Module 9: Logistic Regression

Rebecca C. Steorts

Agenda

- 1986 Challenger explosion
- Binary structure of the response data
- Background: exponential families, generalized linear models (GLMs), logistic regression
- Example on exponenital familes and GLMs
- Bayesian logistic regression
- Returning to the Challenger case study

The Challenger Case Study

On 28 January 1986, the Space Shuttle Challenger broke apart, 73 seconds into flight. All seven crew members died. The cause of the disaster was the failure of an o-ring on the right solid rocket booster.

O-rings

- O-rings help seal the joints of different segments of the solid rocket boosters.
- We learned after this fatal mission that o-rings can fail at extremely low temperatures.

Loading the Faraway Package

```r
# Load data from space shuttle missions
library(faraway)
data("orings")
orings[1,] <- c(53,1)
head(orings)
```

```
##   temp damage
## 1   53      1
## 2   57      1
## 3   58      1
## 4   63      1
## 5   66      0
## 6   67      0
```

Space Shuttle Missions

The 1986 crash of the space shuttle Challenger was linked to failure of o-ring seals in the rocket engines.

Data was collected on the 23 previous shuttle missions, where the following variables were collected:

- temperate for each mission
- damage to the number of o-rings (failure versus non-failure)

Plot

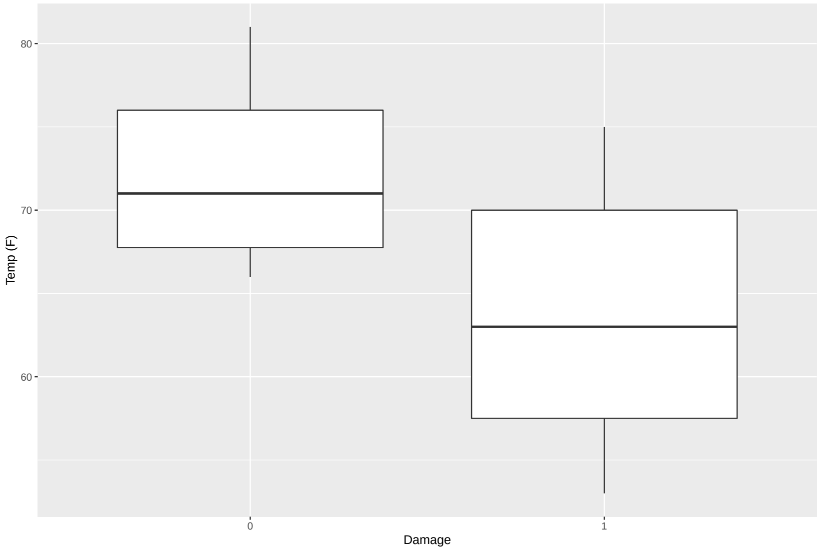
```r
library(ggplot2)
geom_boxplot(outlier.colour="black", outlier.shape=14,
             outlier.size=2, notch=FALSE)
```

```
## geom_boxplot: outlier.colour = black, outlier.fill = NUI
## stat_boxplot: na.rm = FALSE, orientation = NA
## position_dodge2
```

```r
damage <- as.factor(orings$damage)
temp <- orings$temp
head(damage)
```

```
## [1] 1 1 1 1 0 0
## Levels: 0 1
```

Boxplot of temperature versus o-ring failure



Linear models

Why is **linear regression** not appropriate for this data?

Beyond Linear Models

While linear models are useful, they are limited when

1. the range of y_i is restricted (e.g., binary or count)
2. the variance of y_i depends on the mean

Generalized linear models (GLMs) extend the linear model framework to address both of these issues.

Motivations and goals

- We will revisit not just the challenger data, but other missions to understand the relationship between o-ring failure and temperature.
- In order to understand this case study, we first need to learn about exponential families, generalized linear models, and logistic regression.

Background

We need to introduce:

- exponential families
- generalized linear models
- and logistic regression

Exponential Families

Any density that can be written in the form of equation 1 is called an **exponential family**.

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1}$$

where θ and ϕ are the **natural and dispersion parameters**, respectively and a, b, c are functions.

Connection to GLMs

In a GLM, pdfs or pmfs can be shown to be an exponential family using equation~1.

When doing this, it's important to identify the parameters of the exponential family, namely:

$$\theta,\ \phi,\ a(\phi),\ b(\theta),\ c(y, \phi).$$

Our overall goal is to estimate $\mu = E[Y \mid X]$.

Connection to GLMs

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \qquad (2)$$

- The natural parameter θ is used to govern the shape of the density $Y \mid X$. Thus, μ depends on θ.
- The dispersion parameter ϕ is assumed known.
- For GLM's, $\eta = \beta^T X = \beta_1 X_1 + \ldots \beta_p X_p$.

Our goal is to model a transformation of the mean μ by a function of X:

$$g(\mu) = \eta(X).$$

Generalized Linear Models

Given covariates X and an outcome Y, a **generalized linear model** is defined by three components:

1. a **random component**, which specifies a distribution for $Y \mid X$.
2. a **systematic component** that relates the parameter η to the covariates X
3. a **link function** that connects the random and systematic components

Exponential Families and GLMs

We assume $\mu = E[Y \mid X]$ and our goal is to estimate μ.

- The **systematic component** relates η to X.

In a GLM,

$$\eta = \beta^T X = \beta_1 X_1 + \ldots \beta_p X_p$$

The **link component** connects the **random** and **systematic components**, via a link function g.

The link function provides a connection between $\mu = E[Y \mid X]$ and η.

Exponential Families and GLMs

Let's look at an example to solidify our knowledge of exponential families and GLM's.

Bernoulli Example

Suppose $Y \in \{0, 1\}$ and

$$Y \mid X \overset{iid}{\sim} \text{Bernoulli}(p).$$

Show that $Y \mid X$ is in the exponential family, and provide the respective parameters. Also, identify the link function g.

Bernoulli Solution

Note that:

$$
\begin{aligned}
f(y) &= p^y(1-p)^{1-y} && (3)\\
&= \exp\{y \log p + (1-y)\log(1-p)\} && (4)\\
&= \exp\{y \log(\frac{p}{1-p}) + \log(1-p) + 0\} && (5)
\end{aligned}
$$

Bernoulli Solution

$$f(y) = \exp\{y \log(\frac{p}{1-p}) + \log(1-p) + 0\} \tag{6}$$

- The natural parameter is $\theta = \log \dfrac{p}{1-p}$.
- The mean is $\mu = p$, which implies that $p = e^\theta/(1 + e^\theta)$.
- This implies $b(\theta) = -\log(1-p) = -\log(1+e^\theta)$.
- There is no dispersion parameter, so $a(\phi) = 1$ and $c(y, \phi) = 0$.

Bernoulli Solution

$$f(y) = \exp\{y \log(\frac{p}{1-p}) + \log(1-p) + 0\} \tag{7}$$

The link function is

$$g(\mu) = \log(\frac{\mu}{1-\mu})$$

such that we model

$$\log(\frac{\mu}{1-\mu}) = \text{logit}(\mu) = \beta^T X.$$

This is known as **logistic regression**, which is a GLM with the **logit link**.

Challenger Case Study

Let's return to the case study of the challenger, where

- The response is the damage to the o-ring (in each shuttle launch).
- The covariate is the temperature (F) in each shuttle launch.

Notation and Setup

- Let p_i be the probability that an o-ring i fails.
- The corresponding **odds of failure** is

$$\frac{p_i}{1 - p_i}.$$

Notation and Setup

- The probability of failure p_i is between $[0, 1]$
- The odds of failure is any real number.

Logistic Regression

The response

$$Y_i \mid p_i \sim \text{Bernoulli}(p_i) \tag{8}$$

for $i = 1, \ldots, n$.

The logistic GLM writes that the logit of the probability p_i as linear function of the predictor variable(s) x_i:

$$\text{logit}(p_i) := \log(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 x_i. \tag{9}$$

Interpretation of Co-efficients

- The regression coefficients β_0, β_1 are directly related to the log odds $\log(\frac{p_i}{1-p_i})$ and not p_i.
- For example, the intercept β_0 is the $\log(\frac{p_i}{1-p_i})$ for observation i when the predictor takes a value of 0.
- The slope β_1 refers to the change in the expected log odds of failure of an o-ring for a decrease in temperature.

Intuition of Model

We assume our 23 data points are **conditionally independent**.

$$\Pr(\text{failure} = 1) = \frac{\exp\{\beta_0 + \beta_1 \times \text{temp}\}}{1 + \exp\{\beta_0 + \beta_1 \times \text{temp}\}}$$

$$\text{failure}_1, \ldots, \text{failure}_{23} \mid \beta_0, \beta_1, \text{temp}_1, \ldots, \text{temp}_{23} \tag{10}$$

$$\sim \prod_i \left(\frac{\exp\{\beta_0 + \beta_1 \times \text{temp}_i\}}{1 + \exp\{\beta_0 + \beta_1 \times \text{temp}_i\}} \right)^{\text{failure}_i} \tag{11}$$

$$\times \left(\frac{1}{1 + \exp\{\beta_0 + \beta_1 \times \text{temp}_i\}} \right)^{\text{1-failure}_i} \tag{12}$$

Exercise

Assume that $\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$.

Show that

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{e^{\beta_0+\beta_1 x_i} + 1}.$$

This shows that logit function guarantees that the probability p_i lives in $[0, 1]$.

Logistic Regression

Recall that

$$Y_i \mid p_i \sim \text{Bernoulli}(p_i) \tag{13}$$

for $i = 1, \ldots, n$.

$$\text{logit}(p_i) := \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i. \tag{14}$$

Note: This is the logistic GLM that we saw earlier. To perform logistic regression in `R`, you can use the `glm` function with the logit link.

Bayesian Logistic Regression

How can we build minimal Bayesian prior knowledge?

Priors on β_0 and β_1

Conjugate priors do not exists on β_0 and β_1.

We will consider the following weakly informative priors:

$$\beta_0 \sim \text{Normal}(0, 1000) \tag{15}$$
$$\beta_1 \sim \text{Normal}(0, 1000) \tag{16}$$
$$\tag{17}$$

Posterior sampling

Since we cannot find the posterior in closed form, we will resort to MCMC to approximate inference regarding β_0, β_1.

We can do this easily using the `logitMCMC` function in the `MCMCpack` R package.

This package implements a random walk Metroplis algorithm.

The random walk metropolis algorithm

- We saw the random walk metropolis algorithm in Module 6, slide 31.
- For a different review of this method, there is a nice explanation of it here: https://www.youtube.com/watch?v=U561HGMWjcw
- We don't need to code it up, as someone has written a nice package in R, but you could do this on your own if you wanted to.

Posterior sampling

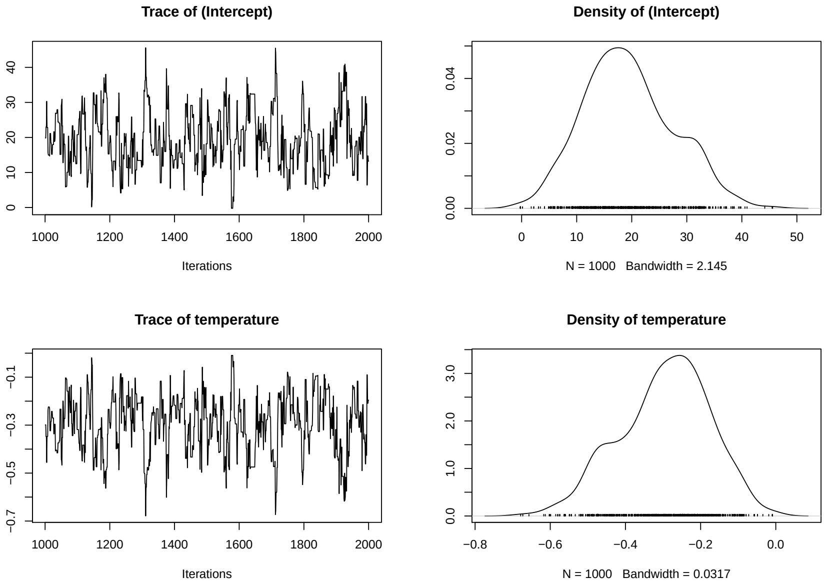
```
library(MCMCpack)
```

```
## Loading required package: coda

## Loading required package: MASS

## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)

## ## Copyright (C) 2003-2021 Andrew D. Martin, Kevin M. Qu

## ##
## ## Support provided by the U.S. National Science Foundat

## ## (Grants SES-0350646 and SES-0350613)
## ##
```

```
failure <- orings$damage
temperature <- orings$temp
output <- MCMClogit(failure~temperature,
                    mcmc=1000, b0=0, B0=0.001)
```

Traceplots

```
plot(output)
```



Summary

```r
summary(output)
```

```
##
## Iterations = 1001:2000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                Mean     SD Naive SE Time-series SE
## (Intercept) 19.4239 8.1171 0.256684        0.88555
## temperature -0.2955 0.1191 0.003765        0.01309
##
## 2. Quantiles for each variable:
##
##                2.5%     25%     50%     75%    97.5%
## (Intercept)  5.3608 13.6196 18.7297 24.4156 36.08274
## temperature -0.5441 -0.3734 -0.2853 -0.2108 -0.09241
```

Simulating Posterior Prediction

Given a certain temperature, we can simulate the results of future space shuttle launches using the posterior predictive distribution.

Suppose that on launch day, it's 80 degrees (F).

How would we simulate a predictive probability that a o-ring would fail?

Simulating Posterior Prediction

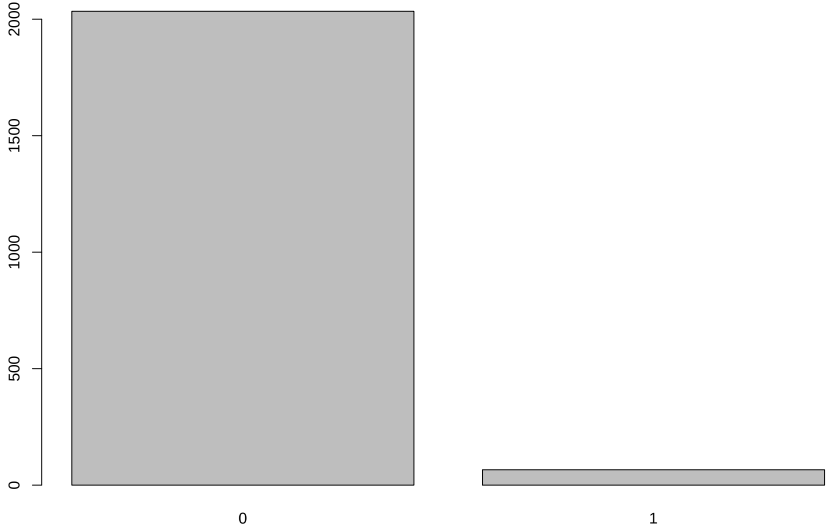
```r
library(boot)
```

```
##
## Attaching package: 'boot'

## The following objects are masked from 'package:faraway'
##
##     logit, melanoma
```

```r
temp <- 80
fail.prob <- inv.logit(output[,1]+ temp*output[,2])
y.pred <- rbinom(2100, size=1, prob=fail.prob)
```

Simulating Posterior Prediction

```
barplot(table(y.pred))
```



Your Turn

Suppose that it's very cold, 20 F.

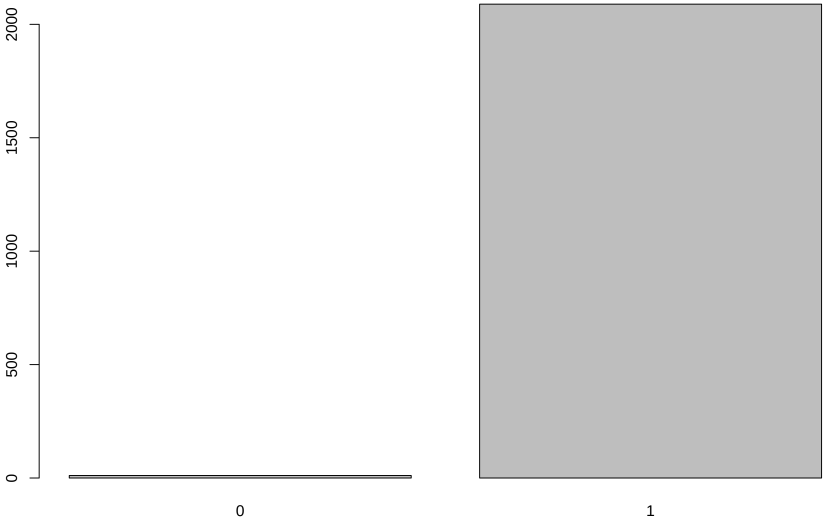
How would we simulate a predictive probability that a o-ring would fail?

Your Turn

```r
library(boot)
temp <- 20
fail.prob <- inv.logit(output[,1]+ temp*output[,2])
y.pred <- rbinom(2100, size=1, prob=fail.prob)
```

Your Turn

```
barplot(table(y.pred))
```



Summary

- 1986 Challenger explosion
- Binary structure of the response data
- Background: exponential families, generalized linear models (GLMs), logistic regression
- Example on exponential families and GLMs
- Bayesian logistic regression
- Returning to the Challenger case study
- What did you learn?

Exponential Families

More info about exponential families can be found here
https://www.cs.princeton.edu/courses/archive/spring09/cos513/s
cribe/lecture11.pdf

Gaussian Example

Suppose

$$Y \mid X \sim \text{Normal}(\mu, \sigma^2).$$

Then

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Show that $Y \mid X$ is in the exponential family, and provide the respective parameters.

Gaussian Solution

$$
\begin{aligned}
f(y) &= (\sqrt{2\pi}\sigma)^{-1}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} && (18)\\
&= \exp\{\log(\sqrt{2\pi}\sigma)^{-1}\}\exp\left\{-\frac{1}{2\sigma^2}(y^2-2y\mu+\mu^2)\right\} && (19)\\
&= \exp\{-\log(\sqrt{2\pi})-\log\sigma\}\exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}\right\} && (20)\\
&= \exp\{\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}-\log(\sqrt{2\pi})-\log\sigma\} && (21)
\end{aligned}
$$

The natural parameter $\theta = \mu$ and $b(\theta) = \theta^2/2$.

The dispersion parameter is $\phi = \sigma$ and $a(\phi) = \sigma^2$.

Finally, $c(y, \phi) = \frac{y^2}{2\sigma^2} - \log\phi - \log 2\pi$.

The link function $g(\mu) = \mu$ such that we model $\mu = \beta^T X$.