Logistic Regression

Nate Wells

Math 243: Stat Learning

October 25th, 2021

Outline

In today's class, we will...

- Implement Logistic Regression in R

Section 1

Applications of Linear Regression

The Unsinkable Example

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

```
## 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, 1~
## $ pclass    <chr> "1st", "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, 1, ~
## $ name      <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen Loraine~
## $ age       <dbl> 29.0000, 2.0000, 30.0000, 25.0000, 0.9167, 47.0000, 63.0000,~
## $ embarked  <chr> "Southampton", "Southampton", "Southampton", "Southampton", ~
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montreal~
## $ room      <chr> "B-5", "C26", "C26", "C26", "C22", "E-12", "D-7", "A-36", "C~
## $ 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)", "(12~
## $ sex       <chr> "female", "female", "male", "female", "male", "male", "femal~
```

- Goal: Determine relationship between survival, sex, and age.

Data Analysis

```
library(skimr)
Titanic %>% select(age, sex, survived) %>% summary()
```

```
##       age              sex              survived
##  Min.   : 0.1667   Length:1313        Min.   :0.000
##  1st Qu.:21.0000   Class :character   1st Qu.:0.000
##  Median :30.0000   Mode  :character   Median :0.000
##  Mean   :31.1942                      Mean   :0.342
##  3rd Qu.:41.0000                      3rd Qu.:1.000
##  Max.   :71.0000                      Max.   :1.000
##  NA's   :680
```

```
Titanic %>% count(sex)
```

```
## # A tibble: 2 x 2
##   sex        n
##   <chr>  <int>
## 1 female   463
## 2 male     850
```

```
Titanic %>% count(survived)
```

```
## # A tibble: 2 x 2
##   survived     n
##      <dbl> <int>
## 1        0   864
## 2        1   449
```

- What are some concerns we may have about variables `sex`, age and `survival`?

Data Analysis

```
library(skimr)
Titanic %>% select(age, sex, survived) %>% summary()
```

```
##       age              sex                survived
##  Min.   : 0.1667   Length:1313        Min.   :0.000
##  1st Qu.:21.0000   Class :character   1st Qu.:0.000
##  Median :30.0000   Mode  :character   Median :0.000
##  Mean   :31.1942                      Mean   :0.342
##  3rd Qu.:41.0000                      3rd Qu.:1.000
##  Max.   :71.0000                      Max.   :1.000
##  NA's   :680
```

```
Titanic %>% count(sex)
```

```
## # A tibble: 2 x 2
##   sex        n
##   <chr>  <int>
## 1 female   463
## 2 male     850
```

```
Titanic %>% count(survived)
```

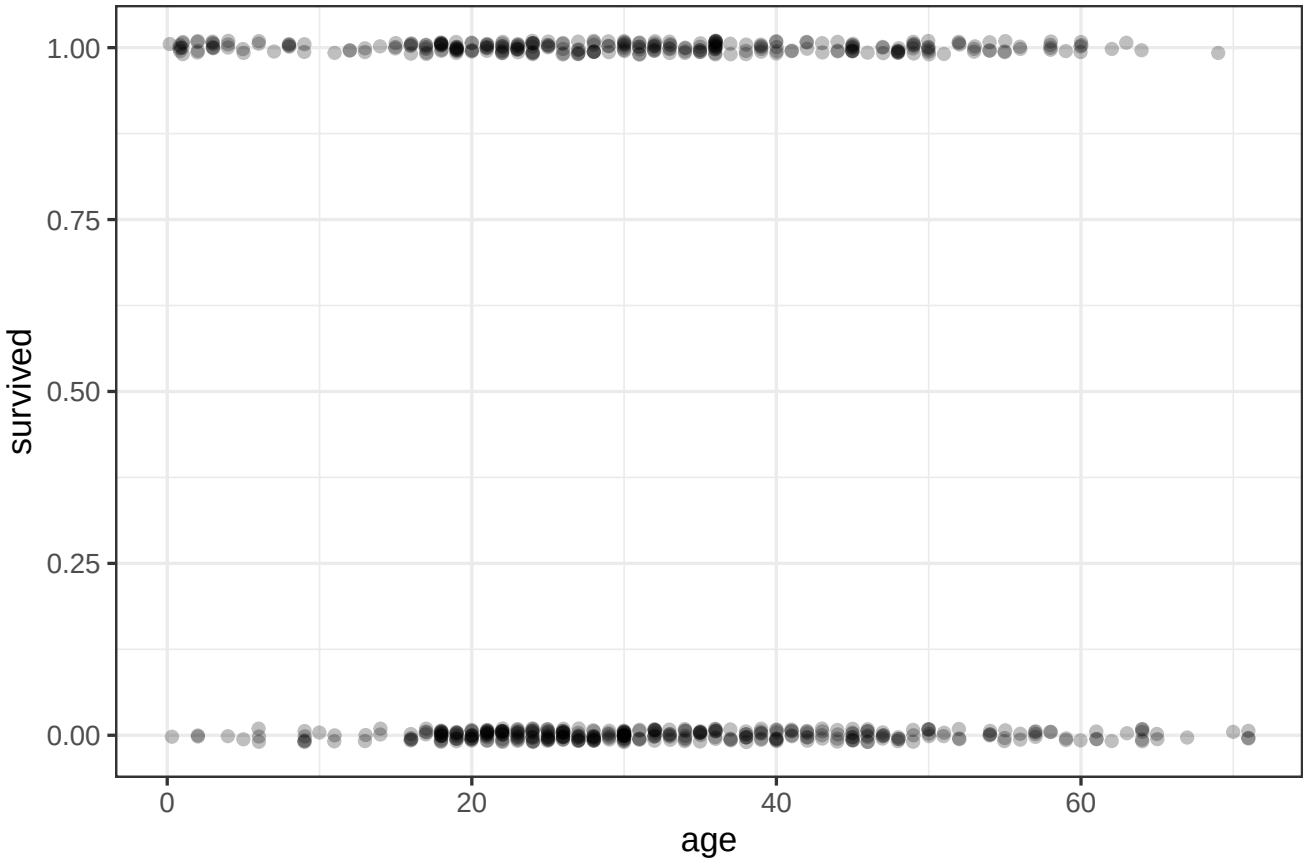
```
## # A tibble: 2 x 2
##   survived     n
##      <dbl> <int>
## 1        0   864
## 2        1   449
```

- What are some concerns we may have about variables `sex`, `age` and `survival`?

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

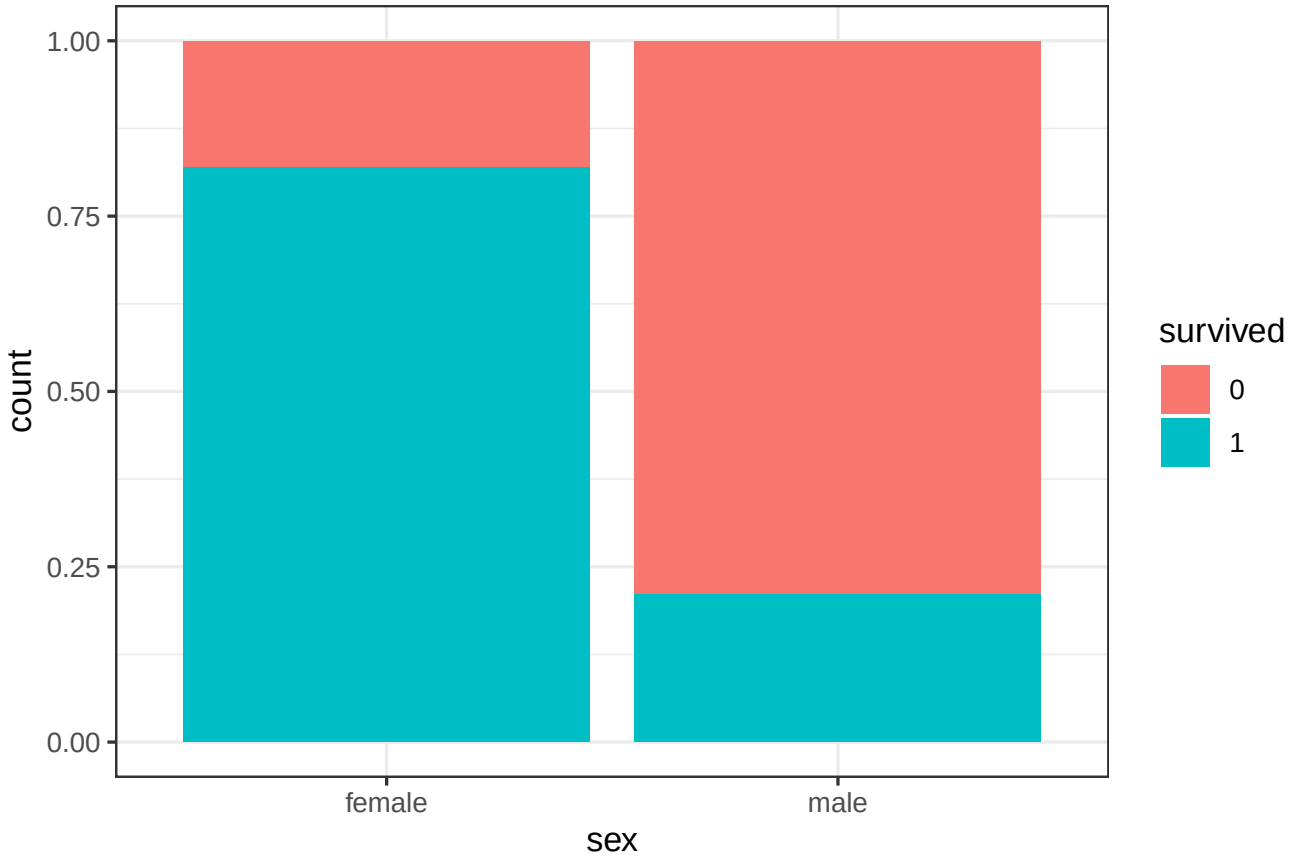
Children first?

- Who survived the Titanic?



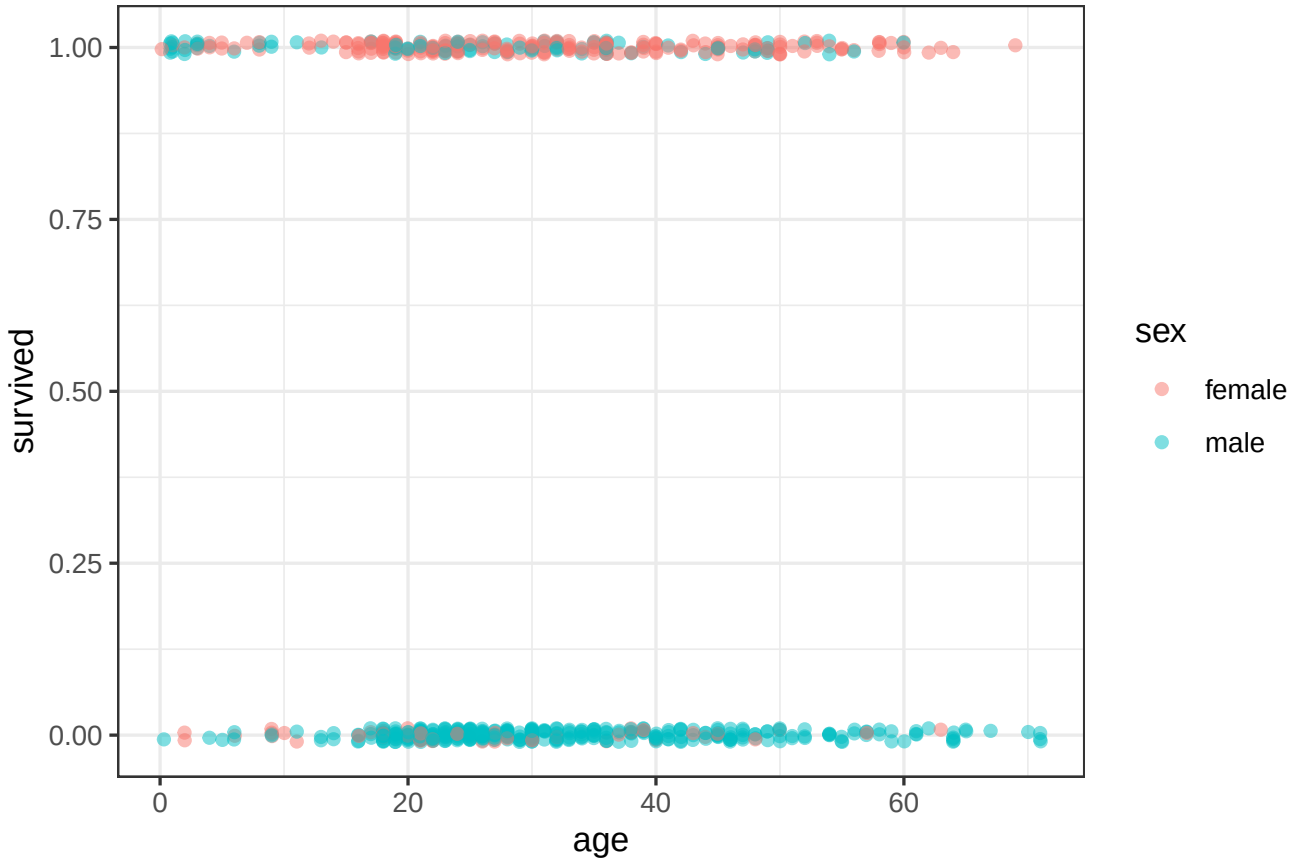
Women First?

- Who survived the Titanic?



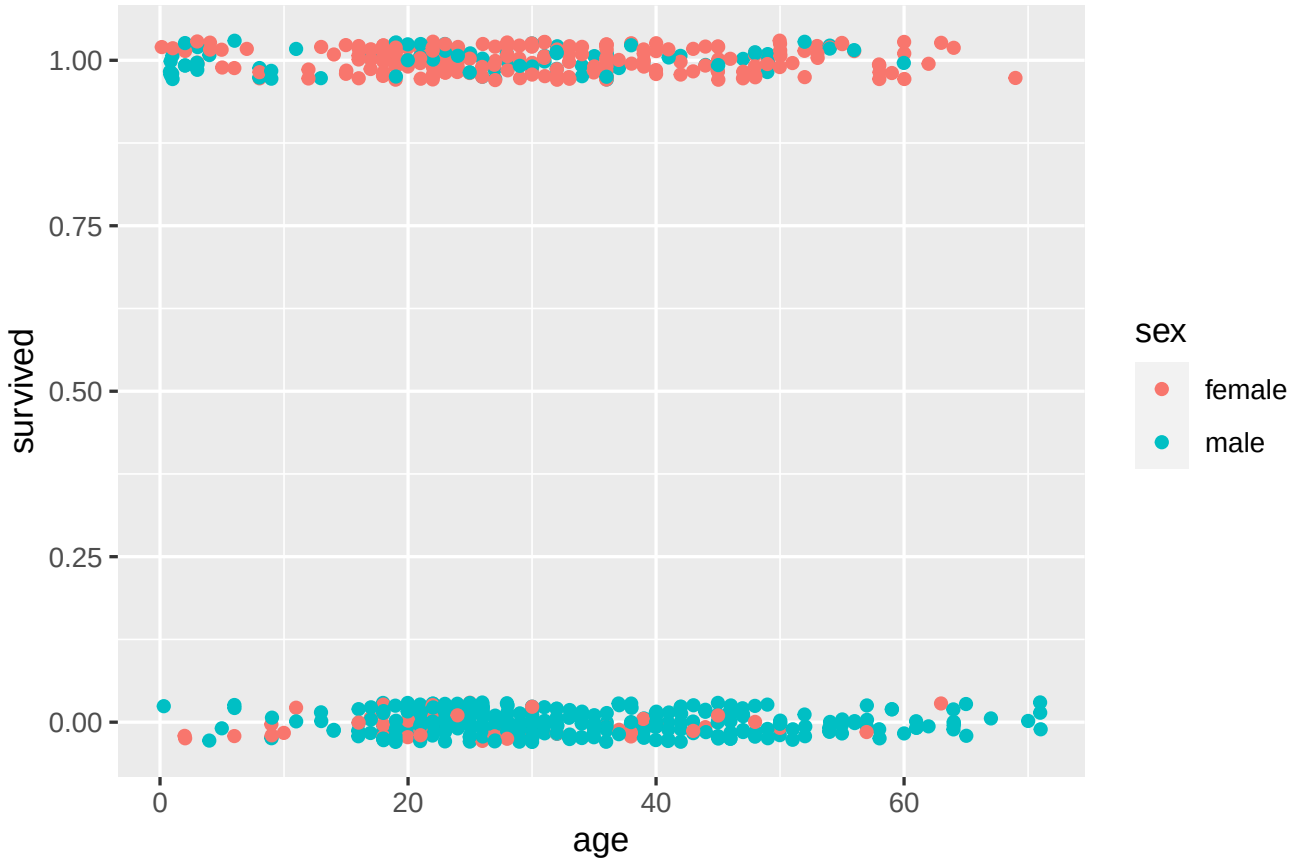
Women and Children First?

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



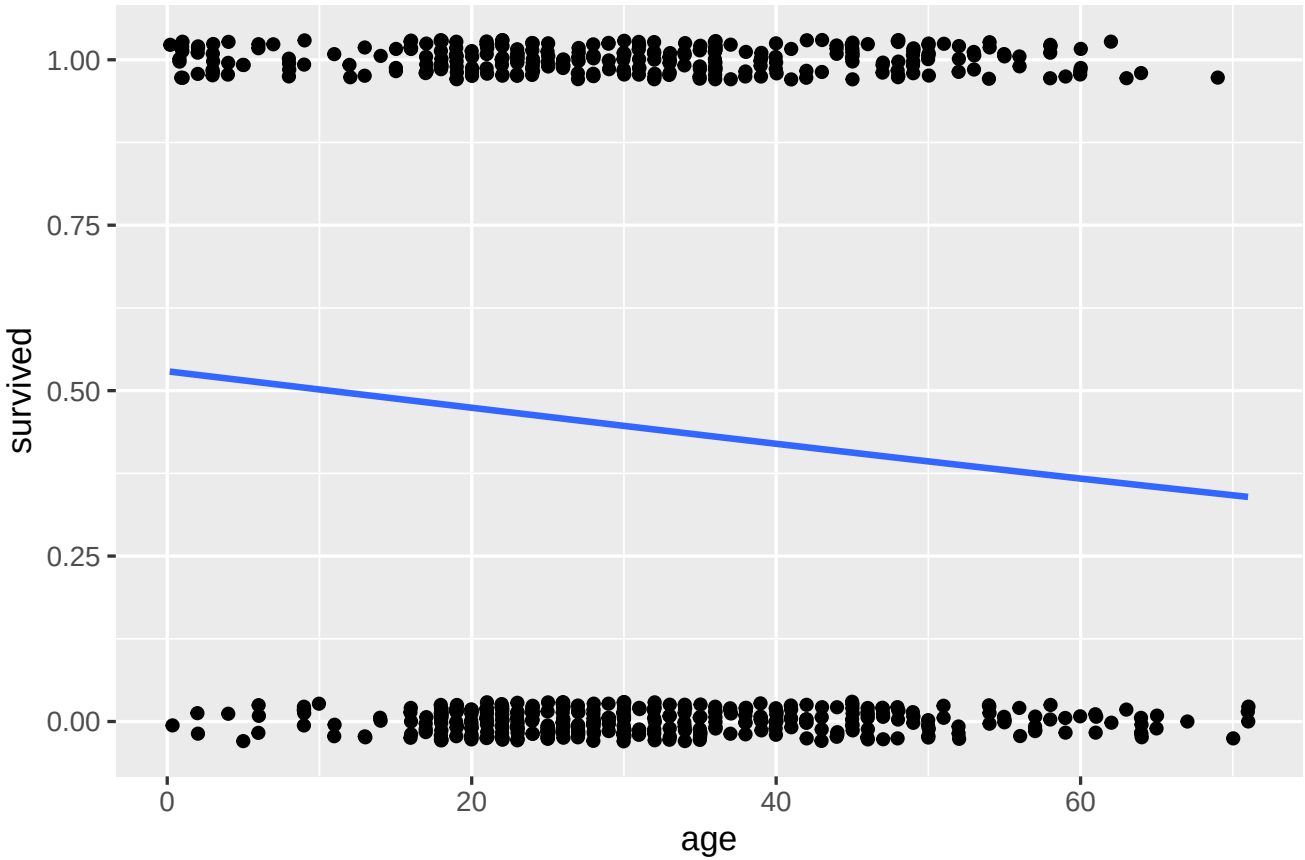
Women and Children First?

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



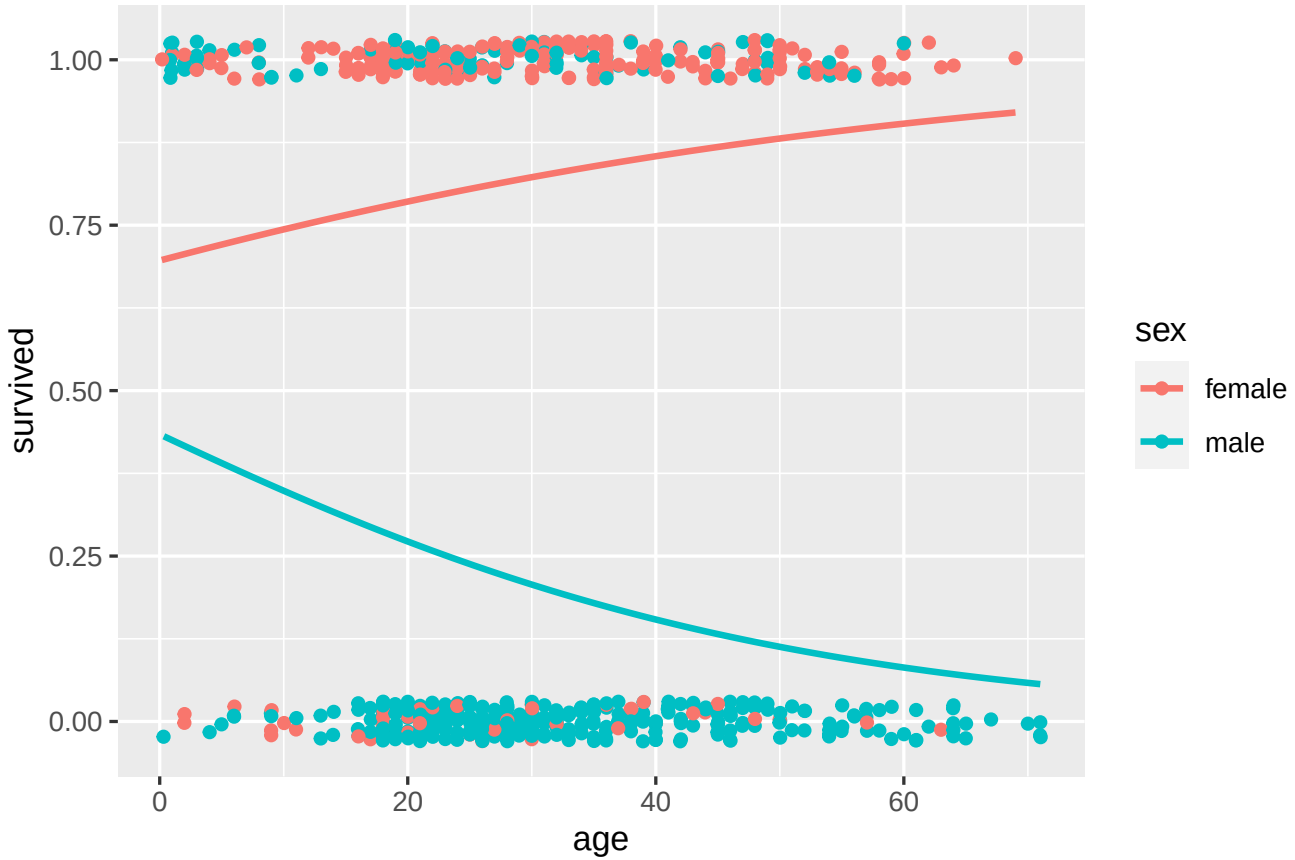
Logistic Model 1

```
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

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

summary(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

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

summary(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

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

summary(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.

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

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

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

Prediction and Classification in R

Suppose we have 10 hypothetical passengers with the following age/sex combinations:

```
passengers
```

```
##    age    sex
## 1   10   male
## 2   14 female
## 3   18   male
## 4   22   male
## 5   26 female
## 6   30   male
## 7   34   male
## 8   38   male
## 9   42 female
## 10  46 female
```

Prediction and Classification in R

What are their survival log odds?

```
predict(logreg, passengers)
```

```
##         1         2         3         4         5         6         7         8
## -1.054862  1.734957 -1.158230 -1.209913  1.579906 -1.313280 -1.364964 -1.416647
##         9        10
##  1.373172  1.321488
```

Prediction and Classification in R

What are their survival log odds?

```
predict(logreg, passengers)
```

```
##         1         2         3         4         5         6         7         8
## -1.054862  1.734957 -1.158230 -1.209913  1.579906 -1.313280 -1.364964 -1.416647
##         9        10
##  1.373172  1.321488
```

Survival probabilities?

```
predict(logreg, passengers, type = "response")
```

```
##         1         2         3         4         5         6         7         8
## 0.2582925 0.8500454 0.2389891 0.2297164 0.8291913 0.2119384 0.2034347 0.1951877
##         9        10
## 0.7978922 0.7894292
```

Prediction and Classification in R

What are their survival log odds?

```
predict(logreg, passengers)
```

```
##         1         2         3         4         5         6         7         8
## -1.054862  1.734957 -1.158230 -1.209913  1.579906 -1.313280 -1.364964 -1.416647
##         9        10
##  1.373172  1.321488
```

Survival probabilities?

```
predict(logreg, passengers, type = "response")
```

```
##         1         2         3         4         5         6         7         8
## 0.2582925 0.8500454 0.2389891 0.2297164 0.8291913 0.2119384 0.2034347 0.1951877
##         9        10
## 0.7978922 0.7894292
```

Classification?

```
ifelse(predict(logreg, passengers, type = "response") >= .5, 1, 0)
```

```
##  1  2  3  4  5  6  7  8  9 10
##  0  1  0  0  1  0  0  0  1  1
```

Confusion Tables

How well does our model do on training data?

```
probs<-predict(logreg, Titanic1, type = "response")
preds<-ifelse(probs >=.5, 1, 0)
conf_log <- table(preds, Titanic1$survived)
conf_log
```

```
##
## preds   0   1
##     0 308  82
##     1  44 199
```

Confusion Tables

How well does our model do on training data?

```
probs<-predict(logreg, Titanic1, type = "response")
preds<-ifelse(probs >=.5, 1, 0)
conf_log <- table(preds, Titanic1$survived)
conf_log
```

```
##
## preds   0   1
##     0 308  82
##     1  44 199
```

Training Error rate:

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Confusion Tables

How well does our model do on training data?

```
probs<-predict(logreg, Titanic1, type = "response")
preds<-ifelse(probs >=.5, 1, 0)
conf_log <- table(preds, Titanic1$survived)
conf_log
```

```
##
## preds   0   1
##     0 308  82
##     1  44 199
```

Training Error rate:

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

```
n <- length(Titanic1$survived)
false_pos <- conf_log[1,2]
false_neg <- conf_log[2,1]
error <- 1/n *(false_pos + false_neg)
error
```

```
## [1] 0.1990521
```

A better confusion matrix

The `confusionMatrix` function in the `caret` package provides a confusion matrix along withe relevant statistics:

```
library(caret)
confusionMatrix(data =   factor(preds) , reference =  factor(Titanic1$survived)  )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 308  82
##          1  44 199
##
##                Accuracy : 0.8009
##                  95% CI : (0.7677, 0.8314)
##     No Information Rate : 0.5561
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5912
##
##  Mcnemar's Test P-Value : 0.0009799
##
##             Sensitivity : 0.8750
##             Specificity : 0.7082
##          Pos Pred Value : 0.7897
##          Neg Pred Value : 0.8189
##              Prevalence : 0.5561
##          Detection Rate : 0.4866
##    Detection Prevalence : 0.6161
##       Balanced Accuracy : 0.7916
##
##        'Positive' Class : 0
##
```

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification

- Type II Error rate: $1 - \text{Sensitivity}$

Specificity: Rate of correct negative identification

- Type I Error rate: $1 - \text{Specificity}$

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification

- Type II Error rate: $1 - \text{Sensitivity}$

Specificity: Rate of correct negative identification

- Type I Error rate: $1 - \text{Specificity}$

By changing our classification cutoff, we can increase sensitivity to the detriment of specificity (or vice versa)

- But the tradeoff is non-linear

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification

- Type II Error rate: $1 - \text{Sensitivity}$

Specificity: Rate of correct negative identification

- Type I Error rate: $1 - \text{Specificity}$

By changing our classification cutoff, we can increase sensitivity to the detriment of specificity (or vice versa)

- But the tradeoff is non-linear
  - Increasing specificity by .1 may decrease sensitivity by .15 when specificity is .8
  - But the same increase in specificity may decrease sensitivity by .25 when specificity is .9.

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification

- Type II Error rate: $1 - \text{Sensitivity}$

Specificity: Rate of correct negative identification

- Type I Error rate: $1 - \text{Specificity}$

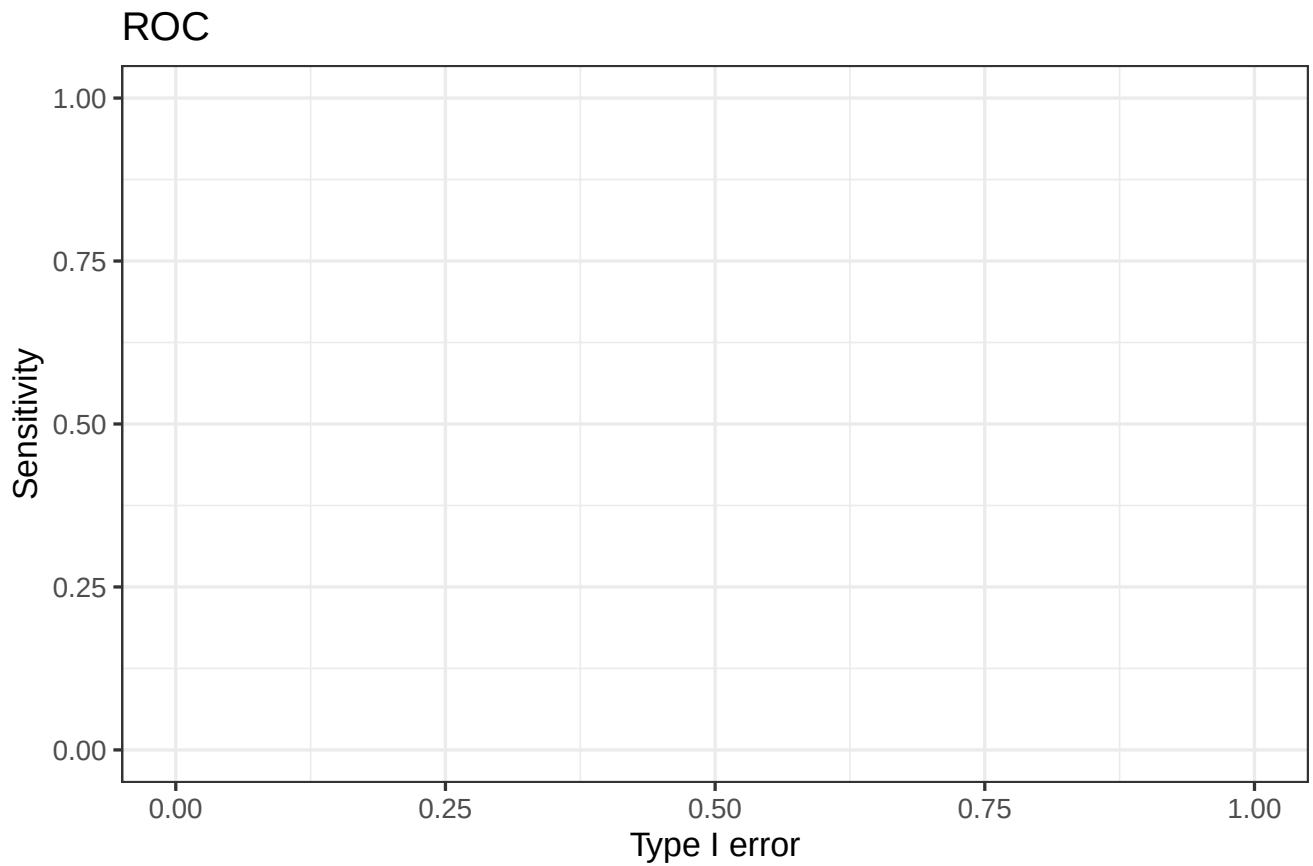
By changing our classification cutoff, we can increase sensitivity to the detriment of specificity (or vice versa)

- But the tradeoff is non-linear
  - Increasing specificity by .1 may decrease sensitivity by .15 when specificity is .8
  - But the same increase in specificity may decrease sensitivity by .25 when specificity is .9.

We measure the relative effect of sensitivity and specificity using an ROC curve

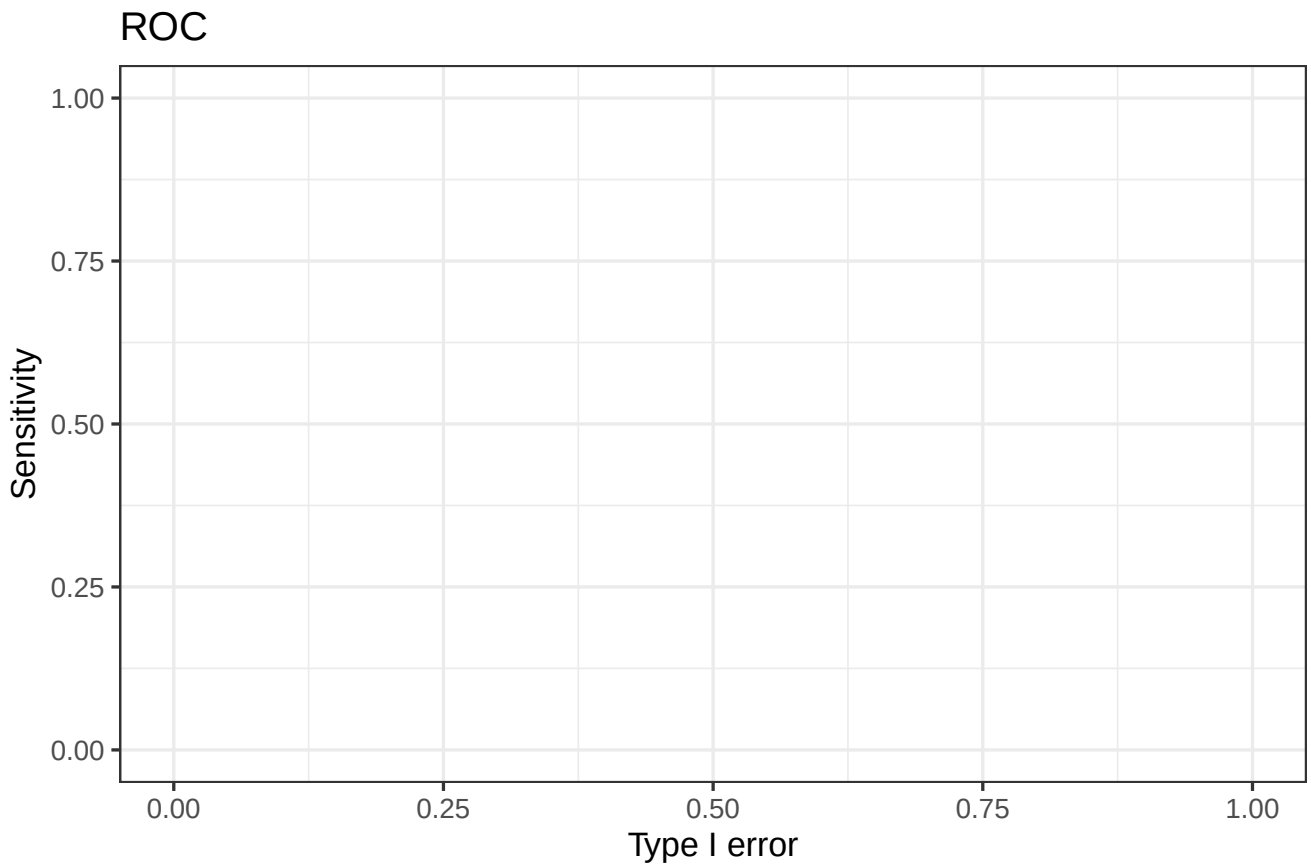
ROC Curves

A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity vs. type I error rate, based on classification probabilities.



ROC Curves

A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity vs. type I error rate, based on classification probabilities.



Poll: For a perfectly accurate model, what is the expected area under the ROC curve?

ROC Curves in R

The `roc` function in the `pROC` package can create ROC curves.

```
library(pROC)
curve <- roc(response = Titanic1$survived, predictor = probs)
plot(curve, legacy.axes=TRUE)
```