Logistic Regression

Rebecca C. Steorts (partial slide adaption from Maria Tacket) some material from Chapter 6 of Roback and Legler text.

Announcements

1. Quiz 2 returned. Excellent job by the class.
2. Homework 5 released last week. Due Thursday at 5 PM.
3. Quiz 3 released next week on GLM theory material. Class on Wednesday will be permitted to use for the quiz.

Computing set up

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(viridis)
library(kableExtra)
library(magrittr)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "90%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Topics

- Identify Bernoulli and binomial random variables
- Write GLM for binomial response variable
- Interpret the coefficients for a logistic regression model

Notes based on Chapter 6 Roback and Legler (2021) unless noted otherwise.

Basics of logistic regression

Bernoulli + Binomial random variables

Logistic regression is used to analyze data with two types of responses:

- **Binary**: These responses take on two values success ($Y = 1$) or failure ($Y = 0$), yes ($Y = 1$) or no ($Y = 0$), etc.

$$P(Y = y) = p^y(1 - p)^{1-y} \qquad y = 0, 1$$

- **Binomial**: Number of successes in a Bernoulli process, n independent trials with a constant probability of success p.

$$P(Y = y) = \binom{n}{y} p^y(1 - p)^{n-y} \qquad y = 0, 1, \ldots, n$$

In both instances, the goal is to model p the probability of success.

Logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- The response variable, $\log\left(\frac{p}{1-p}\right)$, is the log(odds) of success, i.e. the logit
- Use the model to calculate the probability of success

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

- When the response is a Bernoulli random variable, the probabilities can be used to classify each observation as a success or failure

Logistic vs linear regression model

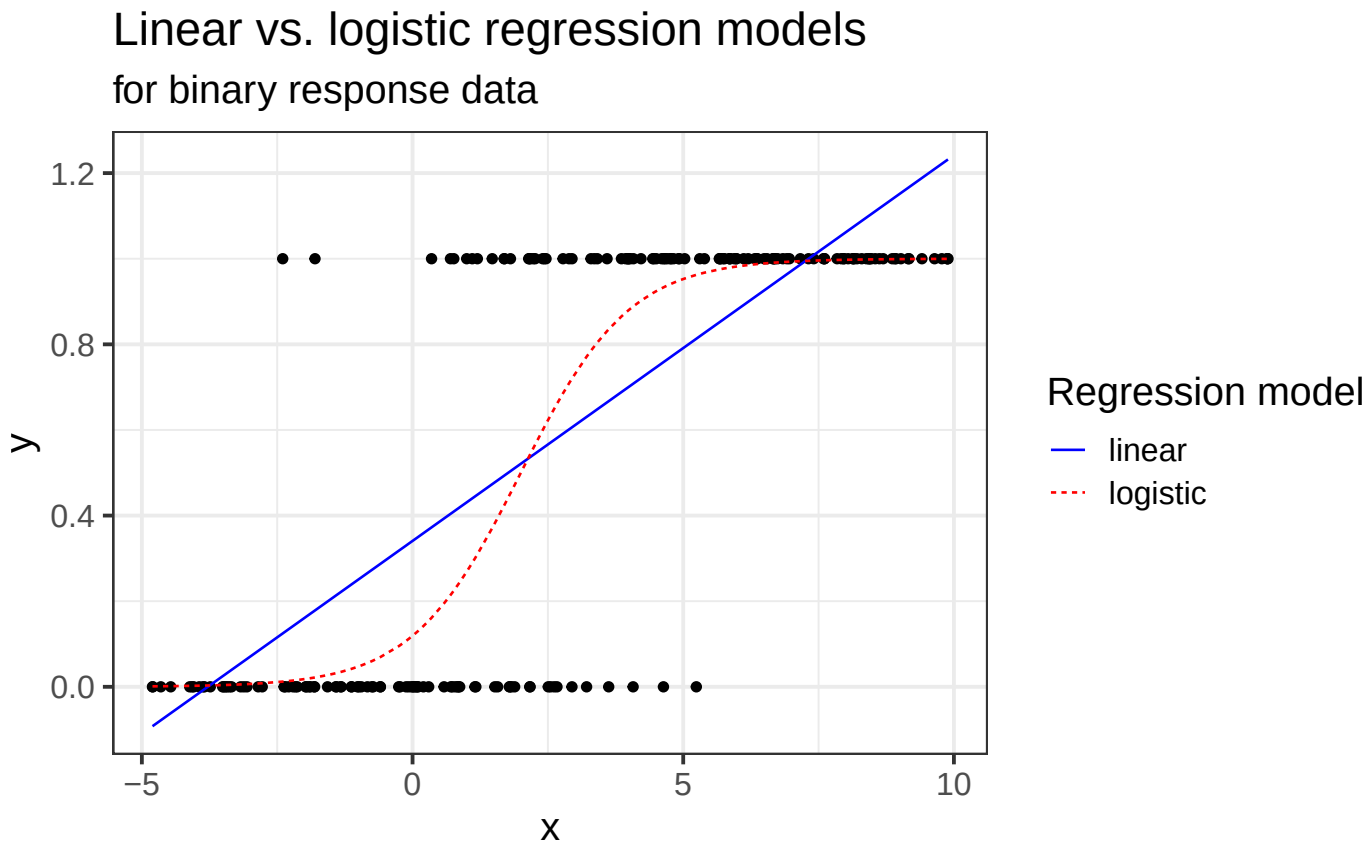


Figure 1: Graph from BMLR Chapter 6

Logit link

Bernoulli and Binomial random variables can be written in one-parameter exponential family form,
$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

Bernoulli

$$f(y;p) = e^{y\log(\frac{p}{1-p})+\log(1-p)}$$

Binomial

$$f(y;n,p) = e^{y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}}$$

Logit link

Bernoulli and Binomial random variables can be written in one-parameter exponential family form,
$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

Bernoulli

$$f(y;p) = e^{y\log(\frac{p}{1-p})+\log(1-p)}$$

Binomial

$$f(y;n,p) = e^{y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}}$$

They have the same canonical link $b(p) = \log\left(\frac{p}{1-p}\right)$

Assumptions for logistic regression

The following assumptions need to be satisfied to use logistic regression to make inferences

1. **Binary response**: The response is dichotomous (has two possible outcomes) or is the sum of dichotomous responses
2. **Independence**: The observations must be independent of one another
3. **Variance structure**: Variance of a binomial random variable is $np(1-p)$ ($n = 1$ for Bernoulli) , so the variability is highest when $p = 0.5$
4. **Linearity**: The log of the odds ratio, $\log\left(\frac{p}{1-p}\right)$, must be a linear function of the predictors $x_1, \ldots, x_p$

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

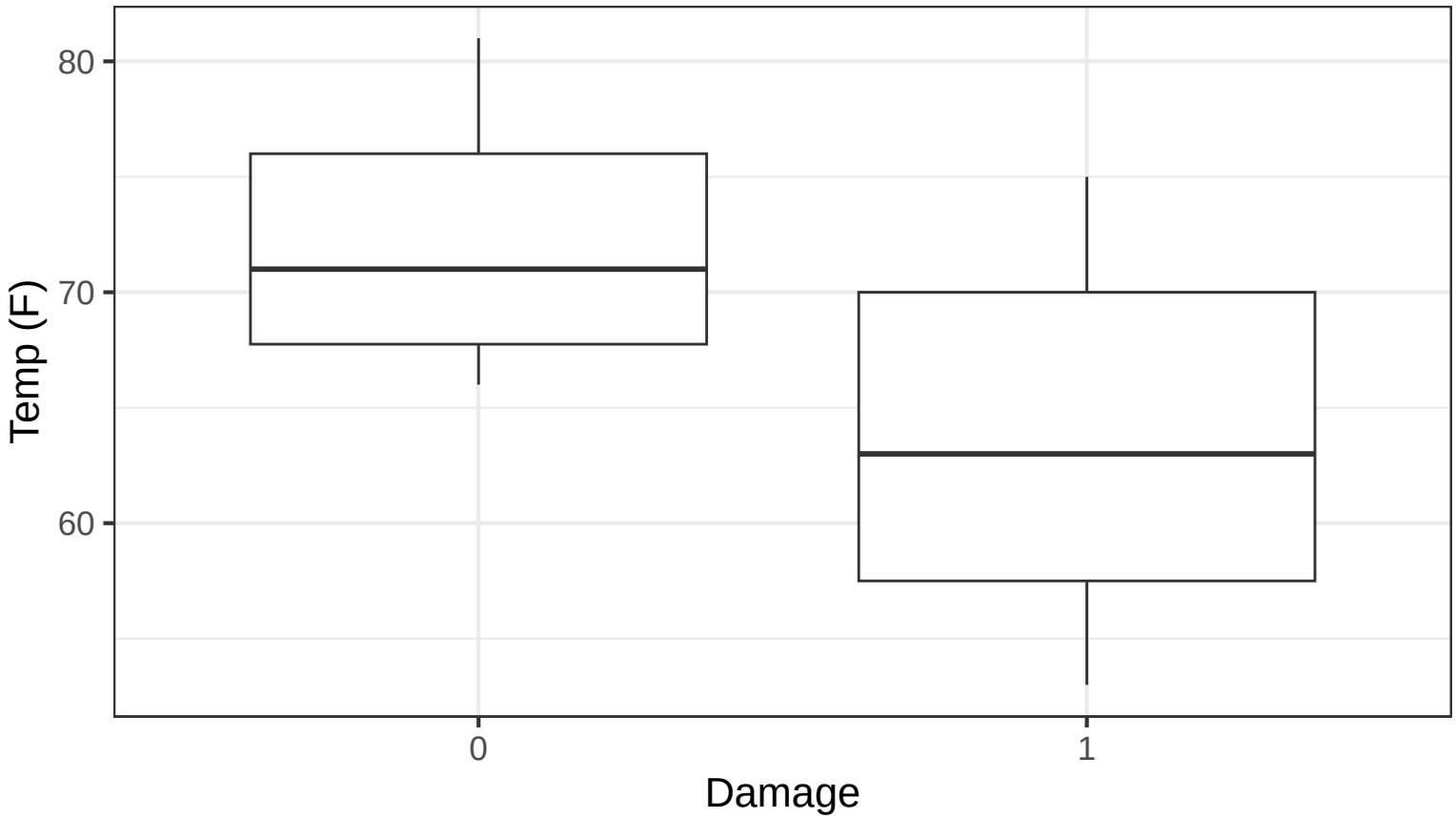
```
library(ggplot2)
geom_boxplot(outlier.colour="black", outlier.shape=14,
             outlier.size=2, notch=FALSE)
```

```
## geom_boxplot: outliers = TRUE, outlier.colour = black, o
## stat_boxplot: na.rm = FALSE, orientation = NA
## position_dodge2
```

```
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

In order to understand this case study, we first need to learn about exponential families, generalized linear models, and logistic regression. We will consider this more formally than we did earlier.

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

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

- 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.

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

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{6}$$

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

- The natural parameter is $\theta = \log\frac{p}{1-p}$.
- 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

$$f(y) = \exp\{y \log(\frac{p}{1-p}) + \log(1-p) + 0\} \tag{8}$$

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

$$Y_i \mid p_i \sim \text{Bernoulli}(p_i) \tag{9}$$

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:

$$\text{logit}(p_i) := \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i. \tag{10}$$

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**.

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

$$\text{failure}_1, \ldots, \text{failure}_{23} \mid \beta_0, \beta_1, \text{temp}_1, \ldots, \text{temp}_{23} \tag{11}$$

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

$$\times \left(\frac{1}{1 + \exp\{\beta_0 + \beta_1 \times \text{temp}_i\}} \right)^{\text{1-failure}_i} \tag{13}$$

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

$$Y_i \mid p_i \sim \text{Bernoulli}(p_i) \tag{14}$$

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

$$\text{logit}(p_i) := \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i. \tag{15}$$

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.

Challenger Data Exploration

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

Background

- The Challenger exploded 73 second after liftoff on January 28th, 1986 and claimed all seven lives on board.
- Engineers that manufactured the large boosters that launched the rocket were aware of the possible failures that could happen during cold temperatures.

Data from Previous Launches

- The main concern in launching the Challenger was the evidence that the large O-rings sealing the several sections of the boosters could fail in cold temperatures.
- The "fail" column in the data set below records how many O-rings experienced failures during that particular launch.
- The "temp" column lists the outside temperature at the time of launch.
- On the day of the explosion, the outside temperature was 31 degrees.

Challenger Data Exploration

```
head(Challeng)
```

```
##           temp pres fail n erosion blowby damage
## 4/12/81     66   50    0 6       0      0      0
## 11/12/81    70   50    1 6       1      0      4
## 3/22/82     69   50    0 6       0      0      0
## 11/11/82    68   50    0 6       0      0      0
## 4/4/83      67   50    0 6       0      0      0
## 6/18/83     72   50    0 6       0      0      0
```

Logistic Model

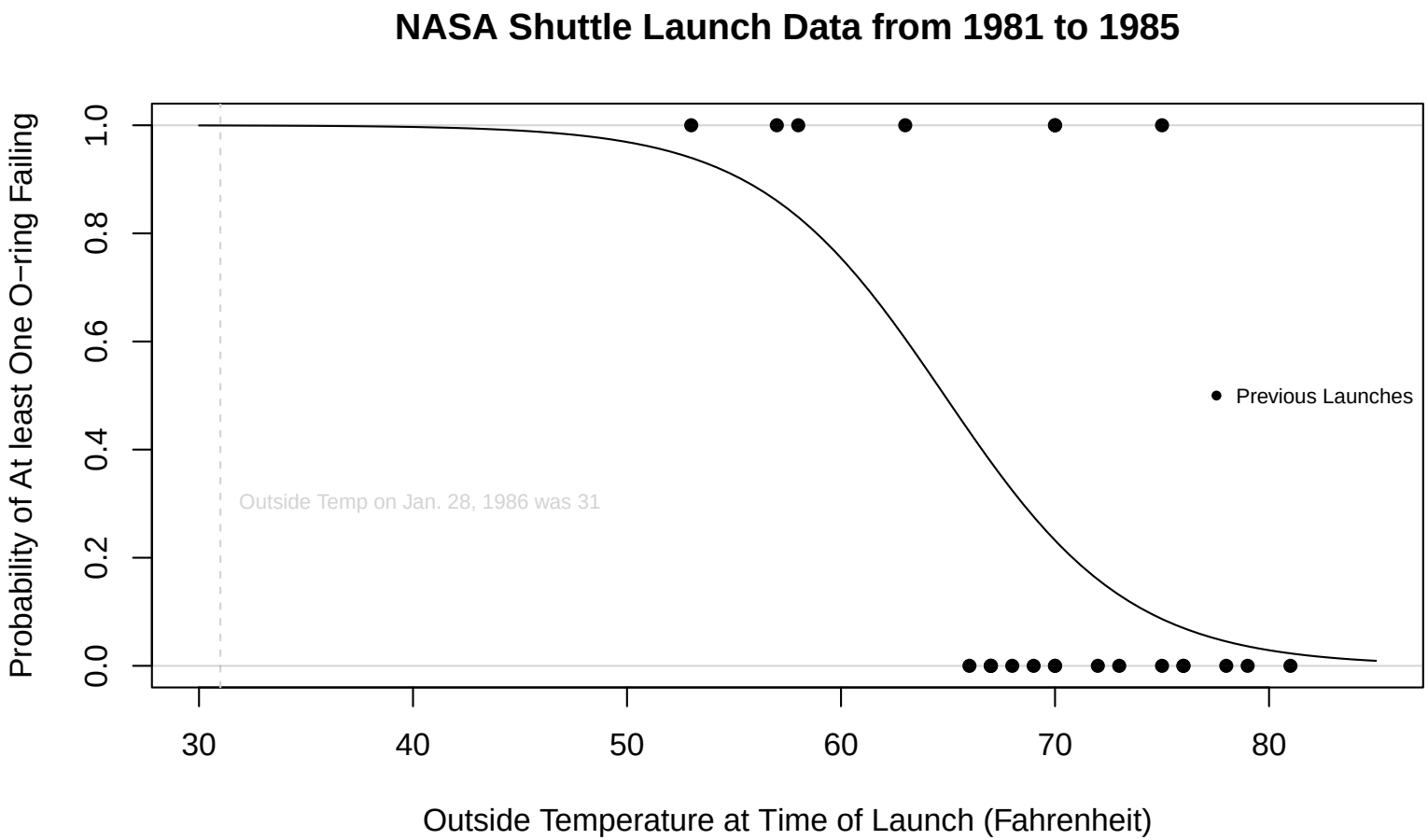
The probability of at least one o-ring failing during a shuttle launch based on the known outside temperature at the time of launch is given by the following logistic regression model:

$$P(Y_i = 1 \mid x_i) = \frac{\exp\{\beta_0 + \beta_1 \times x_i\}}{1 + \exp\{\beta_0 + \beta_1 \times x_i\}} = \pi_i$$

- $Y_i = 1$: denotes at least one o-ring failing for the given launch
- $Y_i = 0$: no failures
- x_i: denotes the outside temperature

Visualize the Model

Plot showing how much colder it was on the day of the Challenger launch (31 degrees, shown by the vertical dashed gray line) compared to all 23 previous shuttle launches (black dots in the graph).



Is the coefficient of temperature statistically significant?

```
chall.glm <- glm(fail>0 ~ temp,
                 data=Challeng, family=binomial)
tidy(chall.glm, conf.int = T) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	15.043	7.379	2.039	0.041	3.331	34.342
temp	-0.232	0.108	-2.145	0.032	-0.515	-0.061

1. $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$
2. $Z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.232 - 0}{0.108} = -2.145$ (using exact values)
3. $P(|Z| > |-2.145|) = 0.032 > 0.05.$
4. Yes, it is statistically significant.

What are plausible values for the coefficient of `temp`?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	15.043	7.379	2.039	0.041	3.331	34.342
temp	-0.232	0.108	-2.145	0.032	-0.515	-0.061

95% confidence interval for the coefficient of `temp`

$$\hat{\beta}_1 \pm z^* \times SE(\hat{\beta}_1)$$

where $z^* \sim N(0, 1)$

$$-0.232 \pm 1.96 \times 0.108 = (-\mathbf{0.444}, -\mathbf{0.020})$$

Interpretation of coefficient estimates

- Since the temperature being zero is not really realistic for this model, the value of e^{β_o} is not interpretable.
- However, the value of

$$e^{\beta_1} = e^{-0.232} = 0.79.$$

shows that the odds of the o-rings failing for a given launch decreases by a factor of 0.79 for every 1 degree increase in temperature.

- Said differently, the odds of an o-ring failure during launch decreases by $21\%(1 - 0.79)$ for every 1 degree increase in temperature.
- From the reverse perspective, every 1 degree decrease in temperature increases the odds of a failed o-ring by a factor of $e^{0.232} = 1.26$.

Illustration

For a temperature of 31 degrees, what would the predictor probability of at least one o-ring failure be?

Mathematically, this is

$$P(Y_i = 1 \mid x_i) = \frac{\exp(15.043 - 0.232 \times 31)}{1 + \exp(15.043 - 0.232 \times 31)}$$

```
pred <- predict(chall.glm,
                data.frame(temp=31), type='response')
```

This shows that $\hat{\pi}_i \approx 0.99961$.

Interpretation of coefficient estimates

The Challenger shuttle was launched at a temperature of 31 degrees. By waiting until 53 degrees, the odds of failure would have been decreased by a factor of $e^{-0.232(53-31)} = 0.006$, which is a 99.4% reduction in the odds of an o-ring failure!

```
(pred <- predict(chall.glm,
                 data.frame(temp=31), type='response'))
```

```
##         1
## 0.9996088
```

This illustrates that an o-ring failure was very likely to happen at such a cold temperature.

References

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.