Lecture 12: Gelman Hill Ch 5

Logistic Regression

When the research goal is model or predict a binary outcome, modifications need to be made to the linear model framework that we have previously used.

Before, we had the following specification:

$$y_i = X_i\underline{\beta} + \varepsilon_i,$$

where $\varepsilon \sim N(0, \sigma^2)$ or equivalently $y_i \sim N(X\underline{\beta}, \sigma^2)$.

Logistic Regression with a single predictor

One of the classical datasets used for regression involved o-rings on the challenger space shuttle.

https://www.space.com/31732-space-shuttle-challenger-disaster-explained-infographic.html

From a NASA press release...

> On January 28, 1986, as the Space Shuttle Challenger broke up over the Atlantic Ocean 73 seconds into its flight, Allan McDonald looked on in shock – despite the fact that the night before, he had refused to sign the launch recommendation over safety concerns.
>
> McDonald, the director of the Space Shuttle Solid Rocket Motor Project for the engineering contractor Morton Thiokol, was concerned that below-freezing temperatures might impact the integrity of the solid rockets' O-rings.

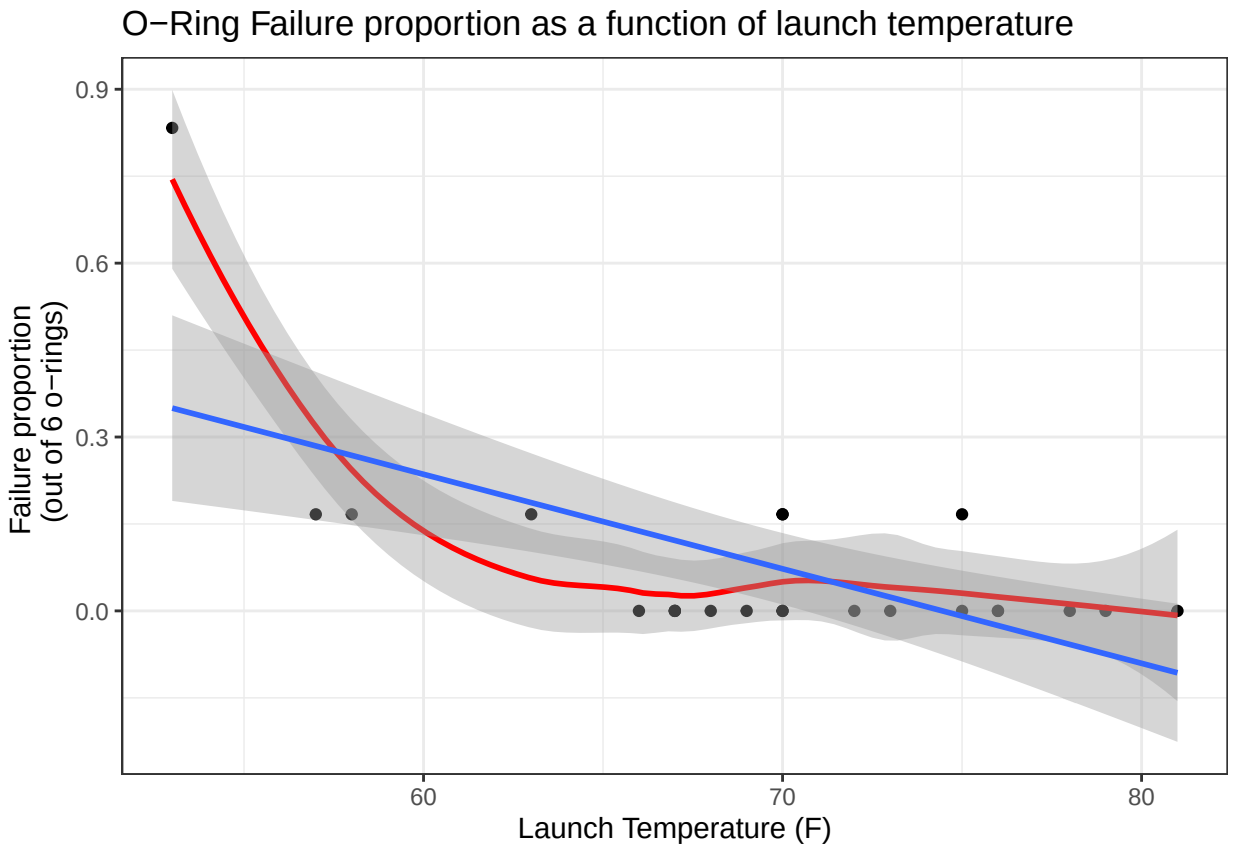
The `orings` dataset is available in the `faraway` package in R. The data contains the number of o-rings that failed on previous launches (out of a total of 6) along with the temperature on the day of the launch.

```
library(faraway)
data(orings)
orings_updated <- orings %>% mutate(non_damage = 6 - damage,
                                    failure_proportion = damage/6)
kable(orings_updated, digits = 2)
```

temp	damage	non_damage	failure_proportion
53	5	1	0.83
57	1	5	0.17
58	1	5	0.17
63	1	5	0.17
66	0	6	0.00
67	0	6	0.00
67	0	6	0.00
67	0	6	0.00
68	0	6	0.00
69	0	6	0.00
70	1	5	0.17
70	0	6	0.00
70	1	5	0.17
70	0	6	0.00
72	0	6	0.00
73	0	6	0.00
75	0	6	0.00
75	1	5	0.17
76	0	6	0.00
76	0	6	0.00
78	0	6	0.00
79	0	6	0.00
81	0	6	0.00

The temperature on the morning of the crash was 31F. We are going to build a model to explain the impact of temperature on the failure probability of the o-rings.

```
orings_updated %>% ggplot(aes(y=failure_proportion, x = temp)) + geom_point() +
  geom_smooth(method = 'loess', color = 'red') +
  ggtitle('O-Ring Failure proportion as a function of launch temperature') +
  ylab('Failure proportion \n (out of 6 o-rings)') + xlab('Launch Temperature (F)') +
  theme_bw() + geom_smooth(method = 'lm')
```



Using the `glm` function we can fit a logistic regression model. Note in this case the response is a data frame with successes and failures.

```
challenger_glm <- glm(cbind(damage, non_damage) ~ temp,data = orings_updated, family = binomial)
display(challenger_glm)
```

```
## glm(formula = cbind(damage, non_damage) ~ temp, family = binomial,
##     data = orings_updated)
##             coef.est coef.se
## (Intercept) 11.66     3.30
## temp        -0.22     0.05
## ---
##   n = 23, k = 2
##   residual deviance = 16.9, null deviance = 38.9 (difference = 22.0)
```

```
temp_tibble <- tibble(temp = seq(30,85, by = .1)) %>%
  mutate(pred = predict(challenger_glm, ., type = 'response'))

orings_updated %>% ggplot(aes(y=failure_proportion, x = temp)) + geom_point() +
  geom_smooth(method = 'loess', color = 'red') +
  ggtitle('O-Ring Failure proportion as a function of launch temperature') +
  ylab('Failure proportion \n (out of 6 o-rings)') + xlab('Launch Temperature (F)') +
  geom_line(aes(y = pred, x=temp), data = temp_tibble, inherit.aes = F) +
  theme_bw()
```