For π vs. Temp, our binary model is a better fit. In the case where we need to model the expected number of failures vs. $Temp$, we realize that our binomial model has sufficient information, because it accounts for the percentage of $O.ring$ failures on each flight.

However, if we assume $Y = 0$ iff $X = 0$, then $p*$ of the binary can be compared to the binomial using $p* = 1 - (1 - p(t))^6$. If we utilize this relationship, we see that the two models are comparably well fit. More tests will be needed. Note: that relationship between $p*(t)$ and our estimated $p(t)$ from the binomial model makes the assumption that all of the failure rates are the same probability as the group failure rate.

Both models are found on each plot below.

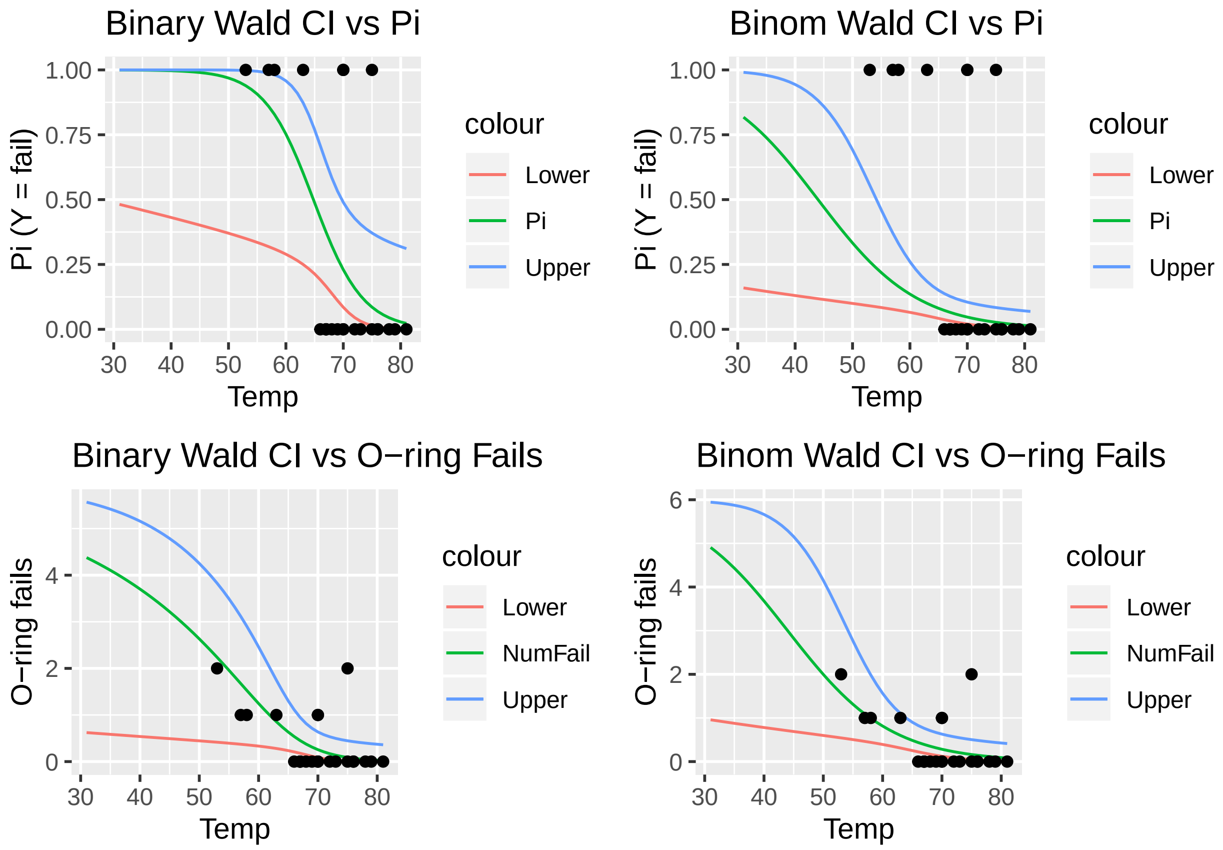
```
xTemp <- seq(31, 81, 1)

pi.hat <- data.frame(Temp = xTemp, Binary = predict(mod.binary.Ha, list(Temp = xTemp), type = 
pi.chart <- ggplot(pi.hat, aes(x=Temp)) +
  geom_line(aes(y=Binary, color='Binary')) +
  geom_line(aes(y=Binomial, color='Binomial')) +
  geom_point(mapping = aes(x = Temp, y = failure), data = challenger.data) + ylab("Pi") +
  ggtitle(paste('Binary and Binomial Models vs Pi'))

num.fail <- data.frame(Temp = xTemp, Binary =1-(1 - predict(mod.binary.Ha, list(Temp = xTemp),
n.chart <- ggplot(num.fail, aes(x=Temp)) +
  geom_line(aes(y=Binary*6, color='Binary')) +
  geom_line(aes(y=Binomial*6, color='Binomial')) +
  geom_point(mapping = aes(x = Temp, y = O.ring), data = challenger.data) + ylab("Number of O-
  ggtitle(paste('Binary and Binomial Models vs Number of O-ring Failures'))

grid.arrange(pi.chart, n.chart,ncol = 2)
```



Answer 3(c) We included the 95% Wald confidence interval bands for π on the plot and the count plot, for both binary and binomial models. By doing so we observe that for the π plot, the binary model and confidence intervals perform better than the binomial model. However, the values are comparable acorss the number of failures models.

The lack of data and the need to extrapolate far outside of the dataset create uncertainty, because the variance at the lower temperatures is much higher due lack of data. There is much more data in the space of higher temperatures.

```
predicted_lm <- predict(object = mod.binary.Ha, newdata = list(Temp = xTemp), type = "link", s

fit <- predicted_lm$fit
se <- predicted_lm$se

pi.hat <- data.frame(Temp = xTemp,
                     Pi = predict(mod.binary.Ha, list(Temp = xTemp), type = "response"),
                     Upper = exp(fit + qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit + qnorm(p =
                     Lower = exp(fit - qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit - qnorm(p =
by.pi <- ggplot(pi.hat, aes(x=Temp)) + geom_line(aes(y=Lower, color='Lower')) + geom_line(aes(
num.fail <- data.frame(Temp = xTemp,
                       NumFail = 1-(1-predict(mod.binary.Ha, list(Temp = xTemp), type = "respon
                       Upper = 1 - (1 - exp(fit + qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit +
                       Lower =1 - (1- exp(fit - qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit - q
by.fail <-ggplot(num.fail, aes(x=Temp)) + geom_line(aes(y=Lower*6, color='Lower')) +  geom_line

## Here are the same plots, but for the total number of failures, using the Binomial model
predicted_lm2 <- predict(object = mod.binom.H0, newdata = list(Temp = xTemp), type = "link", se

fit <- predicted_lm2$fit
se <- predicted_lm2$se

pi.hat <- data.frame(Temp = xTemp,
                     Pi = predict(mod.binom.H0, list(Temp = xTemp), type = "response"),
                     Upper = exp(fit + qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit + qnorm(p = 1
                     Lower = exp(fit - qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit - qnorm(p = 1
bo.pi <- ggplot(pi.hat, aes(x=Temp)) + geom_line(aes(y=Lower, color='Lower')) + geom_line(aes(y

num.fail <- data.frame(Temp = xTemp,
                       NumFail = predict(mod.binom.H0, list(Temp = xTemp), type = "response"),
                       Upper = exp(fit + qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit + qnorm(p =
                       Lower = exp(fit - qnorm(p = 1 - 0.05/2) * se) / (1 + exp(fit - qnorm(p =
bo.fail <- ggplot(num.fail, aes(x=Temp)) + geom_line(aes(y=Lower*6, color='Lower')) + geom_line

grid.arrange(by.pi, bo.pi, by.fail, bo.fail, ncol = 2, nrow = 2)
```



The variance at the lower temperature is much higher due to lack of data. There is much more data in the space of higher temperatures.

(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.

```
alpha = 0.05
predict.data <- data.frame(Temp = 31)
linear.pred <- predict(object = mod.binary.Ha, newdata = predict.data, type = "link", se  = TRU
pi.hat <- exp(linear.pred$fit) / (1+exp(linear.pred$fit))

CI.lin.pred <- linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2)) * linear.pred$se
CI.pi <- exp(CI.lin.pred)/(1+exp(CI.lin.pred))
# data.frame(predict.data, pi.hat, lower = CI.pi[1], upper= CI.pi[2])

linear.pred2 <- predict(object = mod.binom.H0, newdata = predict.data, type = "link", se  = TRU
pi.hat2 <- exp(linear.pred2$fit) / (1+exp(linear.pred2$fit))

CI.lin.pred2 <- linear.pred2$fit + qnorm(p = c(alpha/2, 1-alpha/2)) * linear.pred2$se
CI.pi2 <- exp(CI.lin.pred2)/(1+exp(CI.lin.pred2))
# data.frame(predict.data, pi.hat2, lower = CI.pi2[1], upper= CI.pi2[2])
CI.preds <- data.frame(Model = c('Binary','Binomial'), Lower = c(CI.pi[1],CI.pi2[1]),Pi.Hat = c
mod_stargazer(CI.preds,  summary = F, title = "Wald CI at 31 degrees", digits = 5) #type = "te
```

Table 5: Wald CI at 31 degrees

	Model	Lower	Pi.Hat	Upper
1	Binary	0.48161	0.99961	1.00000
2	Binomial	0.15960	0.81777	0.99066

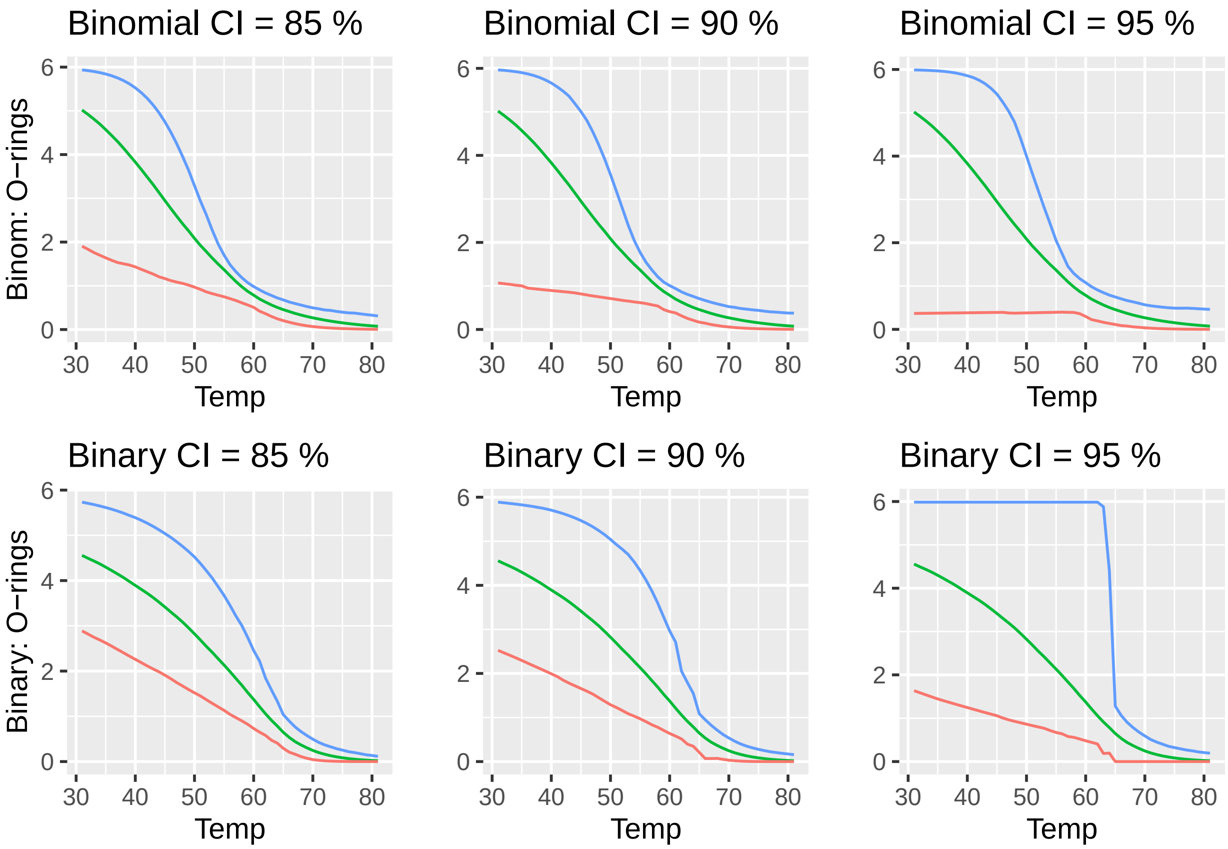
The probability of failure is 0.9996088 with a 95% Wald CI of [0.4816106, 0.9999999]. For our binary logisitic regression, our assumptions include the dependent variable to be binary, the observations to be indepdendent, and linearity of the indepedent variables. Lack of multicollinearity is requied as well, but in this case, we only have one independent variable. The binomial logistic regression requires that we assume each failure is independent, which contributes to the very low confidence lower-bound: without data, there is more opportunity for uncertainty across six different models.

Answer 3 (e): Parametric Boostrapping using 'percentile' intervals Below we have modeled the binomial logistic for three different levels of α using "percentile" interval parametric boostrapping 1,000 times, following the procedure outlined in Dalal et al. The data.frame has the intervals using this methodology for 31 and 72 degrees.

Additionally, using the property of $p* = 1 - (1 - p(t))^6$, we also estimated the binary model below.

```
n = 1000
temp.range <- seq(31, 81, 1)
b.range <- seq(31, 81, 1)
VAL <- data.frame(row.names = b.range)
VAL.binary <- data.frame(row.names = b.range)
wgts <- rep.int(6,23)
## Bootstrapping
for (i in seq(n)){
  ## Resampling
  samp.t <- challenger.data[sample(nrow(challenger.data),replace = T,size = 23),c('Temp','O.ri
  ## Binomial
  binom.mod <- glm(O.ring/6 ~ Temp, family = binomial (link=logit), weights = wgts, data = samp
  temp.binom <- predict(binom.mod, newdata = data.frame(Temp = temp.range),type = 'response')
  VAL <- data.frame(VAL, t = temp.binom)
  ## Binary
  binar.mod <- glm(failure ~ Temp, family = binomial (link=logit), data = samp.t, maxit = 100,
  temp.binar <- predict(binar.mod, newdata = data.frame(Temp = temp.range),type = 'response')
  VAL.binary <- data.frame(VAL.binary, t = temp.binar)
}
#Interval Confidence by Percentiles: Binomial
alpha = 0.05
new_data = t(apply(VAL, 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2)),  na.rm = TRUE))
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = ne
ci.95 <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sh

alpha = 0.1
new_data = t(apply(VAL, 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2)),  na.rm = TRUE))
```

```
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = new
ci.90 <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sho

## Here's the required confidence intervals at 90% for Binomial, saved to dataframe
CI.table = data.frame(Temperature = c('31','72'), Binom = bs.ci[c('31','72'),]*6)

alpha = 0.15
new_data = t(apply(VAL, 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2)),  na.rm = TRUE))
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = new
ci.85 <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sho

## Same thing, but for binary model
alpha = 0.05
new_data = t(apply((1-(1-VAL.binary)^(1/6)), 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2))
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = new
ci.95b <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sh

alpha = 0.1
new_data = t(apply((1-(1-VAL.binary)^(1/6)), 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2))
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = new
ci.90b <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sh
  geom_line(aes(y=lower*6, color='lower'), show.legend = F) + geom_line(aes(y=upper*6, color='u

## These are the other required confidence intervals at 90% for Binary
CI.table = data.frame(CI.table, Binar = bs.ci[c('31','72'),]*6)
alpha = 0.15
new_data = t(apply((1-(1-VAL.binary)^(1/6)), 1, quantile, probs = c(alpha/2,0.5,1 - (alpha/2))
bs.ci = data.frame(Temp = as.numeric(row.names(new_data)), lower = new_data[,1], num.fail = new

ci.85b <- ggplot(bs.ci, aes(x=Temp)) + geom_line(aes(y=num.fail*6, color='Mean.N.Failures'), sh
  geom_line(aes(y=lower*6, color='lower'), show.legend = F) + geom_line(aes(y=upper*6, color='u

grid.arrange(ci.85, ci.90, ci.95, ci.85b, ci.90b, ci.95b, ncol = 3, nrow = 2)
```



```r
## Display the charts and table
cat(paste(capture.output(stargazer(CI.table[,c(1,3,5,7,9)], summary = F, title = "Bootstrapped
```

Table 6: Bootstrapped 90 percent Confidence Intervals

	Temperature	Binom.lower	Binom.upper	Binar.lower	Binar.upper
31	31	1.069	5.964	2.524	5.887
72	72	0.035	0.482	0.010	0.400

Answer 3 (f) In order to determine if a quadratic term is needed in the model for the temperature, we test for the effect of temperature in both the binary and binomial models, compare the statistical significance (through a Chi squared `anova()` test), and then check for parsimony using deviance (through the `AIC` calculation). The addition of the variable is neither significant nor more parsimonious, so we do not include it, preserving our few degrees of freedom.

```r
# For the Binary Model
mod.binary.TT <-glm(formula= failure ~ Temp + I(Temp^2) , family = binomial (link=logit), data=
# summary(model.3.a.HA)
anova(mod.binary.Ha, mod.binary.TT, test = 'Chisq')
```

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

```
print(paste("Difference in Deviances (via AIC):", AIC(mod.binary.Ha)-AIC(mod.binary.TT)))
```

```
## [1] "Difference in Deviances (via AIC): -1.07350981601074"
```

```
# For the Binomial Model
mod.binom.TT <-glm(formula= O.ring/6 ~Temp +  I(Temp^2) , family = binomial (link=logit), data=
# summary(mod.binom.TT)
anova(mod.binom.H0,mod.binom.TT, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/6 ~ Temp
## Model 2: O.ring/6 ~ 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
```

```
print(paste("Difference in Deviances (via AIC):", AIC(mod.binom.H0)-AIC(mod.binom.TT)))
```

```
## [1] "Difference in Deviances (via AIC): -1.5052995921303"
```

Answer 4 To fit a linear regression model with the same set of explanatory variables in our final model, several CLM assumptions must be held true–some that aren't required in the GLM. The model has to be linear in parameters with random sampling of observations. The conditional mean should be zero and there should be no multi-collinearity. The model should have no autocorrelation, and the error terms should be normally distributed.

From the Residuals vs Fitted plot, it clears shows that the linear relationship assumptions is violated. There is a linear patten between the fitted values and the residuals. The Normal Q-Q plot shows that the residuals follow a straight line with several data points deviate from the pattern. The Scale-Location graph shows the variance of the residuals which indicates that heteroscedasticity exist in our dataset. The residuals vs Leverage plot shows that we have several extreme values that might influence the regression result.

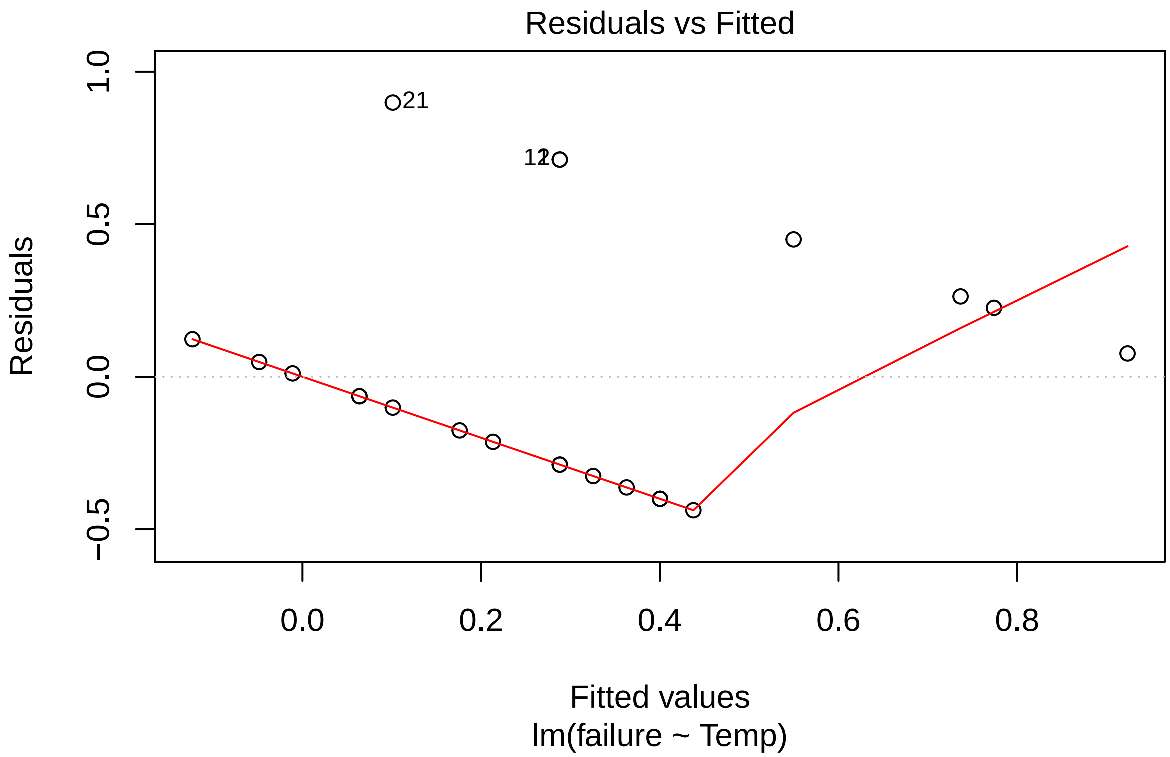
The plot results show that our linear regression model violated the CLM assumption and a binary logistic regression will be a more suitable model for our case.

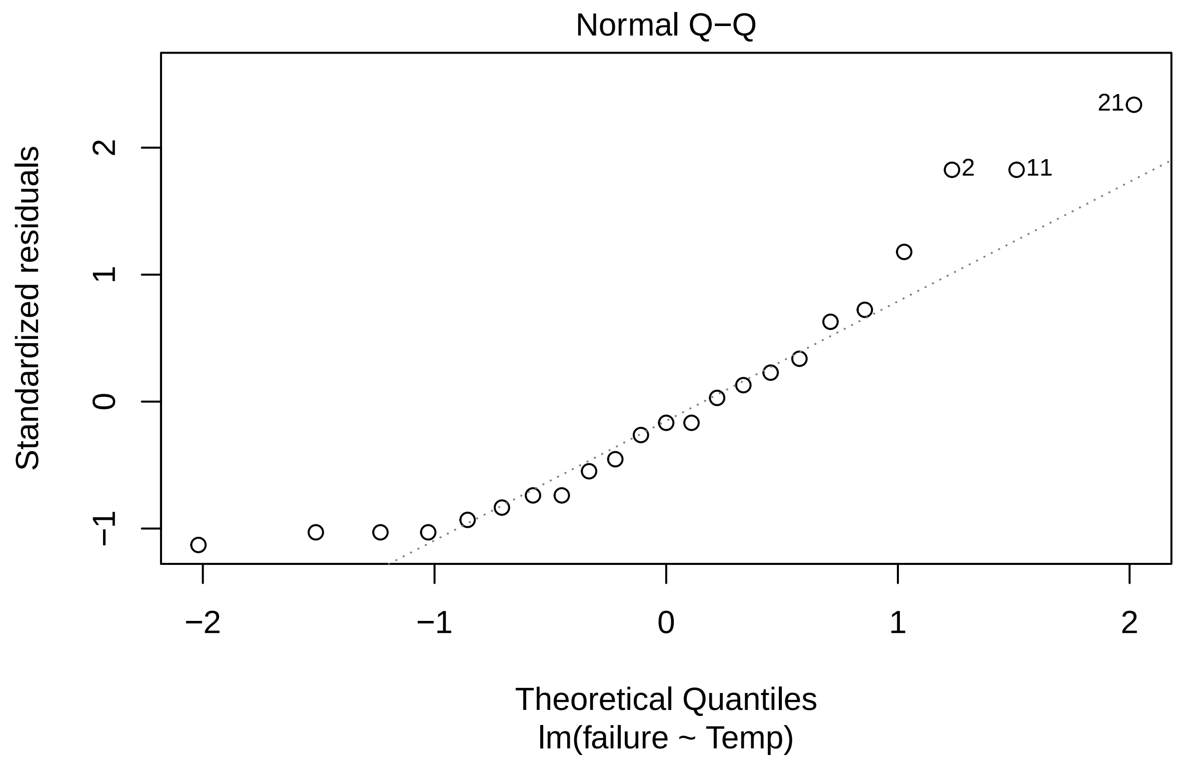
```
challenger.lm <- lm(failure~ Temp, data = challenger.data)
mod_stargazer(challenger.lm, title = 'Modeling using OLS', no.space = T)
```

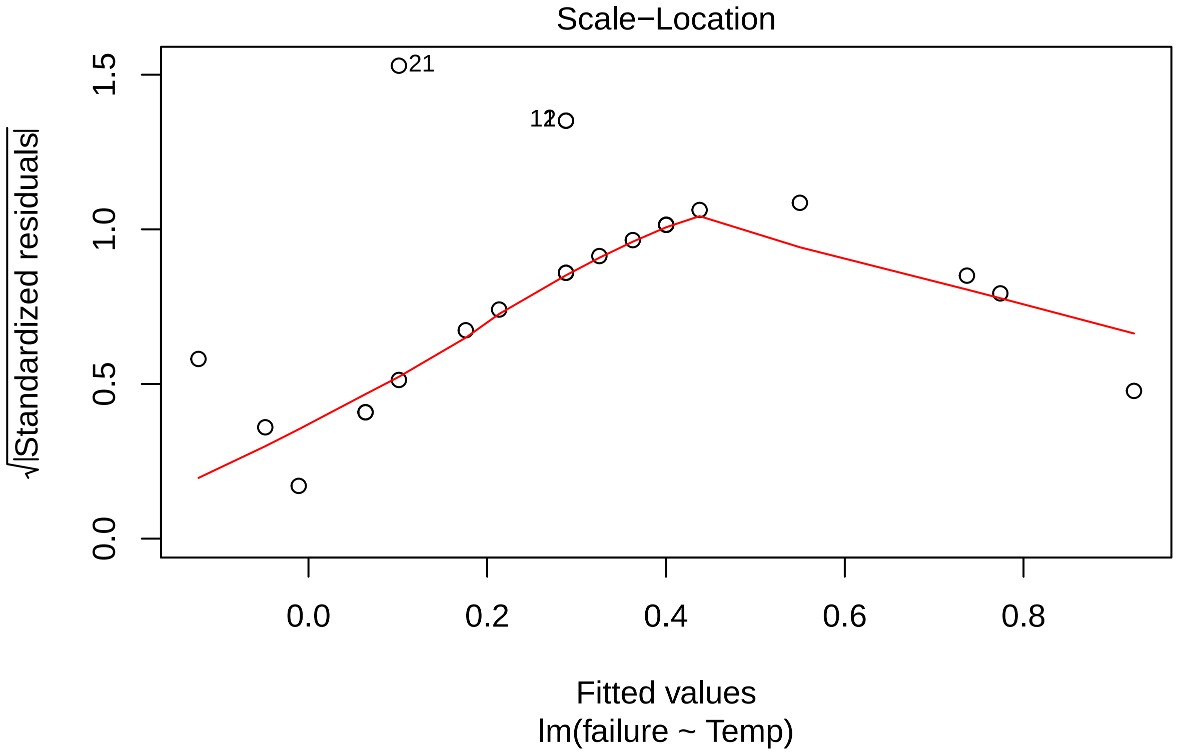
```
plot(challenger.lm)
```

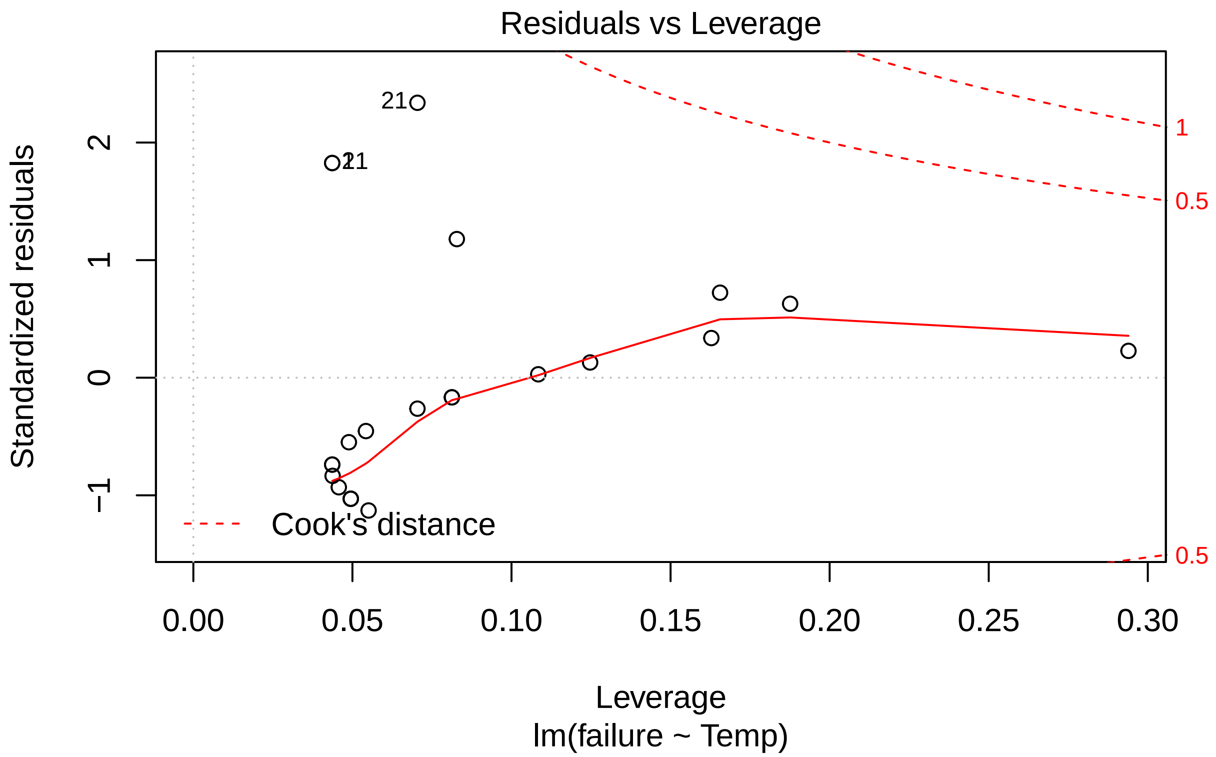
Table 7: Modeling using OLS

	Dependent variable:
	failure
Temp	-0.037^{***}
	(0.012)
Constant	2.905^{***}
	(0.842)
Observations	23
R^2	0.314
Adjusted R^2	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




Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
21
2
11
Theoretical Quantiles
lm(failure ~ Temp)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
21
12
Fitted values
lm(failure ~ Temp)



Answer 5: The Final Model Selection Based on our analysis, our final model will be the binary model with *Temp* as the only explanatory variable. However, we see benefit in looking at the explanatory variable both linearly and after having undergone a log transformation. The log transformation will also work if we change temperature systems from imperial to metric.

```
# stargazer(mod.binary.Ha, type = 'text')
mod.binary.Ha2 <-glm(formula=failure ~ log(Temp), family = binomial (link=logit), data=challeng
mod_stargazer(mod.binary.Ha,mod.binary.Ha2, title = 'Final Model vs Final Model log', no.space
```

Table 8: Final Model vs Final Model log

	Dependent variable:	
	failure	
	(1)	(2)
Temp	−0.232**	
	(0.108)	
log(Temp)		−15.797**
		(7.420)
Constant	15.043**	65.866**
	(7.379)	(31.316)
Observations	23	23
Log Likelihood	−10.158	−10.034
Akaike Inf. Crit.	24.315	24.068
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
print(paste("A one degree decrease in temperature increases the odds of O-rings failing by ",ro
```

[1] “A one degree decrease in temperature increases the odds of O-rings failing by 26 % with a confidence interval of increasing 67 % to 6 %.”

```
print(paste("A 1% decrease in temperature increases the odds of O-rings failing by ",round((1/e
```

[1] “A 1% decrease in temperature increases the odds of O-rings failing by 17 %, with a confidence interval of increasing 43 % to 4 %.”

```
print(paste("A one degree decrease in temperature increases the probability of O-rings failing
```

[1] “A one degree decrease in temperature increases the probability of O-rings failing by 21 % with a confidence interval of increasing 40 % to 6 %.”

```
print(paste("A 1% decrease in temperature increases the probability of O-rings failing by ",rou
```

[1] “A 1% decrease in temperature increases the probability of O-rings failing by 15 % with a confidence interval of increasing 30 % to 4 %.” We choose this model because the data are more likely to be independently distributed. We do assume a linear relationship between the transformed response in terms of the $logit()$ link function and the explanatory variable. $logit(\pi) = \beta_0 + \beta_1 * Temp$. While we did take the log transformation of $Temp$, its explanatory usefulness is greater than its benefit to deviance reduction, relative to the benefit of interpretation of the model. The log-transform is slightly more beneficial. The homogeneity of variance does not need to be satisfied, so we do need to check for heteroskedasticity.

Interpretation Given the intended audience and the questions we are trying to answer, we find it most beneficial to look at temperature decreases (rather than increases), as it affects O-ring performance.

Thus, we take the reciprocal of the exponentiated coefficient for $Temp$, with our increment being one degree Fahrenheit. Using $\Delta OR_{Temp} = \frac{1}{e^{1 \times \beta_{degree}}}$ we find that a one degree decrease in temperature makes the estimated odds of O-rings failing 26% larger with a 95% confidence interval of 6% to 67% larger

Similarly, through interpretation of the slightly more parsimonious log-transformed $Temp$ variable, $\Delta OR_{log(Temp)} = \frac{1}{e^{0.01 \times \beta_{log}}}$ produces a similar result, but interpreted as a percentage change: A 1% decrease in temperature makes the estimated odds of O-rings failing 17% larger, with a 95% confidence interval of a 4% to 43% larger.

Using $\Delta\pi_{Temp} = \frac{e^{1 \times \beta_{degree}}}{1+e^{1 \times \beta_{degree}}}$, we find the , we can interpret the effect on likelihood of O-ring failure: A one degree decrease in temperature increases the probability of O-rings failing by 21% with a 95% confidence interval of increasing failing 6% to 40% For the log transformed data,

$\Delta\pi_{Temp} = \frac{e^{0.01 \times \beta_{log}}}{1+e^{0.01 \times \beta_{log}}}$ yields a 1 percent decrease in temperature increases the probability of O-rings failing by 15 percent with a 95% confidence interval of decreasing 4% to 30%.

This model predicts that, at 31 degrees Fahrenheit, as was the temperature on the morning of the launch, there was a near 100% probability of an O-ring failing.