Logistic Regression

Nate Wells

Math 243: Stat Learning

September 30th, 2020

Outline

In today's class, we will...

- Discuss Logistic Regression for Classification
- Implement Logistic Regression in R

Section 1

Logistic Regression

Classificaiton Problems

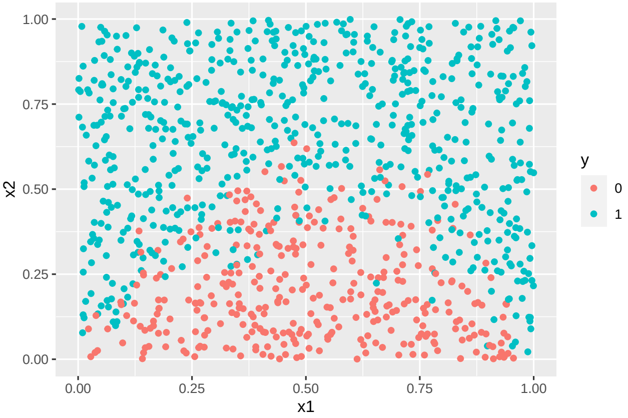
Suppose Y is a categorical variable with levels $A_1, A_2, \ldots, A_k$.

Goal: Build a model f to classify an observation into levels $A_1, A_2, \ldots, A_k$ based on the values of several predictors $X_1, X_2, \ldots, X_p$ (quantitative or categorical)

$$\hat{Y} = f(X_1, X_2, \ldots, X_p) \qquad \text{where } f \text{ take values in } \{A_1, \ldots, A_k\}$$

Classification Regions

Any classification model will divide predictor space into unions of regions, where each point in a region will be classified in the same way.



Different models will have different geometries for classification boundaries.

The Bayes Classifier and KNN

The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

The Bayes Classifier and KNN

The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

- In practice, the conditional probabilities are not known.

The Bayes Classifier and KNN

The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

- In practice, the conditional probabilities are not known.
- But we can approximate them using *KNN*:

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

Why not use KNN always?

1. KNN has very low training time (basically none), but often large test time (especially for large K)

Why not use KNN always?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.

Why not use KNN always?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.

Why not use KNN always?

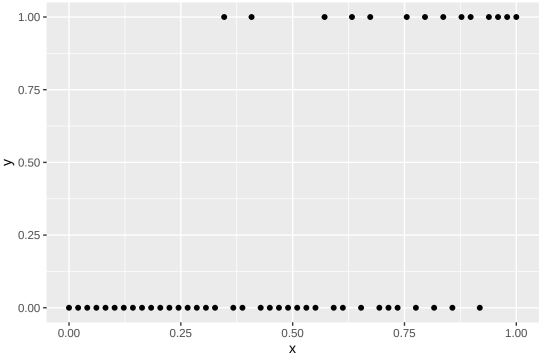
1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.
4. KNN suffers from the "curse of dimensionality". For fixed K and large p, adding more predictors increases bias and variance.

Why not use KNN always?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.
4. KNN suffers from the "curse of dimensionality". For fixed K and large p, adding more predictors increases bias and variance.
5. KNN requires large sample sizes (compared to alternatives)

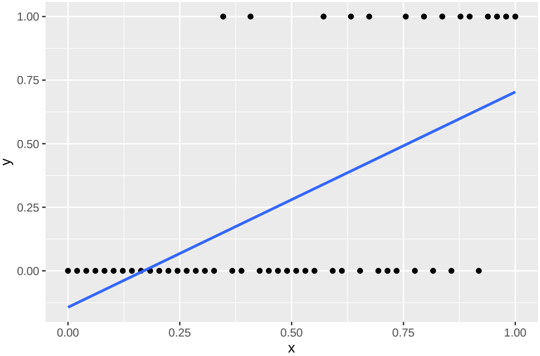
Alternatives

Suppose Y is a binary categorical variable with 1 quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



Alternatives

Suppose Y is a binary categorical variable with 1 quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



Linear model: $p(X) = \beta_0 + \beta_1 X$

Predict 1 if $\hat{Y} \geq 0.5$, and 0 otherwise.

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.
2. Too inflexible (enormous bias)

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.
2. Too inflexible (enormous bias)
3. Cannot easily extend to categorical Y with more than 2 levels.

Odds

Suppose an event occurs with probability p. The odds of the event occurring is

$$\text{odds} = \frac{p}{1-p}$$

Odds

Suppose an event occurs with probability p. The odds of the event occurring is

$$\text{odds} = \frac{p}{1-p}$$

- If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
- If $p = .5$, then $\text{odds} = 1$ (or even odds).

Odds

Suppose an event occurs with probability p. The odds of the event occurring is

$$\text{odds} = \frac{p}{1-p}$$

- If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
- If $p = .5$, then $\text{odds} = 1$ (or even odds).

For extremely likely or unlikely events, odds can be astronomical.

Odds

Suppose an event occurs with probability p. The odds of the event occurring is

$$\text{odds} = \frac{p}{1-p}$$

- If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
- If $p = .5$, then $\text{odds} = 1$ (or even odds).

For extremely likely or unlikely events, odds can be astronomical.

- "The possibility of successfully navigating an asteroid field is approximately 3,720 to 1"

Odds

Suppose an event occurs with probability p. The odds of the event occurring is

$$\text{odds} = \frac{p}{1-p}$$

- If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
- If $p = .5$, then $\text{odds} = 1$ (or even odds).

For extremely likely or unlikely events, odds can be astronomical.

- "The possibility of successfully navigating an asteroid field is approximately 3,720 to 1"

Instead, we consider log odds:

$$\text{log odds} = \ln \frac{p}{1-p} = \ln p - \ln(1-p)$$

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*

Solving for odds:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

Logistic Regression

Suppose Y is binary categorical, and that the log odds of $Y = 1$ is linear in X.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*

Solving for odds:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

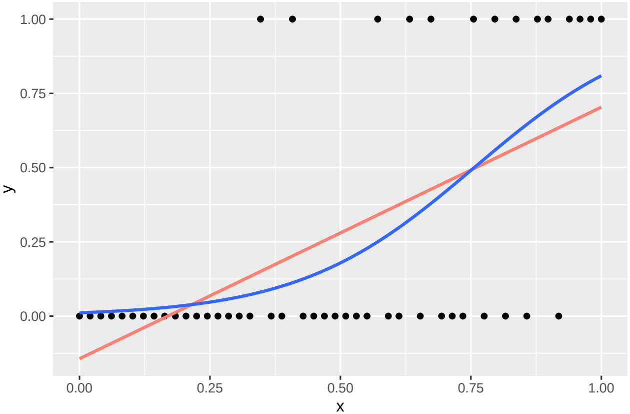
Solving for $p(X)$:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

The Logistic Curve

The conditional probability $p(X)$ takes the form of a logistic curve:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



Multiple Logistic Regression

Nothing stops us from modeling Y based on more than 1 predictor.

Multiple Logistic Regression

Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Multiple Logistic Regression

Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

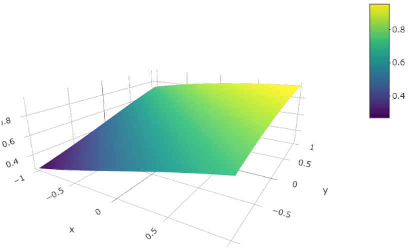
$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Multiple Logistic Regression

Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$



Interactive graphic on schedule page of course website

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

1. For historical reasons

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

1. For historical reasons
2. Due to its relative simplicity

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to...

1. Create spam filters

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to...

1. Create spam filters
2. Forecast election results

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to. . .

1. Create spam filters
2. Forecast election results
3. Investigate health outcomes based on comorbidities

Section 2

Creating Logistic Models

The Maximum Likelihood Method

Assume that the log-odds of $Y = 1$ is indeed linear in X, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- We need to estimate the parameters β_0, β_1 based on training data.

The Maximum Likelihood Method

Assume that the log-odds of $Y = 1$ is indeed linear in X, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- We need to estimate the parameters β_0, β_1 based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.

The Maximum Likelihood Method

Assume that the log-odds of $Y = 1$ is indeed linear in X, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- We need to estimate the parameters β_0, β_1 based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
- But it turns out the method of Maximum Likelihood is preferable, since it allows us to relax some conditions on residuals.

The Maximum Likelihood Method

Assume that the log-odds of $Y = 1$ is indeed linear in X, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- We need to estimate the parameters β_0, β_1 based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
- But it turns out the method of Maximum Likelihood is preferable, since it allows us to relax some conditions on residuals.

The likelihood function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

The Maximum Likelihood Method

Assume that the log-odds of $Y = 1$ is indeed linear in X, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- We need to estimate the parameters β_0, β_1 based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
- But it turns out the method of Maximum Likelihood is preferable, since it allows us to relax some conditions on residuals.

The likelihood function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- The goal is to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so as to maximize ℓ

The Unsinkable Example

The `Titanic` data set contains information on passengers of the *Titanic*

```
## Warning: 2 parsing failures.
## row  col          expected actual              file
##  37 name delimiter or quote      M 'data/titanic.csv'
##  37 name delimiter or quote        'data/titanic.csv'

## Rows: 1,313
## Columns: 11
## $ row.names <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17...
## $ pclass    <chr> "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st", "...
## $ survived  <dbl> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, ...
## $ name      <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen Lora...
## $ age       <dbl> 29.0000, 2.0000, 30.0000, 25.0000, 0.9167, 47.0000, 63.00...
## $ embarked  <chr> "Southampton", "Southampton", "Southampton", "Southampton...
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montr...
## $ room      <chr> "B-5", "C26", "C26", "C26", "C22", "E-12", "D-7", "A-36",...
## $ ticket    <chr> "24160 L221", NA, NA, NA, NA, NA, "13502 L77", NA, NA, NA...
## $ boat      <chr> "2", NA, "(135)", NA, "11", "3", "10", NA, "2", "(22)", "...
## $ sex       <chr> "female", "female", "male", "female", "male", "male", "fe...
```

What relationship can we discover between survival, sex, and age?

Data Processing

```
summary(Titanic)
```

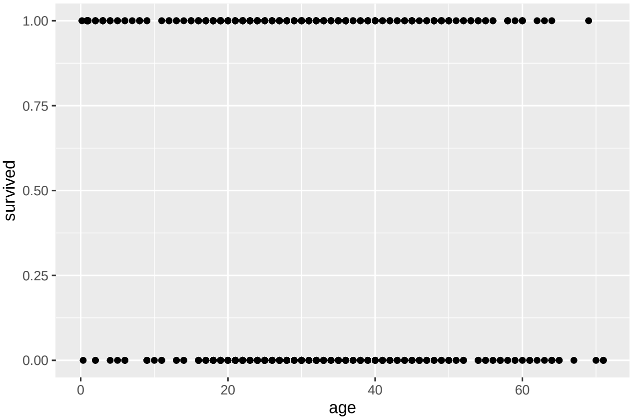
```
##    row.names      pclass            survived        name
##  Min.   :   1   Length:1313        Min.   :0.000   Length:1313
##  1st Qu.: 329   Class :character   1st Qu.:0.000   Class :character
##  Median : 657   Mode  :character   Median :0.000   Mode  :character
##  Mean   : 657                      Mean   :0.342
##  3rd Qu.: 985                      3rd Qu.:1.000
##  Max.   :1313                      Max.   :1.000
##
##       age          embarked          home.dest            room
##  Min.   : 0.1667   Length:1313       Length:1313        Length:1313
##  1st Qu.:21.0000   Class :character   Class :character   Class :character
##  Median :30.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :31.1942
##  3rd Qu.:41.0000
##  Max.   :71.0000
##  NA's   :680
##     ticket             boat               sex
##  Length:1313        Length:1313        Length:1313
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

What do we do about those NA's?

```
library(tidyr)
Titanic1<-Titanic %>% drop_na(age)
```

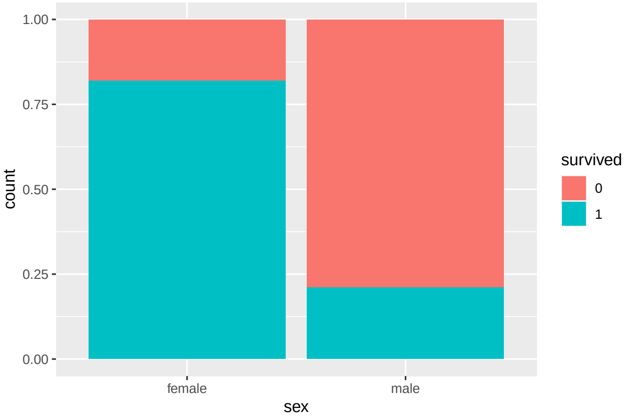
Children first?

```
Titanic1 %>% ggplot( aes( x = age, y = survived))+ geom_point()
```



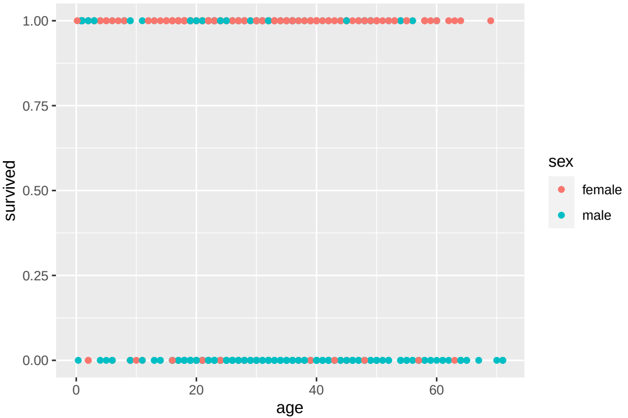
Women First?

```
Titanic1 %>%  mutate(survived = as.factor(survived)) %>%
  ggplot( aes( x = sex, fill = survived))+
  geom_bar(position = "fill")
```



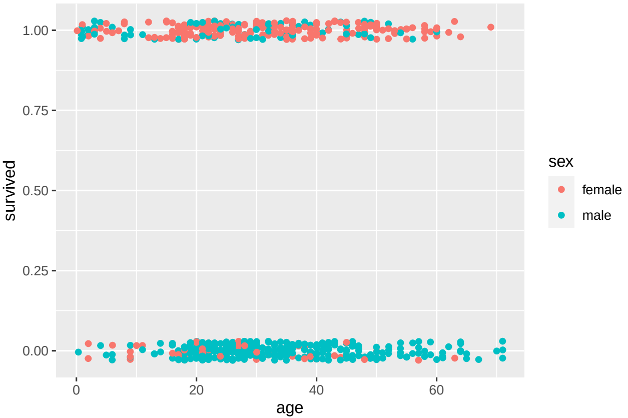
Women and Children First?

```
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex))+ geom_point()
```



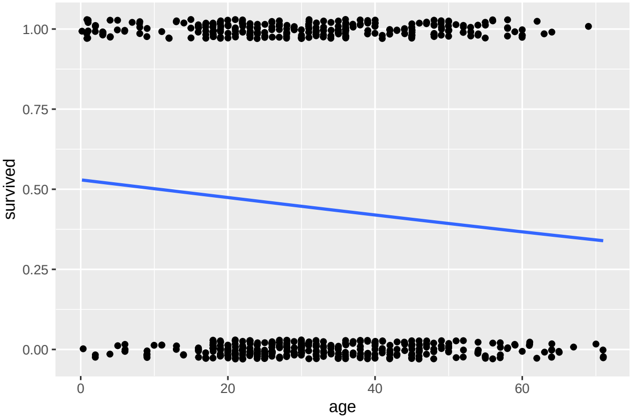
Women and Children First?

```
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex))+ geom_jitter(height =
```



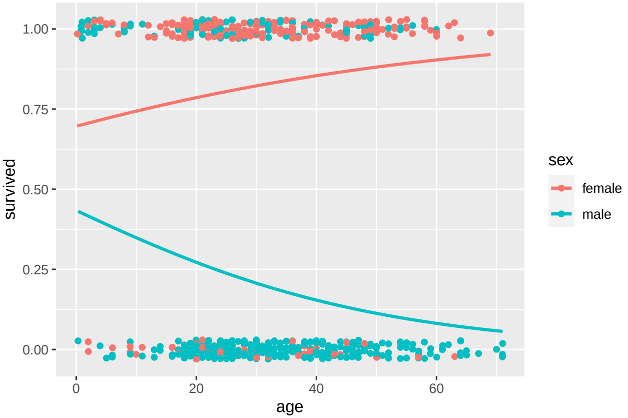
Logistic Model 1

```
Titanic1 %>% ggplot( aes( x = age, y = survived ))+
  geom_jitter(height = 0.03) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F)
```



Logistic Models 2 and 3

```
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex ))+
  geom_jitter(height = 0.03) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F)
```



R code for Logistic Models

```
simple_logreg <- glm(survived ~ age, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##                 Estimate  Std. Error    z value   Pr(>|z|)
## (Intercept)  0.11719513 0.187746466  0.6242202 0.53248299
## age         -0.01102924 0.005492735 -2.0079686 0.04464663
```

R code for Logistic Models

```
simple_logreg <- glm(survived ~ age, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##                Estimate  Std. Error    z value   Pr(>|z|)
## (Intercept)  0.11719513 0.187746466  0.6242202 0.53248299
## age         -0.01102924 0.005492735 -2.0079686 0.04464663
```

$\ln \frac{p(\text{Age})}{1-p(\text{Age})} = 0.11 - 0.01 \cdot \text{Age}$

R code for Logistic Models

```
simple_logreg <- glm(survived ~ age, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##                Estimate  Std. Error    z value   Pr(>|z|)
## (Intercept)  0.11719513 0.187746466  0.6242202 0.53248299
## age         -0.01102924 0.005492735 -2.0079686 0.04464663
```

$\ln \frac{p(\text{Age})}{1-p(\text{Age})} = 0.11 - 0.01 \cdot \text{Age}$

Since $e^{0.011} = 1.01106$, increasing age by 1 year decreases survival probability by 1.106%

R code for Multiple Logistic Models

```
simple_logreg <- glm(survived ~ age + sex, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  1.9158497 0.278035089   6.890676 5.552794e-12
## age         -0.0129209 0.006863803  -1.882469 5.977237e-02
## sexmale     -2.8415031 0.209063920 -13.591552 4.494495e-42
```

R code for Multiple Logistic Models

```
simple_logreg <- glm(survived ~ age + sex, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##               Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  1.9158497 0.278035089   6.890676 5.552794e-12
## age         -0.0129209 0.006863803  -1.882469 5.977237e-02
## sexmale     -2.8415031 0.209063920 -13.591552 4.494495e-42
```

$\ln \frac{p(X)}{1-p(X)} = 1.91 - 0.012 \cdot \text{Age} - 2.85 \cdot \text{Male}$

R code for Multiple Logistic Models

```
simple_logreg <- glm(survived ~ age + sex, data = Titanic1, family = "binomial")

summary(simple_logreg)$coefficients
```

```
##               Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  1.9158497 0.278035089   6.890676 5.552794e-12
## age         -0.0129209 0.006863803  -1.882469 5.977237e-02
## sexmale     -2.8415031 0.209063920 -13.591552 4.494495e-42
```

$\ln \frac{p(X)}{1-p(X)} = 1.91 - 0.012 \cdot \text{Age} - 2.85 \cdot \text{Male}$

What is the survival probability for a male child of age 5?

Classification using Logistic Regression

Develop a classification scheme based on the linear regression model.

Classification using Logistic Regression

Develop a classification scheme based on the linear regression model.

$$\hat{Y} = \begin{cases} 1, & \text{if } p(X) \geq 1 - p(X), \\ 0, & \text{otherwise.} \end{cases}$$

Classification using Logistic Regression

Develop a classification scheme based on the linear regression model.

$$\hat{Y} = \begin{cases} 1, & \text{if } p(X) \geq 1 - p(X), \\ 0, & \text{otherwise.} \end{cases}$$

$$\hat{Y} = \begin{cases} 1, & \text{if odds } \geq 1, \\ 0, & \text{if odds } < 1 \end{cases}$$

Classification using Logistic Regression

Develop a classification scheme based on the linear regression model.

$$\hat{Y} = \begin{cases} 1, & \text{if } p(X) \geq 1 - p(X), \\ 0, & \text{otherwise.} \end{cases}$$

$$\hat{Y} = \begin{cases} 1, & \text{if odds } \geq 1, \\ 0, & \text{if odds } < 1 \end{cases}$$

$$\hat{Y} = \begin{cases} 1, & \text{if log odds } \geq 0, \\ 0, & \text{if log odds } < 0 \end{cases}$$