

## Problem 1

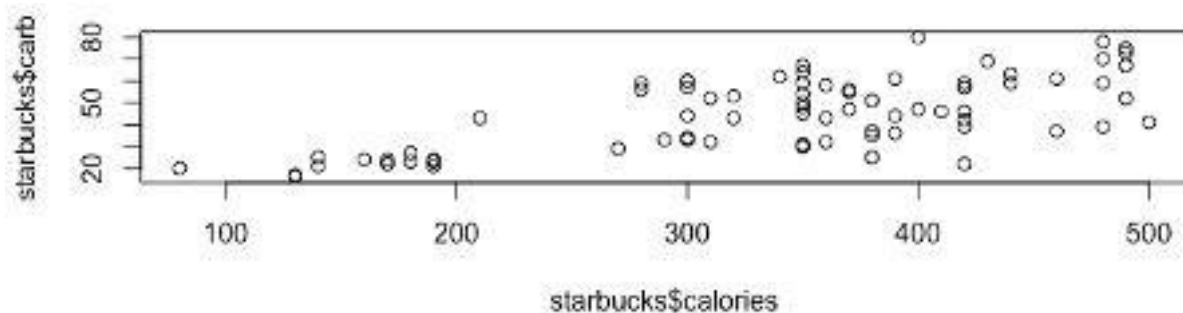
**1.1 Create a scatterplot of this data with calories on the x-axis and carbohydrate grams on the y-axis, and describe the relationship you see.**

**Code:**

```
> if(!require('openintro')) {  
+ install.packages('openintro')  
+ library(openintro)  
+ }  
> if(!require('lattice')) {  
+ install.packages('lattice')  
+ library(lattice)  
+ }  
> head(starbucksDF)
```

```
plot(starbucks$calories,starbucks$carb)
```

As you can see here, as the calories increase, the carbohydrates increase. I predict for this reason, there will be a positive correlation between calories and carbohydrates.



**1.2 In the scatterplot you made, what is the explanatory variable? What is the response variable? Why might you want to construct the problem in this way?**

The explanatory variable in this scatterplot is the calories because it is the dependent variable (x). The response variable is the carbohydrates (y-axis) because it is responding to the change in calories. We are using the calories (the explanatory variable) to predict what the carbohydrates will be (the response variable).

### 1.3 Fit a simple linear regression to this data, with carbohydrate grams as the dependent variable and the calories as the explanatory variable. Use the lm() function.

Code:

```
plot(starbucks$calories,starbucks$carb)
starbucks.lm <- lm(starbucksDF$carb ~ starbucksDF$calories, data = starbucks)
summary(starbucks.lm)
# with(starbucksDF, plot(starbucksDF$calories, starbucksDF$carb)) with(starbucksDF,
plot(starbucksDF$calories, starbucksDF$carb, xlab='Calories', ylab='Carbs',
main='Starbucks Calorie & Carb Linear Regression')) abline(starbucks.lm,col="pink")
```

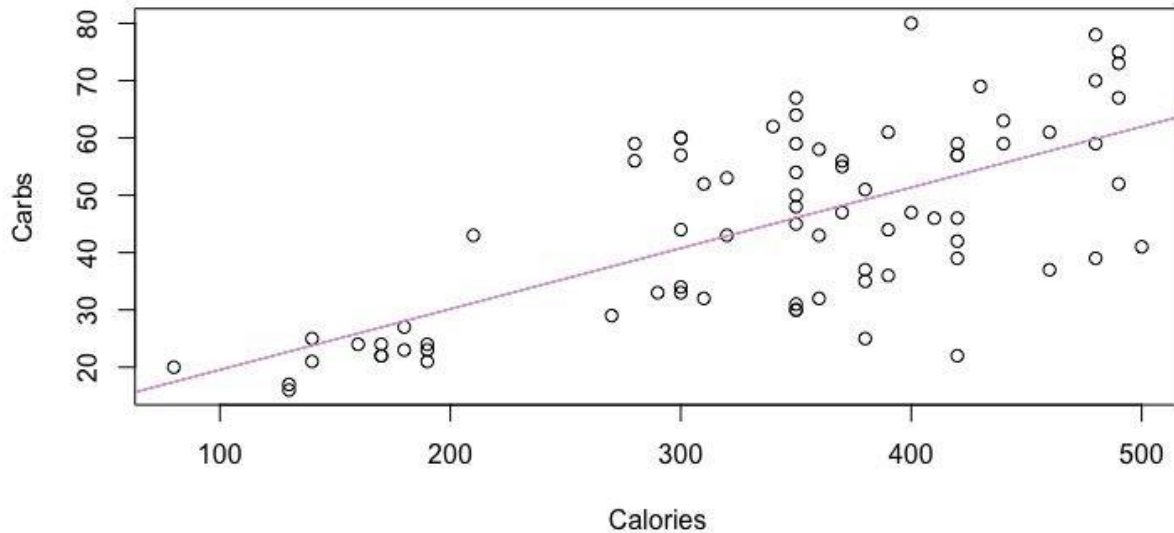
```
Call:
lm(formula = starbucksDF$carb ~ starbucksDF$calories, data = starbucks)

Residuals:
    Min       1Q   Median       3Q      Max
-31.477  -7.476  -1.029   10.127   28.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.94356    4.74600    1.884  0.0634 .
starbucksDF$calories  0.10603    0.01338    7.923 1.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.29 on 75 degrees of freedom
Multiple R-squared:  0.4556,    Adjusted R-squared:  0.4484
F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

### Starbucks Calorie & Carb Linear Regression



#### 1.4 Write the fitted model out using mathematical notation. Interpret the slope and the intercept parameters.

Fitted model using mathematical notation:

$$y = 8.94 + 0.106x$$

- **Slope interpretation:** For every additional calorie, carbohydrates increase by .106 on Starbucks' menu.
- **Intercept interpretation:** When there are 0 calories, the expected value of carbohydrates in Starbucks' menu is 8.94. However, it's important to note that the standard error is 4.76.

#### 1.5 Find and interpret the value of $R^2$ for this model.

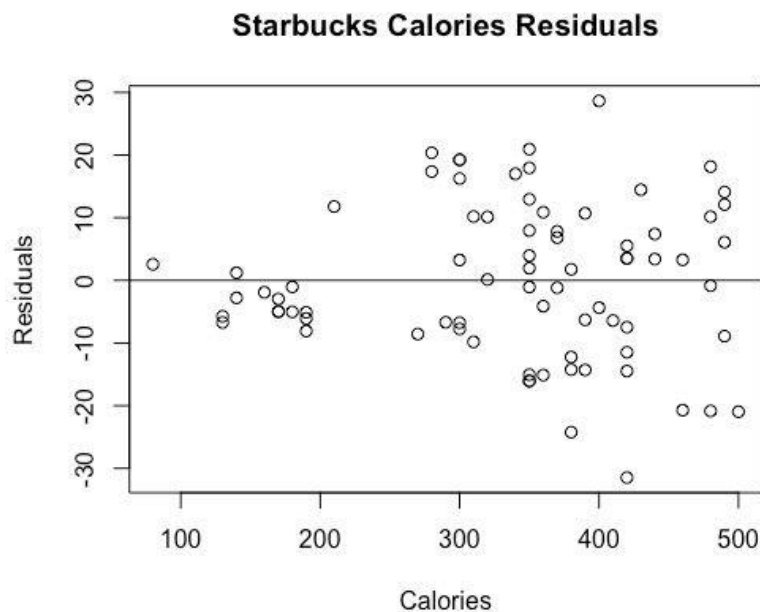
Multiple R-squared: 0.4556, Adjusted R-squared: 0.4484

Because  $R^2$  measures how close each data point fits to the regression line.  $R^2$  usually gives one a good understanding on how "good of a prediction" your regression line is. It also shows us the account of variation in  $y$ , in this case the amount of variation in carbohydrates, that is accounted for in the regression. In this case, 44.8% of the variation in carbohydrates is accounted for by its regression on calories.

**1.6 Create a residual plot. The ggplot2 function fortify can help a lot with this. Describe what you see in the residual plot. Does the model look like a good fit?**

**Code:**

```
# with(starbucksDF, plot(starbucksDF$calories, starbucksDF$carb)) with(starbucksDF,
plot(starbucksDF$calories, starbucksDF$carb, xlab='Calories', ylab='Carbs',
main='Starbucks Calorie & Carb Linear Regression')) abline(starbucks.lm,col="blue")
starbucks.res = resid(starbucks.lm)
plot(starbucksDF$calories, starbucks.res,
      ylab="Residuals", xlab="Calories",
      main="Starbucks Calories Residuals")
abline(0, 0)
```



It's clear within this residual plot that there is no real symmetrical similarity between the residuals and the plotting shows a large difference between the regression line and the actual data point. Additionally, one can observe that as the calories increase, the accuracy between the carbs predictions becomes less and less accurate (this can be seen with the data points getting farther and farther away from the line). This indicates that the regression model above is not as accurate as one would like it to be.

**Problem 2**

**2.1 Convert the Eth, Sex, and Lrn variables to binary variables. One way to do this is with the function ifelse(). You should construct them so that**

- 1. Eth = 1 if the student is not aboriginal and Eth = 0 if the student is aboriginal;**
- 2. Sex = 1 if the student is male and Sex = 0 if the student is female;**
- 3. Lrn = 1 if the student is a slow learner and Lrn = 0 if the student is an average learner.**

**Code:**

```
install.packages("tidyverse")
library(tidyverse)
library(tidyverse)
newdata <- data_frame(Eth = c(1,1,1,0,0), Sex =
c(0,1,0,1,0),
Lrn = c(0,0,1,1,0))
data("absenteeism")
head(absenteeism)
```

```
absenteeism <- absenteeism %>%
mutate(eth = ifelse(eth == "N", 1, 0),
sex = ifelse(sex == "M", 1, 0),
lrn = ifelse(lrn == "SL", 1, 0))
```

**2.2 Fit a linear model to the data with Days as the dependent variable and the three variables mentioned in (1) as explanatory variables.**

**Code:**

```
absenteeism.lm <- lm(absenteeism$days ~ absenteeism$eth + absenteeism$sex +
absenteeism$lrn, data = absenteeism)
summary(absenteeism.lm)
```

```

> summary(model)

Call:
lm(formula = days ~ eth + sex + lrn, data = absenteeism)

Residuals:
    Min       1Q   Median       3Q      Max
-22.190 -10.078  -4.928   5.768  59.914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.932     2.570   7.365 1.32e-11 ***
eth          -9.112     2.599  -3.506 0.000609 ***
sex           3.104     2.637   1.177 0.241108
lrn           2.154     2.651   0.813 0.417732
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.67 on 142 degrees of freedom
Multiple R-squared:  0.08933, Adjusted R-squared:  0.07009
F-statistic: 4.643 on 3 and 142 DF, p-value: 0.003967

```

**2.3 Write the fitted model out using mathematical notation. Interpret all of the fitted values in context.**

**2.3**

Mathematical equation:

$$Y = 18.932 - 9.112x_1 + 3.104x_2 + 2.154x_3$$

$x_1 = \text{eth}$

$x_2 = \text{sex}$

$x_3 = \text{lrn}$

For those that do come from Australia (are aboriginal), who are women, and are average learners, these students are expected to miss an average of 18.392 days of school. On the contrary, if everything is on the opposite end (students are from Australia, are males, and are slow learners), these students are expected to miss around 14 days of school.

**2.4**

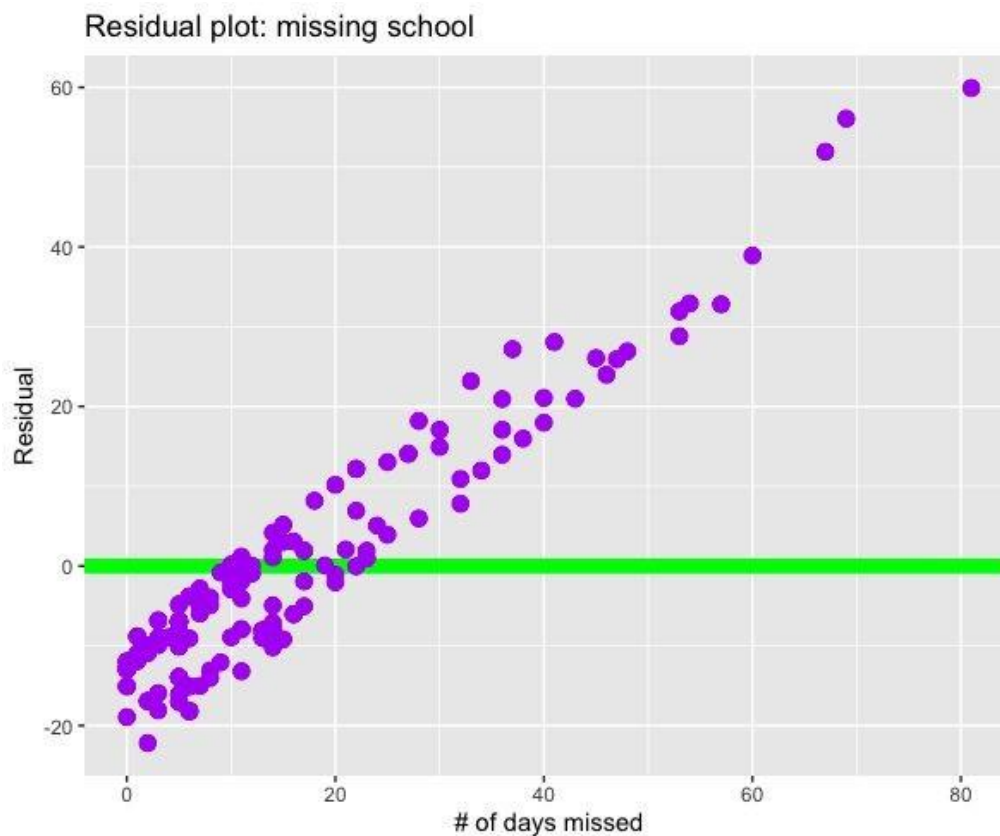
Intercept- 18.392

Multiple R-squared: 0.08933, Adjusted R-squared: 0.07009

As described above, the intercept describes how many days would be missed if all X1, X2, and X3 variables were 0 (aka are not aboriginal, female, and average learners), the days missed would be 18.392. Additionally, 7.009% of the variance in the school days missed can be explained by the multiple linear regression model.

## 2.5 Create a residual plot. Describe what you see in the residual plot. Does the model look like a good fit?

```
dat <- fortify(model)
ggplot(data = dat) +
  geom_hline(yintercept = 0, color = "green", size = 3) +
  geom_point(aes(x = absenteeism$days, y = .resid), size = 3, color = "purple") + labs(x =
"# of days missed", y = "Residual", title = "Residual plot: missing school")
```



As one can see, as the days at school that are missed increase, the residual value and the difference between the regression line (predicted value) and actual data points continue to become larger and larger, and farther and farther away from the line. This indicates that this multiple linear regression model is not accurate/reliable and that we need to reconsider and do further research on what is causing these big changes.

**2.6 Below is some data on new children in the school system. Predict the number of days each student will be absent, and display these predictions with the new data in a table.**

```
library(tidyverse)
newdata <- data_frame(Eth = c(1,1,1,0,0),
                      Sex = c(0,1,0,1,0),
                      Lrn = c(0,0,1,1,0))
```

Code:

```
newdata <- data_frame(eth = c(1,1,1,0,0),
                    sex = c(0,1,0,1,0),
                    lrn = c(0,0,1,1,0))
newdata <- newdata %>%
  mutate(Days_Missed = predict(model, newdata = newdata))
view(newdata)
```

	eth	sex	lrn	Days_Missed
1	1	0	0	9.819607
2	1	1	0	12.923862
3	1	0	1	11.973764
4	0	1	1	24.190261
5	0	0	0	18.931848



### Problem 3

**3.1 Each row of the data represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.**

```
# A tibble: 23 x 2
  temp damage
  <int> <int>
1     53     5
2     57     1
3     58     1
4     63     1
5     66     0
6     67     0
7     67     0
8     67     0
9     68     0
10    69     0
# ... with 13 more rows
> |
```

As one can see in the above table, the most damage occurs at 53 degrees whereas, anything above 63 degrees has no damage at all.

**3.3 Using orings2, fit a logistic regression to the data using the glm function.**

**Code:**

```
fit <- glm(fail ~ temp, data=orings2, family='binomial')
summary(fit)
```

**Call:**

```
glm(formula = fail ~ temp, family = "binomial", data = orings2)
```

**Deviance Residuals:**

```
   Min      1Q  Median      3Q      Max
-1.2646 -0.3395 -0.2472 -0.1299  3.0216
```

**Coefficients:**

```
            Estimate Std. Error z value Pr(>|z|) (Intercept)
11.66299           3.29616  3.538 0.000403 ***
temp            -0.21623    0.05318 -4.066 4.77e-05 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 76.745 on 137 degrees of freedom  
Residual deviance: 54.759 on 136 degrees of freedom AIC:  
58.759

Number of Fisher Scoring iterations: 6

### 3.4 Write out the logistic model using the point estimates of the model parameters.

Logistical model

$$P = 11.66 - .021x$$

P = the probability of the rocket getting damaged

### 3.5 Interpret the coefficient for temperature: How does an increase of the temperature by 1 degree affect the odds that the O-ring in the shuttle will be damaged?

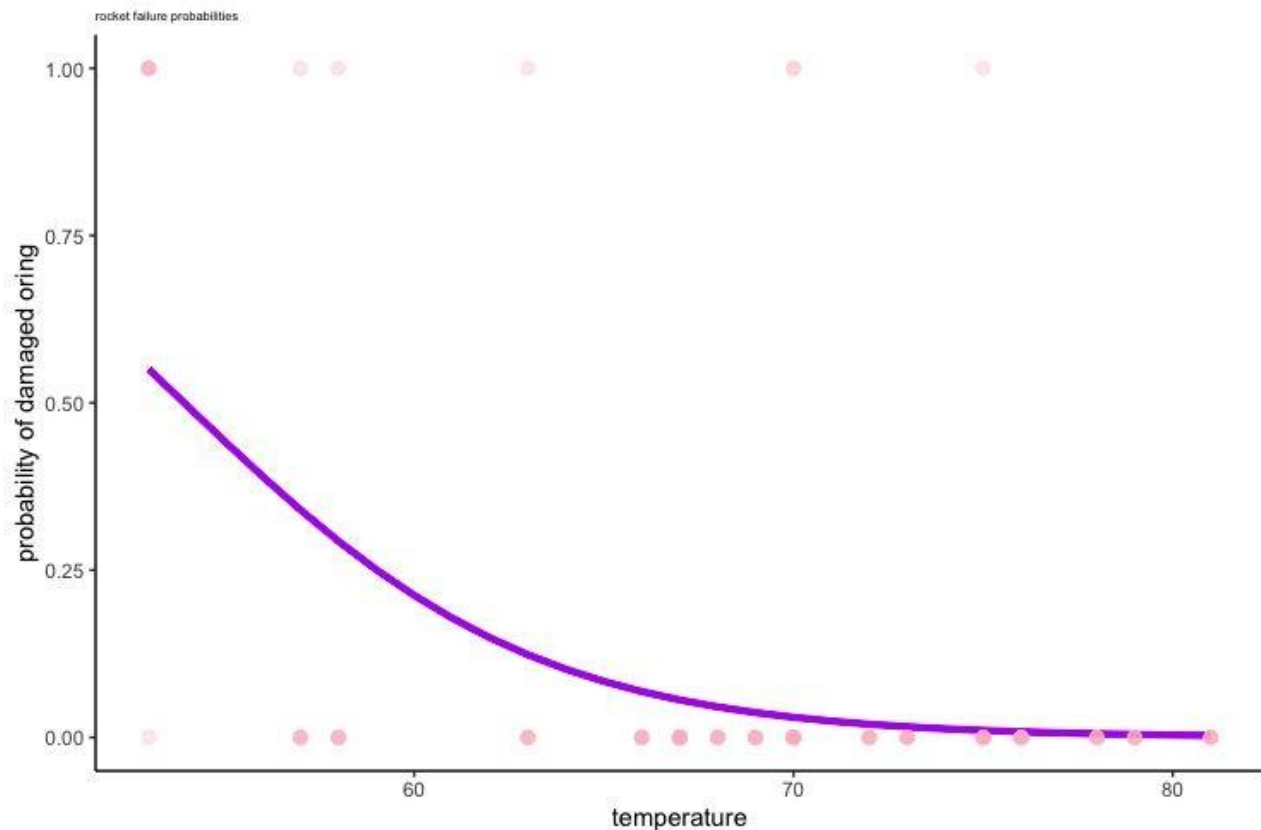
For every 1 unit increase in a degree in temperature, the log(odds of the O-ring getting damaged) decreases by 21%.

### 3.6 After the Challenger Explosion in January of 1986, an investigation was done to determine the cause. The investigation found that the explosion was caused by the failure of the O-rings due to the low temperature during the shuttle launch. Based on the model, do you think the investigation was correct? Explain.

Yes I think that it is certainly possible that the Challenger rocket was extremely damaged and failed because of low temperature. As one can observe in the logistic regression, the likelihood of a rocket failing is much higher when there is lower temperature, therefore, the conclusion that the Challenger failed because of lower temperature is supported by this logistic regression.

3.7 Predict the probability of a damaged O-ring for each temperature value from 53-81 (i.e. 53:81). Create a plot showing the observed data (in orings2) as points ( $x$ = temperature,  $y$  = fail) and draw a line through the predicted probabilities at each temperature value from 53-81. Describe what you see in the plot.

Code:



In the figure above, one can see that once again as mentioned earlier, as the temperature increases, the probability of the failure becomes to be at around 0%. On the other hand, in the 50 degree Fahrenheit range, the probability of failure is a little bit above 50% whereas everything above near the higher 60 degree range has close to a 0% chance of failing.

### My R-Script:

```
knitr::opts_chunk$set(echo = TRUE)
```

```
# Clear the console
cat("\014")
if(!require('openintro')) {
  install.packages('openintro')
  library(openintro)
}
```

```

if(!require('lattice')) {
  install.packages('lattice')
  library(lattice)
}
starbucksDF <- starbucks

head(starbucksDF)

plot(starbucks$calories,starbucks$carb)
starbucks.lm <- lm(starbucksDF$carb ~ starbucksDF$calories, data = starbucks)
summary(starbucks.lm)
# with(starbucksDF, plot(starbucksDF$calories, starbucksDF$carb))
with(starbucksDF, plot(starbucksDF$calories, starbucksDF$carb, xlab='Calories', ylab='Carbs',
main='Starbucks Calorie & Carb Linear Regression'))
abline(starbucks.lm,col="pink")

starbucks.res = resid(starbucks.lm)
plot(starbucksDF$calories, starbucks.res,
  ylab="Residuals", xlab="Calories",
  main="Starbucks Calories Residuals")
abline(0, 0)

data("absenteeism")
head(absenteeism)

install.packages("dplyr")
library(dplyr)
absenteeism <- absenteeism %>%
  mutate(eth = ifelse(eth == "N", 1, 0),
    sex = ifelse(sex == "M", 1, 0),
    lrn = ifelse(lrn == "SL", 1, 0))

model <- lm(days ~ eth + sex + lrn , data = absenteeism)
summary(model)

dat <- fortify(model)
ggplot(data = dat) +
  geom_hline(yintercept = 0, color = "green", size = 3) +
  geom_point(aes(x = absenteeism$days , y = .resid), size = 3, color = "purple") +
  labs(x = "# of days missed", y = "Residual", title = "Residual plot: missing school")

newdata <- tibble(eth = c(1,1,1,0,0),
  sex = c(0,1,0,1,0),
  lrn = c(0,0,1,1,0))

```

```
newdata <- newdata %>%
  mutate(Days_Missed = predict(model, newdata = newdata))
View(newdata)
```

```
knitr::opts_chunk$set(echo = TRUE)
```

```
# Clear the console
cat("\014")
if(!require('openintro')) {
  install.packages('openintro')
  library(openintro)
}
if(!require('lattice')) {
  install.packages('lattice')
  library(lattice)
}
```

```
data("orings")
data.frame(orings)
head(orings)
```

```
library(openintro)
data("orings")
orings2 <- NULL
for(i in 1:nrow(orings)){
  new <- data.frame(temp = orings$temp[i], # for each row in orings,
                   fail = rep(c(1,0), c(orings$damage[i], # create 6 new rows:
                                       6-orings$damage[i]))) # 1 for each launch
  orings2 <- rbind(orings2, new)
}
```

```
fit <- glm(fail ~ temp, data=orings2, family='binomial')
summary(fit)
```

```
install.packages("ggplot2")
install.packages("colorspace")
library(ggplot2)
install.packages("tidyverse")
```

```
install.packages("remotes")
remotes::install_github("CSAFE-ISU/csafethemes")
```

```
library(dplyr)
newdata <- data.frame(temp = 53:81)
newdata <- newdata %>%
  mutate(pred_prob = predict(fit, newdata = newdata, type = "response"))
ggplot(data = orings2) +
  geom_line(data = newdata, aes(x= temp , y= pred_prob), color = "purple",
    size = 1.5) +
  geom_point(aes(x = temp, y = fail), size = 2.5, color = "pink", alpha = .3) +
  theme_classic() +
  labs(x = "temperature", y = "probability of damaged oring",
    title = "rocket failure probabilities") +
  theme(plot.title = element_text(size = rel(.5)))
```