Logistic Regression

Nate Wells

Math 243: Stat Learning

October 25th, 2021

Outline

In today's class, we will. . .

- Discuss further theory of logistic regression
- Implement logistic regression in R

Section 1

Logistic Regression Theory

Summary

- In a classification problem, we are interested a categorical response variable Y.

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j|X)$$

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j|X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

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

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j|X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

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

- This implies the conditional probability is logistic:

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

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j|X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

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

- This implies the conditional probability is logistic:

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

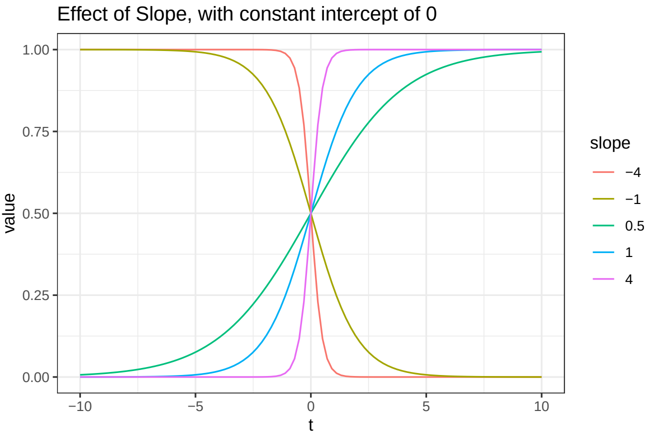
- To classify, we assign a test observation the value 1 if

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} \geq 0.5$$

Effect of Coefficients in Logistic Model

- Consider a logistic regression model for a binary categorical variable Y based on a single predictor X.

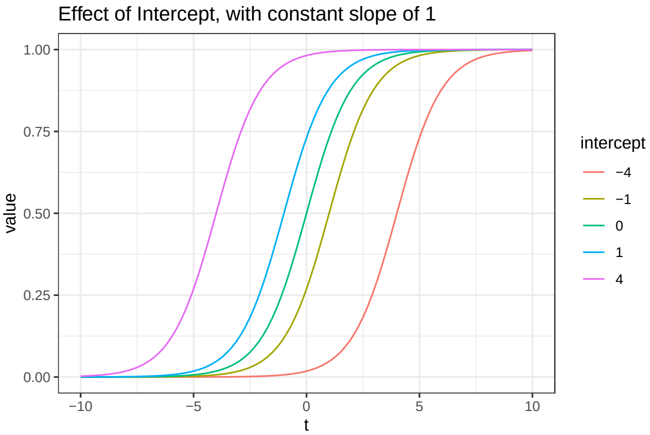
$$\ln \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X \qquad p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



Effect of Coefficients in Logistic Model

- Consider a logistic regression model for a binary categorical variable Y based on a single predictor X.

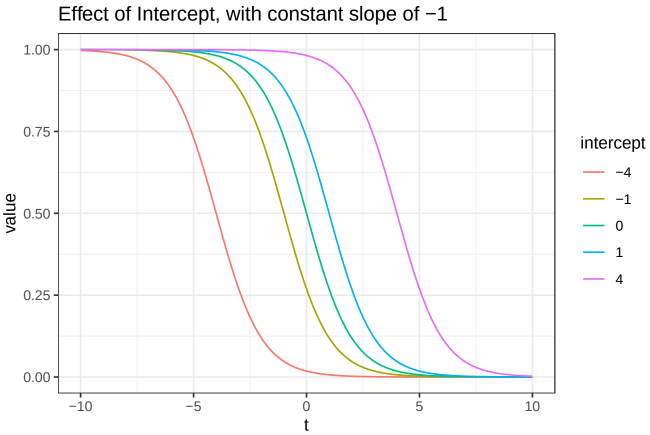
$$\ln \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X \qquad p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



Effect of Coefficients in Logistic Model

- Consider a logistic regression model for a binary categorical variable Y based on a single predictor X.

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



Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression
  - And in practice, residuals tend not to be approximately Normally distributed

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

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

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression
  - And in practice, residuals tend not to be approximately Normally distributed
- Instead, we use the method of **Maximum Likelihood** (ML)

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

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

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

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

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

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

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

  - View ℓ as a function of parameters $\beta_0, \ldots, \beta_p$ for **fixed** observations $x_1, \ldots, x_n$.

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

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

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

  - View ℓ as a function of parameters $\beta_0, \ldots, \beta_p$ for **fixed** observations $x_1, \ldots, x_n$.
- The goal is to choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so as to maximize ℓ

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

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

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

  - View ℓ as a function of parameters $\beta_0, \ldots, \beta_p$ for **fixed** observations $x_1, \ldots, x_n$.
- The goal is to choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so as to maximize ℓ
  - How? (Calculus or numeric methods, or R!)

Section 2

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)
```

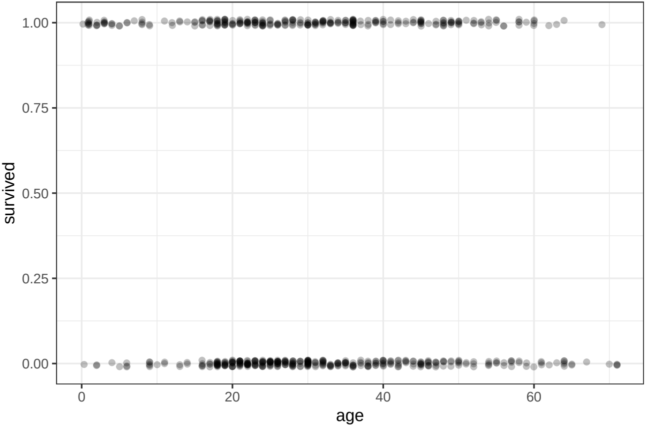
```
## # A tibble: 2 x 2
##   survived     n
##      <dbl> <int>
## 1        0   864
## 2        1   449
```

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

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

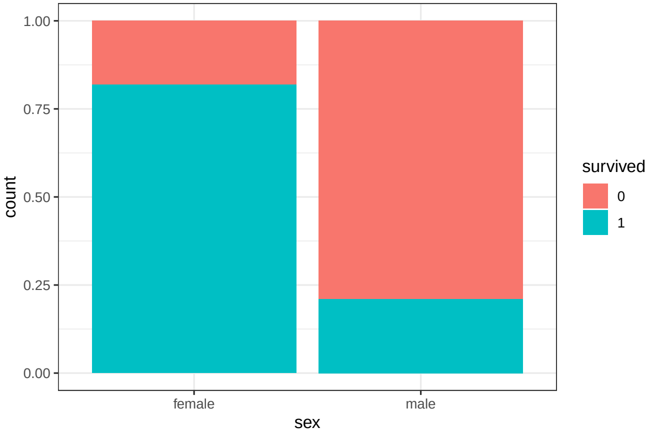
Children first?

- Who survived the Titanic?



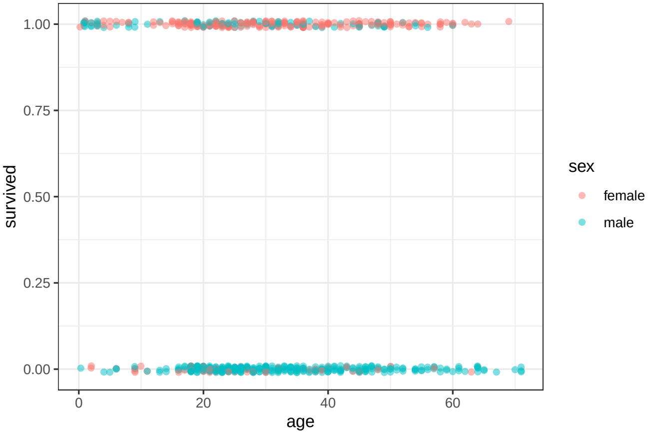
Women First?

- Who survived the Titanic?



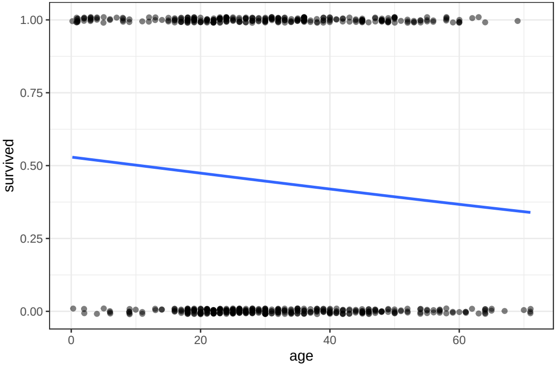
Women and Children First?

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



Logistic Model 1

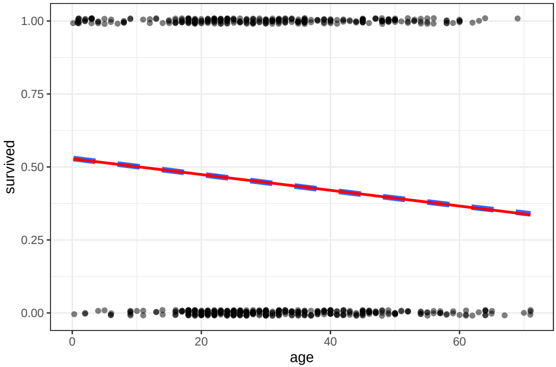
```
Titanic1 %>% ggplot( aes( x = age, y = survived ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F)
```



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

VS Linear Model

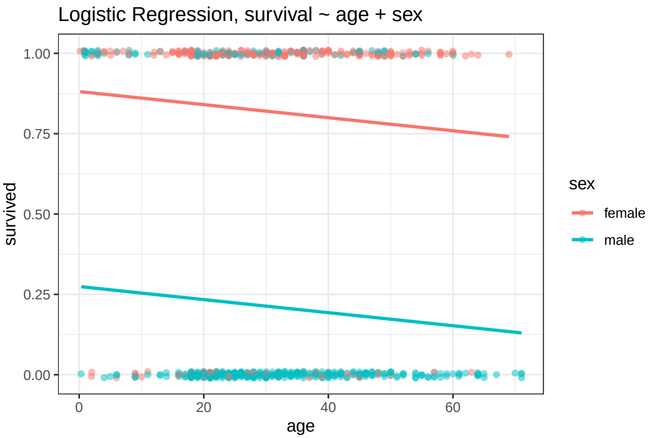
```
Titanic1 %>% ggplot( aes( x = age, y = survived ))+
  geom_jitter(height = .01, alpha = .5)+theme_bw()+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = F,size = 2,linetype
  geom_smooth(method = "lm", se = F, color  = "red")
```



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

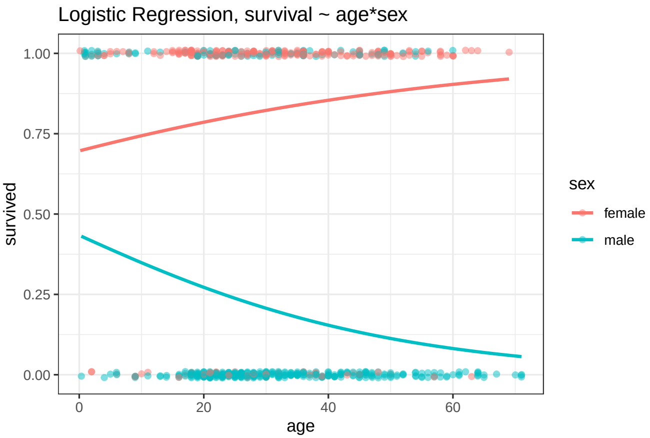
Logistic Model 2:

```
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

```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
```

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
```

- 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

```
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

```
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

```
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`:

```r
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?

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
##
```