STAT 380 – Logistic Regression Part 1 (Lecture 12)

IDEA: In this lecture, we are going to discuss a classification technique called logistic regression. The version that we are going to study is appropriate for a binary response variable. “Binary’ means that the response variable has exactly 2 possible values such as yes/no, pass/fail, survive/die, success/failure, 1/0).

NOTE: Our goal is to model the probability, p, that the response takes on the level we are interested in studying. We usually refer to this level as the “success” even though the outcome may not necessarily be a positive outcome. For example, we might want to model the probability that a driver is in an accident.

Notation: For a binary random variable, we define the probability of success, p_i, as:

$$p_i = P(\, y_i = 1 \mid x_i \,)$$

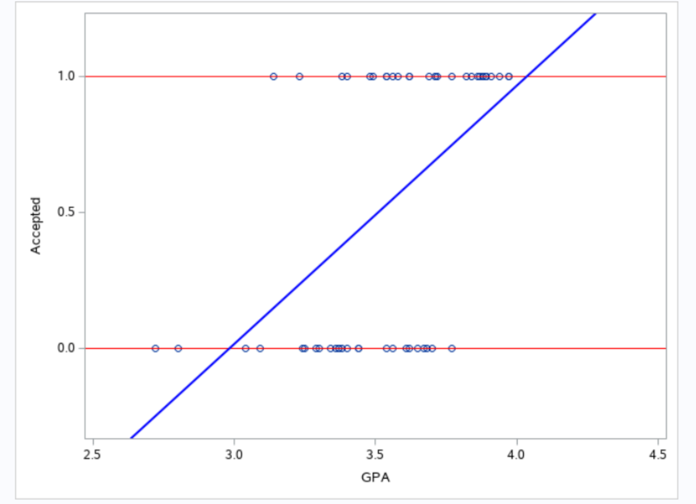
where $P(\, y_i = 1 \mid x_i \,)$ represents the probability that the i^{th}response (y_i) takes on the outcome of interest (i.e., is a “success”) *given* the i^{th} observation’s input value (x_i). A “success” will be determined by the language used in the problem.

NOTE: When the response is binary, the estimated probabilities can be used to classify a future/new observation as a success or failure. If the predicted probability is larger than a chosen threshold (say 0.5), then we predict success. Otherwise, we predict failure.

EXAMPLE 1: The least-squares regression model we have studied does not provide meaningful estimates of p_i. The following picture depicts a dataset in which medical students are classified as being accepted to medical school (Accepted = 1) or not (Accepted = 0) as a function of their GPA’s. Explain why using

$$p_i = \beta_0 + \beta_1 x_i$$

to model the probability of acceptance p_i is a poor choice.



NOTE: In order to understand the logistic regression model, we must clarify some terminology.

EXAMPLE 2: Suppose that the probability of rain on Monday is 0.75. What are the odds of rain on Monday?

Definition: The probability of success p_i, where $p_i = P(\,y_i = 1 \mid x_i\,)$, is a probability that takes on values $0 < p_i < 1$. Then, the odds that $y_i = 1$ (also called the odds of success) is given by:

$$odds_i = \frac{\#successes}{\#failures} = \frac{\#successes/n}{\#failures/n} = \frac{Probability\ of\ Success}{Probability\ of\ Failure} = \frac{p_i}{1-p_i}$$

and the log odds (or log(odds)) is given by:

$$\log\,(odds_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where log indicates the natural logarithm (i.e., base e logarithm). The transformation from p_i to log(odds) is called the logistic or logit transformation.

NOTE: To illustrate the behavior, consider the following table and answer the following:

p_i (Fract.)	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{9}{10}$	$\frac{19}{20}$
p_i (Dec.)	0.05	0.10	0.20	0.25	0.50	0.75	0.80	0.90	0.95
Odds (Fract.)	$\frac{1}{19}$	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{1}$	$\frac{3}{1}$	$\frac{4}{1}$	$\frac{9}{1}$	$\frac{19}{1}$
$\log(odds)$	-2.94	-2.20	-1.39	-1.10	0	1.10	1.39	2.20	2.94

Probability can take on values in which interval? In other words, $\qquad < p_i <$

Odds can take on values in which interval? In other words, $\qquad < odds_i <$

$\log(odds)$ can take on values in which interval? In other words, $\qquad < \log odds_i <$

Definition: In the case of a binary response, with probability of success p, the logistic regression model for the probability of *success*, p, takes several equivalent forms:

Logit Form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}$$

Odds Form:

$$\frac{p_i}{1-p_i} = e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}}$$

Probability Form:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}}}{1 + e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}}}$$

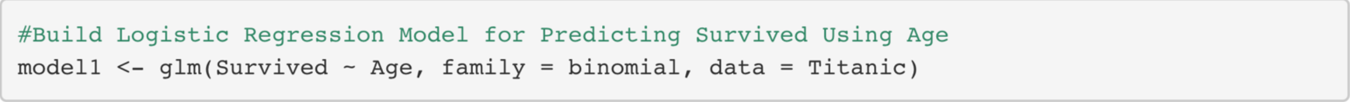
EXAMPLE 3: This problem uses L12_titanic3.csv. The dataset contains information about passengers on the ill-fated maiden voyage of the Titanic. Specifically, the dataset contains the following information:

Variable Name	Variable Meaning	Notes
Survived	Whether the passenger survived	Values include: Yes No
PClass	Purchased ticket class	Values include: 1 = First class 2 = Second class 3 = Third class
Sex	Passenger's sex	
Age	Passenger's age	
Siblings	Number of siblings or spouses aboard	
Parch	Number of parents or children aboard	
Fare	Passenger's fare	

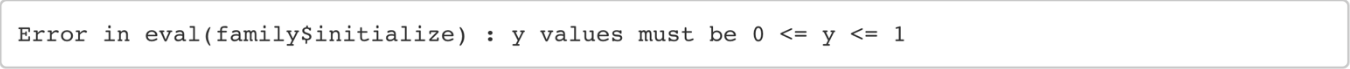
The goal of this problem is to predict whether a passenger is going to survive using the other variables in the dataset.

a. What is a "success" based on the language used in the storyline?

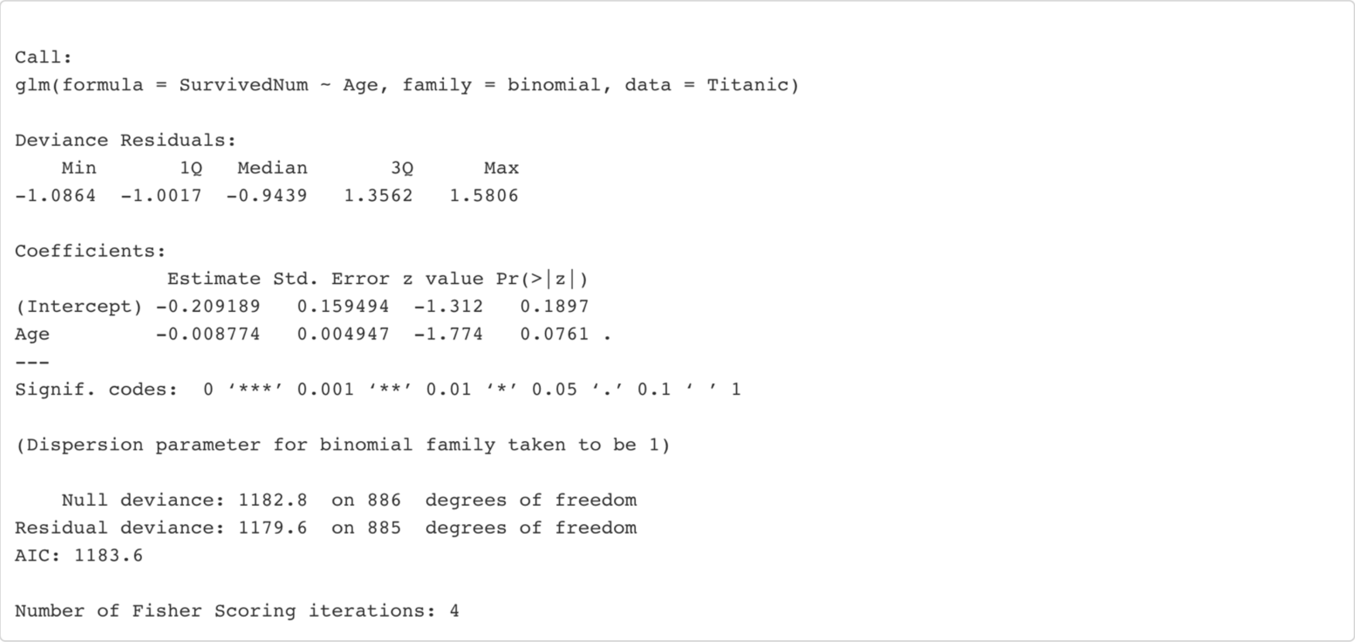
b. Suppose we try to build a logistic regression model using the glm() function (and specify family = binomial) for predicting Survived using Age. Based on the error message, solve the issue and write the estimated equation.



```
#Build Logistic Regression Model for Predicting Survived Using Age
model1 <- glm(Survived ~ Age, family = binomial, data = Titanic)
```



```
Error in eval(family$initialize) : y values must be 0 <= y <= 1
```



```
Call:
glm(formula = SurvivedNum ~ Age, family = binomial, data = Titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0864  -1.0017  -0.9439   1.3562   1.5806

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.209189   0.159494  -1.312   0.1897
Age         -0.008774   0.004947  -1.774   0.0761 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 886  degrees of freedom
Residual deviance: 1179.6  on 885  degrees of freedom
AIC: 1183.6

Number of Fisher Scoring iterations: 4
```

Logit Form:

Probability Form:

c. Another way to solve the error message produced in Part b. is to convert Survived to a factor; however, you must use care when taking this approach. To illustrate, consider the following methods that convert to the response to a factor. What is the impact on the estimated equation?

```
#Build logistic regression used Survived as a factor
model1_factor <- glm(as.factor(Survived) ~ Age, family = binomial, data = Titanic)
summary(model1_factor)
```

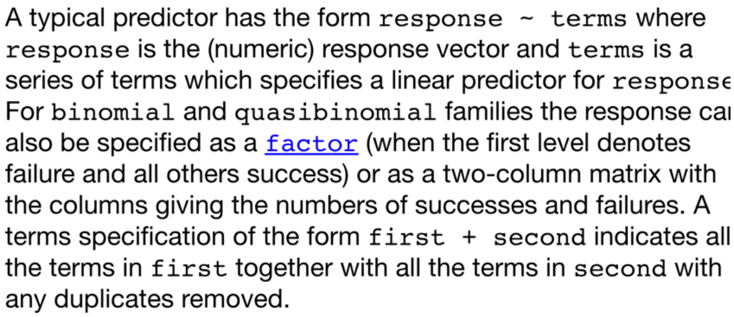
```

Call:
glm(formula = as.factor(Survived) ~ Age, family = binomial, data = Titanic)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.0864  -1.0017  -0.9439   1.3562   1.5806  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.209189   0.159494  -1.312   0.1897  
Age         -0.008774   0.004947  -1.774   0.0761 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 886  degrees of freedom
Residual deviance: 1179.6  on 885  degrees of freedom
AIC: 1183.6

Number of Fisher Scoring iterations: 4
```

```
#What happens if Survived took values of "Did" and "Did Not"? Do answers change?
Titanic <- 
  Titanic %>% mutate(Survived2 = ifelse(Survived == "Yes", "Did", "Did Not"),
                     Survived2 = as.factor(Survived2))

model1_factor2 <- glm(Survived2 ~ Age, family = binomial, data = Titanic)
summary(model1_factor2)
```

```

Call:
glm(formula = Survived2 ~ Age, family = binomial, data = Titanic)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5806  -1.3562   0.9439   1.0017   1.0864  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) 0.209189   0.159494   1.312   0.1897  
Age         0.008774   0.004947   1.774   0.0761 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 886  degrees of freedom
Residual deviance: 1179.6  on 885  degrees of freedom
AIC: 1183.6

Number of Fisher Scoring iterations: 4
```

d. Which of these models (model1_factor or model1_factor2) is predicting the probability that a person survived the sinking the of the Titanic? How do you know?

From help file documentation for glm function:



A typical predictor has the form `response ~ terms` where `response` is the (numeric) response vector and `terms` is a series of terms which specifies a linear predictor for `response`. For `binomial` and `quasibinomial` families the response can also be specified as a factor (when the first level denotes failure and all others success) or as a two-column matrix with the columns giving the numbers of successes and failures. A terms specification of the form `first + second` indicates all the terms in `first` together with all the terms in `second` with any duplicates removed.

```
levels(as.factor(Titanic$Survived))
```

```
[1] "No"  "Yes"
```

```
levels(as.factor(Titanic$Survived2))
```

```
[1] "Did"     "Did Not"
```

SUMMARY: When building a logistic regression model, our initial attempt produced an error that the response is not a numeric value between 0 and 1. We can solve this by using the as.factor() function OR by converting the response to an indicator variable. The second method (using an indicator) is preferred because it is easy to specify which level of the response is the "success".

e. For the logistic regression model predicting the odds of surviving based on a person's age, interpret the coefficient for age.

$\hat{\beta}_1 = -0.0088$

General interpretation of slope: As *x* increases by 1 unit, we expect y to change by the slope, on average. Here, "*y*" is the logit (or log odds of success).

Context specific: As a person's age increases by 1 year, we expect the log odds of surviving to decrease by 0.0088, on average.

ISSUE: We do not think on log scale.

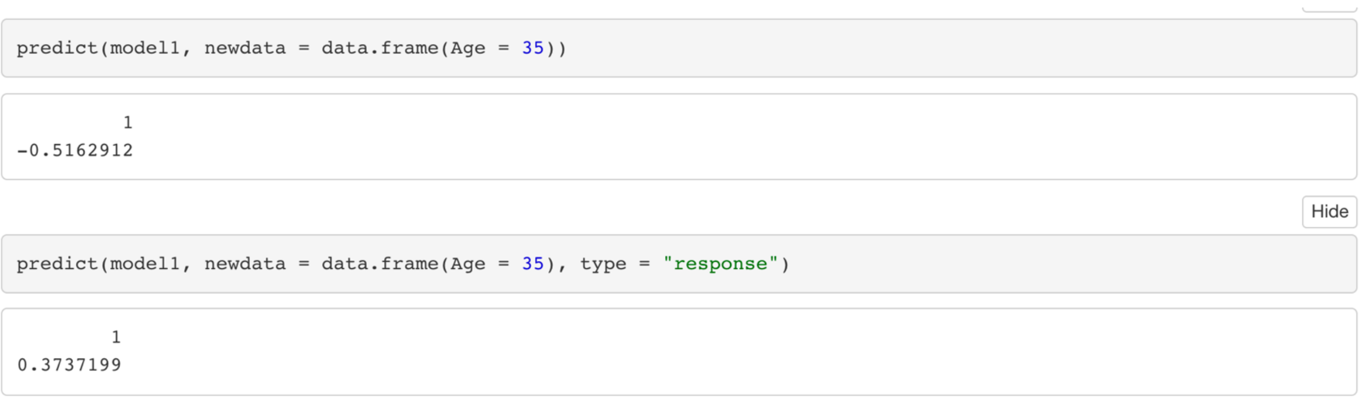
Notation: Use $Odds_X$ and $Odds_{X+1}$ to denote the odds when age takes on values of X and $X + 1$, respectively. Find an expression for the odds ratio, $\frac{Odds_{X+1}}{Odds_X}$.

e. (Continued)

f. Manually predict the probability of surviving for a person who is 35 years old.

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_{\{i}}\right) = -0.2092 - 0.0088x_{i,age}$$

g. Repeat Part f. using the the `predict` function. You must know whether the result is the linear predictor of the log odds or the predicted probability.



```
predict(model1, newdata = data.frame(Age = 35))
```

```
         1 
-0.5162912 
```

```
predict(model1, newdata = data.frame(Age = 35), type = "response")
```

```
        1 
0.3737199 
```

h. Based on the estimated probability would you predict that the 35-year old person survives or not?

IDEA: Establish a threshold such that we predict $Y = 1$ if the predicted (estimated) probability is greater than the threshold and predict $Y = 0$ otherwise.

i. Write code to obtain the predicted probabilities for all people in the dataset and make a prediction as to whether each person survives or not based on a threshold of 0.38. Name the resulting vector of predictions pred_surv.

j. Complete the following table. For the “Type” column, identify each as correct/incorrect AND true positive/true negative/false positive/false negative.

Row Num	Survived	Age	Predicted Prob. of Surviving	Predicted Value of Surviving	Type
1	No	22	0.4007780	Yes	
2	Yes	38	0.3675797	No	
3	Yes	26	0.3923793	Yes	
4	Yes	35	0.3737199	No	
5	No	35	0.3737199	No	