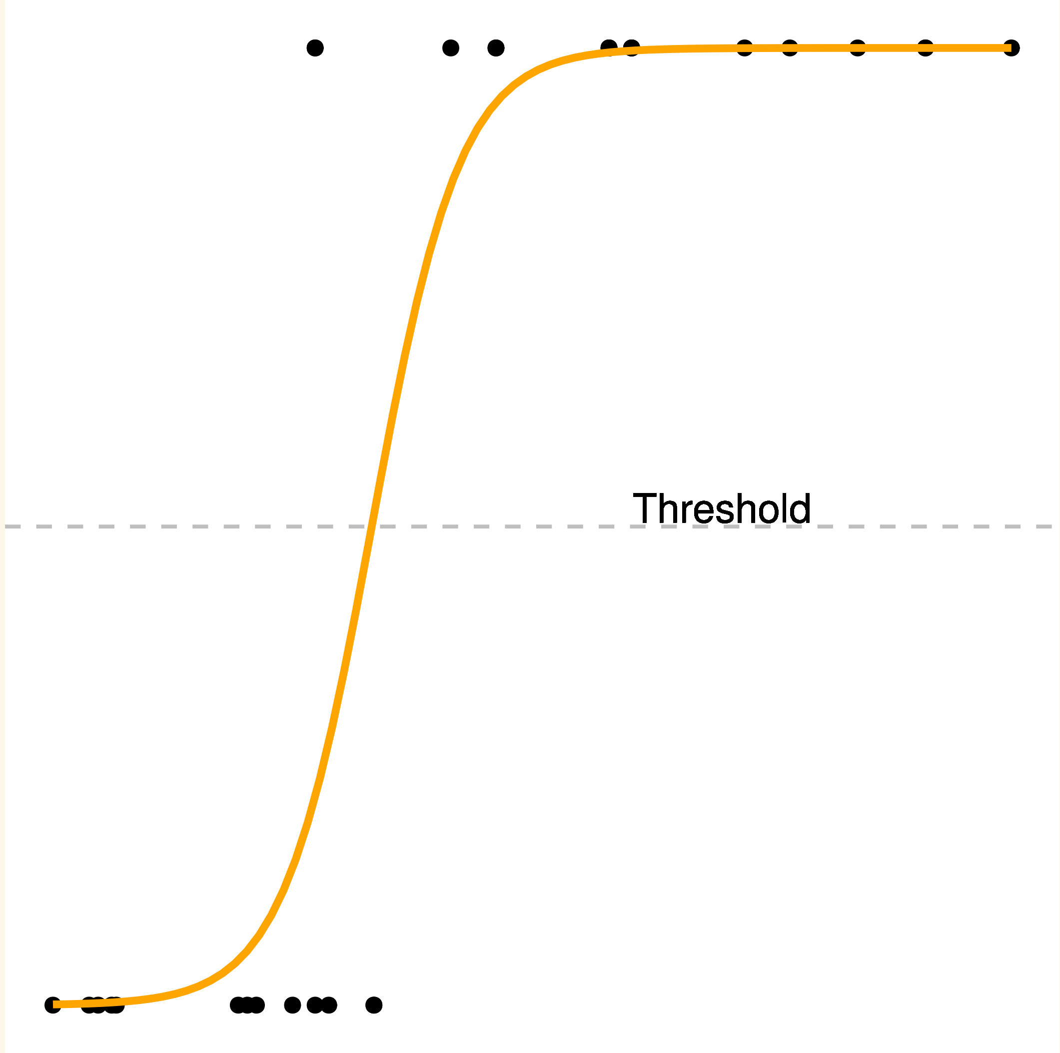
Logistic regression for binary responses: Model

Stat 230

May 09 2022

Overview



Today:

- Binary responses and Bernoulli distribution
- logistic and logit functions
- odds, log odds and odds ratio
- logistic regression model
- interpretation

Example: Donner party (Case study 20.1)

- Response: `Status` either `Died` or `Survived`
- Explanatory: `Age` and `Sex`

```
library(Sleuth3)
donner <- case2001
head(donner)
```

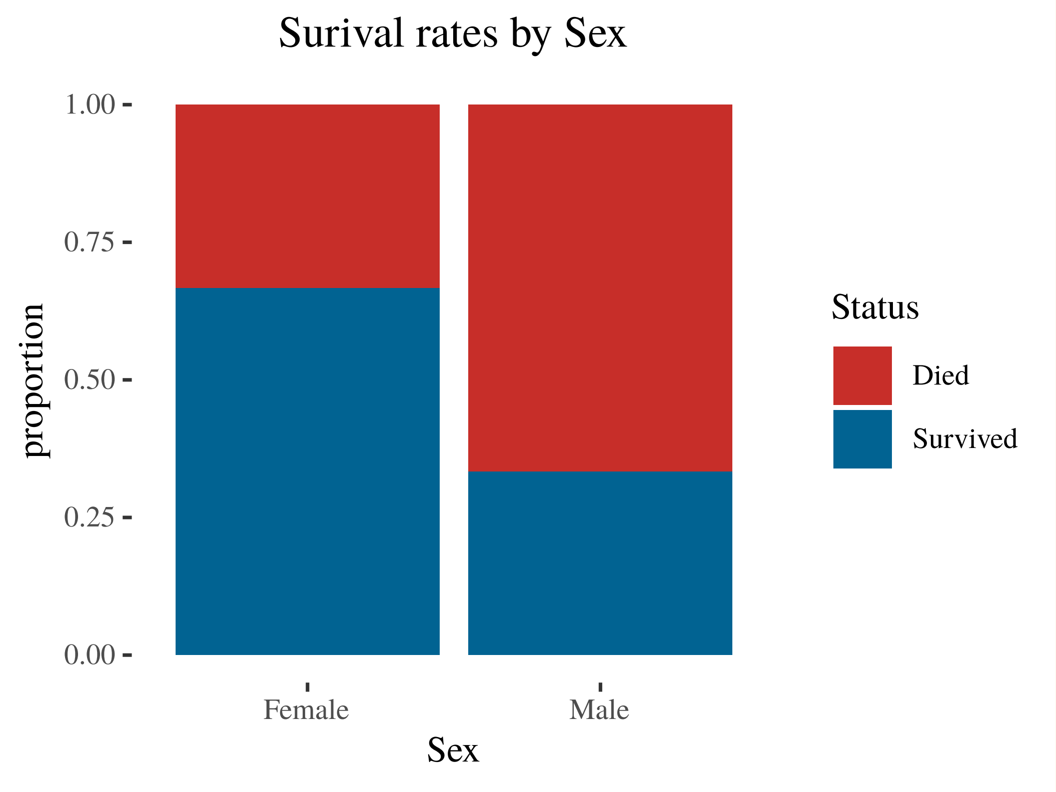
```
  Age    Sex   Status
1  23   Male     Died
2  40 Female Survived
3  40   Male Survived
4  30   Male     Died
5  28   Male     Died
6  40   Male     Died
```

EDA for associations between:

- `age` and `status`?
- `sex` and `status`?

Example: Donner party

```r
library(ggplot2)
ggplot(donner, aes(fill = Status, x = Sex)) +
 geom_bar(position="fill") +
 labs(y="proportion", title="Surival rates by Sex")
```



Stacked bar graph shows survival `status` conditioned on `sex`.

- `Females` had higher survival rates than `males`.

Example: Donner party

```r
library(dplyr)
donner %>%
 group_by(Sex) %>% # for each Sex group
 summarize(mean(Status == "Survived")) # proportion who survived
```

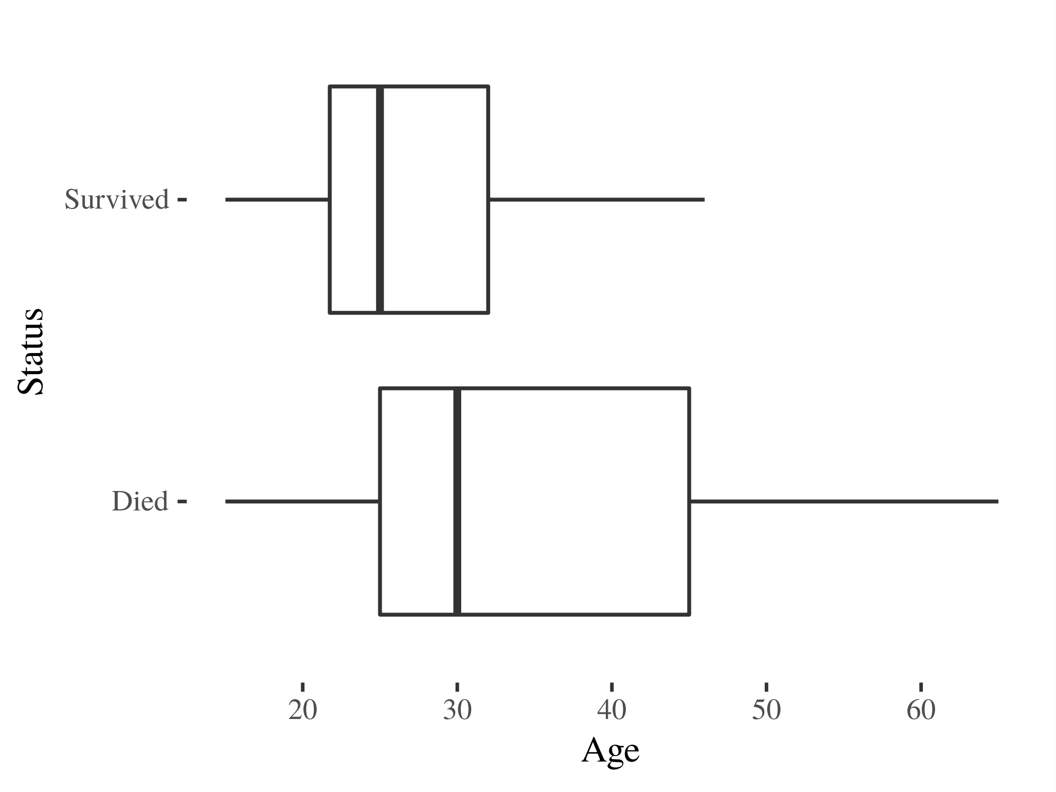
```
# A tibble: 2 × 2
  Sex    `mean(Status == "Survived")`
  <fct>                         <dbl>
1 Female                        0.667
2 Male                          0.333
```

Survival rates conditioned on `sex`

- $2/3$ of `females` survived while only $1/3$ of `males` did.

Example: Donner party

```r
library(ggplot2)
ggplot(donner, aes(x = Status, y = Age)) +
 geom_boxplot() +
 coord_flip()
```



Age distribution conditioned on status

- People who survived have lower median age than people who died.

Example: Donner party

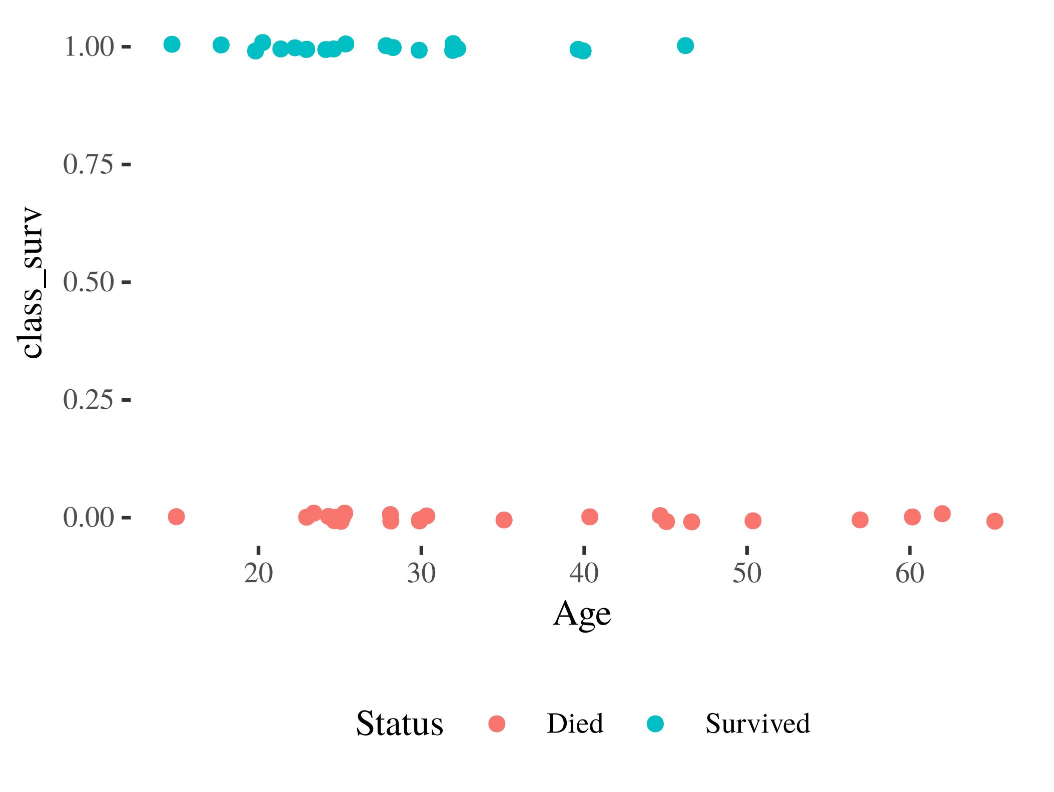
```
donner %>%
 group_by(Status) %>% # for each status group
 summarize(mean(Age), sd(Age), median(Age)) # get summary stats
```

```
# A tibble: 2 × 4
  Status    `mean(Age)` `sd(Age)` `median(Age)`
  <fct>           <dbl>     <dbl>         <dbl>
1 Died             35.5     14.3             30
2 Survived         27.2      8.00            25
```

- `Age` distribution statistics conditioned on `status`
  - The mean `age` of people who `survived` is 27.2 years compared to 35.5 for people who `died`.
- **Problem:** this EDA has `Status` as the "given" (conditioning) variable
- But we want to model `Status` given `Age`.

Example: Donner party

```
library(dplyr)
# recode Status as a binary (0 or 1) response:
donner$class_surv <- recode(donner$Status, Survived = 1, Died = 0)
ggplot(donner, aes(x = Age, y = class_surv)) +
 geom_jitter(aes(color=Status), height = .01)
```




```
mean(donner$class_surv)
[1] 0.4444444
```

- The mean of a binary variable gives the proportion of 1 's
- As age increases, proportion who survived decreases

Want to find a function to predict survival given age

The Bernoulli distribution

- A probability model for a random trial that has two possible outcomes: **success or failure**
- A Bernoulli random variable Y
  - $Y = 1$ if a success occurred
  - $Y = 0$ if a failure occurred
- π is the probability of success:

$$\pi = P(Y = 1) = P(\text{ success }), \quad 1 - \pi = P(Y = 0) = P(\text{ failure })$$

- Shorthand notation: $Y \sim \mathrm{Bern}(\pi)$

The Bernoulli distribution

- The expected value, or mean, of Y is equal to

$$E(Y) = \mu = \pi$$

- The standard deviation of Y is equal to

$$SD(Y) = \sigma = \sqrt{\pi(1 - \pi)}$$

- The expected value measures the "long run" average value that we would see from Y if we were to repeat the random trial many, many times.
- The standard deviation tells us how these values of Y will vary over these repeated trials.

The logistic model form

- Our Bernoulli responses are modeled as a function of predictors $X_i = x_{1,i}, \ldots, x_{p,i}$ through the probability of success:

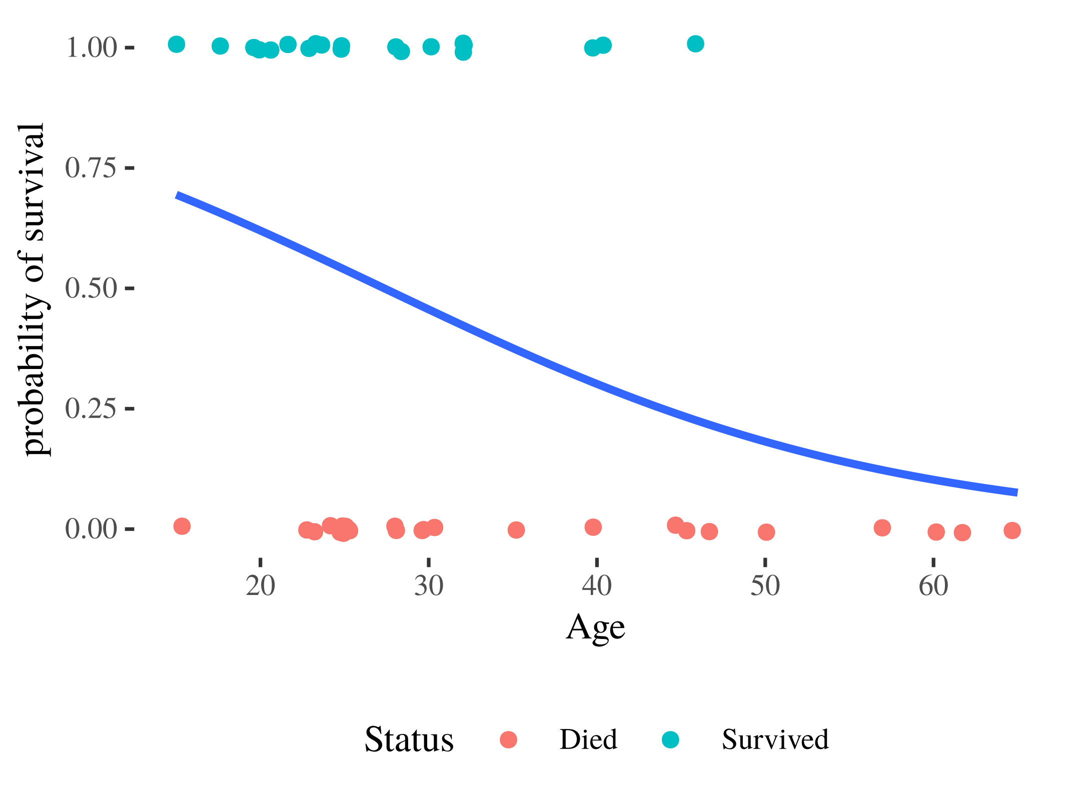
$$Y_i \mid X_i \overset{\text{indep.}}{\sim} \text{Bern}(\pi\left(X_i\right))$$

- Need to find a function $f()$ that takes in any number and produces a probability between 0 and 1 :

$$\pi\left(X_i\right) = f\left(x_{1,i}, \ldots, x_{p,i}\right)$$

Example: Donner party model

```
ggplot(donner, aes(x = Age, y = class_surv)) +
  geom_jitter(aes(color=Status), height = .01) +
# Add estimated logistic curve for the prob of survival.
  geom_smooth(method="glm",
              method.args = list(family=binomial),
              se=FALSE) +
 labs(y = "probability of survival") +
theme(legend.position = "bottom")
```



Fitting logistic regression in R

- The basic syntax for a logistic model fit is

```
glm(y ~ x1 + x2, family = binomial, data= )
```

- The variable y can be either form:
  - y can be binary $0/1$ coded response where 1 is a "success"
  - y can be a factor variable with two levels. The second level is what R will call a "success"

Example: Donner party model

- The model form logistic regression of survival status on age:

$$\eta = \beta_0 + \beta_1 Age$$

- The second level of Status will be what R defines as a "success"
- So, π is the probability of survival

```r
table(donner$Status) # second level = Survived
```

```
    Died Survived 
      25       20 
```

Example: Donner party model

```
donner_glm1 <- glm(Status ~ Age , family=binomial, data=donner)
library(broom)
tidy(donner_glm1)
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   1.82      0.999       1.82  0.0688
2 Age          -0.0665    0.0322     -2.06  0.0391
```

- The estimated value of η is

$$\hat{\eta} = 1.82 - 0.0665 \text{ Age}$$

- What is the estimated probability of survival? of death?
- What are these values for a 20 year old?

Example: Donner party model

- the estimated probability of survival is

$$\hat{\pi}(\text{ age }) = \frac{e^{1.82-0.0665 \text{ Age}}}{1+e^{1.82-0.0665 \text{ Age}}}$$

- the estimated probability of death is

$$1-\hat{\pi}(\text{ age }) = 1-\frac{e^{1.82-0.0665 \text{ Age}}}{1+e^{1.82-0.0665 Age}}$$

- the estimated probability of survival for a 20 year old is

$$\hat{\pi}(\text{ age } = 20) = \frac{e^{1.82-0.0665(20)}}{1+e^{1.82-0.0665(20)}} = 0.62$$

- the estimated probability of death for a 20 year old is

$$1-\hat{\pi}(\text{ age } = 20) = 1-\frac{e^{1.82-0.0665(20)}}{1+e^{1.82-0.0665(20)}} = 0.38$$

Example: Donner party model

```
tidy(donner_glm1, conf.int=TRUE)
```

```
# A tibble: 2 × 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)   1.82      0.999       1.82  0.0688 -0.00599    3.99  
2 Age          -0.0665    0.0322     -2.06  0.0391 -0.140     -0.0102
```

How does a one year increase in age affect survival status?

The logistic model: interpretation

- To interpret the model, "solve for" η
- The inverse of the logistic function is called the **logit function**:

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} = \log\left(\frac{\pi\left(X_i\right)}{1 - \pi\left(X_i\right)}\right)$$

- **Interpretation**: a one unit increase in x_1 is associated with an additive β_1 change in the logit function, holding other terms fixed.
- But what does this mean?

Odds of success

- odds of success:

$$odds = \frac{\pi(X)}{1 - \pi(X)}$$

- e.g If $\pi = 0.6$, then odds of success is $0.6/0.4 = 1.5$.
- for every 6 successes, we see 4 failures.

Odds Ratio

- odds ratio for A vs. B :

$$\frac{\text{odds of success for A}}{\text{odds of success for B}} = \frac{\frac{\pi(A)}{1-\pi(A)}}{\frac{\pi(B)}{1-\pi(B)}}$$

- e.g. If $\pi(A) = 0.75$ and $\pi(B) = 0.6$, then the odds ratio is 2 . odds ratio for A vs. B $= \frac{0.75/0.25}{0.6/0.4} = \frac{3}{1.5} = 2$
- Odds of success for group A is 2 times the odds of success in group B.

Interpretation

- logit $=\log$ odds of success

$$\eta_i = \log\left(\frac{\pi\left(X_i\right)}{1-\pi\left(X_i\right)}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

- The odds of success equals

$$\text{odds}(x_1, \ldots, x_p) = \frac{\pi(X)}{1-\pi(X)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

- What happens if we increase x_1 by one unit, holding other predictors fixed?

Interpretation

$$\text{odds}(x_1+1,\dots,x_p) = e^{\beta_0+\beta_1(x_1+1)+\dots+\beta_p x_p} = e^{\beta_0+\beta_1 x_1+\dots+\beta_p x_p} \times e^{\beta_1}$$

- Increasing x_1 by one unit has a multiplicative change of e^{β_1} in the odds of success.
- The multiplicative change of e^{β_1} is also called the odds ratio for a one unit increase in x_1

$$\frac{\text{odds of success for } x_1+1}{\text{odds of succes for } x_1} = \frac{\text{odds}(x_1+1,\dots,x_p)}{\text{odds}(x_1,\dots,x_p)} = e^{\beta_1}$$

Interpretation

- What if we have a predictor that is logged?

$$\text{odds}(x_1, \ldots, x_p) = e^{\beta_0 + \beta_1 \log(x_1) + \cdots + \beta_p x_p} = e^{\beta_0} x_1^{\beta_1} e^{\beta_2 x_2 + \cdots + \beta_p x_p}$$

- Changing x_1 by a factor of m :

$$\text{odds } (m x_1, \ldots, x_p) = e^{\beta_0} (m x_1)^{\beta_1} e^{\beta_2 x_2 + \cdots + \beta_p x_p} = e^{\beta_0} x_1^{\beta_1} e^{\beta_2 x_2 + \cdots + \beta_p x_p} \times m^{\beta_1}$$

results in a multiplicative change of m^{β_1} in the odds of success.

Example: Donner party model

```r
donner_glm1 <- glm(Status ~ Age , family=binomial, data=donner)
library(broom)
tidy(donner_glm1)
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   1.82      0.999       1.82  0.0688
2 Age          -0.0665    0.0322     -2.06  0.0391
```

- The estimated log odds of survival is

$$\operatorname{logit}(\hat{\pi}(age)) = \log \frac{\hat{\pi}(\text{ age })}{1-\hat{\pi}(age)} = 1.82 - 0.0665 \text{ Age}$$

- What is the estimated odds of survival? of death?
- What are these values for a 20 year old?

Example: Donner party model

- the estimated odds of survival is

$$\widehat{\text{odd (age}} \text{)} = \frac{\hat{\pi}(\text{ age })}{1 - \hat{\pi}(\text{ age })} = e^{1.82} e^{-0.0665 \text{Age}}$$

- the estimated odds of death is

$$\text{odds.}\widehat{\text{death}} \text{ (age)} = \frac{1 - \hat{\pi}(\text{ age })}{\hat{\pi}(\text{ age })} = e^{-1.82} e^{0.0665 \text{ Age}}$$

- the estimated odds of survival for a 20 year old is

$$\text{odds } \widehat{\text{(age}} = 20) = \frac{\hat{\pi}(\text{ age } = 20)}{1 - \hat{\pi}(\text{ age } = 20)} = e^{1.82} e^{-0.0665(20)} = 1.632$$

- the estimated odds of death for a 20 year old is

$$\text{odds.dea}\widehat{\text{th(age}} = 20) = \frac{1 - \hat{\pi}(\text{ age } = 20)}{\hat{\pi}(\text{ age } = 20)} = e^{-1.82} e^{0.0665(20)} = 0.613$$

Fitting logistic model in R

`glm` model attributes:

- `fitted(my.glm)` gives $\hat{\pi}(X)$ for each case in your data
- `predict(my.glm)` gives $\log \frac{\hat{\pi}(X)}{1-\hat{\pi}(X)}$ for each case in your data.
  - Add `newdata=` to get predicted log-odds for new data.
- `predict(my.glm, type = "response")` gives $\hat{\pi}(X)$ for each case in your data.
  - Add `newdata=` to get predicted log-odds for new data.

Example: Donner party model

Get the predicted probability of survival:

```
new_ages <- data.frame(Age = c(20, 45))
predict(donner_glm1, newdata = new_ages, type="response")
        1         2 
0.6198974 0.2363774 
```

with the default type, we get log odds survival estimates:

```
new_ages <- data.frame(Age = c(20, 45))
predict(donner_glm1, newdata = new_ages) # log odds
         1          2 
 0.4891127 -1.1726443 
```

```
exp(predict(donner_glm1, newdata = new_ages)) # odds
        1         2 
1.6308685 0.3095473 
```

Example: Donner party model

```
tidy(donner_glm1, conf.int=TRUE)
```

```
# A tibble: 2 × 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)   1.82      0.999       1.82  0.0688 -0.00599    3.99
2 Age          -0.0665    0.0322     -2.06  0.0391 -0.140     -0.0102
```

How does a one year increase in age affect survival status?

Example: Donner party model

A one year increase in age is associated with a $e^{-0.0665} = 0.936$ multiplicative decrease on the odds of survival (95% CI $0.87, 0.99$)

```
exp(-0.0665) # factor change
[1] 0.9356629
exp(c(-0.140, -0.0102)) # factor change CI
[1] 0.8693582 0.9898518
```

A one year increase in age is associated with a decrease on the odds of survival by 6.4%(95% CI 1% to 13%).

```
100*(exp(-0.0665) - 1) # percent change
[1] -6.433708
100*(exp(c(-0.140, -0.0102)) - 1) # % change CI
[1] -13.064176  -1.014816
```

Example: Donner party model

`broom` package: add `exponentiate=TRUE` to get exponentiatied estimated effects and confidence intervals

- but SE, test stat and p-values are untouched!

```
tidy(donner_glm1, conf.int=TRUE, exponentiate = TRUE)
```

```
# A tibble: 2 × 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)    6.16     0.999       1.82  0.0688    0.994    54.1
2 Age            0.936    0.0322     -2.06  0.0391    0.870     0.990
```

Example: Donner party model

```r
tidy(donner_glm1, conf.int=TRUE)
```

```
# A tibble: 2 × 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)  1.82      0.999        1.82  0.0688 -0.00599    3.99  
2 Age         -0.0665    0.0322      -2.06  0.0391 -0.140     -0.0102
```

How does a one year increase in age affect the odds of death?

$$\widehat{\text{odds.death}}(\text{age}) = \frac{1-\hat{\pi}}{\hat{\pi}} = \frac{1}{e^{1.82}e^{-0.0665\text{ Age}}} = e^{-1.82}e^{0.0665\text{ Age}}$$

Example: Donner party model

A one year increase in age is associated with a $e^{0.0665} = 1.069$ multiplicative increase on the odds of death $(95\%$ CI $1.01, 1.15)$

```
exp(0.0665) # factor change in odds of death
[1] 1.068761
```

```
exp(c(0.140, 0.0102)) # factor change CI
[1] 1.150274 1.010252
```

Example: Donner party model

A one year increase in age is associated with an increase in the odds of death by 6.9% (95% CI 1.0% to 15.0%)

```
100*(exp(0.0665) - 1) # percent change
[1] 6.876096
```

```
100*(exp(c(0.140, 0.0102)) - 1) # % change CI
[1] 15.02738  1.02522
```