# Lab 1 VHE

*Victoria Eastman*

*September 19, 2018*

## Notes from Async videos

- always good to write out the model after it's been estimated (ie logit(pi) = 0.5 + 5good + 3frank etc.)

### Initial EDA

Problem statement:

```
# Import libraries
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(Hmisc))

setwd("/home/victoriaeastman/berkeley/w271/w271_lab1")
data <- read.csv("challenger.csv")
```

```
glimpse(data)
```

```
## Observations: 23
## Variables: 5
## $ Flight   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Temp     <int> 66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 5...
## $ Pressure <int> 50, 50, 50, 50, 50, 50, 100, 100, 200, 200, 200, 200,...
## $ O.ring   <int> 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 2, 0, 0, 0, 0,...
## $ Number   <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6,...
```

```
describe(data)
```

```
## data
##
##  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
## ---------------------------------------------------------------------------
```

```r
# I'm curious about the value counts for o-ring failures
table(data$O.ring)
```

```
##
##  0  1  2
## 16  5  2
```

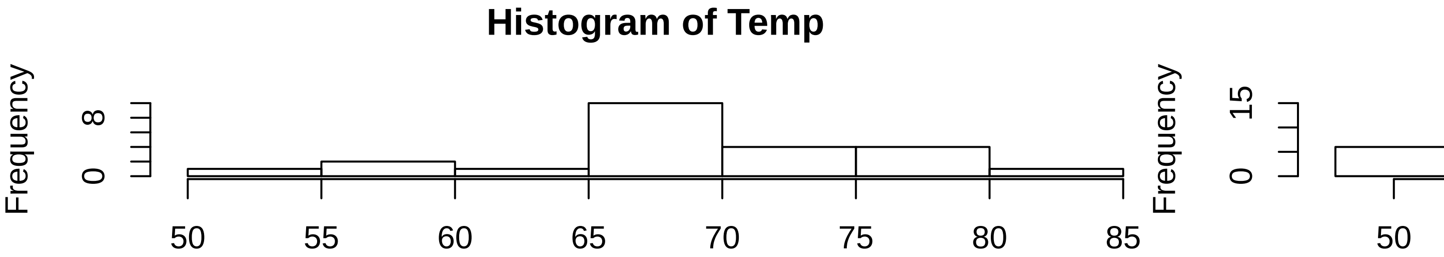
Initial findings:

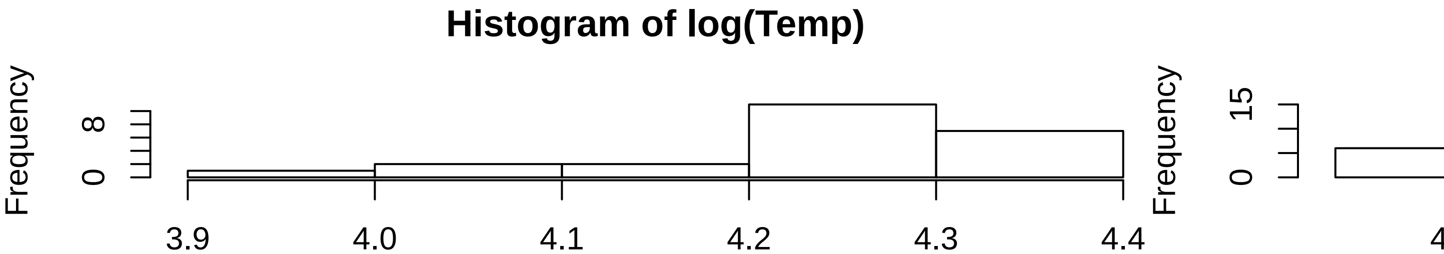
- 23 data points with no missing values for any variables
- Dependent variable, O.ring, is categorical and takes three values: 0, 1, and 2 representing the number of o-ring failures on space launches. The mean value is 0.3913 which means the data is skewed towards 0 o-ring failures. Futher investigation shows there were 2 flights with 2 o-ring failures, 5 with 1 failure, and 16 with no failures.
- The explanatory variables are as follows:
  - Temp: temperature at launch (degrees F)
  - Pressure: Combustion pression (psi)

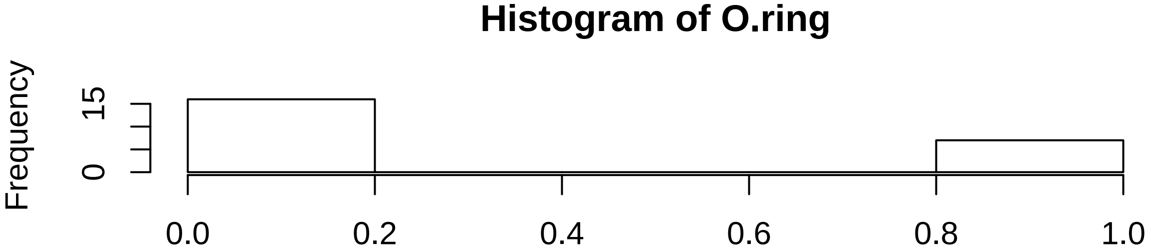
The goal of this study is to estimate a logistic regression so we are going to recategorize the O.ring variable as 0 for no failures and 1 for *at least 1* failure.

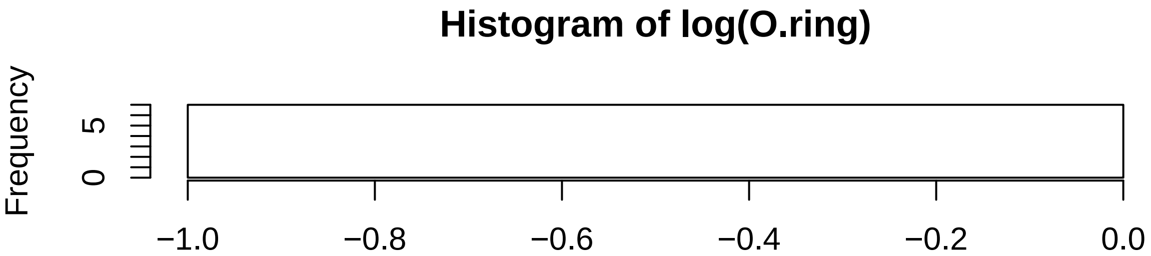
```r
# Change the O.ring variable
data$O.ring[data$O.ring >= 1] = 1
```

```r
# histogram of explanatory variables
for (i in 2:4){
  par(mfrow = c(2,1))
  hist(as.numeric(data[,i]), main=paste0("Histogram of ", colnames(data)[i]), xlab=NA)
  hist(as.numeric(log(data[,i])), main=paste0("Histogram of log(", colnames(data)[i], ")"), xlab=NA)
  #hist(as.numeric(data[,i]^2), main=paste0("Histogram of log(", colnames(data)[i], ")"), xlab=NA)
}
```


Histogram of Temp

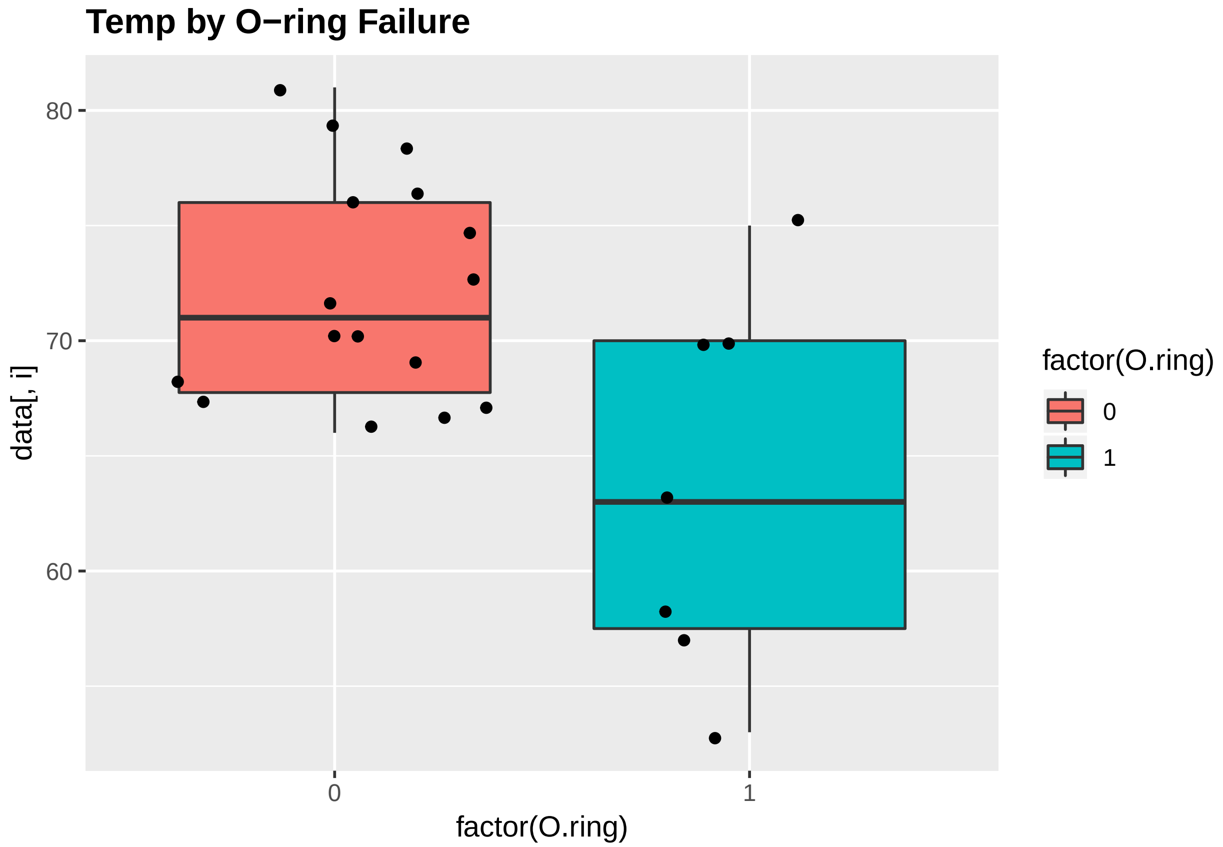

Histogram of log(Temp)

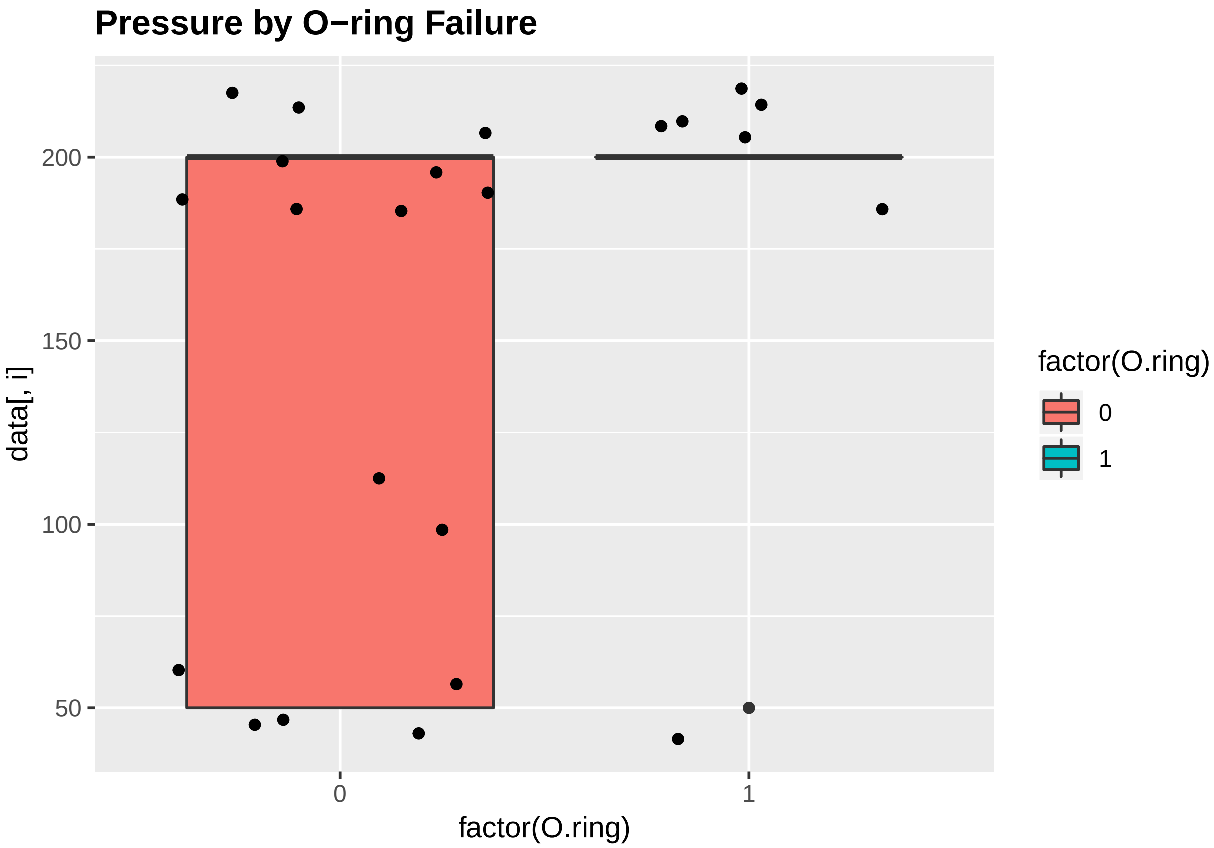

Histogram of O.ring


Histogram of log(O.ring)

The distribution of the temperature variable is fairly close to a normal distribution and does not appear to become closer to a normal distribution after a log transformation. Neither the Pressure or O.ring variables are close to a normal distribution and are not improved by a log transformation. Thus it seems like the variables should be left in their un-tranformed state.

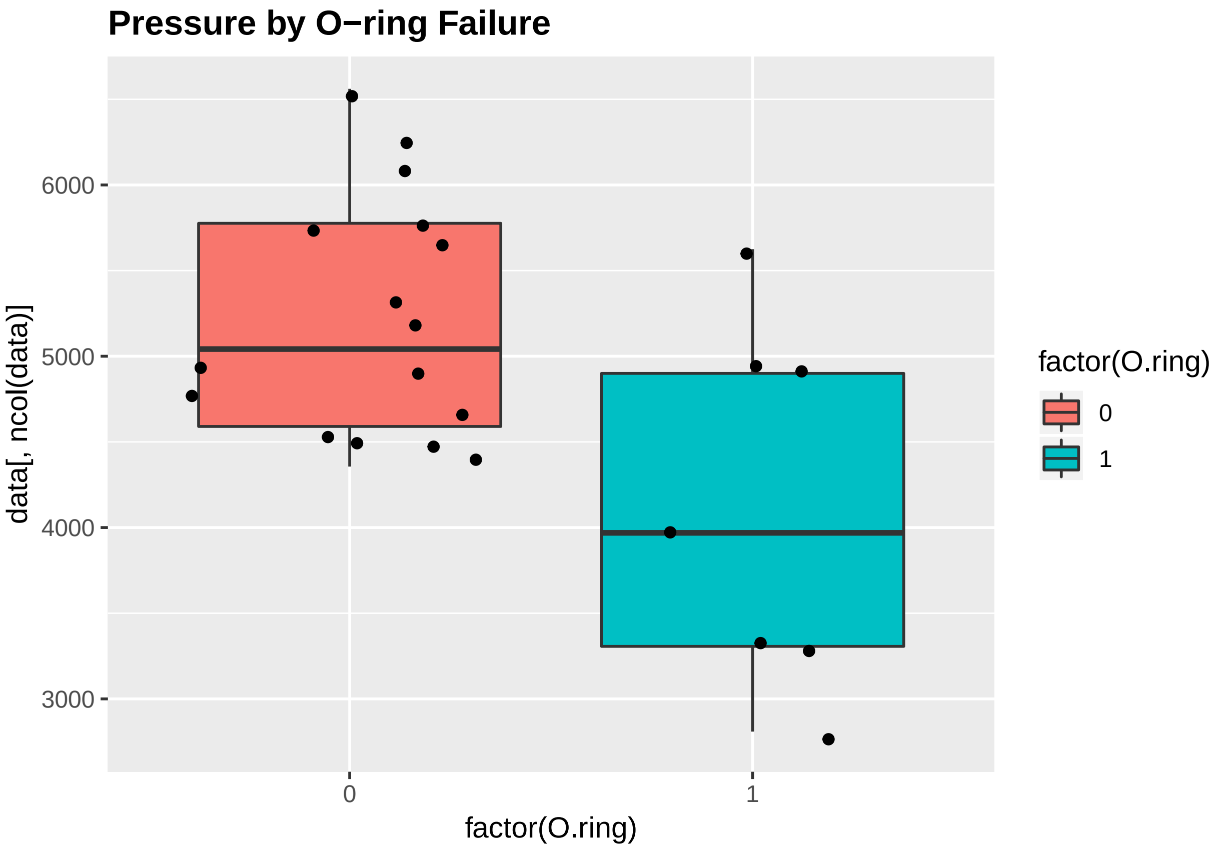
```
for (i in 2:3){
  print(ggplot(data, aes(factor(O.ring), data[,i])) +
          geom_boxplot(aes(fill = factor(O.ring))) +
          geom_jitter() +
          ggtitle(paste0(colnames(data)[i], " by O-ring Failure")) +
          theme(plot.title = element_text(lineheight=1, face="bold")))
}
```





The first box-plot clearly shows

```
data$tempsqr = data$Temp^2
# Visualize interaction variables
ggplot(data, aes(factor(O.ring), data[,ncol(data)])) +
          geom_boxplot(aes(fill = factor(O.ring))) +
          geom_jitter() +
          ggtitle(paste0(colnames(data)[i], " by O-ring Failure")) +
          theme(plot.title = element_text(lineheight=1, face="bold"))
```


Pressure by O−ring Failure
data[, ncol(data)]
6000
5000
4000
3000
0
1
factor(O.ring)
factor(O.ring)
0
1