# W271 Lab 1 – Investigation of the 1989 Space Shuttle Challenger Accident

Lingyao Meng and Devin Robison

### Part 1 (25 points)

Conduct a thorough EDA of the data set, including univariate, bivariate and trivariate analysis. 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.

```r
setwd("~/Documents/Berkeley/W271/W271_lab1")
#load the packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(psych)
```

```
##
## Attaching package: 'psych'
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
#load and summarize data
data <- read.csv("challenger.csv")
describe(data)
```

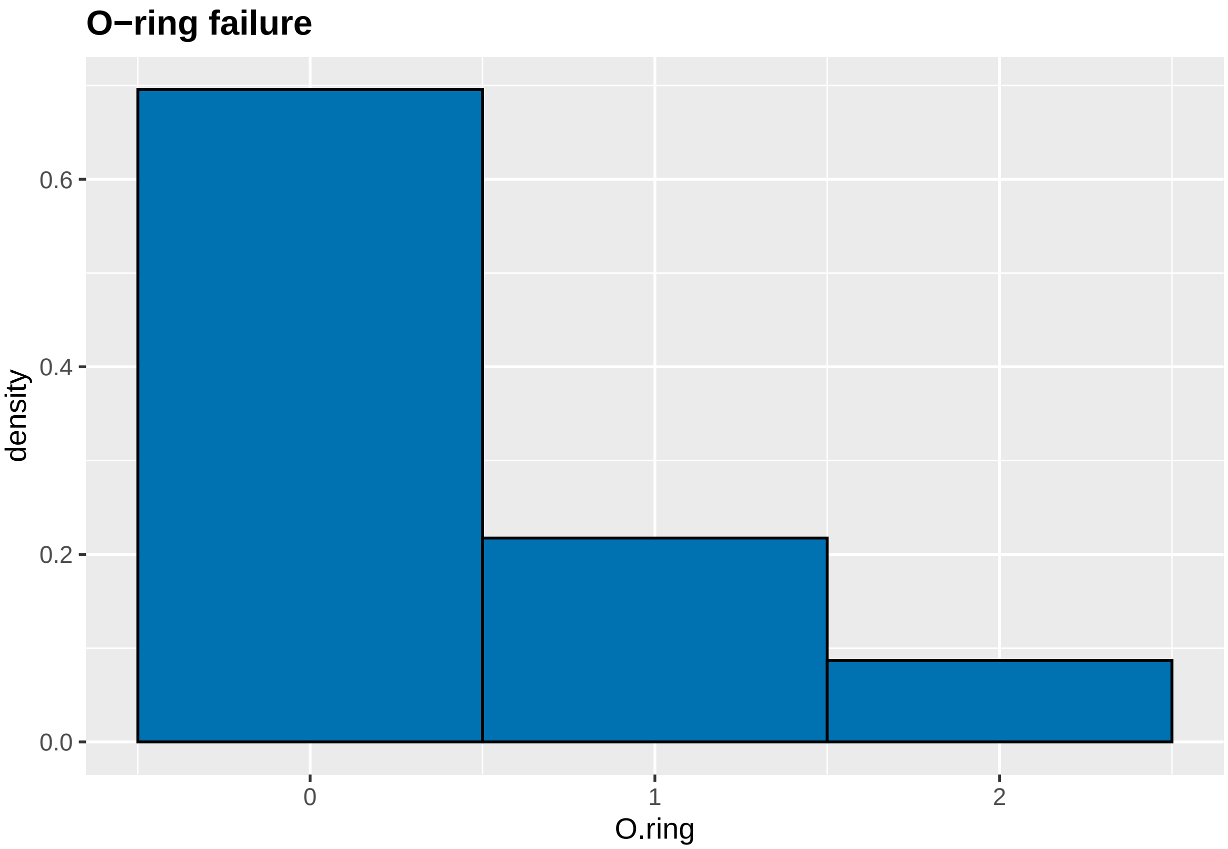
```
##          vars  n   mean    sd median trimmed  mad min max range  skew kurtosis
## Flight      1 23  12.00  6.78     12   12.00 8.90   1  23    22  0.00    -1.36
## Temp        2 23  69.57  7.06     70   70.00 5.93  53  81    28 -0.57    -0.27
## Pressure    3 23 152.17 68.22    200  157.89 0.00  50 200   150 -0.69    -1.50
## O.ring      4 23   0.39  0.66      0    0.26 0.00   0   2     2  1.31     0.39
## Number      5 23   6.00  0.00      6    6.00 0.00   6   6     0   NaN      NaN
##             se
## Flight    1.41
## Temp      1.47
## Pressure 14.23
## O.ring    0.14
## Number    0.00
```

The initial inspection shows that there is no missing value in any avariable. *O.ring* denotes the number of O-ring failures in a flight while *Number* denotes the total number of O-rings, which is a constant 6 for all flights. *O.ring* is the response variable of our interest. The potential explanatory variables are the launch temperature, denoted by *Temp* and the pressure for leak test, denoted by *Pressure*, both of which are numerical variables.

Next we performed the univariate analysis for *O.ring*, *Temp* and *Pressure*.

###Univariate analysis of O-ring failure

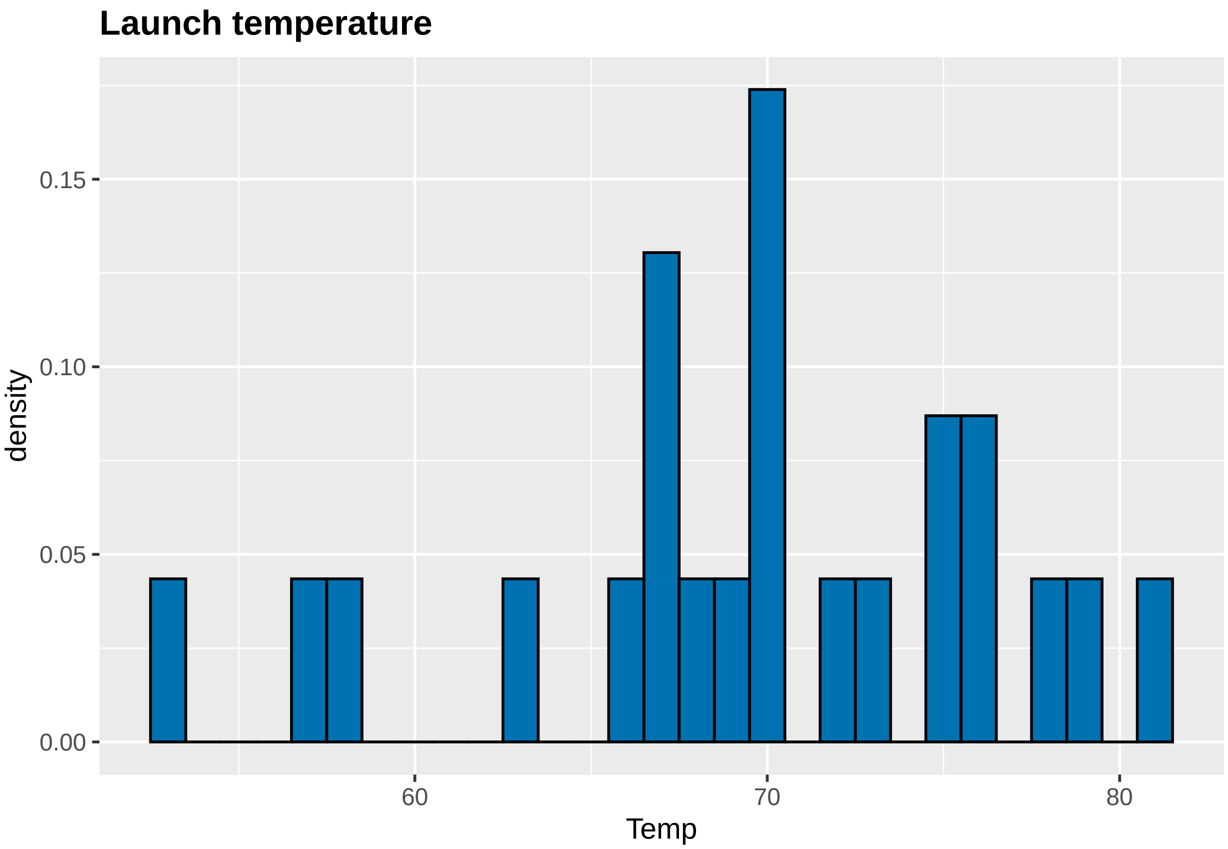
```
ggplot(data, aes(x = O.ring)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("O-ring failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



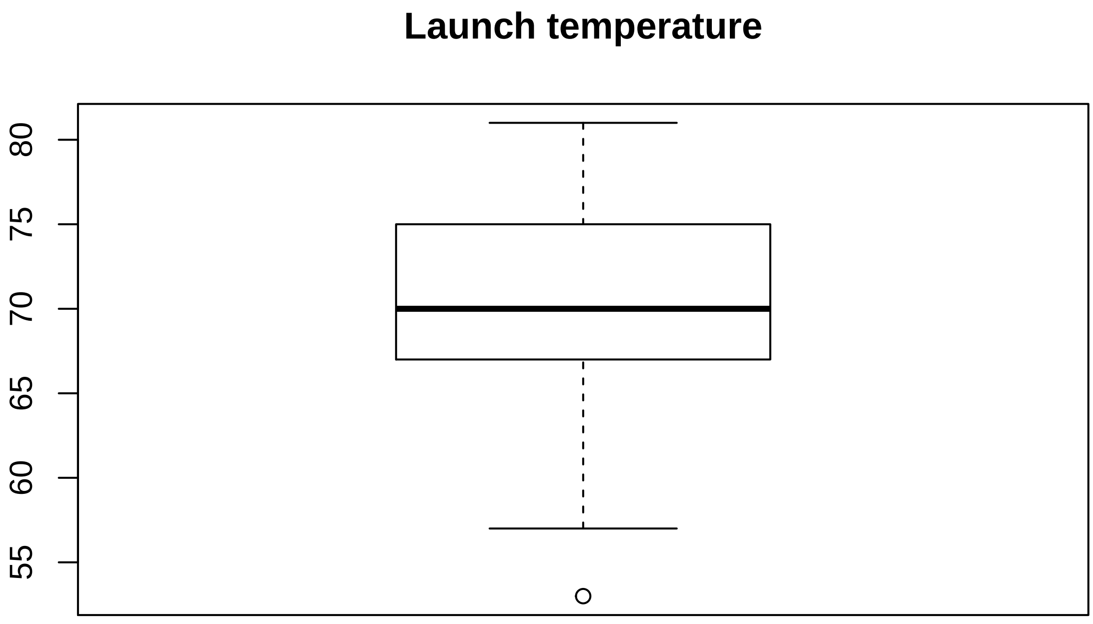
From the histogram above we can tell that about 70% of the flights had no O-ring failure, 20% had only 1 failure and 10% had 2 failures.

###Univariate analysis of launch temperature

```
ggplot(data, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Launch temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



```
boxplot(data$Temp, main = "Launch temperature")
```



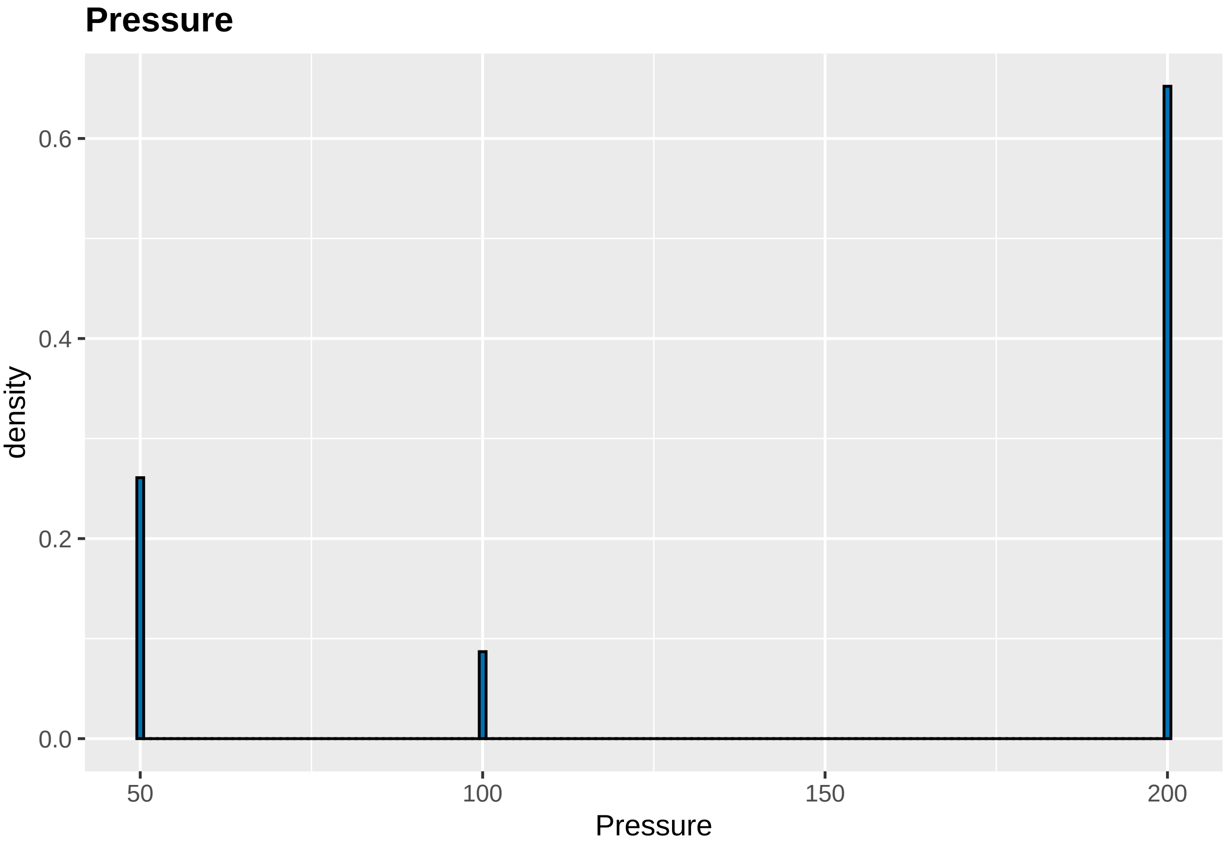
```
outlier <- data %>% filter(Temp < 55)
outlier
```

```
##   Flight Temp Pressure O.ring Number
## 1     14   53      200      2      6
```

Visually, the distribution of *Temp* is slightly right skewed. From the data summary, we can also see that the median (70.0 F) is a bit larger than the mean (69.57 F). From the boxplot, we observed an outlier with the temperature lower than 55 F. After checking, we found that the outlier had 2 O-ring failures.

###Univariate analysis of pressure for leak test

```r
ggplot(data, aes(x = Pressure)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```

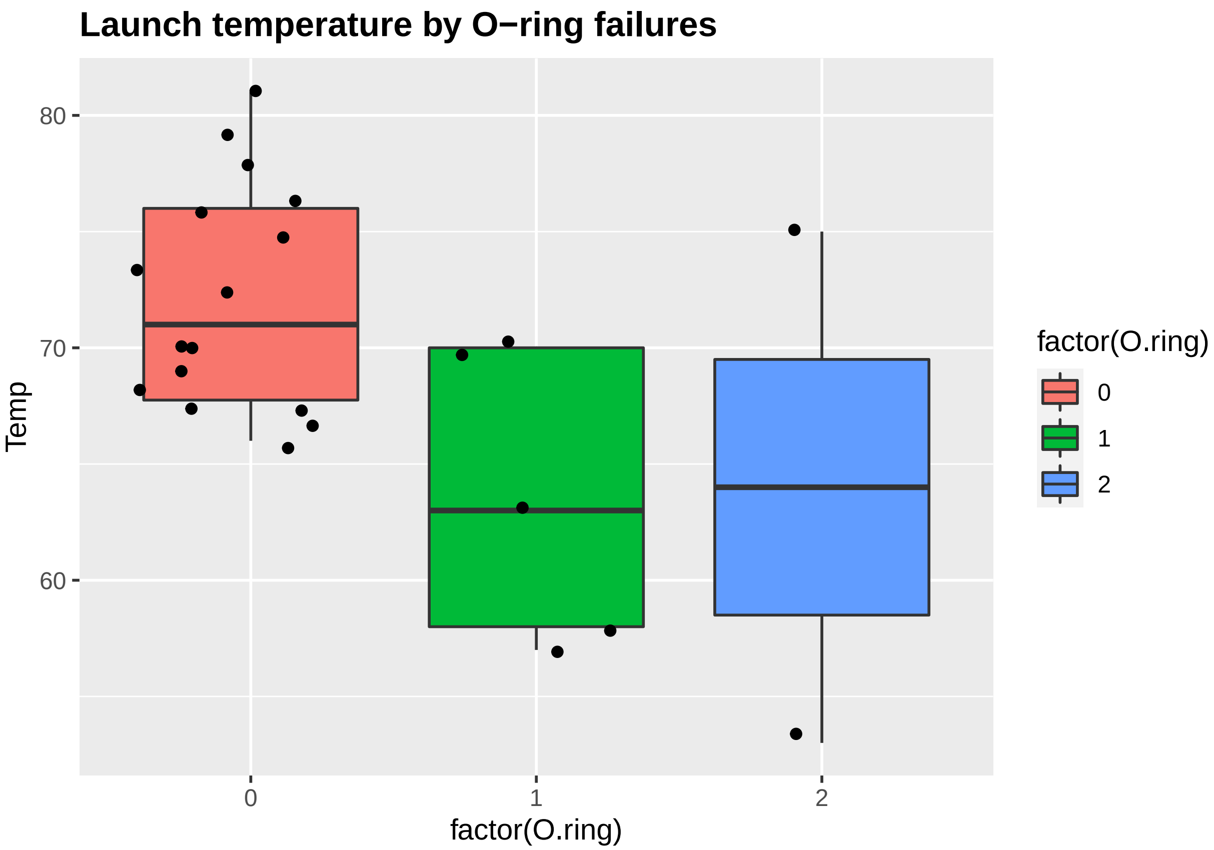


According the Dalal's paper, the pressure used for leak test was originally 50 psi, then increased to 100 psi and finally 200 psi, in the sequence of flights launching. Based on the analysis on *Pressure*, we found that over 60% of flights used 200 psi pressure for leak test, less than 10% used 100 psi and less than 30% used 50 psi.

In order to elucidate the potential cause for the O-ring device to fail, we performed the bivariate analysis between the response variable and each of the explanatory variables.

###Bivariate analysis of O-ring failure Vs. Launch temperature

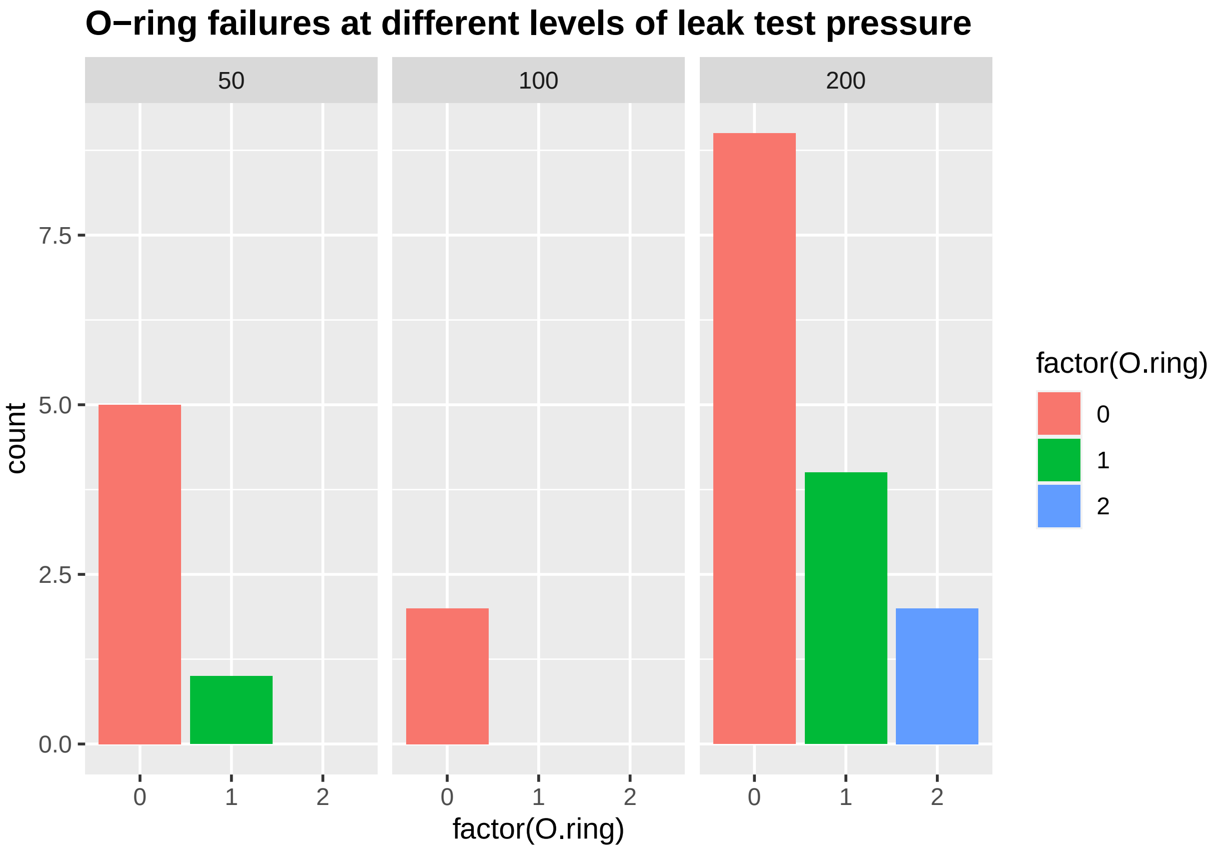
```r
ggplot(data, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  geom_jitter() +
  ggtitle("Launch temperature by O-ring failures") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



We firsly grouped launch temperature by the number of O-ring failures and ploted each group using boxplot. Apparently, the flights with 0 O-ring failure were launched under higher temperature, than that for the flights with 1 or 2 failures. However, it worths to notice that the data size of flights with 1 or 2 O-ring failures is smaller than that of 0 failure flights. Especially, there are only 2 flights with 2 O-ring failures.

###Bivariate analysis of O-ring failure Vs. Pressure for leak test

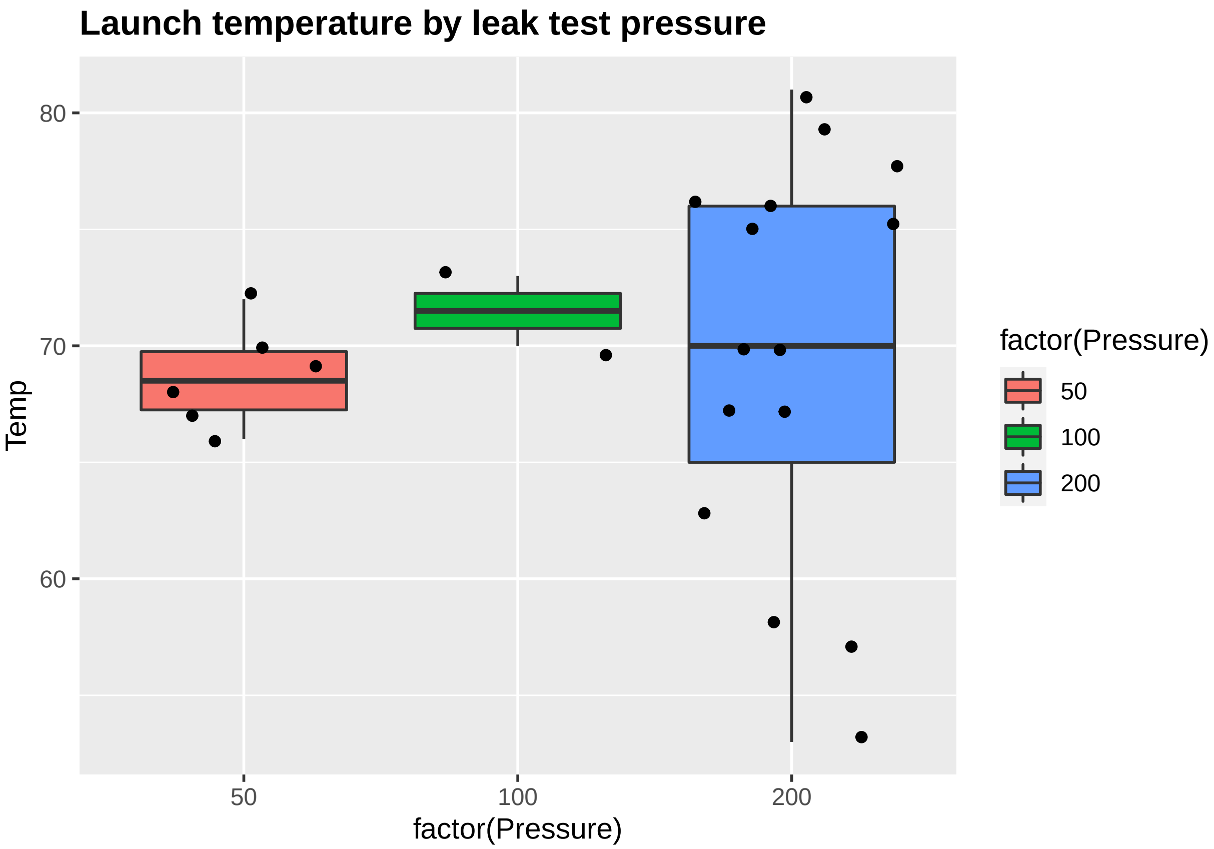
```
ggplot(data, aes(x = factor(O.ring), fill = factor(O.ring))) +
  geom_bar() +
  facet_wrap(~Pressure) +
  ggtitle("O-ring failures at different levels of leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



In the bivariate analysis between O-ring failures and leak test pressure, 0, 1 and 2 O-ring failures were counted, respectively, at different levels of pressure. From the corresponding plot, we cannot tell obvious correlation between *Pressure* and *O.ring*. Further analysis is to be conducted for elucidating potential correlation.

###Bivariate analysis of Launch temperature Vs. Pressure for leak test

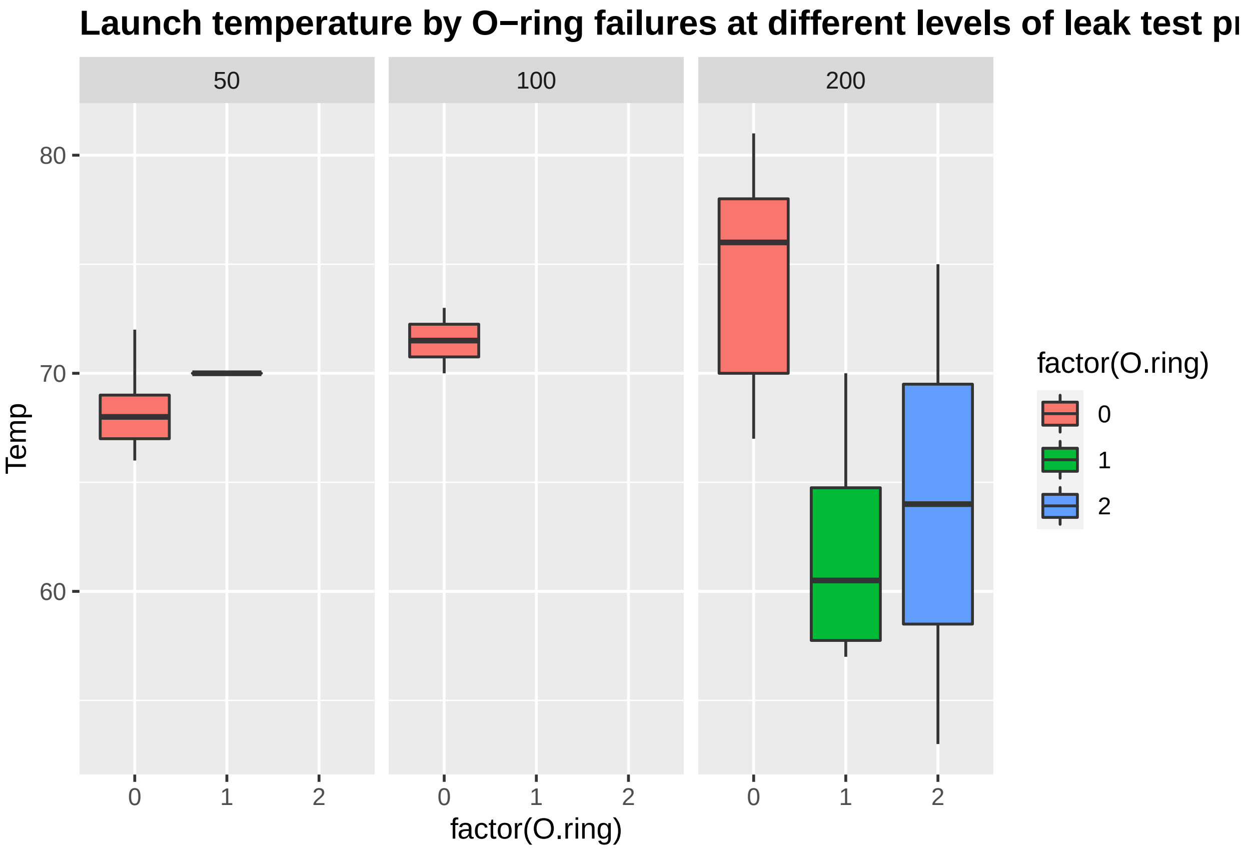
```
ggplot(data, aes(factor(Pressure), Temp)) +
  geom_boxplot(aes(fill = factor(Pressure))) +
  geom_jitter() +
  ggtitle("Launch temperature by leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



Bivariate analysis between launch temperature and pressure for leak test was also performed to check for potential dependence between the explanatory variables. The boxplot of temperature, grouped by pressure, seems to show that when 100 psi pressure was used, the launch temperature is higher than that when 50 or 200 psi was used. However, we cannot ensure that two variables are dependent solely based on this observation because there are only two data points for 100 psi.

###Trivariate analysis

```r
ggplot(data, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  facet_wrap(~Pressure) +
  ggtitle("Launch temperature by O-ring failures at different levels of leak test pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



Trivariate analysis was performed by examing launch temperature by O-ring failures at three different levels of leak test pressure. Similar correlation between *Temp* and *O-ring* was observed when the pressure of 200 psi was used. However, the plot at pressure levels of 50 or 100 psi didn't provide much useful information due to the small data size.

**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.

By using logistic regression to estimate the probability an O-ring will fail, the authors assumed the number of failed O-rings in a given launch to be a binomial variable. In other words, the response variable of the logistic regression was assumed to have the binomial distribution. One of the assumptions for a process to be modeled by binomial distribution is that the trials are independent of each other. In the O-ring case, each O-ring is a trial, so it's necessary to assume that each O-ring is independent for each launch to ensure the validity of model used. However, this assumption is not necessarily true. For instance, the 6 primary O-rings locate at 2 rocket motors. It's possible for the O-rings locating at the same motor to have more similar probabilities to fail. If so, the assumption of independence doesn't hold any more and the logistic regression model used here is invalid. To check on this, the authors fit another model using a binary response to indicate whether there was an incident in a given launch. The second model doesn't require independence of each O-ring. In fact, the second model was quite close to the original model, which alleviated the authors' concerns about independence.

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

From the exploratory data analysis, we found some correlation between *Temp* and *O-ring* while the correlation between *Pressure* and *O-ring* was not very obvious. However, we still want to include *Pressure* as an explanatory variable for the first logistic regression model. For a given launch $i$, we denote the probability for an O-ring to fail as $\pi_i$, launch temperature as $t_i$ and leak test pressure as $p_i$. The first model has the following equation:

$$logit\left(\pi_i\right) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 t_i + \beta_2 p_i$$

This model was fit and estimated using the *glm* function:

```
mod1 <- glm(O.ring/Number ~ Temp + Pressure, weights = Number,
               family = binomial (link = logit), data = data)
summary(mod1)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure, family = binomial(link = logit),
##     data = data, 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|)
## (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
```

```
cbind(Estimate = coef(mod1), confint(mod1))
```

```
## Waiting for profiling to be done...
```

```
##                 Estimate        2.5 %      97.5 %
## (Intercept)  2.520194641 -4.322926283  9.77264497
## Temp        -0.098296750 -0.194071699 -0.01356289
## Pressure     0.008484021 -0.004346403  0.02885221
```

```
c.temp <- -5
exp(c.temp*coef(mod1)['Temp'])
```

```
##    Temp
## 1.63474
```

The coefficient of $t_i$ was estimated to be -0.0983 with the 95% Wald confidence interval of -0.1941 to -0.0136, indicating that the decrease on the launch temperature would cause the increase on the odds for an O-ring to fail. Specifically, a decrease of 5 F would increase the odds for failure by around 63%. The coefficient of $p_i$ was estimated to be 0.0085 while 0 was included in the 95% Wald confidence interval, indicating that leak test pressure may not be an important factor for explaining O-ring failure.

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

Because the Wald interval usually has lower true coverage than the cofidence level, we performed likelihood ratio test using the *Anova* function to judge the importance of the explanatory varialbes in the first model.

```
Anova(mod1, 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
```

For the test of *Temp* with $H_0 : \beta_1 = 0$ vs. $H_\alpha : \beta_1 \neq 0$, we obtained the LRT statistic of 5.184 with a p-value of 0.0228. Using the Type I Error rate $alpha = 0.05$, we would reject the null hypothesis and claim that there is marginal evidence that *Temp* is important to be included in the model, given that *Pressure* is in the model. For the test of *Pressure* with $H_0 : \beta_2 = 0$ vs. $H_\alpha : \beta_2 \neq 0$, we obtained the LRT statistic of 1.541 with a p-value of 0.2145. Using the Type I Error rate $alpha = 0.05$, we could not reject the null hypothesis. Therefore, there is a lack of evidence to claim that *Pressure* is important to be included in the model, given that *Temp* is in the model.

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

The authors fit a model using both *Temp* and *Pressure* and then fit another model using only *Temp*. By comparing the residual deviances of two models, they found that keeing only *Temp* in the model just increased the residual deviance by 1.54, which was not significant, indicating that *Pressure* may had a very weak effect. This is consistent with our LRT results in the above section. However, the apparently weak effect of *Pressure* might be due to limited data for 50 and 100 psi. Assume a more comprehensive dataset available, removing *Pressure* from the model would cause serious information loss.

**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 $\pi$ is the probability of an O-ring failure. Complete the following:

(a) Estimate the model.

```r
mod2 <- glm(O.ring/Number ~ Temp, weights = Number,
               family = binomial (link = logit), data = data)
summary(mod2)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp, family = binomial(link = logit),
##     data = data, 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
```

```r
cbind(Estimate = coef(mod2), confint(mod2))
```

```
## Waiting for profiling to be done...
```

```
##               Estimate      2.5 %     97.5 %
## (Intercept)  5.0849772 -1.0102633 11.1854755
## Temp        -0.1156012 -0.2122262 -0.0244701
```

(b) Construct two plots: (1) $\pi$ 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°.

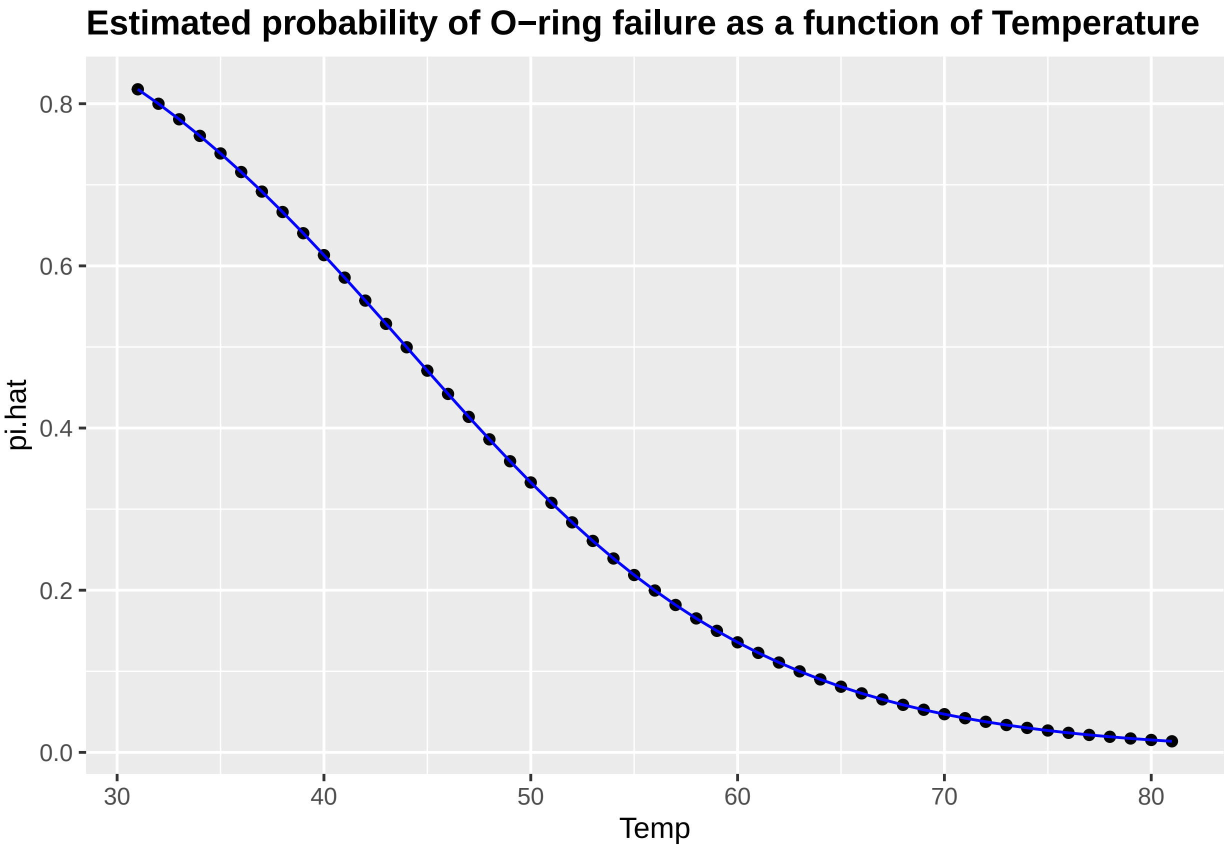
Given our logit model from (a), and required temperature range, we compute $\pi_i$ at each temperature as:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 \cdot t_i}}{1 + e^{\beta_0 + \beta_1 \cdot t_i}}$$

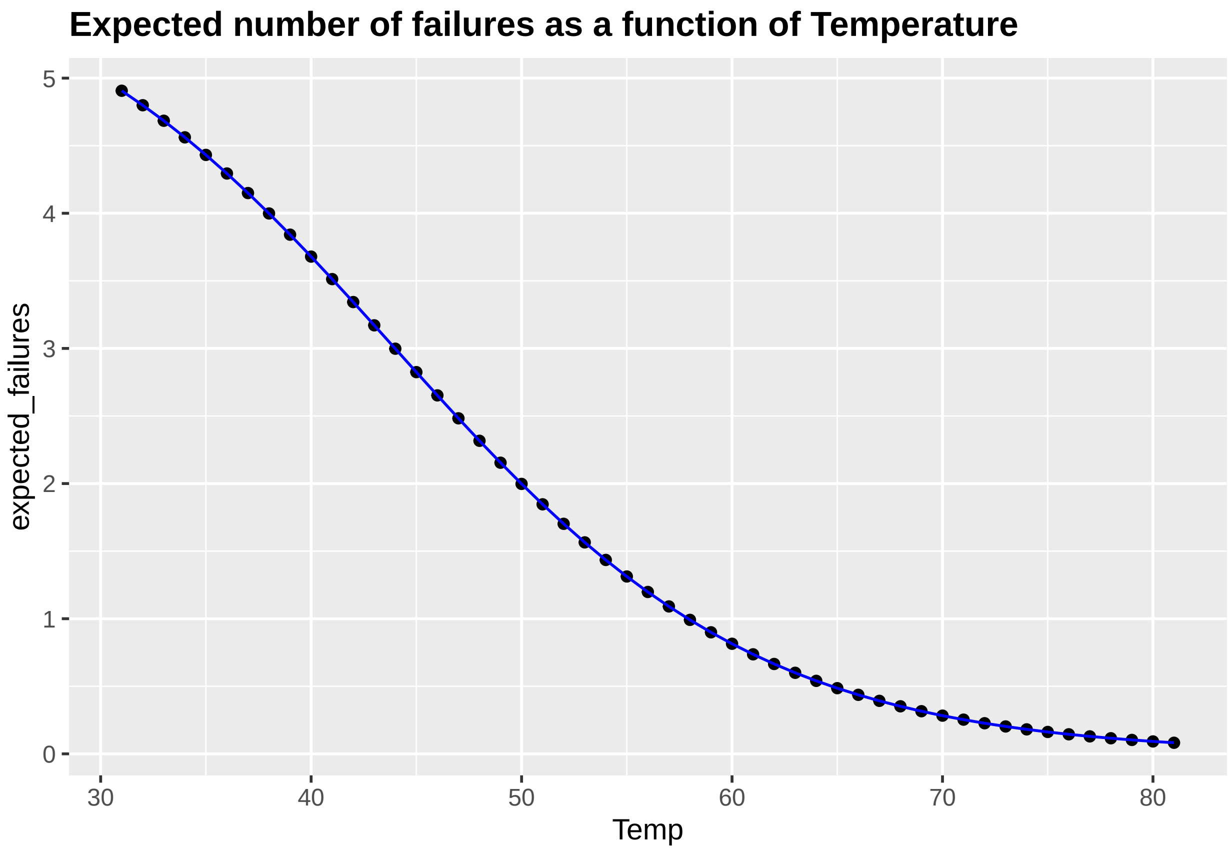
Next, we use our assumption of independence of O-ring failures to compute the expected number of failures at a given temperature $t_i$ as:

$$E[binom(n, \pi_i)] = n \cdot \pi_i = 6 \cdot \pi_i$$

```
results_df <- data.frame(Temp=seq(from = 31, to = 81, by = 1))
params = predict(object=mod2, newdata = results_df, type = 'link', se = TRUE)
results_df['pi.hat'] = exp(params$fit) / (1 + exp(params$fit))

# Note that results_df['pi'] is the probability of failure at the corresponding temperature.
# Since we've assumed all our failures are independent, then our expected number of failures i
#  of E[bin(n, p)]
#     = n * p
#     = 6 * results_df['pi']
results_df['expected_failures'] = 6*results_df['pi.hat']

ggplot(data=results_df, aes(Temp, pi.hat)) +
  geom_point() +
  geom_line(color='blue') +
  ggtitle("Estimated probability of O-ring failure as a function of Temperature") +
  theme(plot.title = element_text(lineheight=1, face='bold'))
```



```
ggplot(data=results_df, aes(Temp, expected_failures)) +
  geom_point() +
  geom_line(color='blue') +
  ggtitle("Expected number of failures as a function of Temperature") +
  theme(plot.title = element_text(lineheight=1, face='bold'))
```



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

We build the 95% Wald confidence interval bounds, using the fact that our confidence interval for $logit(\pi_i)$ as:

$$\alpha = 0.05\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i \pm Z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i)}$$

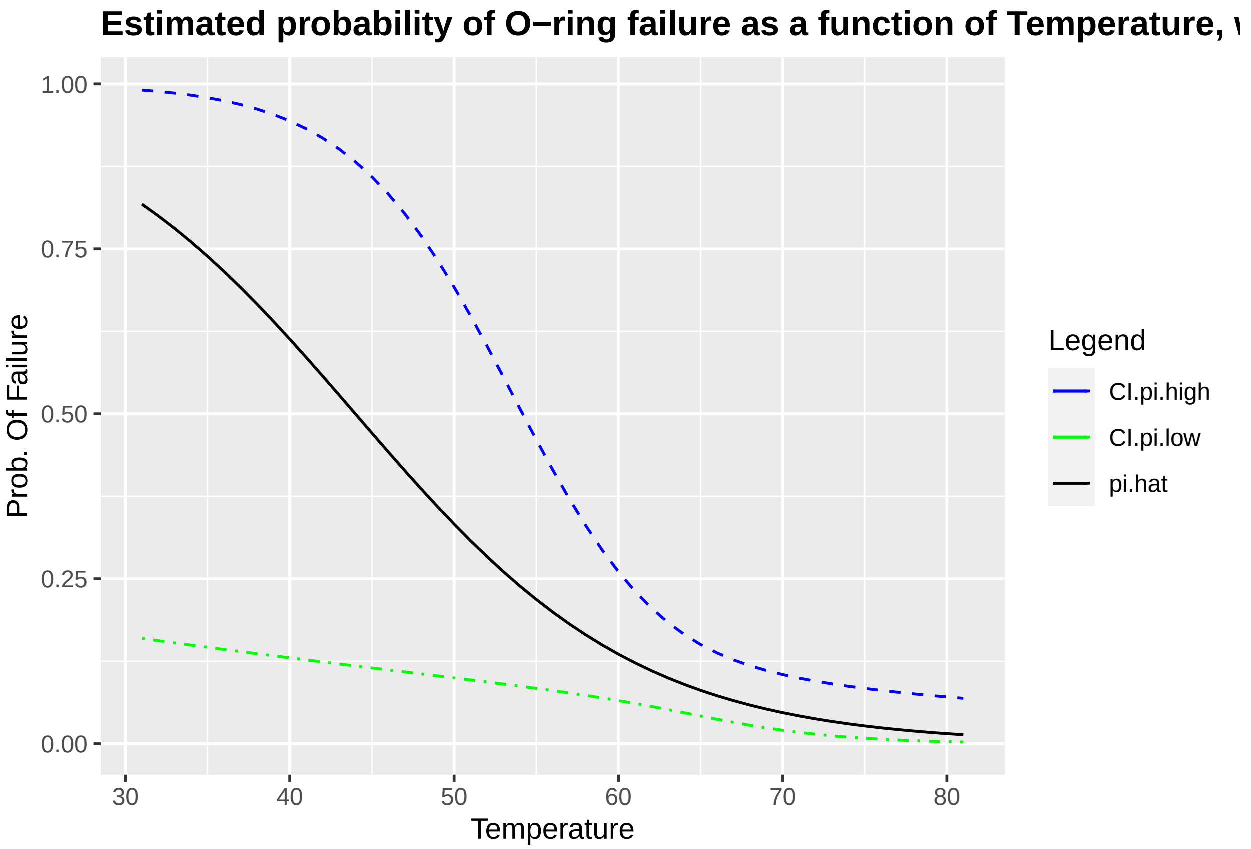
Then constructing a confidence interval for $\pi_i$ as:

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i \pm Z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i)}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i \pm Z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 \cdot t_i)}}}$$

Bounds for lower temperatures are significanlty wider, because we have much fewer samples at lower temperatures, and no samples below $53°F$.

```
alpha = 0.05
CI.preds.low = params$fit + (qnorm(p=c(alpha/2)) * params$se.fit)
CI.preds.high = params$fit + (qnorm(p=c(1-alpha/2)) * params$se.fit)

results_df['CI.pi.low'] = exp(CI.preds.low) / (1 + exp(CI.preds.low))
results_df['CI.pi.high'] = exp(CI.preds.high) / (1 + exp(CI.preds.high))
```

```
colors <- c("CI.pi.low" = "green", "CI.pi.high" = "blue", "pi.hat" = "black")

ggplot(data=results_df, aes(x=Temp), group=colors,) +
  geom_line(aes(y=pi.hat, color='pi.hat')) +
  geom_line(aes(y=CI.pi.high, color='CI.pi.high'), linetype='dashed') +
  geom_line(aes(y=CI.pi.low, color='CI.pi.low'), linetype='dotdash') +
  labs(x = "Temperature", y = "Prob. Of Failure", color = "Legend") +
  scale_color_manual(values = colors) +
  ggtitle("Estimated probability of O-ring failure as a function of Temperature, with Confiden
  theme(plot.title = element_text(lineheight=1, face='bold'))
```



(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.

Based on our model, the probability of an O-ring failure at 31°F is:

```
# Compute probability, above.
params = predict(mod2, newdata=data.frame(Temp=31), type = 'link', se = TRUE)
pi.hat = exp(params$fit) / (1 + exp(params$fit))
pi.hat
```

```
##         1
## 0.8177744
```

$$\pi_i = \frac{e^{\beta_0+\beta_1\cdot 31}}{1+e^{\beta_0+\beta_1\cdot 31}} = 0.81778$$

Since its not specified, we'll look at both the Wald and LRT confidence intervals for this temperature.

```r
library(mcprofile)
# Compute Wald CI @ Temp = 31.
alpha=0.05
logit_pred = pi.hat + qnorm(c(alpha, 1-alpha/2))*params$se.fit
CI.pi.pred = exp(logit_pred) / ( 1 + exp(logit_pred))
paste("Wald CI:")
```

```
## [1] "Wald CI:"
```

```r
as.numeric(CI.pi.pred)
```

```
## [1] 0.1374890 0.9816626
```

```r
# Compute LR @ Temp = 31
K = matrix(data = c(1, 31), nrow=1, ncol=2)
mc.ci.profile = mcprofile(object=mod2, CM=K)
mc.ci.logit = confint(object=mc.ci.profile, level=0.95)
mc.ci = exp(mc.ci.logit$confint) / (1 + exp(mc.ci.logit$confint))
paste("MC CI:")
```

```
## [1] "MC CI:"
```

```r
as.numeric(mc.ci)
```

```
## [1] 0.1418508 0.9905217
```

$$Wald.CI = (0.13749, 0.98166) MCprofile.CI = (0.14185, 0.99052)$$

(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

```r
compute_glm = function(X, t) {
  # Get pi estimates for sample
  pi.logit.sample = exp(mod2$coefficients[1] + mod2$coefficients[2] * X)
  pi.sample = pi.logit.sample / (1 + pi.logit.sample)

  # Sample binomial for simulated errors.
  y = rbinom(n=length(X), size=6, prob=pi.sample)

  mod.sample <- glm(formula = y/data$Number ~ X, family = binomial(link = logit), weights=data$
  pi.logit.star = exp(mod.sample$coefficients[1] + mod.sample$coefficients[2] * t)
  pi.hat.star = pi.logit.star / (1 + pi.logit.star)

  return(pi.hat.star)
}

bootstrap_estimate = function(temp) {
  samples = 23
  n_samples=2000

  x = sample(data$Temp, samples*n_samples, replace=TRUE)
  xmatrix = matrix(data=unlist(x), nrow=samples, ncol=n_samples)

  save.results<-apply(X = xmatrix, MARGIN = 2, FUN = compute_glm, t = temp)

  return(save.results)
}

results = bootstrap_estimate(temp=31)
```

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

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

```
quantile(results, probs=c(0.05, 0.95), na.rm=TRUE)
```

```
##        5%       95%
## 0.0999591 0.9921663
```

```
results = bootstrap_estimate(temp=72)
```

```
## Warning: glm.fit: algorithm did not converge

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

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

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

## Warning: glm.fit: algorithm did not converge

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

```
quantile(results, probs=c(0.05, 0.95), na.rm=TRUE)
```

```
##         5%        95%
## 0.01008823 0.06913274
```

Based on this boostrap analysis, we obtain

$$bootstrap.CI(Temp = 31) = (0.12509, 0.99232) bootstrap.CI(Temp = 72) = (0.00986, 0.06962)$$

Which is slightly more conservative than either our WALD or LRT CI's, but quite similar.

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

```
mod3 <- glm(O.ring/Number ~ Temp + I(Temp^2), weights = Number,
               family = binomial (link = logit), data = data)

Anova(mod2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/Number
##      LR Chisq Df Pr(>Chisq)
## Temp    6.144  1    0.01319 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mod3, test='LR')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/Number
##           LR Chisq Df Pr(>Chisq)
## Temp       0.71878  1     0.3965
## I(Temp^2)  0.49470  1     0.4818
```

Based on the likelihood ratio test, the quadratic term is not significant for the model (P value ~0.4818).

**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.

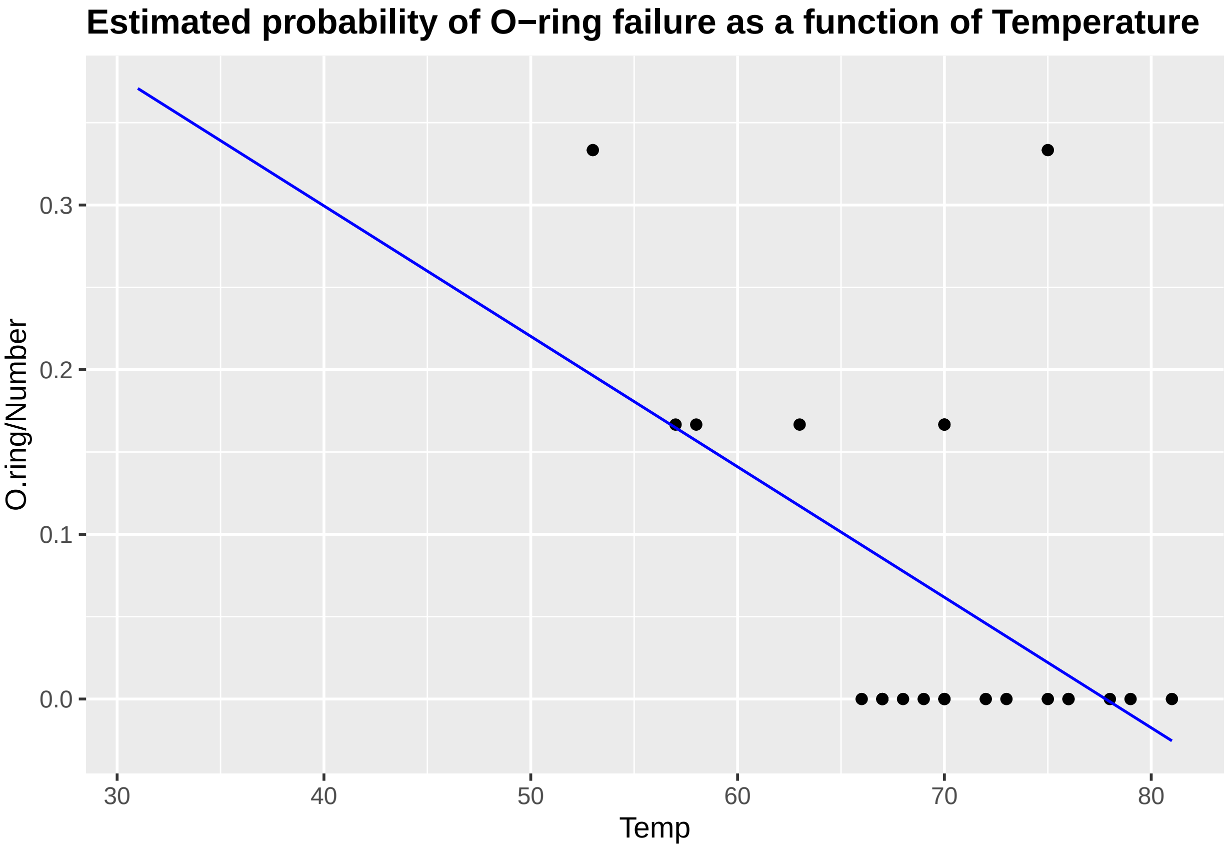
First, we should begin by verifying that our temperature data satisfies the basic requirements for an OLS model. Since we are only using temperature as our explanitory variable, it will take the form of a Simple Linear Regression (SLR) model. SLR model's have a set of requirements that we'll want to verify (See Wooldridge 2.5)

## TODO

$$y = \beta_0 + \beta_1 \cdot temp$$

SLR1 requires that our response variable o.ring/Number, is linear as a function of temperature or at least well approximated. As can be seen from the overlay graph, this assumption is dubious given the existing data set, and even if correct is likely unduely influenced by outliers.

```
mod4 = lm(formula = O.ring/data$Number ~ Temp, data=data)
results_df['pi.hat.linear'] = predict(mod4, newdata=results_df)

ggplot(data=results_df) +
  geom_point(data=data, aes(Temp, O.ring/Number)) +
  geom_line(aes(Temp, pi.hat.linear), color='blue') +
  ggtitle("Estimated probability of O-ring failure as a function of Temperature") +
  theme(plot.title = element_text(lineheight=1, face='bold'))
```



SLR 2 requires our response and explanatory variables are independent and identically distributed is also not entirely obvious, as discussed in part 2 (book question 4.a).

SLR 3 requires that all of our explanatory variables are non-identical, which is true.

SLR 4 Assumes that the mean of our residual values is zero. Examining the 'residuals vs fitted' plot below, we can see that this is likely not an entirely reasonable assumption, with the residuals appearing to follow more of a quadratic form. Its difficult to say for certain though, as we don't not have many low temperature data points. Additionally, given that our data set is relatively small, it may not be appropriate to rely asymptotic (CLT) assumptions. As we're asked to rely on the same set of explanatory variables for this model as those for our binary logistic regression, violation of SLR 4 may be an indicator that our OLS model is not appropriate.

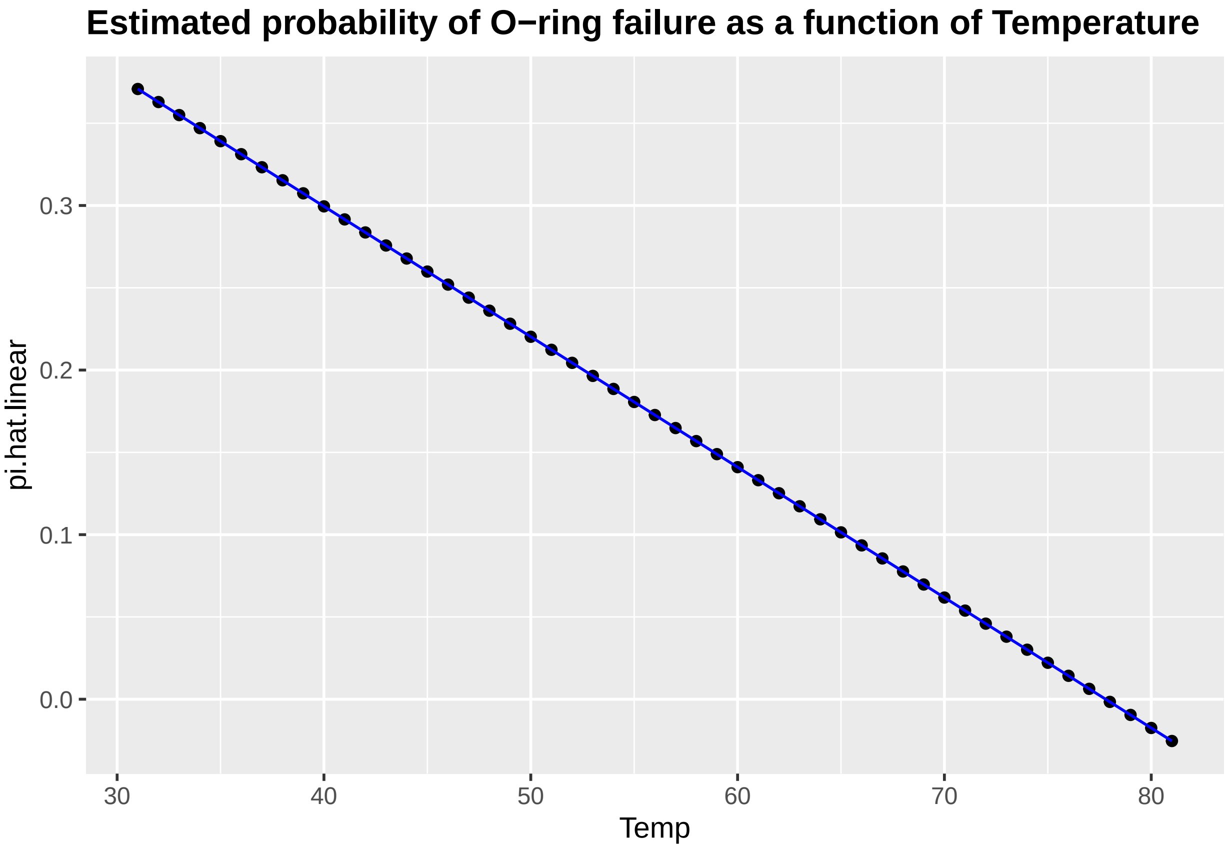
SLR 5 Homoskedacticity assumption: residual variance is uncorrelated with our explanatory variables. Examining our scale-location plot, we would like to see a relatively flat line if our homoskedacticity assumption holds. This does not appear to be the case, but this ends up being an indicator that we need to rely on techniques that produce robust standard errors for our model analysis.

```
library(lmtest)
```

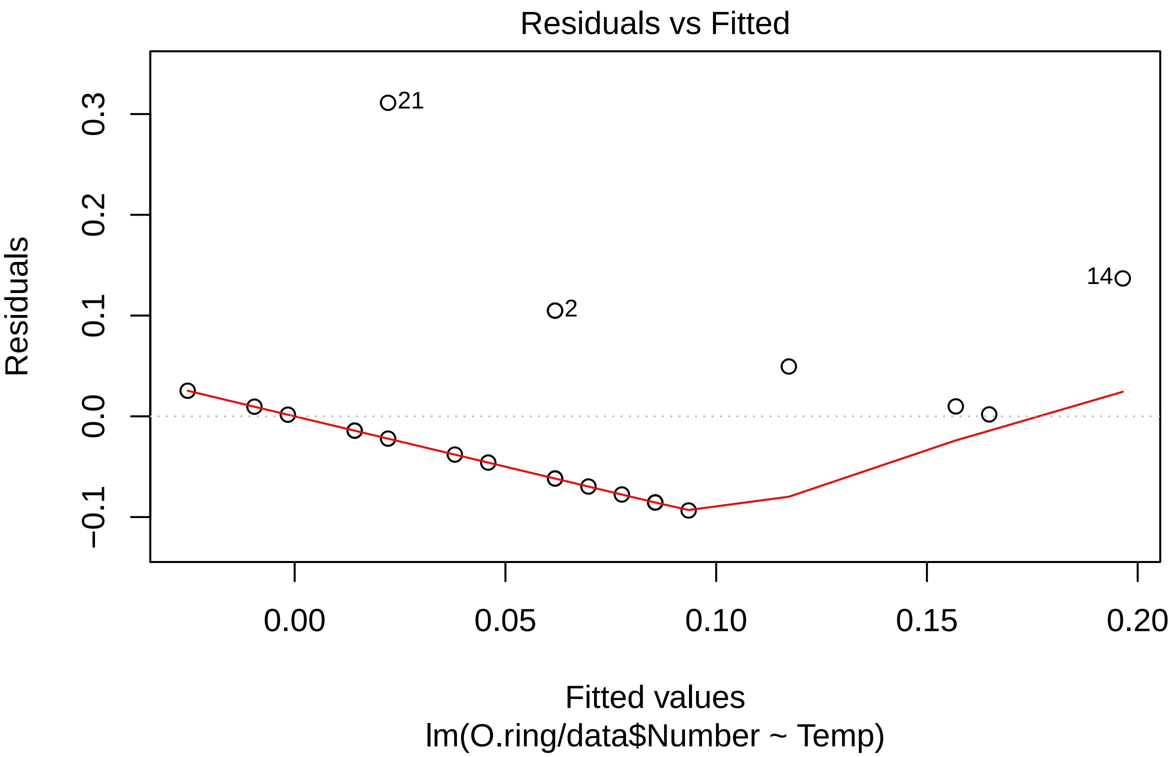
```
## Loading required package: zoo
```

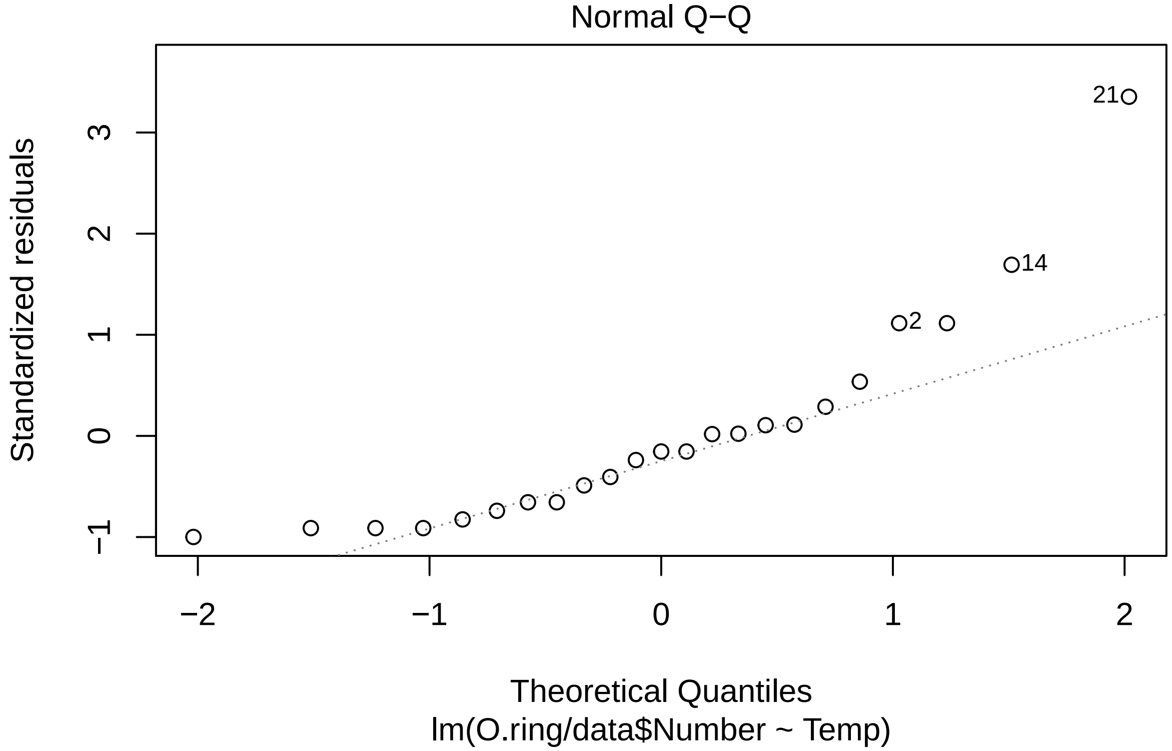
```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

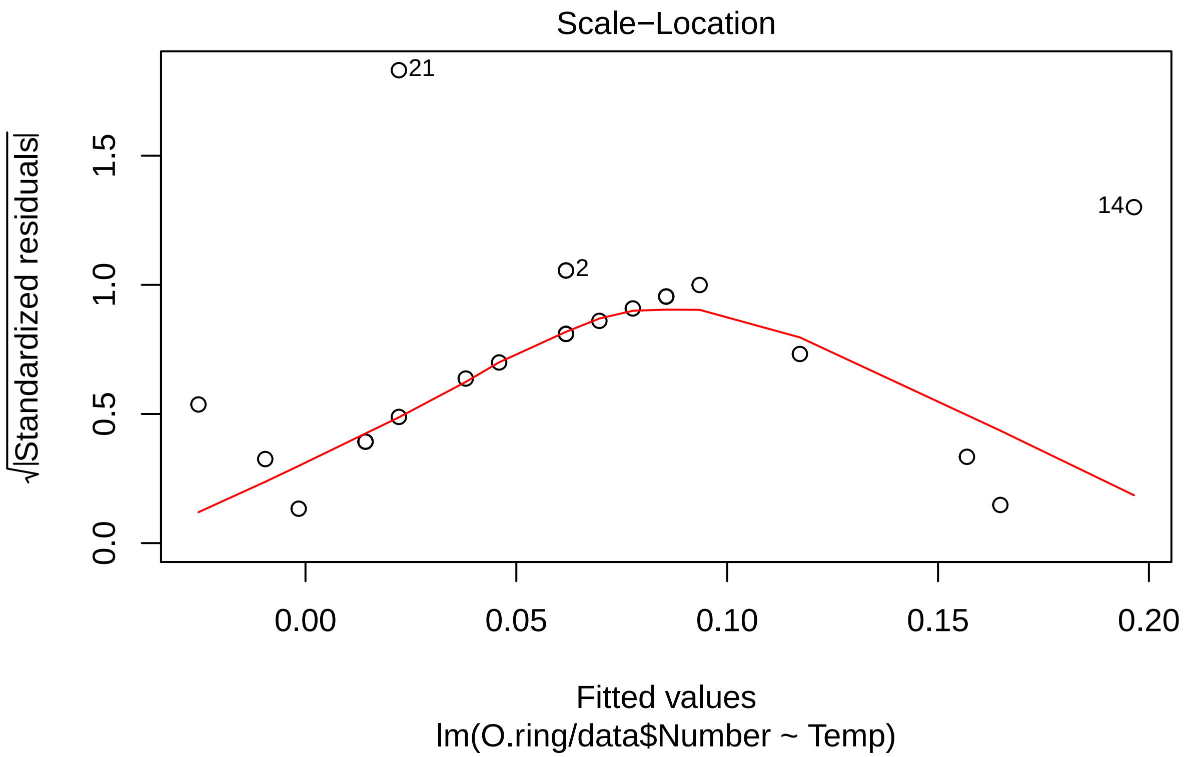
```
library(sandwich)
ggplot(data=results_df, aes(Temp, pi.hat.linear)) +
  geom_point() +
  geom_line(color='blue') +
  ggtitle("Estimated probability of O-ring failure as a function of Temperature") +
  theme(plot.title = element_text(lineheight=1, face='bold'))
```

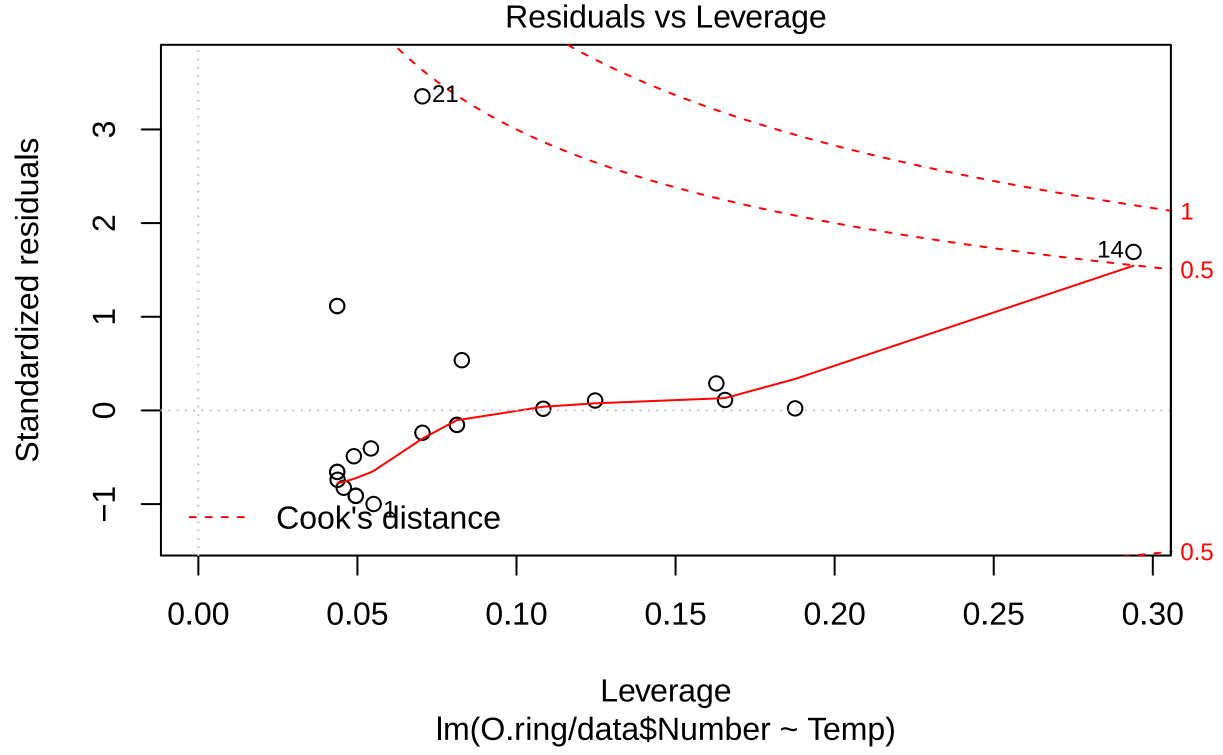


```
# TODO
#ggplot(data=data) +
#  geom_point(aes(mod4$fitted.values, mod4$residuals)) +
#  geom_abline(intercept = 0, slope=0, color='red') +
#  ggtitle("Estimated probability of O-ring failure as a function of Temperature") +
#  theme(plot.title = element_text(lineheight=1, face='bold'))

plot(mod4)
```


Residuals vs Fitted
Residuals
0.3
0.2
0.1
0.0
−0.1
21
2
14
0.00
0.05
0.10
0.15
0.20
Fitted values
lm(O.ring/data$Number ~ Temp)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
21
14
2
−2
−1
0
1
2
Theoretical Quantiles
lm(O.ring/data$Number ~ Temp)





```
# Based on our Heteroskedacticity observations, we will want to use robust standard errors for
coeftest(mod4, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6164021  0.2419719  2.5474  0.01875 *
## Temp        -0.0079233  0.0034482 -2.2978  0.03195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, while analysis does correctly indicate that Temp is significant, with a P value $0.03195 < 0.05$, it is not appropriate for our use case.

For this application, it is more appropriate to choose the binary logistic regression model. As observed in our discussion, the assumption that a linear relationship for probability of failure as a function of temperature is not obvious and even if it were a valid assumption, the output of the model itself is invalid across the range of temperatures of interest. Its prediction of a 30 percent chance of failure does not align well with our logistic regression models, and the linear model begins to predict negative probabilities at 78°F.

### 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.