w271 Lab 1: Investigation of the 1989 Space Shuttle Challenger Accident

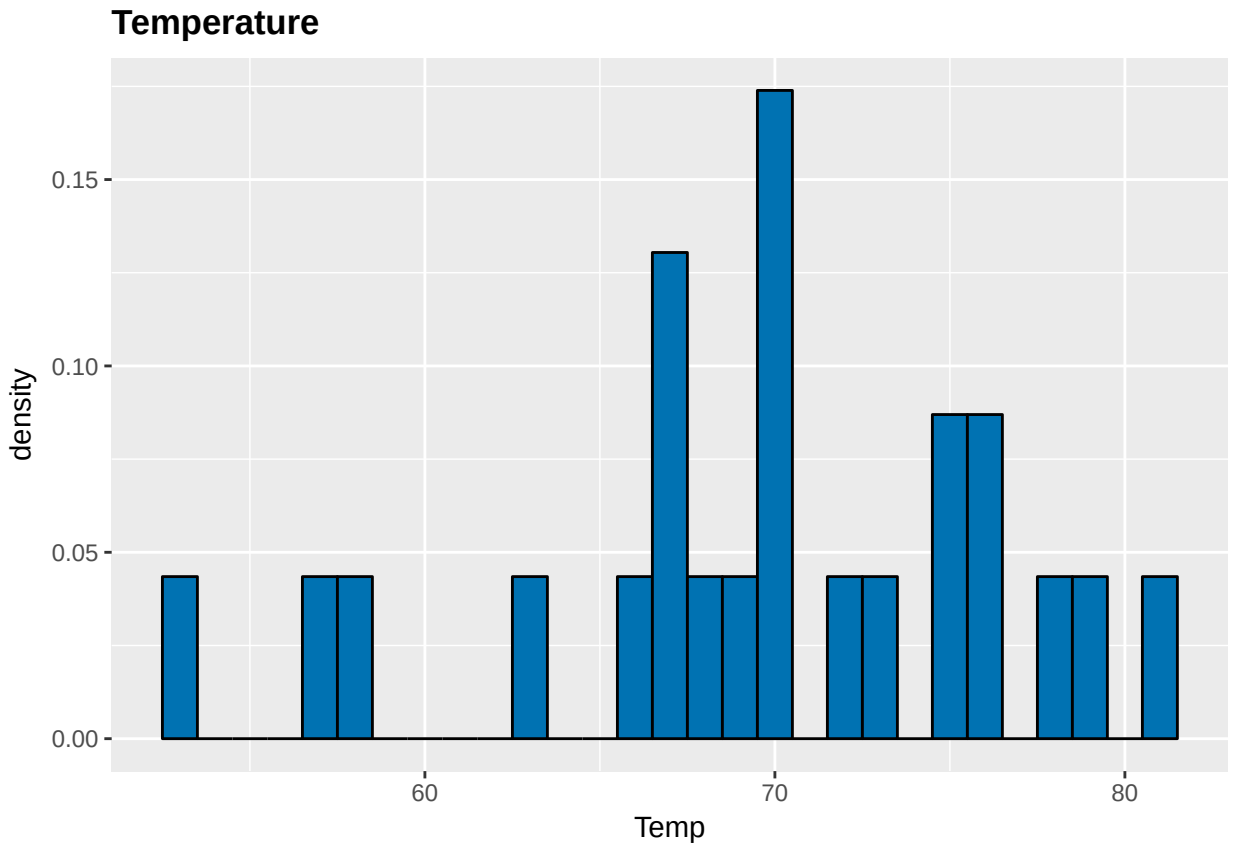
Jessica Hays Fisher, Alice Lam, Marshall Ratliff, Paul Varjan

5/20/2018

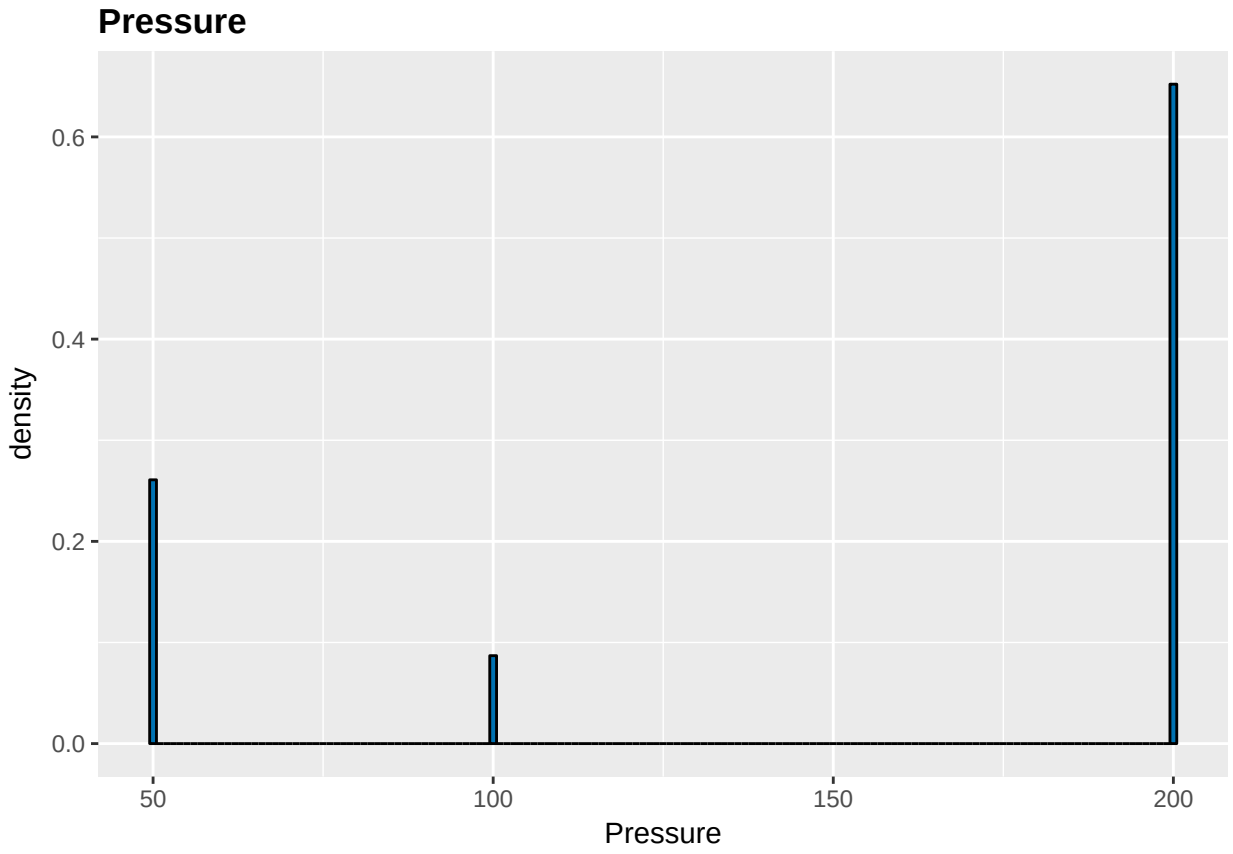
Exploratory Data Analysis

A First Look at the Individual Factors

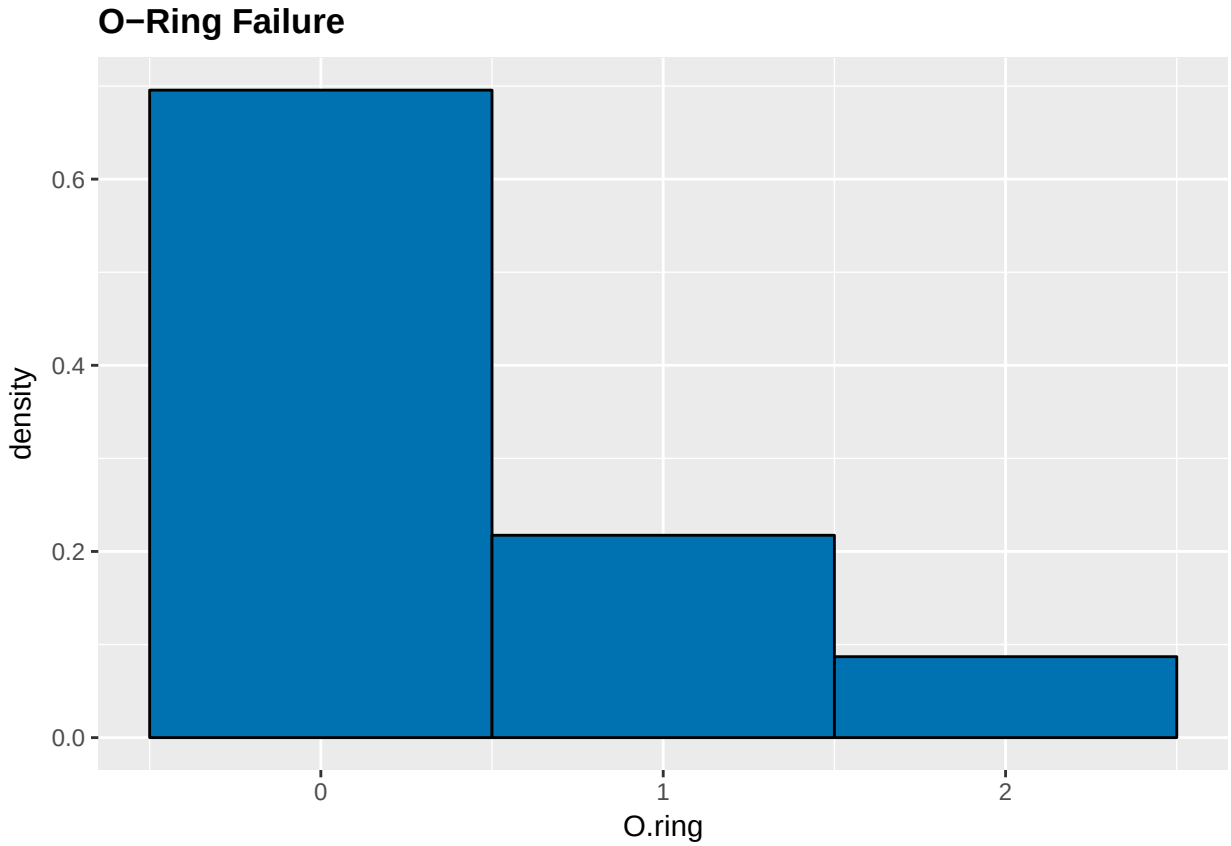
```
ggplot(df, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
ggplot(df, 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"))
```




```
ggplot(df, 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"))
```



Basic Summary Data

```
summary(df)
```

```
##      Flight          Temp          Pressure         O.ring
##  Min.   : 1.0   Min.   :53.00   Min.   : 50.0   Min.   :0.0000
##  1st Qu.: 6.5   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.:0.0000
##  Median :12.0   Median :70.00   Median :200.0   Median :0.0000
##  Mean   :12.0   Mean   :69.57   Mean   :152.2   Mean   :0.3913
##  3rd Qu.:17.5   3rd Qu.:75.00   3rd Qu.:200.0   3rd Qu.:1.0000
##  Max.   :23.0   Max.   :81.00   Max.   :200.0   Max.   :2.0000
##      Number
##  Min.   :6
##  1st Qu.:6
##  Median :6
##  Mean   :6
##  3rd Qu.:6
##  Max.   :6
```

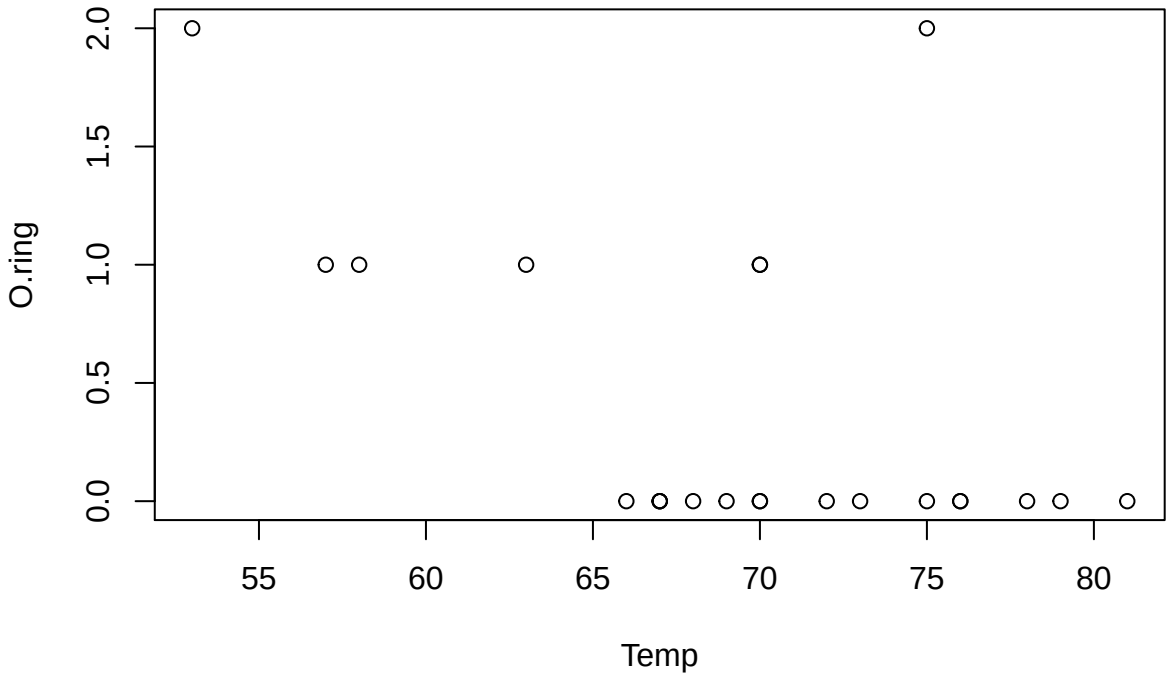
```
describe(df)
```

```
## df
##
```

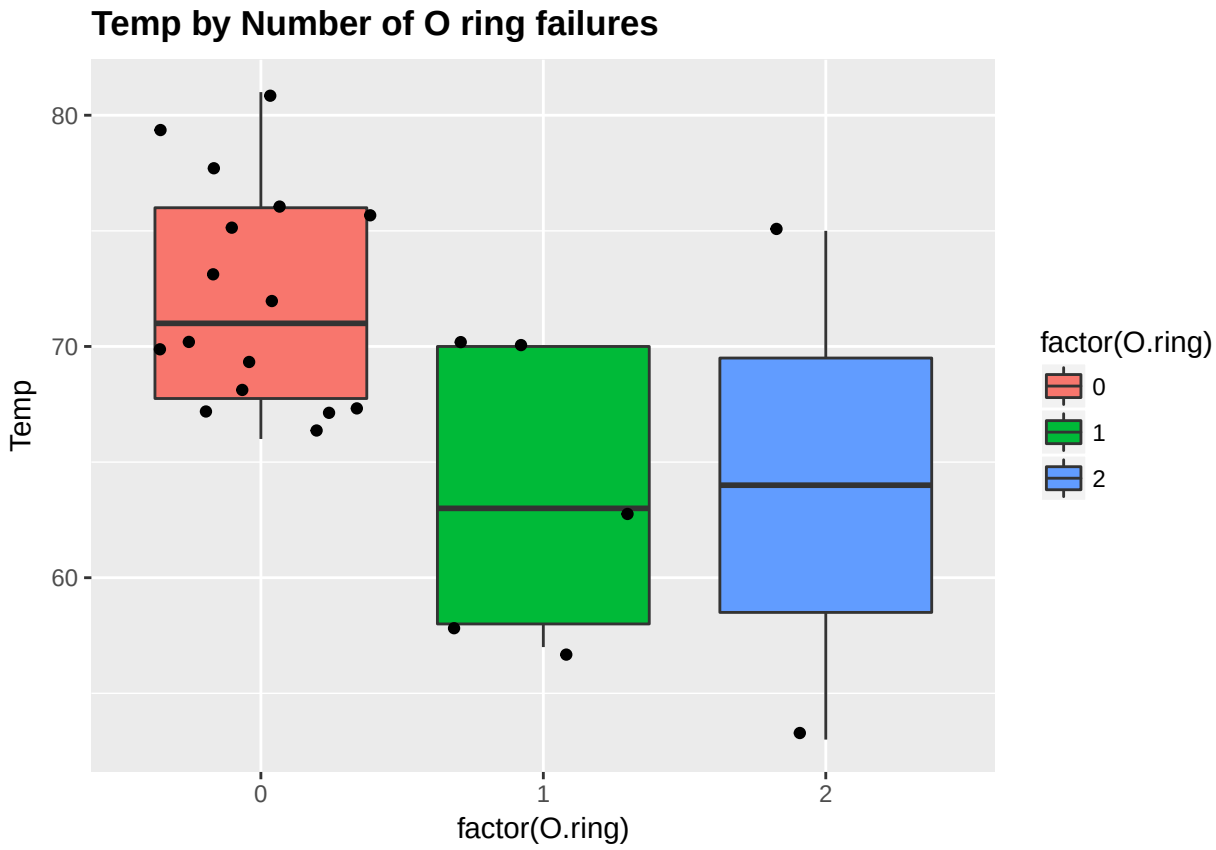
```
##  5  Variables      23  Observations
## ---------------------------------------------------------------------------
## Flight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       23        0       23        1       12        8      2.1      3.2
##      .25      .50      .75      .90      .95
##      6.5     12.0     17.5     20.8     21.9
##
## lowest :  1  2  3  4  5, highest: 19 20 21 22 23
## ---------------------------------------------------------------------------
## Temp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       23        0       16    0.992    69.57    7.968     57.1     59.0
##      .25      .50      .75      .90      .95
##     67.0     70.0     75.0     77.6     78.9
##
## Value          53    57    58    63    66    67    68    69    70    72
## Frequency       1     1     1     1     1     3     1     1     4     1
## Proportion 0.043 0.043 0.043 0.043 0.043 0.130 0.043 0.043 0.174 0.043
##
## Value          73    75    76    78    79    81
## Frequency       1     2     2     1     1     1
## Proportion 0.043 0.087 0.087 0.043 0.043 0.043
## ---------------------------------------------------------------------------
## Pressure
##        n  missing distinct     Info     Mean      Gmd
##       23        0        3    0.706    152.2    67.59
##
## Value          50   100   200
## Frequency       6     2    15
## Proportion 0.261 0.087 0.652
## ---------------------------------------------------------------------------
## O.ring
##        n  missing distinct     Info     Mean      Gmd
##       23        0        3    0.654   0.3913   0.6087
##
## Value          0     1     2
## Frequency     16     5     2
## Proportion 0.696 0.217 0.087
## ---------------------------------------------------------------------------
## Number
##        n  missing distinct     Info     Mean      Gmd
##       23        0        1        0        6        0
##
## Value       6
## Frequency  23
## Proportion  1
## ---------------------------------------------------------------------------
```

Relationships Between Time Series

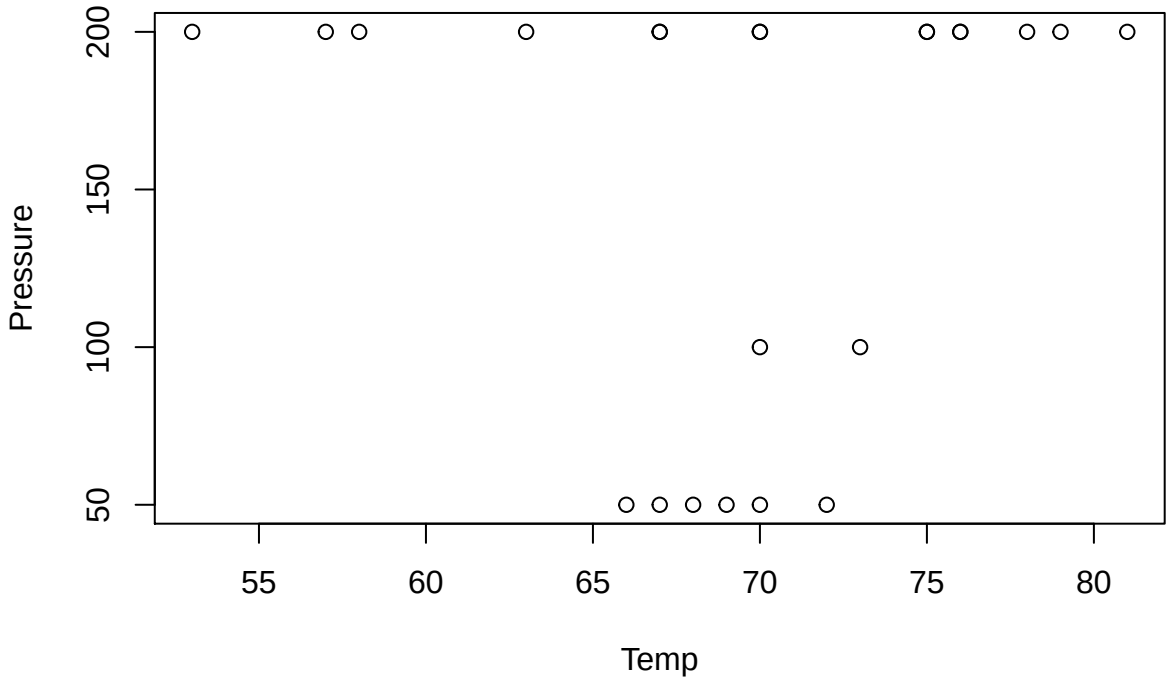
```
plot(O.ring ~ Temp, data = df)
```




```
ggplot(df, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  geom_jitter() +
  ggtitle("Temp by Number of O ring failures") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```

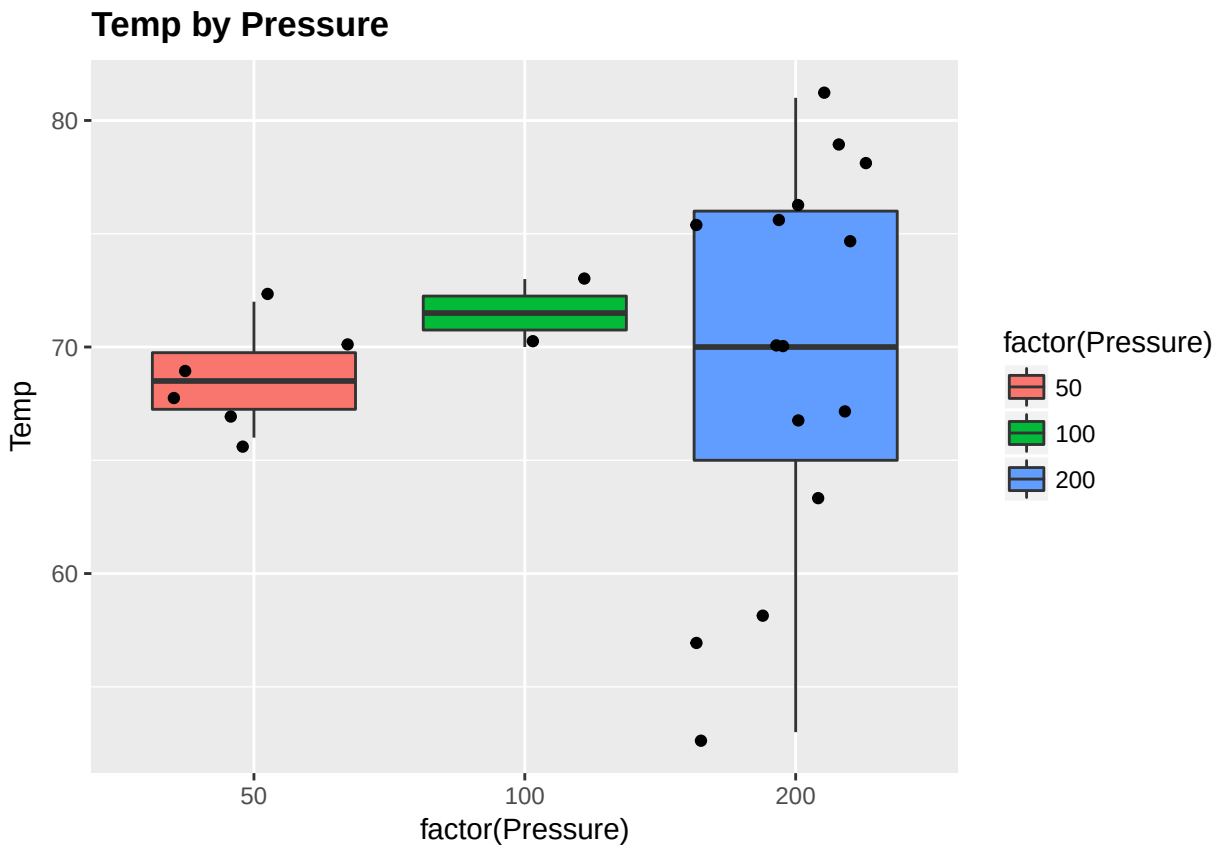



```
plot(Pressure ~ Temp, data = df)
```

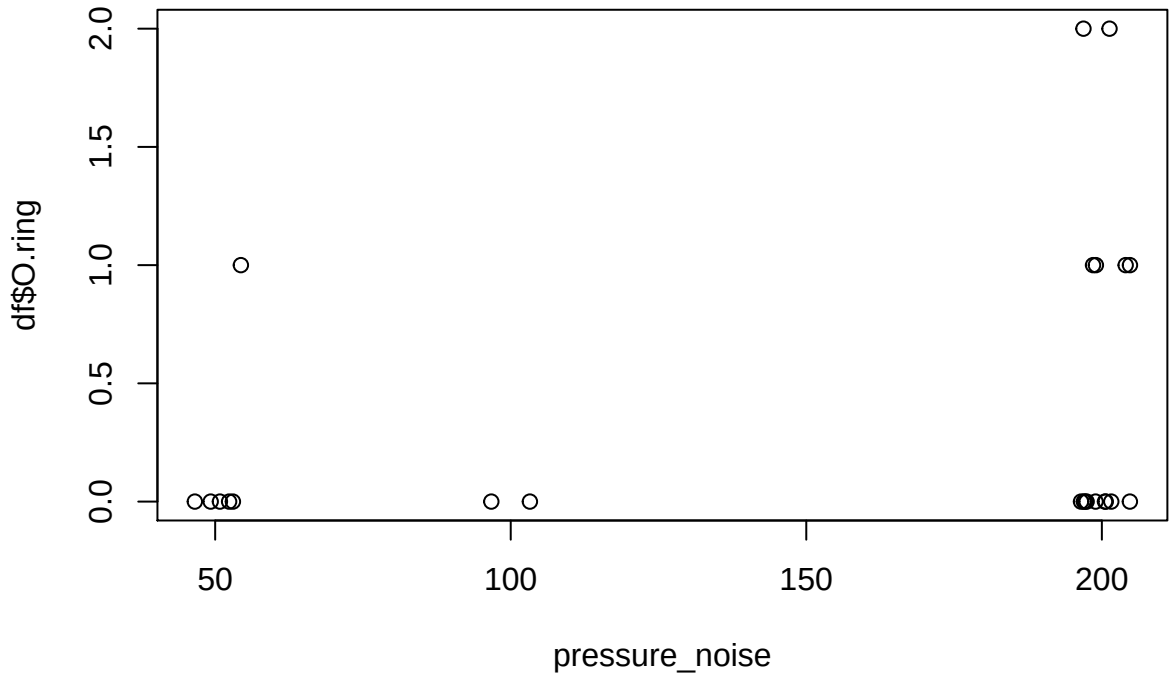



```
ggplot(df, aes(factor(Pressure), Temp)) +
  geom_boxplot(aes(fill = factor(Pressure))) +
  geom_jitter() +
```

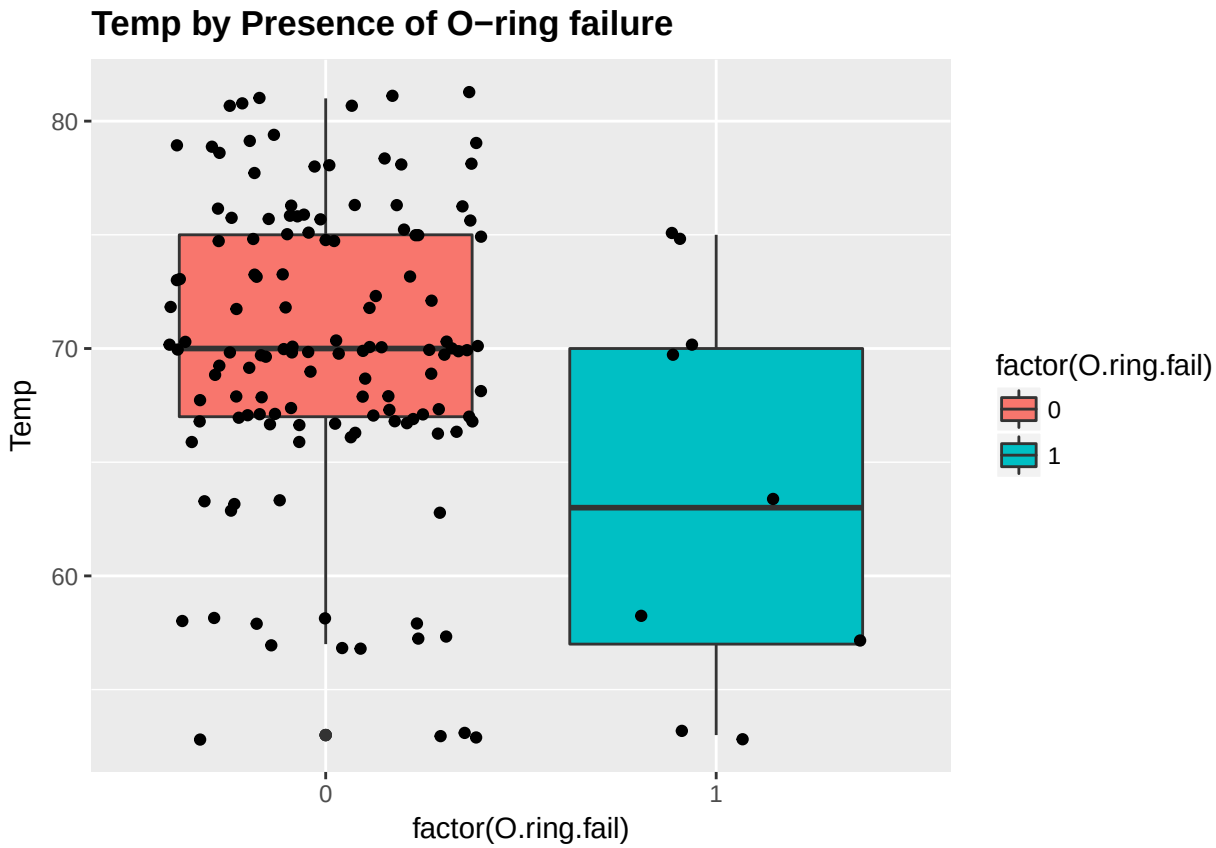
```
  ggtitle("Temp by Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
noise <- runif(length(df$Pressure), min=-5, max = 5)
pressure_noise <- df$Pressure + noise
plot(pressure_noise, df$O.ring)
```




```
ggplot(df2, aes(factor(O.ring.fail), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring.fail))) +
  geom_jitter() +
  ggtitle("Temp by Presence of O-ring failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```


Temp by Presence of O−ring failure
80
70
60
Temp
0
1
factor(O.ring.fail)
factor(O.ring.fail)
0
1

Answer to questions 4 and 5 on Chapter 2 (page 129 and 130) of Bilder and Loughin's *"Analysis of Categorical Data with R"*

4)

a. The authors assume that for each trial, the probability of failure for each of the 6 O-rings is independent. This is necessary to validate the use of the binomial distribution for the probability of failure. The binomial distribution assumes that the success/failure of each trial is independent, and in this case trials correspond to different O-rings in the same test. If the binomial distribution is not accurate, then this means the interpretation of the logistic regression implying the odds of success/failure for each O-ring is invalid. Conceivably, the failure of one O-ring may contribute to some structural damage that causes other O-rings to fail, violating the independence assumption. On the other side, the success of the primary O-ring may diminish the likelihood of failure of the second O-ring, if it does not experience the same conditions. Furthermore, there may be other variables that influence the quality of the O-rings or their likelihood of failure, for example related to their production, that violates the independence of O-rings on a given flight or different flights.

b. Base model of probability of single O.ring failures modeled on linear relationship of temperature and pressure. In df, we model each observation outcome using the count of O-ring failures, in the `O.ring` var, over the total number of O-rings, in the `Number` var which is always 6, as a binomial random variable. Similarly we check that this is the same as counting each O-ring as its own observation representing a bernoulli random variable with probability representing probability of its failure. These are indeed identical.

```
model1.binom <- glm(O.ring/Number ~ Temp + Pressure, data = df, family = binomial,weights=Numb
model1.bern <- glm(O.ring.fail ~ Temp + Pressure, data = df2, family = binomial)
summary(model1.binom)
```

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

```
summary(model1.bern)
```

```
##
## Call:
## glm(formula = O.ring.fail ~ Temp + Pressure, family = binomial,
##     data = df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7940  -0.3670  -0.2500  -0.2162   2.8127
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486822   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: 66.540  on 137  degrees of freedom
## Residual deviance: 58.856  on 135  degrees of freedom
## AIC: 64.856
##
## Number of Fisher Scoring iterations: 6
```

Thus, we stick with the first as our model1:

```
model1 <- glm(O.ring/Number ~ Temp + Pressure, data = df, family = binomial,weights=Number)
summary(model1)
```

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

c. We perform likelihood ratio tests using this model as our alternative hypothesis and the two reduced models setting the coeffs for temp and pressure respectively to zero, then conducting the ANOVA tests using the chi-squared distribution as follows:

```
ha <- model1
h0 <- glm(O.ring/Number ~ Pressure, data = df, family = binomial, weights = Number)
anova(h0, ha, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Pressure
## Model 2: O.ring/Number ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     21.730
## 2        20     16.546  1   5.1838   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
h0 <- glm(O.ring/Number ~ Temp, data = df, family = binomial, weights = Number)
anova(h0, ha, test = "Chisq")
```

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

Thus we see that the inclusion of `Temp` in the model is significant at the alpha=0.05 level, whereas the inclusion of `Pressure` is not even marginally significant.

d. Given the lack of statistical significance of the pressure variable in the model here it certainly validates the authors decision to remove this variable from the model, however it is also reasonable to suggest that further testing may have still been warranted. It is important to keep in mind that we are assuming that the relationship with pressure is linear here, but some transformation may be relevant here, e.g. a log transformation or a translation given the note in the paper that the puddy covers pressure of 50 PSI and thus it may be that only additional pressure should be considered relevant to O-ring failure.

5)

a. The model on `Temp` alone corresponds to the second `h0` model above:

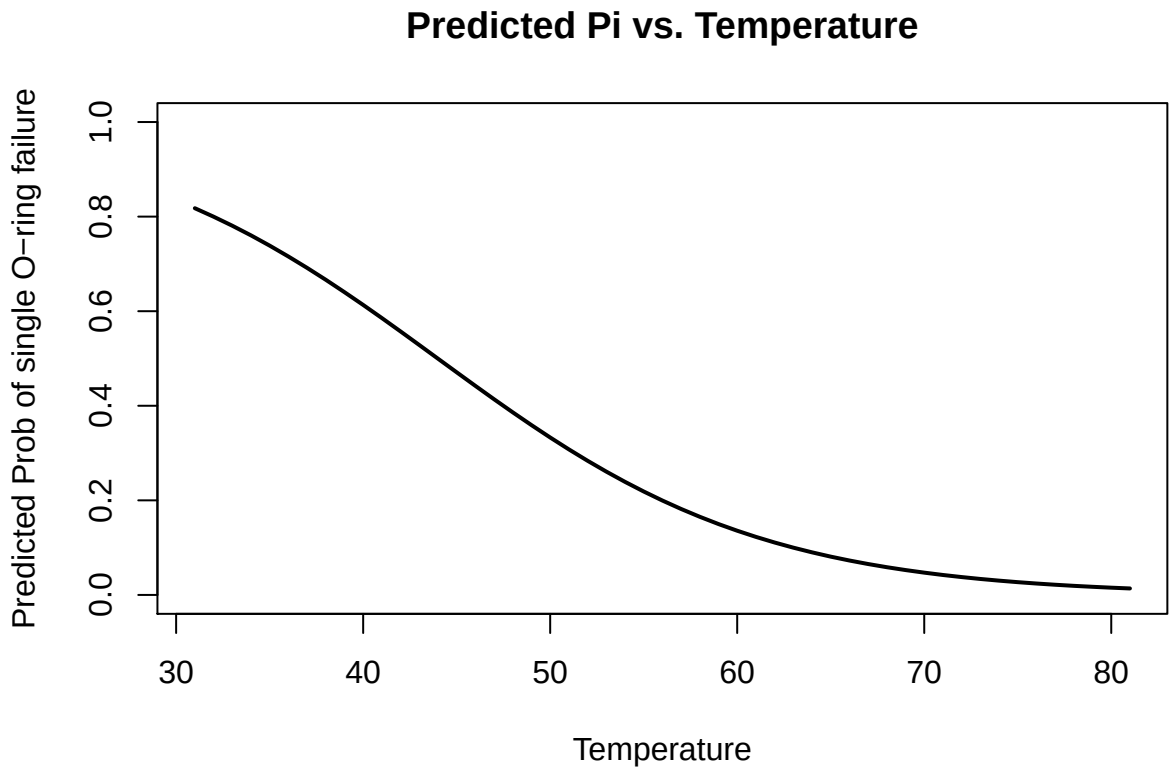
```
model2 <- h0
summary(model2)
```

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

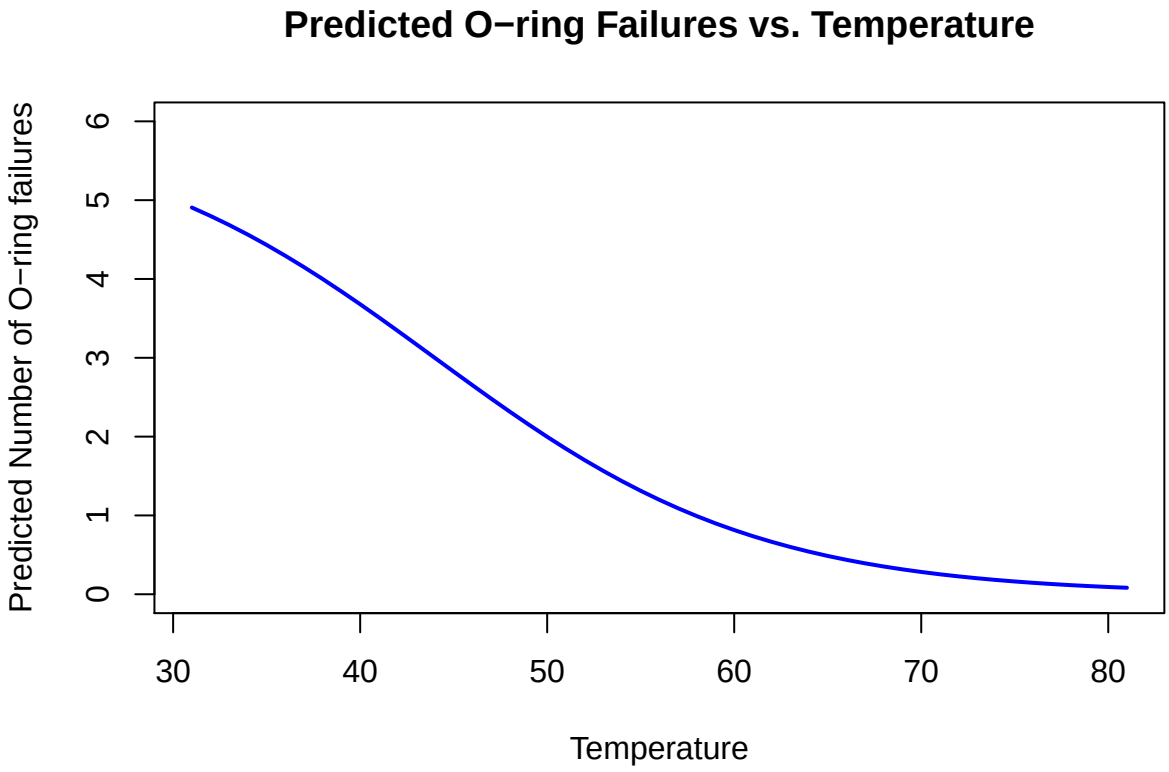
Using only a linear predictor on the `Temp` variable for the log-odds of yields an intercept of 5.085 and a coefficient for `Temp` of -0.116, which is significant at the 0.05 level.

b. (Jessica) Plot

```r
#pi vs. Temp  ## QUESTION - the books says to plot _pi_ vs temp, not pi hat. I assume they're
newdf <- data.frame(Temp = seq(from = 31, to = 81, by = 1))

lp.hat <- predict.glm(model2, newdata = newdf, type = "link", se.fit = TRUE)
lp.hat.mean <- lp.hat$fit
pi.hat <- exp(lp.hat.mean) / (1 + exp(lp.hat.mean))

plot(newdf$Temp, pi.hat, ylim = range(c(0,1)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main = "Predicted Pi vs. Temperature", type = 'l', col = 'black', lwd = 2)
```

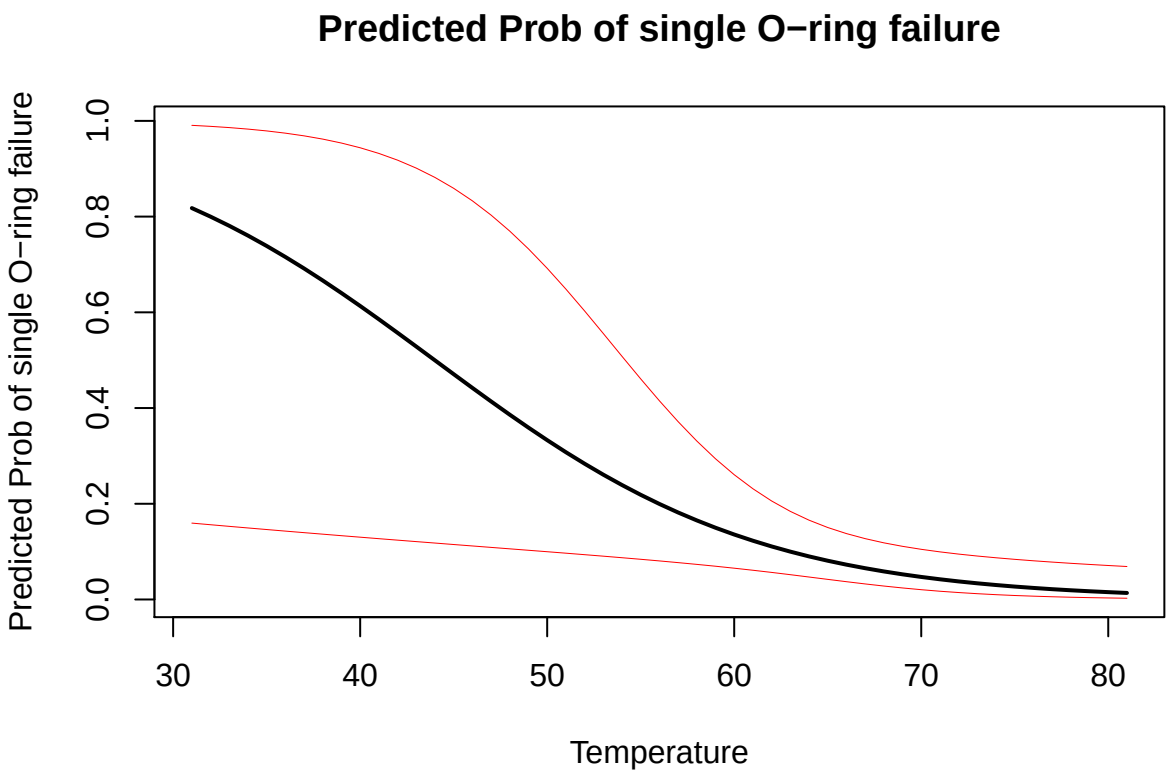



```r
#expected number of failures vs. Temp

plot(newdf$Temp, pi.hat * 6, ylim = range(c(0,6)),
     xlab = "Temperature", ylab = "Predicted Number of O-ring failures",
     main = "Predicted O-ring Failures vs. Temperature", type = 'l', col = 'blue', lwd = 2)
```

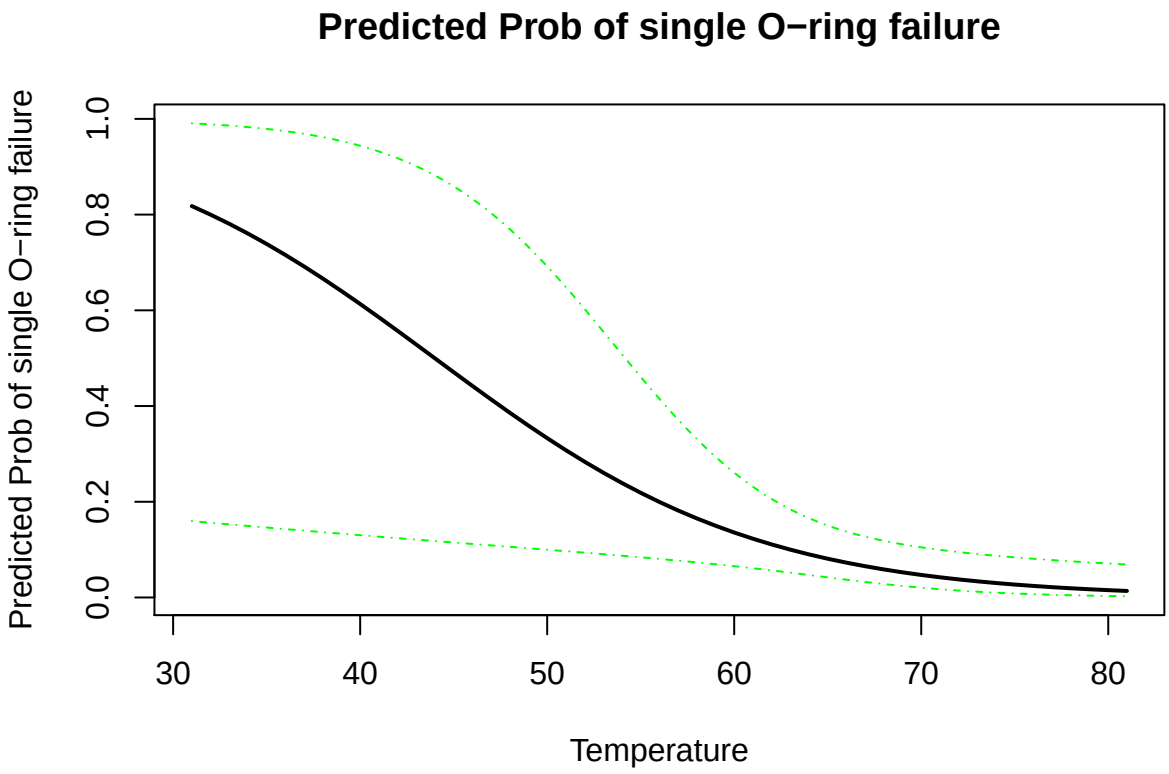


c. (Jessica) Plot. The bands are wider for lower temperature because there are very few observations in this region.

```
#jeff's way
lp.hat.lci <- lp.hat$fit - 1.96 * lp.hat$se.fit
lp.hat.uci <- lp.hat$fit + 1.96 * lp.hat$se.fit


pi.hat.lci <- exp(lp.hat.lci) / (1 + exp(lp.hat.lci))
pi.hat.uci <- exp(lp.hat.uci) / (1 + exp(lp.hat.uci))

### Plot predicted probabilities
plot(newdf$Temp, pi.hat, ylim = range(c(pi.hat.lci, pi.hat.uci)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main= "Predicted Prob of single O-ring failure", type = 'l', col = 'black', lwd = 2)
lines(newdf$Temp, pi.hat.lci, col = 'red', lwd = 0.5)
lines(newdf$Temp, pi.hat.uci, col = 'red', lwd = 0.5)
```




```
#book way
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)
}

plot(newdf$Temp, pi.hat, ylim = range(c(pi.hat.lci, pi.hat.uci)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main= "Predicted Prob of single O-ring failure", type = 'l', col = 'black', lwd = 2)
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = model2, alpha = 0.05)$lower, c
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = model2, alpha = 0.05)$upper, c
```



d. (Alice) Key assumption being made here is that there is a linear relationship between the temperature and the log-likelihood of O-ring failure. It is possible that either assumption is invalid, i.e. the logit is not the proper link-function for this relationship or there is a nonlinear relationship between temperature and the logit of the probability of O-ring failure.

e.(Marshall) Bootstrap

f. We include the quadratic term on temperature and run a LRT using the chi-squared distribution to determine if its inclusion is statistically significant, as follow:

```
model3 <- glm(O.ring/Number ~ Temp + I(Temp^2), data = df, family = binomial, weights = Number)
summary(model3)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + I(Temp^2), family = binomial,
##     data = df, 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
```

```
ha <- model3
anova(h0, ha, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Temp
## Model 2: O.ring/Number ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     18.086
## 2        20     17.592  1   0.4947   0.4818
```

The quadratic term addition to the model is not statistically significant, suggesting that either it shouldn't be included or some other variable transformations or terms should be conducted/tested first.

3. In addition to the questions in Question 4 and 5, answer the following questions:

a. Interpret the main result of your final model in terms of both odds and probability of failure

b. With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case. Why? Or, why not?