Logistic Regression

Prof. Maria Tackett

STA 199

class: middle center

Click for PDF of slides

Introduction

Multiple regression allows us to relate a numerical response variable to one or more numerical or categorical predictors.

We can use multiple regression models to understand relationships, assess differences, and make predictions.

But what about a situation where the response of interest is categorical and binary?

Introduction

Multiple regression allows us to relate a numerical response variable to one or more numerical or categorical predictors.

We can use multiple regression models to understand relationships, assess differences, and make predictions.

But what about a situation where the response of interest is categorical and binary?

- spam or not spam
- malignant or benign tumor
- survived or died
- admitted or or not admitted

STA 199

Titanic

On April 15, 1912 the famous ocean liner *Titanic* sank in the North Atlantic after striking an iceberg on its maiden voyage. The dataset **`titanic.csv`** contains the survival status and other attributes of individuals on the titanic.

- **`survived`**: survival status (1 = survived, 0 = died)
- **`pclass`**: passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **`name`**: name of individual
- **`sex`**: sex (male or female)
- **`age`**: age in years
- **`fare`**: passenger fare in British pounds

We are interested in investigating the variables that contribute to passenger survival. Do women and children really come first?

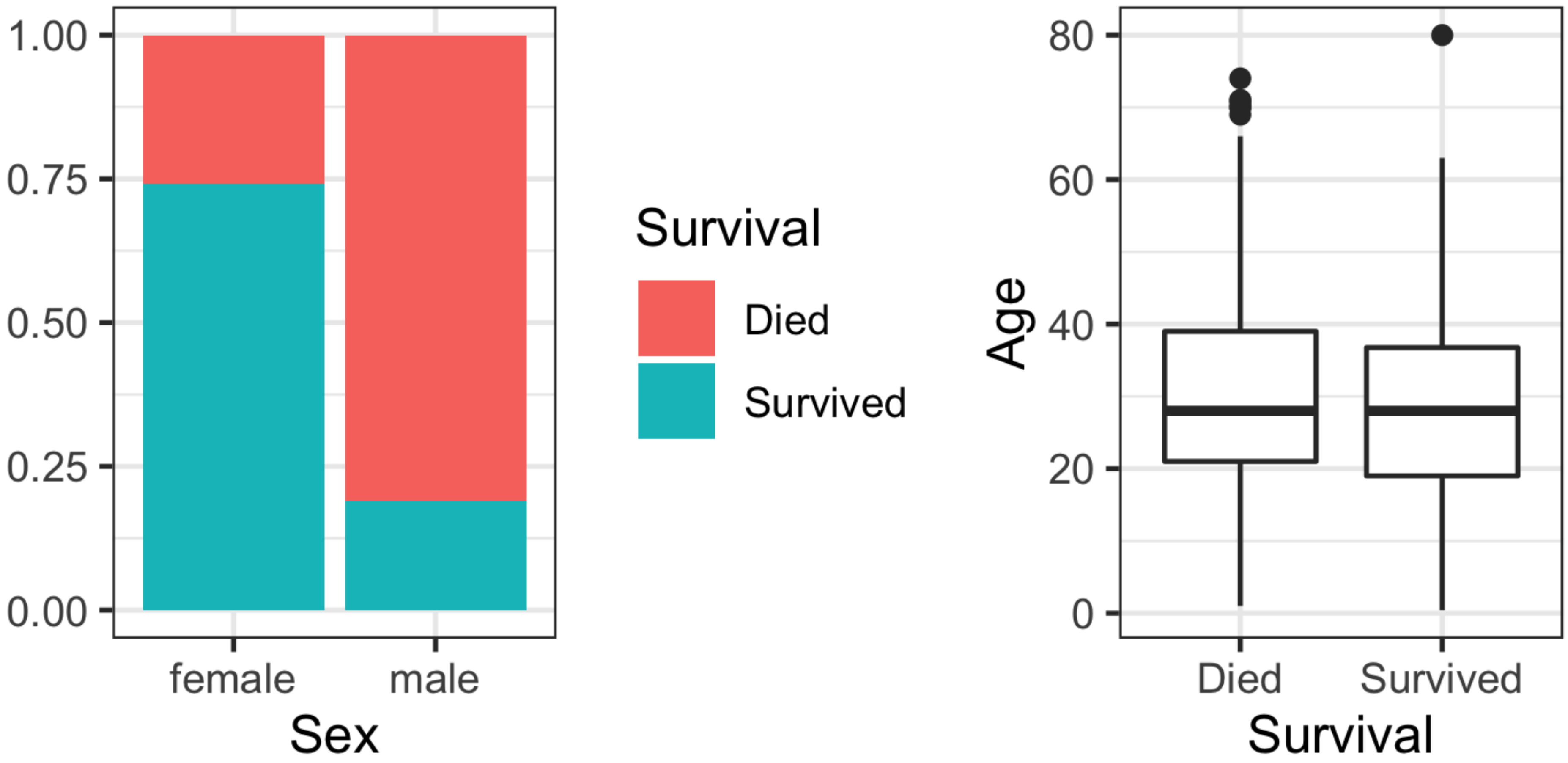
Data and Packages

```
library(tidyverse)
library(broom)
```

```
glimpse(titanic)
```

```
## Rows: 887
## Columns: 7
## $ pclass   <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 
## $ name     <chr> "Mr. Owen Harris Braund", "Mrs. John Bradley (Florence Br
## $ sex      <chr> "male", "female", "female", "female", "male", "male", "ma
## $ age      <dbl> 22, 38, 26, 35, 35, 27, 54, 2, 27, 14, 4, 58, 20, 39, 14,
## $ fare     <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625
## $ died     <dbl> 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 
## $ survived <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 
```

Exploratory Data Analysis



STA 199

The linear model with multiple predictors

- Population model:

$$y = \beta_0 + \beta_1 \ x_1 + \beta_2 \ x_2 + \cdots + \beta_k \ x_k + \epsilon$$

The linear model with multiple predictors

- Population model:

$$y = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + \cdots + \beta_k \, x_k + \epsilon$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1 \, x_1 + b_2 \, x_2 + \cdots + b_k \, x_k$$

The linear model with multiple predictors

- Population model:

$$y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \cdots + \beta_k\ x_k + \epsilon$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1\ x_1 + b_2\ x_2 + \cdots + b_k\ x_k$$

Denote by p the probability of survival and consider the model below.

$$p = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \cdots + \beta_k\ x_k + \epsilon$$

STA 199

The linear model with multiple predictors

- Population model:

$$y = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + \cdots + \beta_k \, x_k + \epsilon$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1 \, x_1 + b_2 \, x_2 + \cdots + b_k \, x_k$$

Denote by p the probability of survival and consider the model below.

$$p = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + \cdots + \beta_k \, x_k + \epsilon$$

Can you see any problems with this approach?

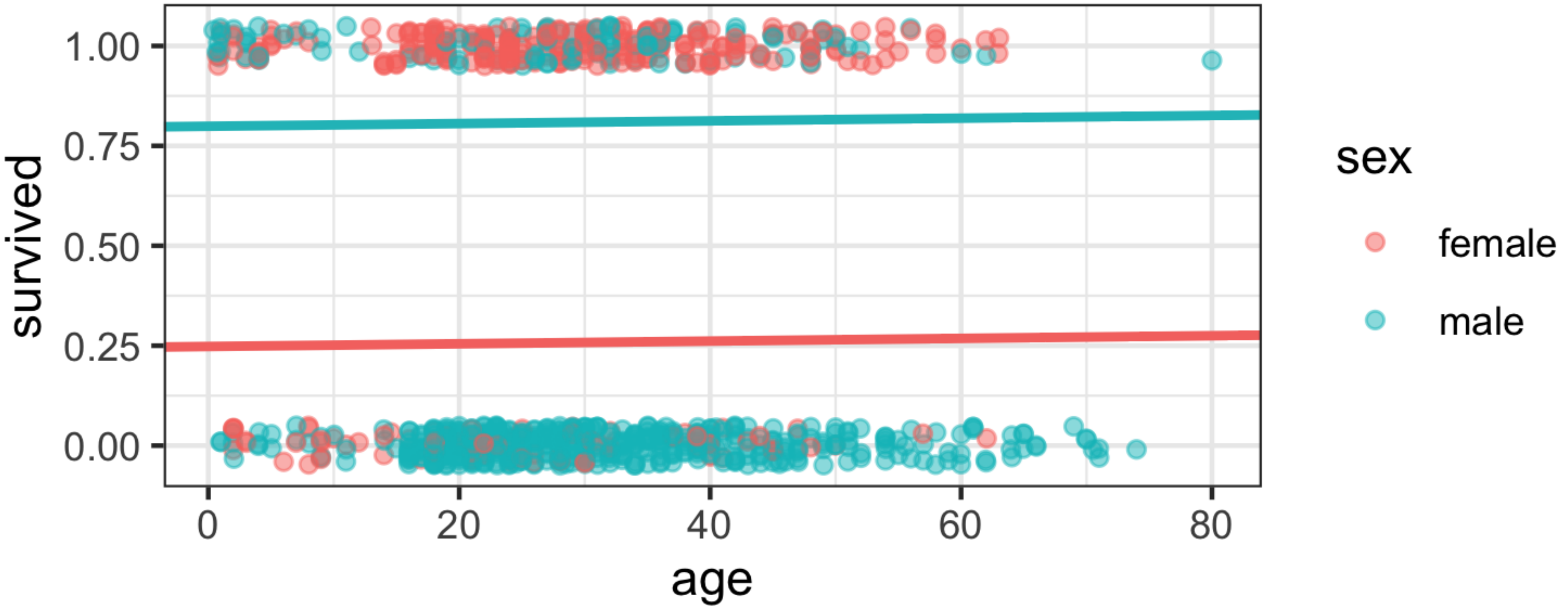
STA 199

Linear Regression?

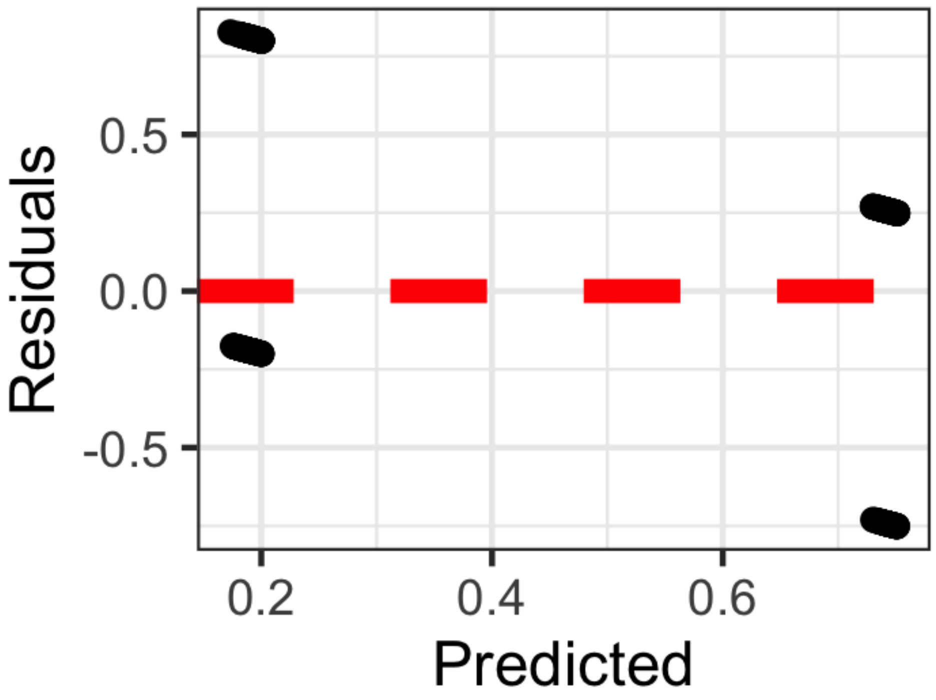
```
lm_survival <- lm(survived ~ age + sex, data = titanic)
tidy(lm_survival)
```

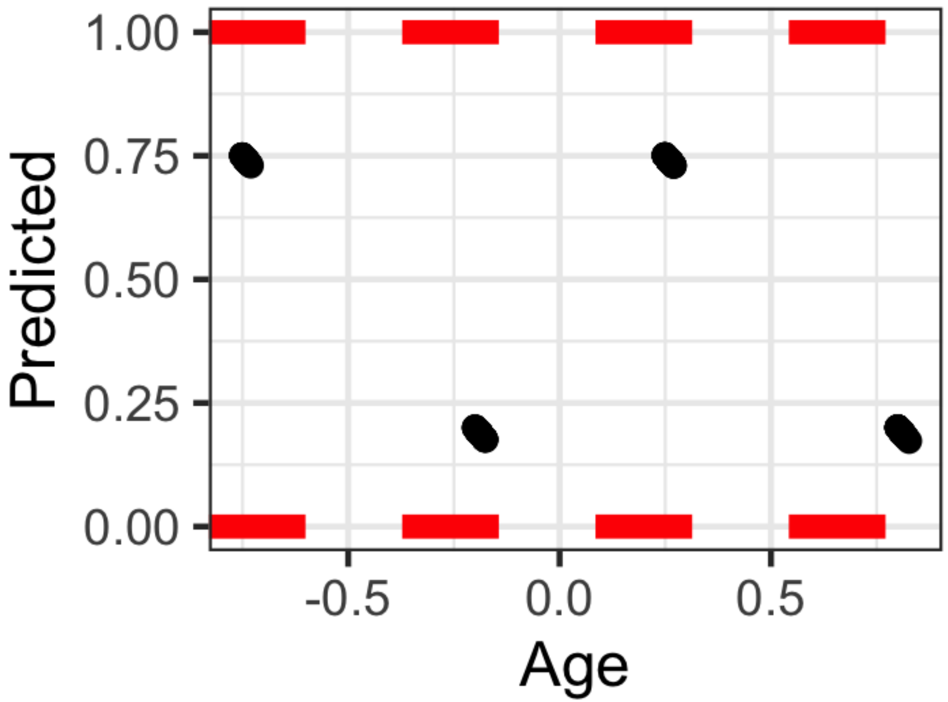
```
## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.752     0.0356      21.1   2.88e-80
## 2 age         -0.000343  0.000979    -0.350 7.26e- 1
## 3 sexmale     -0.551     0.0289     -19.1   3.50e-68
```

Visualizing the Model

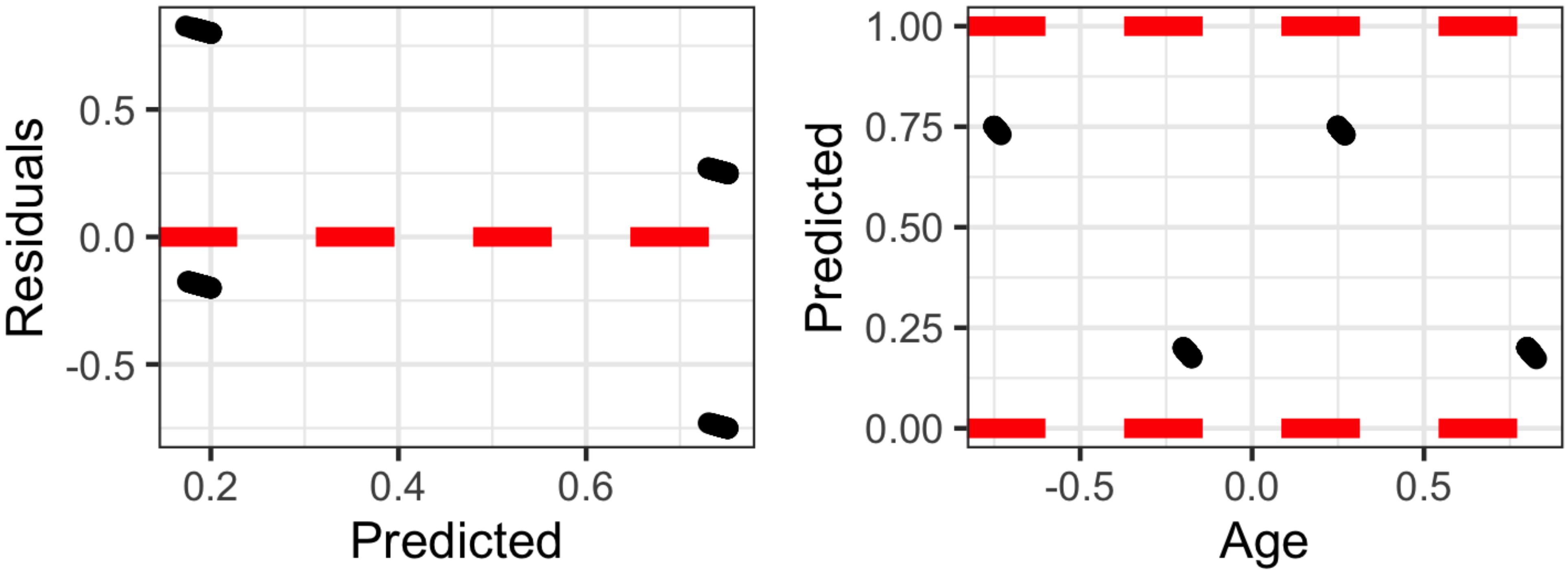


Diagnostics





Diagnostics



This isn't helpful! We need to develop a new tool.

STA 199

Preliminaries

- Denote by p the probability of some event
- The odds the event occurs is $\frac{p}{1-p}$

Preliminaries

- Denote by p the probability of some event
- The **odds** the event occurs is $\frac{p}{1-p}$

Odds are sometimes expressed as X : Y and read X to Y.

It is the ratio of successes to failures, where values larger than 1 favor a success and values smaller than 1 favor a failure.

STA 199

Preliminaries

- Denote by p the probability of some event
- The **odds** the event occurs is $\frac{p}{1-p}$

Odds are sometimes expressed as X : Y and read X to Y.

It is the ratio of successes to failures, where values larger than 1 favor a success and values smaller than 1 favor a failure.

If $P(A) = 1/2$, the odds of A are $\frac{1/2}{1/2} = 1$

STA 199

Preliminaries

- Denote by p the probability of some event
- The **odds** the event occurs is $\frac{p}{1-p}$

Odds are sometimes expressed as X : Y and read X to Y.

It is the ratio of successes to failures, where values larger than 1 favor a success and values smaller than 1 favor a failure.

If $P(A) = 1/2$, the odds of A are $\frac{1/2}{1/2} = 1$

If $P(B) = 1/3$, the odds of B are $\frac{1/3}{2/3} = 0.5$

An **odds ratio** is a ratio of odds.

Preliminaries

- Taking the natural log of the odds yields the **logit** of p

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Preliminaries

- Taking the natural log of the odds yields the logit of p

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The logit takes a value of p between 0 and 1 and outputs a value between $-\infty$ and ∞.

Preliminaries

- Taking the natural log of the odds yields the logit of p

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The logit takes a value of p between 0 and 1 and outputs a value between $-\infty$ and ∞.

The inverse logit (logistic) takes a value between $-\infty$ and ∞ and outputs a value between 0 and 1.

$$\text{inverse logit}(x) = \frac{e^x}{1+e^x}$$

Logistic Regression Model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Logistic Regression Model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Use the inverse logit to find the expression for p.

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}$$

We can use the logistic regression model to obtain predicted probabilities of success for a binary response variable.

Logistic Regression Model

We handle fitting the model via computer using the **`glm`** function.

```
logit_mod <- glm(survived ~ sex + age, data = titanic,
                 family = "binomial")
tidy(logit_mod)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  1.11      0.208       5.34  9.05e- 8
## 2 sexmale     -2.50      0.168     -14.9   3.24e-50
## 3 age         -0.00206   0.00586    -0.351 7.25e- 1
```

STA 199

Logistic Regression Model

And use **`augment`** to find predicted log-odds.

```
pred_log_odds <- augment(logit_mod)
```

The Estimated Logistic Regression Model

```
tidy(logit_mod)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  1.11      0.208       5.34  9.05e- 8
## 2 sexmale     -2.50      0.168     -14.9   3.24e-50
## 3 age         -0.00206   0.00586    -0.351 7.25e- 1
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\ sex - 0.00206\ age$$

$$\hat{p} = \frac{e^{1.11 - 2.50\ sex - 0.00206\ age}}{1 + e^{1.11 - 2.50\ sex - 0.00206\ age}}$$

Interpreting coefficients

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\ sex - 0.00206\ age$$

Interpreting coefficients

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\, sex - 0.00206\, age$$

Holding sex constant, for every additional year of age, we expect the log-odds of survival to decrease by approximately 0.002.

STA 199

Interpreting coefficients

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.11 - 2.50\ sex - 0.00206\ age$$

Holding sex constant, for every additional year of age, we expect the log-odds of survival to decrease by approximately 0.002.

Holding age constant, we expect males to have a log-odds of survival that is 2.50 less than females.

STA 199

Interpreting coefficients

$$\frac{\hat{p}}{1-\hat{p}} = e^{1.11 - 2.50\ sex - 0.00206\ age}$$

STA 199

Interpreting coefficients

$$\frac{\hat{p}}{1-\hat{p}} = e^{1.11 - 2.50\ sex - 0.00206\ age}$$

Holding sex constant, for every one year increase in age, the odds of survival are expected to multiply by a factor of $e^{-0.00206} = 0.998$.

STA 199

Interpreting coefficients

$$\frac{\hat{p}}{1-\hat{p}} = e^{1.11 - 2.50\ sex - 0.00206\ age}$$

Holding sex constant, for every one year increase in age, the odds of survival are expected to multiply by a factor of $e^{-0.00206} = 0.998$.

Holding age constant, the odds of survival for males are $e^{-2.50} = 0.082$ times the odds of survival for females.

STA 199

Classification

- Logistic regression allows us to obtain predicted probabilities of success for a binary variable.
- By imposing a threshold (for example if the probability is greater than 0.50) we can create a classifier.

Classification

- Logistic regression allows us to obtain predicted probabilities of success for a binary variable.
- By imposing a threshold (for example if the probability is greater than 0.50) we can create a classifier.

```
## # A tibble: 2 x 3
##   survived  Died Survived
##      <dbl> <int>    <int>
## 1        0   464       81
## 2        1   109      233
```

Strengths and Weaknesses

Weaknesses

- Logistic regression has assumptions: independence and linearity in the log-odds (some other methods require fewer assumptions)
- If the predictors are correlated, coefficient estimates may be unreliable

Strengths and Weaknesses

Weaknesses

- Logistic regression has assumptions: independence and linearity in the log-odds (some other methods require fewer assumptions)
- If the predictors are correlated, coefficient estimates may be unreliable

Strengths

- Straightforward interpretation of coefficients
- Handles numerical and categorical predictors
- Can quantify uncertainty around a prediction
- Can extend to more than 2 categories (multinomial regression)