Logistic Regression

Nate Wells

Math 243: Stat Learning

October 29th, 2021

Outline

In today's class, we will. . .

- Implement logistic regression in R
- Discuss measurements for assessing classification models

Section 1

Logistic Regression Practice

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.

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.
- Is this primarily an inference or prediction task?

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.
- Is this primarily an inference or prediction task?
    - Can it be neither?

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)
```

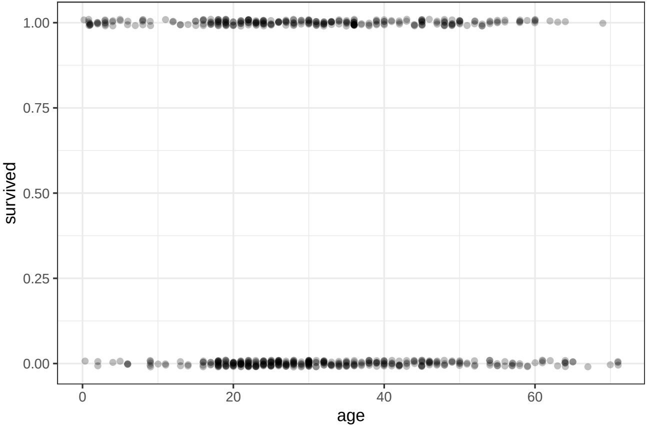
```
## # A tibble: 2 x 2
##   survived     n
##      <dbl> <int>
## 1        0   864
## 2        1   449
```

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

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

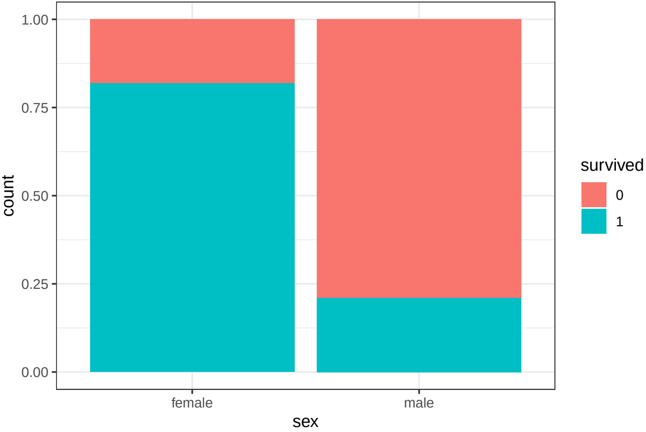
Children first?

- Who survived the Titanic?



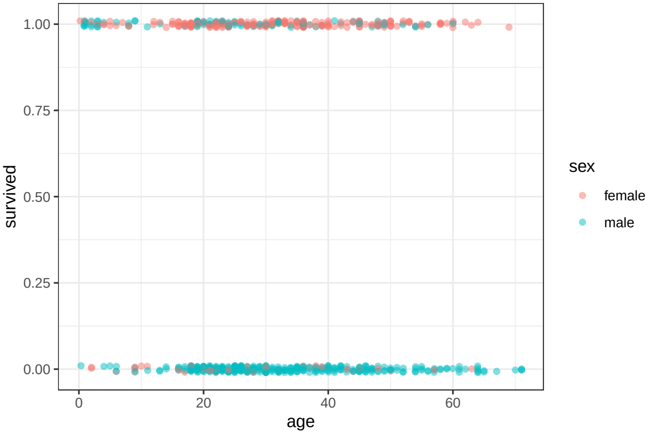
Women First?

- Who survived the Titanic?



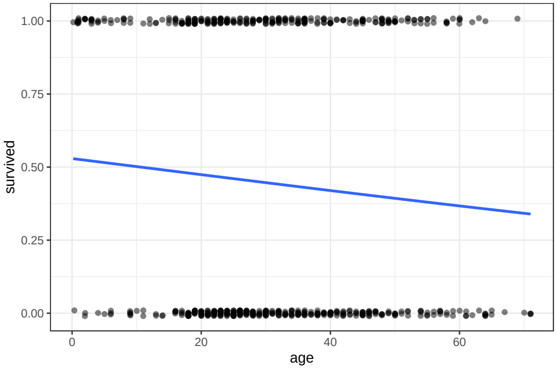
Women and Children First?

```
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()
```



Logistic Model 1

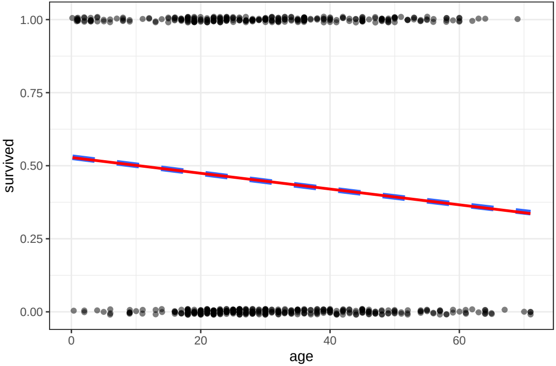
```
Titanic1 %>% ggplot( aes( x = age, y = survived ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F)
```



$$p(X) = \frac{e^{0.117 - 0.01X}}{1 + e^{0.117 - 0.01X}}$$

VS Linear Model

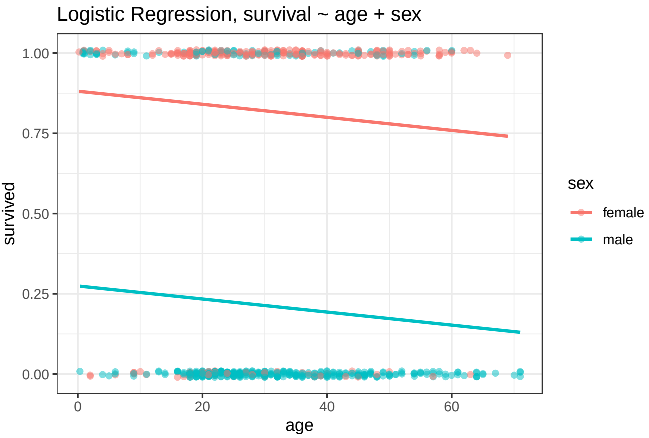
```
Titanic1 %>% ggplot( aes( x = age, y = survived ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F,size = 2,linetype
  geom_smooth(method = "lm", se = F, color  = "red")
```



$$p(X) = 0.528 - 0.003X$$

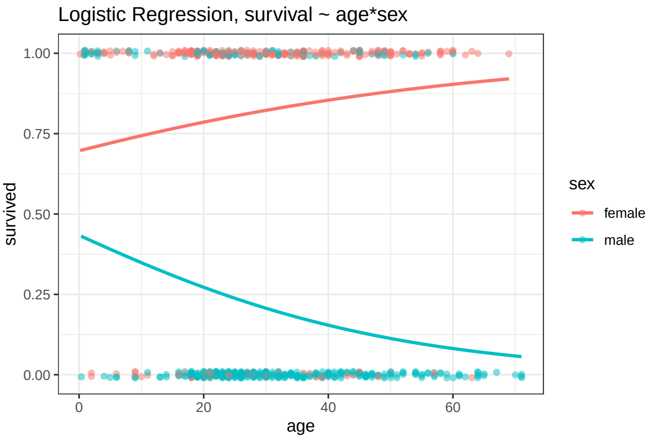
Logistic Model 2:

```
library(moderndive)
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_parallel_slopes(method = "glm", method.args = list(family = "binomial"), se = F)+
  labs(title = "Logistic Regression, survival ~ age + sex")
```



Logistic Model 3:

```
library(moderndive)
Titanic1 %>% ggplot( aes( x = age, y = survived, color = sex ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F)+
  labs(title = "Logistic Regression, survival ~ age*sex")
```



R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

- The logistic model is

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

R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

- The logistic model is

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

- Since

$$e^{-0.011} = 0.989 = 1 - 0.011$$

  increasing age by 1 year decreases survival probability by 1.1% **of the current probability.**

R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

- Where is RSE? R^2? F-stat?

R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

- Where is RSE? R^2? F-stat?
- Logistic regression is from the family of *generalized linear models*
  - GLiMs use *deviance* as metric of model fit.
  - Null deviance measures how well the null model (only intercept) predicts the data
  - Residual deviance measures how well the fitted model predicts the data

R code for Logistic Models

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

```
##
## Call:
## glm(formula = survived ~ age, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -1.0972  -0.9908   1.2502   1.4601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.117195   0.187746   0.624   0.5325
## age         -0.011029   0.005493  -2.008   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 865.47  on 631  degrees of freedom
## AIC: 869.47
##
## Number of Fisher Scoring iterations: 4
```

- Where is RSE? R^2? F-stat?
- Logistic regression is from the family of *generalized linear models*
  - GLiMs use *deviance* as metric of model fit.
  - Null deviance measures how well the null model (only intercept) predicts the data
  - Residual deviance measures how well the fitted model predicts the data
- Fisher Scoring Iterations indicates the number of loops of numeric optimization algorithm

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age + sex`:

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

```
##
## Call:
## glm(formula = survived ~ age + sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -0.7062  -0.6071   0.6452   1.9332
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.915850   0.278035   6.891 5.55e-12 ***
## age         -0.012921   0.006864  -1.882   0.0598 .
## sexmale     -2.841503   0.209064 -13.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 627.45  on 630  degrees of freedom
## AIC: 633.45
##
## Number of Fisher Scoring iterations: 4
```

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age + sex`:

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

```
##
## Call:
## glm(formula = survived ~ age + sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -0.7062  -0.6071   0.6452   1.9332
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.915850   0.278035   6.891 5.55e-12 ***
## age         -0.012921   0.006864  -1.882   0.0598 .
## sexmale     -2.841503   0.209064 -13.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 627.45  on 630  degrees of freedom
## AIC: 633.45
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age + sex`:

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

```
##
## Call:
## glm(formula = survived ~ age + sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -0.7062  -0.6071   0.6452   1.9332
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.915850   0.278035   6.891 5.55e-12 ***
## age         -0.012921   0.006864  -1.882   0.0598 .
## sexmale     -2.841503   0.209064 -13.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 627.45  on 630  degrees of freedom
## AIC: 633.45
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?
- What is the survival probability for a male child of age 5? A female child of age 5?

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age + sex`:

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

```
##
## Call:
## glm(formula = survived ~ age + sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -0.7062  -0.6071   0.6452   1.9332
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.915850   0.278035   6.891 5.55e-12 ***
## age         -0.012921   0.006864  -1.882   0.0598 .
## sexmale     -2.841503   0.209064 -13.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 627.45  on 630  degrees of freedom
## AIC: 633.45
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?
- What is the survival probability for a male child of age 5? A female child of age 5?
- What effect does being male have on survival probability?

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age * sex`:

```
logreg2 <- glm(survived ~ age * sex, data = Titanic1, family = "binomial")
summary(logreg2)
```

```
##
## Call:
## glm(formula = survived ~ age * sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1915  -0.7257  -0.4730   0.6661   2.2390
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.83092    0.36632   2.268   0.0233 *
## age          0.02342    0.01188   1.971   0.0487 *
## sexmale     -1.09657    0.46711  -2.348   0.0189 *
## age:sexmale -0.05935    0.01521  -3.903  9.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 611.19  on 629  degrees of freedom
## AIC: 619.19
##
## Number of Fisher Scoring iterations: 4
```

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age * sex`:

```
logreg2 <- glm(survived ~ age * sex, data = Titanic1, family = "binomial")
summary(logreg2)
```

```
##
## Call:
## glm(formula = survived ~ age * sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1915  -0.7257  -0.4730   0.6661   2.2390
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.83092    0.36632   2.268   0.0233 *
## age          0.02342    0.01188   1.971   0.0487 *
## sexmale     -1.09657    0.46711  -2.348   0.0189 *
## age:sexmale -0.05935    0.01521  -3.903  9.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 611.19  on 629  degrees of freedom
## AIC: 619.19
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age * sex`:

```
logreg2 <- glm(survived ~ age * sex, data = Titanic1, family = "binomial")
summary(logreg2)
```

```
##
## Call:
## glm(formula = survived ~ age * sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1915  -0.7257  -0.4730   0.6661   2.2390
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.83092    0.36632   2.268   0.0233 *
## age          0.02342    0.01188   1.971   0.0487 *
## sexmale     -1.09657    0.46711  -2.348   0.0189 *
## age:sexmale -0.05935    0.01521  -3.903  9.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 611.19  on 629  degrees of freedom
## AIC: 619.19
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?
- What is the survival probability for a male child of age 5? A female child of age 5?

R code for Multiple Logistic Models

- Suppose we fit a logistic model for `survived ~ age * sex`:

```
logreg2 <- glm(survived ~ age * sex, data = Titanic1, family = "binomial")
summary(logreg2)
```

```
##
## Call:
## glm(formula = survived ~ age * sex, family = "binomial", data = Titanic1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1915  -0.7257  -0.4730   0.6661   2.2390
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.83092    0.36632   2.268   0.0233 *
## age          0.02342    0.01188   1.971   0.0487 *
## sexmale     -1.09657    0.46711  -2.348   0.0189 *
## age:sexmale -0.05935    0.01521  -3.903  9.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 611.19  on 629  degrees of freedom
## AIC: 619.19
##
## Number of Fisher Scoring iterations: 4
```

- What is the formula for the logistic model?
- What is the survival probability for a male child of age 5? A female child of age 5?
- What effect does being male have on survival probability?

Section 2

Classification

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?

```
odds<- predict(logreg2, passengers)
odds
```

```
##          1          2          3          4          5          6          7
## -0.6249665  1.1587938 -0.9124210 -1.0561483  1.4398280 -1.3436028 -1.4873301
##          8          9         10
## -1.6310573  1.8145403  1.9082184
```

Prediction and Classification in R

What are their survival log odds?

```
odds<- predict(logreg2, passengers)
odds
```

```
##          1          2          3          4          5          6          7
## -0.6249665  1.1587938 -0.9124210 -1.0561483  1.4398280 -1.3436028 -1.4873301
##          8          9         10
## -1.6310573  1.8145403  1.9082184
```

Survival probabilities?

```
probs <- predict(logreg2, passengers, type = "response")
probs
```

```
##         1         2         3         4         5         6         7         8
## 0.3486527 0.7611135 0.2865047 0.2580462 0.8084280 0.2069182 0.1843228 0.1636856
##         9        10
## 0.8599097 0.8708189
```

Prediction and Classification in R

What are their survival log odds?

```
odds<- predict(logreg2, passengers)
odds
```

```
##          1          2          3          4          5          6          7
## -0.6249665  1.1587938 -0.9124210 -1.0561483  1.4398280 -1.3436028 -1.4873301
##          8          9         10
## -1.6310573  1.8145403  1.9082184
```

Survival probabilities?

```
probs <- predict(logreg2, passengers, type = "response")
probs
```

```
##         1         2         3         4         5         6         7         8
## 0.3486527 0.7611135 0.2865047 0.2580462 0.8084280 0.2069182 0.1843228 0.1636856
##         9        10
## 0.8599097 0.8708189
```

Classification?

```
ifelse(probs >= .5, 1, 0)
```

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

Confusion Matrix

How well does our model do on training data? We'll use several functions from the `yardstick` package.

Confusion Matrix

How well does our model do on training data? We'll use several functions from the `yardstick` package.

- First, we create data frame comparing observed and predicted classes:

Confusion Matrix

How well does our model do on training data? We'll use several functions from the `yardstick` package.

- First, we create data frame comparing observed and predicted classes:

```
probs<-predict(logreg2, Titanic1, type = "response")
preds<-as.factor( ifelse(probs >=.5, 1, 0))
obs <- as.factor(Titanic1$survived)
results <- data.frame(obs, preds, probs)
```

Confusion Matrix

How well does our model do on training data? We'll use several functions from the `yardstick` package.

- First, we create data frame comparing observed and predicted classes:

```
probs<-predict(logreg2, Titanic1, type = "response")
preds<-as.factor( ifelse(probs >=.5, 1, 0))
obs <- as.factor(Titanic1$survived)
results <- data.frame(obs, preds, probs)
```

- And then create a **confusion matrix** using `conf_mat` from `yardstick`

```
library(yardstick)
conf_mat(results, truth  = obs, estimate = preds)
```

```
##           Truth
## Prediction   0   1
##          0 308  82
##          1  44 199
```

Error Measures

- The overall error rate is the proportion of incorrect classifications:

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

Error Measures

- The overall error rate is the proportion of incorrect classifications:

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

- In `yardstick`, the `accuracy` function returns the proportion of correct classifications:

```
accuracy(results, truth = obs, estimate = preds)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.801
```

- Accuracy is the sum of the diagonal elements in the confusion matrix divided by the total number of observations.

Error Measures

- The overall error rate is the proportion of incorrect classifications:

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

- In `yardstick`, the `accuracy` function returns the proportion of correct classifications:

```
accuracy(results, truth = obs, estimate = preds)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.801
```

- Accuracy is the sum of the diagonal elements in the confusion matrix divided by the total number of observations.

- To obtain the error rate, we pull the accuracy estimate and subtract from 1:

```
acc <- accuracy(results, truth = obs, estimate = preds) %>% pull(.estimate)
1 - acc
```

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

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification (i.e. proportion of true positives correctly estimated)

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

Specificity: Rate of correct negative identification (i.e. proportion of true negatives correctly estimated) - Type I Error rate: $1 - \text{Specificity}$

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification (i.e. proportion of true positives correctly estimated)

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

Specificity: Rate of correct negative identification (i.e. proportion of true negatives correctly estimated) - 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 (i.e. proportion of true positives correctly estimated)

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

Specificity: Rate of correct negative identification (i.e. proportion of true negatives correctly estimated) - 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.
- When might we want high specificity? High sensitivity?

Sensitivity and Specificity

Sensitivity: Rate of correct positive identification (i.e. proportion of true positives correctly estimated)

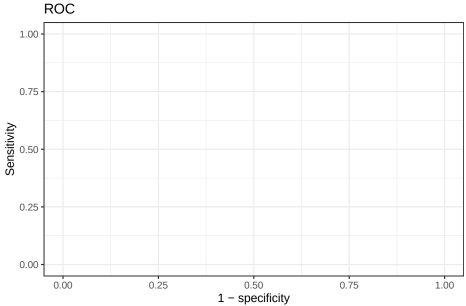
- Type II Error rate: $1 - \text{Sensitivity}$

Specificity: Rate of correct negative identification (i.e. proportion of true negatives correctly estimated) - 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.
- When might we want high specificity? High sensitivity?
- What are the ramifications of changing the classification cutoff vis-a-vis the Bayes' Classifier?

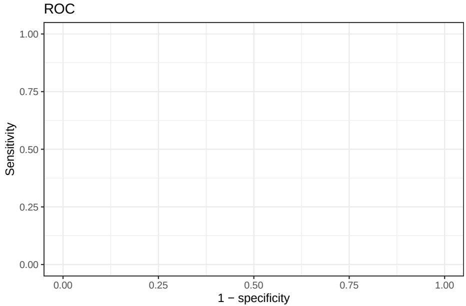
ROC Curves

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



ROC Curves

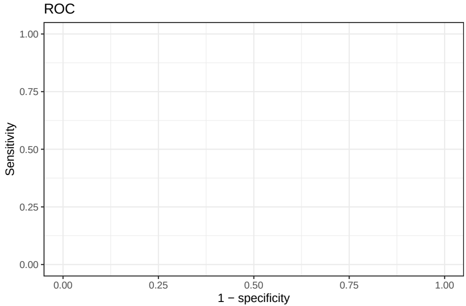
A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity vs. type I error rate, based on classification probabilities.



- What does the ROC curve look like for a perfectly accurate model?

ROC Curves

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

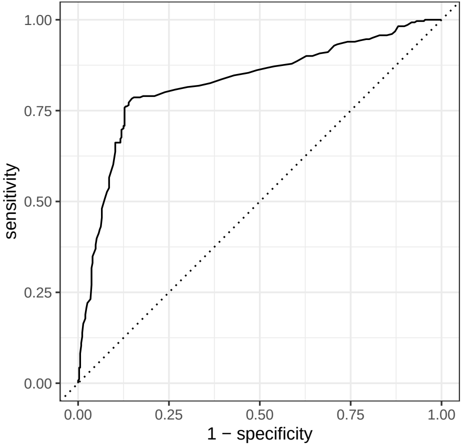


- What does the ROC curve look like for a perfectly accurate model?
- Consider an (uninformed) model that guesses 1 with probability p, regardless of predictors. What point on the ROC plot represents this model?

ROC Curves in R

The `roc_curve` function in the `yardstick` package can create ROC curves.

```
r <- roc_curve(data = results, truth = obs, probs, event_level = "second")
autoplot(r)
```



ROC Curves in R

What threshold corresponds to the "kink" in the ROC curve?

```
r <- r %>% mutate(distance = sqrt(
  (1-specificity)^2 + (1-sensitivity)^2))
```

ROC Curves in R

What threshold corresponds to the "kink" in the ROC curve?

```
r <- r %>% mutate(distance = sqrt(
  (1-specificity)^2 + (1-sensitivity)^2))
```

```
r %>% arrange(distance) %>% head()
```

```
## # A tibble: 6 x 4
##   .threshold specificity sensitivity distance
##        <dbl>       <dbl>       <dbl>    <dbl>
## 1      0.341       0.852       0.783    0.263
## 2      0.325       0.847       0.786    0.263
## 3      0.357       0.855       0.779    0.264
## 4      0.317       0.841       0.786    0.266
## 5      0.365       0.861       0.772    0.267
## 6      0.399       0.869       0.762    0.272
```