Logistic Regression

Sections 4.1 - 4.3

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Lab Questions

From these scenarios determine the type of statistical learning.

a) Regression b) Classification c) Clustering

1. Does the knowledge of math and English test scores allow us to determine the students into at-risk and not-at-risk groups? b
2. A data set involving counts of the numbers of trees of each species in n = 72 sites. A total of 31 species were identified and counted. The goal is to group sample sites together that share similar species compositions as determined by some measure of association. C
3. We want to determine the relationship between the seal strength in grams per inch of a bread wrapper based on the variables sealing temperature, cooling bar temperature, and percent of polyethylene in the stock. a

Classification

- The response variable, Y, is **qualitative** or **categorical**.
- Predicting a qualitative response for an observations can be referred to as **classifying** that observation.
- These methods predict the probability of each of the categories of a qualitative variables, as the basis for making the classification.

Introduction to Logistic Regression

- Logistic regression can be used to model and solve problems when the Y (response) variable is a categorical variable with 2 classes.
- Also called binary classification problems.
- This models the **probability** that Y belongs to one of the two categories.

$Y = 0$ or 1

$P(Y = 1 \mid X)$

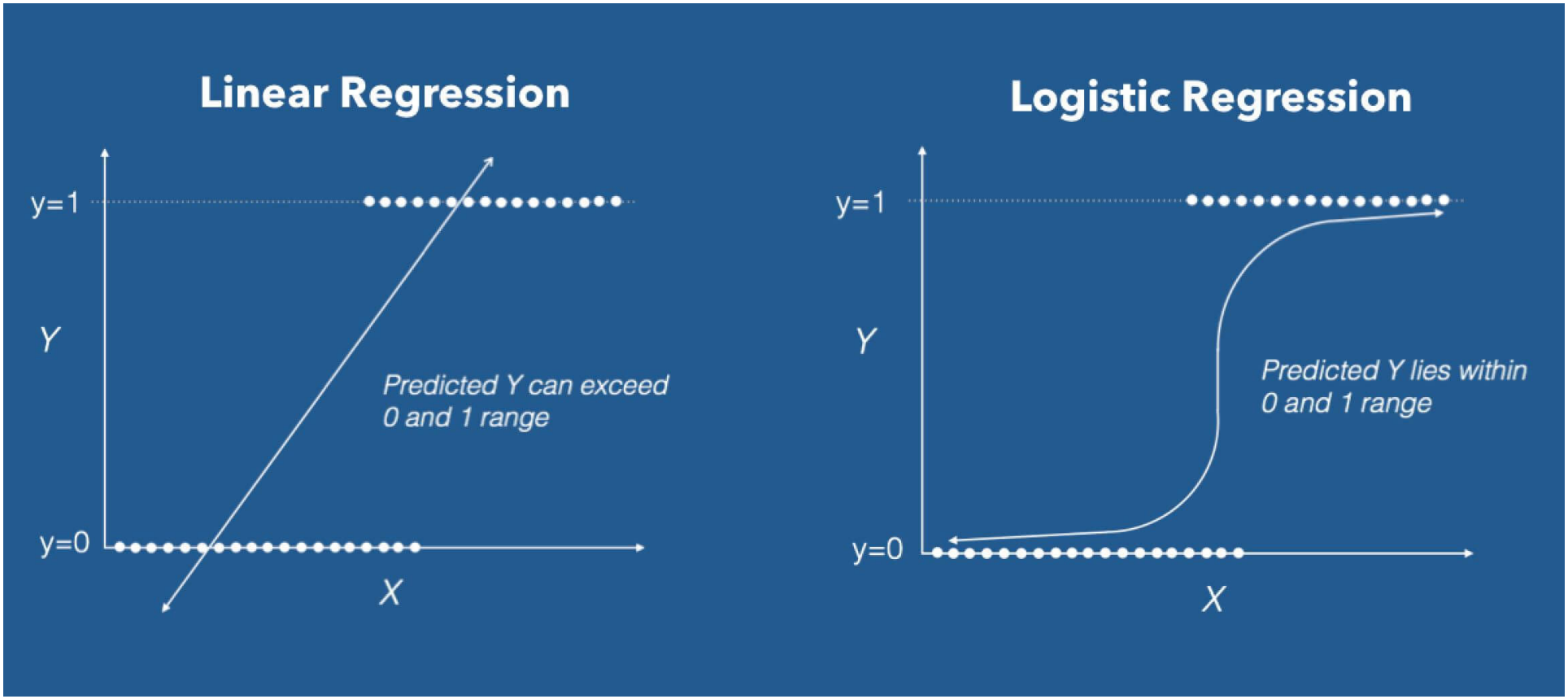
Some real world examples of binary classification problems

You might wonder what kind of problems you can use logistic regression for.
Here are some examples of binary classification problems:

- **Spam Detection** : Predicting if an email is Spam or not
- **Credit Card Fraud** : Predicting if a given credit card transaction is fraud or not
- **Health** : Predicting if a given mass of tissue is benign or malignant
- **Marketing** : Predicting if a given user will buy an insurance product or not
- **Banking** : Predicting if a customer will default on a loan.

Why not Linear Regression?

- The response variable has only 2 possible values, it is desirable to have model that predicts the value either as 0 or one or as a probability score that ranges between 0 and 1.
- If the linear regression is used to predict a binary response, the resulting model may not restrict the predicted Y values within 0 and 1.


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed
0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within
0 and 1 range

The Logistic Model

- Given Y = 0 or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.

The Logistic Model

- Given Y = 0 or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.
- We use a model that gives outputs between 0 and 1 for all values of X. This is called the **logistic function**

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

$$p(x) + p(x)\exp\{\beta_0+\beta_1 x\} = \exp(\beta_0+\beta_1 x)$$

$$p(x) = \exp(\beta_0+\beta_1 x) - p(x)\exp(\beta_0+\beta_1 x)$$

$$p(x) = [1 - p(x)]\exp(\beta_0+\beta_1 x)$$

$$\frac{p(x)}{1-p(x)} = \exp(\beta_0+\beta_1 x)$$

The Logistic Model

- Given Y = 0 or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.
- We use a model that gives outputs between 0 and 1 for all values of X. This is called the **logistic function**

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

- After some manipulation we get

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

The Logistic Model

- Given Y = 0 or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.
- We use a model that gives outputs between 0 and 1 for all values of X. This is called the **logistic function**

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

- After some manipulation we get

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

- Take the logarithm of both sides:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

The left-hand side is called the *log-odds* or *logit*.

The Logistic Model

- Given Y = 0 or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.
- We use a model that gives outputs between 0 and 1 for all values of X. This is called the **logistic function**

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

- After some manipulation we get

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

- Take the logarithm of both sides:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

  The left-hand side is called the *log-odds* or *logit*.

- We use a method called **maximum likelihood** to determine the best coefficients and eventually a good fit.

Maximum Likelihood Method

- It tries to find the value of coefficients (β_0,β_1) such that the predicted probabilities are as close to the observed probabilities as possible.

$P(Y=1|X) = p(x)$ for $y_i = 1$ we try to estimate β_0 & β_1 such that $p(x)$ close to 1

for $y_i = 0$ we try to estimate β_0 and β_1 such that $1 - p(x)$ is close to 1.

$p(x)^{y_i}[1-p(x)]^{1-y_i}$ as close to 1 as possible.

Maximum Likelihood Method

- It tries to find the value of coefficients (β_0,β_1) such that the predicted probabilities are as close to the observed probabilities as possible.
- In other words, for a binary classification (1/0), maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0.

Maximum Likelihood Method

- It tries to find the value of coefficients (β_0,β_1) such that the predicted probabilities are as close to the observed probabilities as possible.
- In other words, for a binary classification (1/0), maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0.
- The likelihood function is written as

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

Maximum Likelihood Method

- It tries to find the value of coefficients (β_0,β_1) such that the predicted probabilities are as close to the observed probabilities as possible.
- In other words, for a binary classification (1/0), maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0.
- The likelihood function is written as

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

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

Maximum Likelihood Method

- It tries to find the value of coefficients (β_0,β_1) such that the predicted probabilities are as close to the observed probabilities as possible.
- In other words, for a binary classification (1/0), maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0.
- The likelihood function is written as

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

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

In R we use the glm(model,family="binomial") function

Generalized Linear Models

Generalized Linear Models (glm) are an extension of the linear model framework, which includes dependent variables which are non-normal also. In general, they possess three characteristics:

1. These models comprise a linear combination of input features.
2. The mean of the response variable is related to the linear combination of input features via a link function.
3. The response variable is considered to have an underlying probability distribution belonging to the family of exponential distributions such as binomial distribution, Poisson distribution, or Gaussian distribution. Practically, binomial distribution is used when the response variable is binary. Poisson distribution is used when the response variable represents count. And, Gaussian distribution is used when the response variable is continuous.

Logistic Regression assumes that the dependent (or response) variable follows a binomial distribution.

Example

- Using the BreastCancer data set in mlbench package. You will have to install the "mlbench" package for this.
- The response, Y is the **Class** this has two categories *malignant* and *benign*.
- We want to use **Cell.shape** as the predictor.
- In order to use this data we have to `clean` this data to use with the glm function.

Cleaning the Data

- Call the data

```
data(BreastCancer, package="mlbench")
summary(BreastCancer)
```

```
     Id              Cl.thickness   Cell.size     Cell.shape  Marg.adhesion
Length:699         1       :145   1       :384   1       :353   1       :407
Class :character   5       :130   10      : 67   2       : 59   2       : 58
Mode  :character   3       :108   3       : 52   10      : 58   3       : 58
                   4       : 80   2       : 45   3       : 56   10      : 55
                   10      : 69   4       : 40   4       : 44   4       : 33
                   2       : 50   5       : 30   5       : 34   8       : 25
                   (Other):117   (Other): 81   (Other): 95   (Other): 63

Epith.c.size  Bare.nuclei   Bl.cromatin  Normal.nucleoli    Mitoses
2       :386   1       :402   2       :166   1       :443     1       :579
3       : 72   10      :132   3       :165   10      : 61     2       : 35
4       : 48   2       : 30   1       :152   3       : 44     3       : 33
1       : 47   5       : 30   7       : 73   2       : 36     10      : 14
6       : 41   3       : 28   4       : 40   8       : 24     4       : 12
5       : 39   (Other): 61   5       : 34   6       : 22     7       :  9
(Other): 66   NA's    : 16   (Other): 69   (Other): 69     (Other): 17

      Class
benign   :458
malignant:241
```

- Create copy with no missing values and remove the id column

```
bc <- BreastCancer[complete.cases(BreastCancer), ]
bc <- bc[,-1] # remove id column
```

Convert the factors to numeric

```
for(i in 1:9) {
bc[, i] <- as.numeric(as.character(bc[, i]))
}
summary(bc)
Cl.thickness      Cell.size        Cell.shape     Marg.adhesion
Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00
Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00
Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83
3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00
Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00

Epith.c.size     Bare.nuclei      Bl.cromatin     Normal.nucleoli
Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00
Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.00
Mean   : 3.234   Mean   : 3.545   Mean   : 3.445   Mean   : 2.87
3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00
Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00

Mitoses             Class
Min.   : 1.000   benign   :444
1st Qu.: 1.000   malignant:239
Median : 1.000
Mean   : 1.603
3rd Qu.: 1.000
Max.   :10.000
```

Summary

```
fit.bc = glm(Class ~ Cell.shape, family = "binomial", data = bc)
summary(fit.bc)

Call:
glm(formula = Class ~ Cell.shape, family = "binomial", data = bc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6383  -0.2219  -0.2219   0.0517   2.7263

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1645     0.3865  -13.36   <2e-16 ***
Cell.shape    1.4727     0.1205   12.22   <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: 884.35  on 682  degrees of freedom
Residual deviance: 267.59  on 681  degrees of freedom
AIC: 271.59
```

$$P(Y = \text{"malignant"} \mid \text{Cell.Shape}) = \frac{e^{-5.1645 + 1.4727 \text{Cell.Shape}}}{1 + e^{-5.1645 + 1.4727 \text{Cell.Shape}}}$$

Interpreting the Predictor

Suppose we have a cell shape of 5.

$$\hat{p}(X) = \frac{\exp^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+\exp^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{\exp^{-5.1645+1.4727\times 5}}{1+\exp^{-5.1645+1.4727\times 5}} = 0.9001597$$

Which means the predicted probability of being malignant given the uniformity of the cell shape (Cell.shape) is 5 is 0.90 (90%).

$Y = \begin{cases} \text{benign} & 0 \\ \text{malignant} & 1 \end{cases}$

ifelse(bc$Class == "malignant", 1, 0)

Interpreting the Predictor

Suppose we have a cell shape of 5.

$$\hat{p}(X) = \frac{\exp^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+\exp^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{\exp^{-5.1645+1.4727\times 5}}{1+\exp^{-5.1645+1.4727\times 5}} = 0.9001597$$

Which means the predicted probability of being malignant given the uniformity of the cell shape (Cell.shape) is 5 is 0.90 (90%).

```
predict.glm(fit.bc,newdata = data.frame(Cell.shape=5),
+             type="response")
1
0.9001648
```

Interpreting the Estimated Parameters

	Coefficient	Std. Error	Z-value	P-value
Intercept	-5.1645	0.3864	-13.36	< 0.0001
Cell.shape	1.4727	0.1205	12.222	< 0.0001

- $\hat{\beta}_1 = 1.4727$ indicates that an increase in Cell.shape is associated with an increase of the probability of **Class**.
- A one unit increase in Cell.shape is associated with an increase in the log odds of Class by 1.4727 units.
- For testing $H_0 : \beta_1 = 0$, this null hypothesis implies that $p(X) = \frac{exp^{\beta_0}}{1+\exp^{\beta_0}}$. Thus if we fail to reject H_0 the probability of **Class** does not depend on Cell.shape. $H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$
- Since p-value for this test is < 0.0001 we reject the null hypothesis and conclude that there is an association between Cell.shape and probability of **Class**.

Lab Questions

Type in the following an run in R.

```
summary(glm(Class ~ Cl.thickness, family =
"binomial",data = bc))
```

4. Is there an association between cell thickness and the probability of being malignant?

a) Yes
b) No

5. What is the probability of being malignant if the cell thickness is 5?

a) 0.458
b) 0.3875
c) 0.542
d) 0.6125

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.11012    0.37894  -13.48   <2e-16 ***
Cl.thickness 0.93042    0.07418   12.54   <2e-16 ***
```

$$P(Y=1 \mid C.T=5) = \frac{e^{-5.11012+0.93042*5}}{1+e^{-5.11012+0.93042*5}}$$

Categorical Predictors

- We will use the data set Titanic that is in the base R.
- We want to determine the probability of survival among gender.
- Again we need to clean this data. Currently this is a contingency table, we want to convert this to raw data. Do the following in R.

```
install.packages("bbl") #package used to convert to raw data
library(bbl) #call the package
x <- as.data.frame(Titanic) #put as a data frame
 #convert to the raw data
titanic = freq2raw(data=x[,1:4], freq=x$Freq)
```

Testing and Training Data Sets

- We need to make sure that we are not overfitting the data.
- Sometimes we want to separate the data set so that we can see how well our model fits the data on a new data set.
- Now that we have 2201 observations in our titanic data frame we can separate 75% training and 25% testing.
- To do this we can

```
set.seed(101) # Set Seed so that same sample can be reproduced
# Now Selecting 75% of data as sample from total 'n' rows
sample <- sample.int(n = nrow(titanic),
                     size = round(.75*nrow(titanic),0),
                     replace = F)
#getting the observations with the random values from the sample
train <- titanic[sample, ]
#eliminating the observations with random values from the sample
test  <- titanic[-sample, ]
```

- The sample array will be $0.75 * 2201 = 1650.75$, rounded to the whole number, 1651 values between 1 to 2201.

Model

The model will be as follows

$$sex = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

$$p(X) = \begin{cases} \frac{\exp^{\beta_0}}{1+\exp^{\beta_0}} & \text{if Male} \\ \frac{\exp^{\beta_0+\beta_1}}{1+\exp^{\beta_0+\beta_1}} & \text{if Female} \end{cases}$$

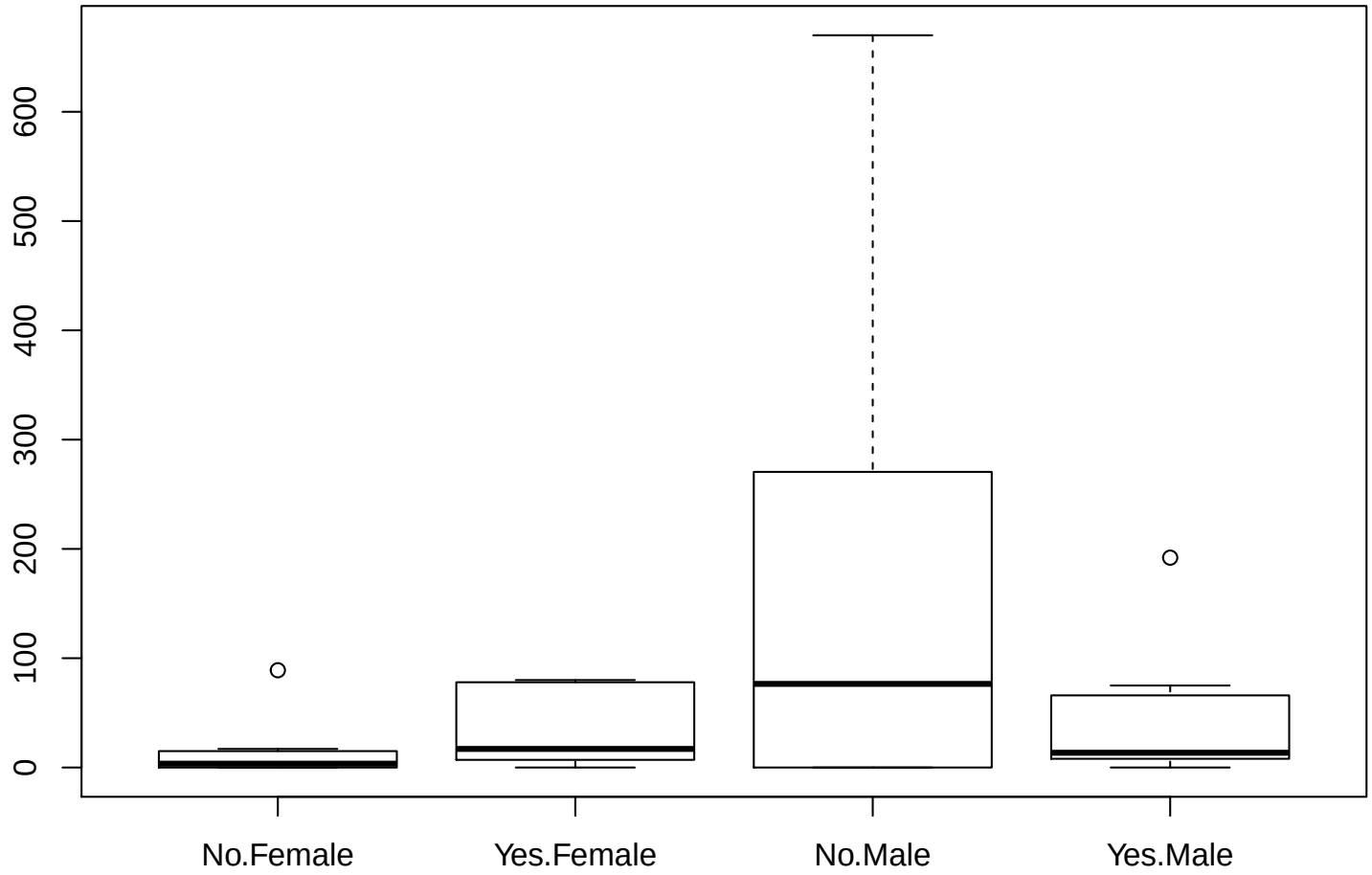
```
titanic.glm = glm(Survived ~ Sex, family = "binomial",
                  data = train)
summary(titanic.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.30452    0.06775  -19.26   <2e-16 ***
SexFemale    2.27107    0.13719   16.55   <2e-16 ***
```

$$p(x) = \begin{cases} \frac{\exp(-1.30452)}{1+\exp(-1.30452)} & \text{if male} \\ \frac{\exp(-1.30452+2.27107)}{1+\exp(-1.30452+2.27107)} & \text{if female} \end{cases}$$

$\hat{P}(\text{Survival} = \text{"Yes"} \mid \text{male}) = 0.2022 \quad (0.212)$

$P(\text{survival} = \text{"Yes"} \mid \text{female}) = 0.7308 \quad (0.732)$

```
boxplot(Freq ~ Survived + Sex, data = x)
```



Confusion Matrix

- A **confusion matrix** is a convenient way to display to observations that are incorrectly assigned to the wrong category.
- The following table is the confusion matrix for the training data.

		True Survive	
		No	Yes
Predicted Survive	No	1034	262
	Yes	93	262

- What percent were correct? What percent were wrong? This last percent is called the training error rate.

$$\%\ \text{correct} = \frac{1034 + 262}{1651} = 0.785$$

$$\%\ \text{error} = \frac{262 + 93}{1651} = 0.215$$

Testing Error Rate

The following table is the confusion matrix for the test data set.

		True Survive	
		No	Yes
Predicted Survive	No	330	105
	Yes	33	82

What is the testing error rate for this model?

$$\%\ \text{error} = \frac{105+33}{550} = 0.251$$

Sensitivity and Specificity

- **Sensitivity** measures the proportion of positives that are correctly identified.
- **Specificity** measures the proportion of negatives that are correctly identified.
- For our example of the testing data

$$\text{Sensitivity} = \frac{82}{105 + 82} = 0.4385$$

$$\text{Specificty} = \frac{330}{330 + 33} = 0.9091$$