Logistic Regression (part I)

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

We are going to review linear (and related) methods for classification:

- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- K-Nearest-Neighbors

Later in the course we'll cover other (nonlinear / nonparameteric) methods for classification.

Introduction

- Pierre Verhulst (1838) talks about the “logistic equation” that he introduced to model the population growth (following Thomas Malthus theory).
- Daniel McFadden (1973)—Nobel Prize in Economics—
- Introduced into software more recently than linear discriminant analysis
- Continued improvement and generalization in the context of the generalized linear model

Introduction

For simplicty ...

- I will focus on a binary response variable Y
- Usually we code the values of Y with 0 and 1
- Also, I will consider one predictor variable X

Keep in mind that logistic regression can also be applied to responses with any number of categories, and with multiple predictors.

Coronary Heart Disease Example

Coronary Heart Disease (CHD)

Coronary Heart Disease Data

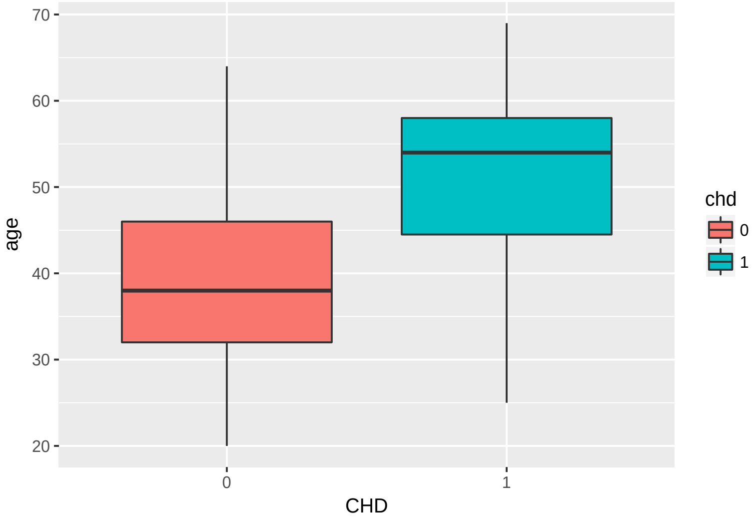
- Famous data set from Hosmer & Lemeshow (2000)
- 100 individuals
- one predictor X: Age (in years)
- response Y: Coronary Heart Disease
  - present $= 1$
  - absent $= 0$
- File: `chd.csv` in github repo

CHD Data Set

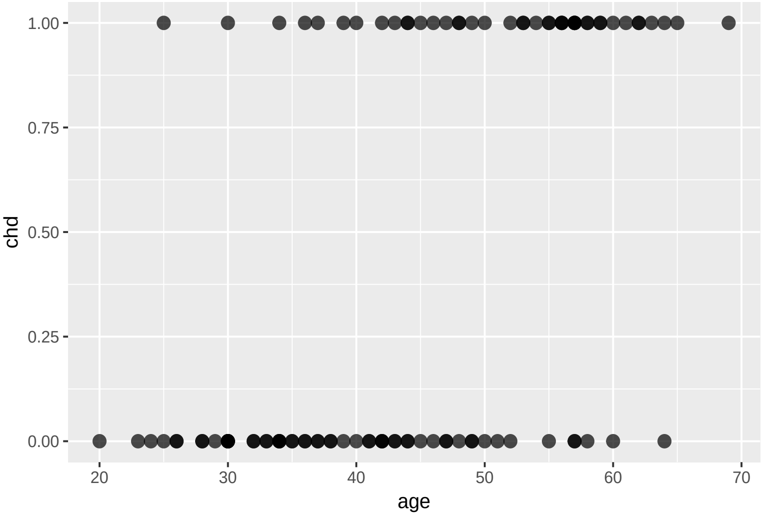
```
##   age chd
## 1  20   0
## 2  23   0
## 3  24   0
## 4  25   0
## 5  25   1
## 6  26   0
```

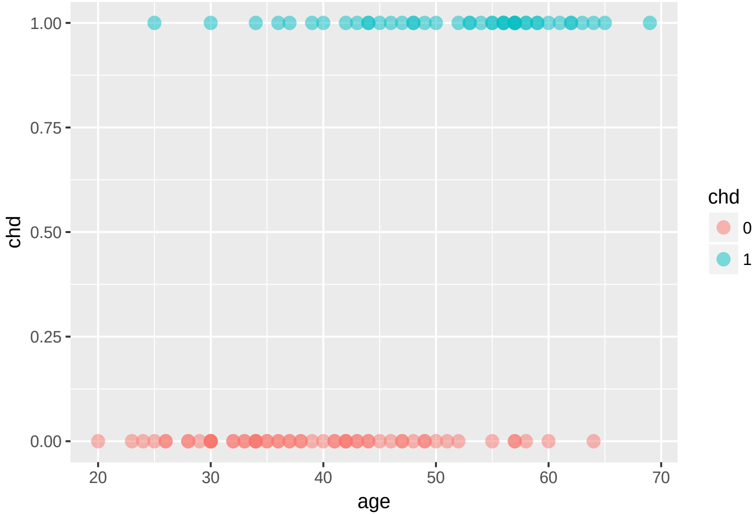
```
summary(dat)
```

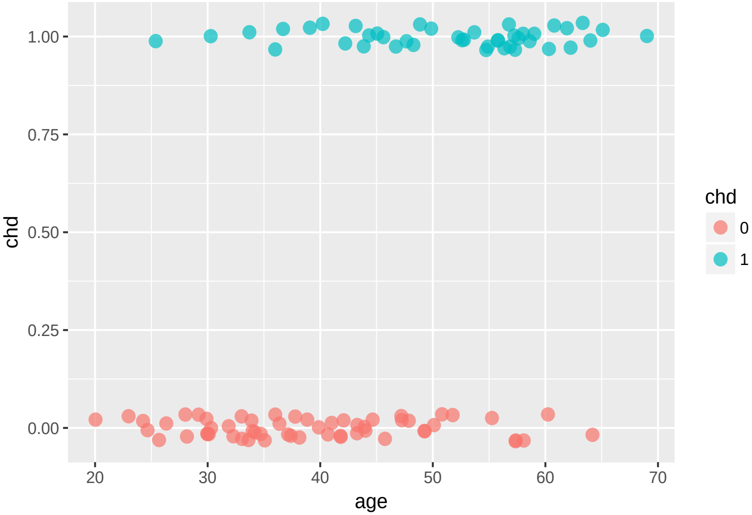
```
##       age             chd      
##  Min.   :20.00   Min.   :0.00  
##  1st Qu.:34.75   1st Qu.:0.00  
##  Median :44.00   Median :0.00  
##  Mean   :44.38   Mean   :0.43  
##  3rd Qu.:55.00   3rd Qu.:1.00  
##  Max.   :69.00   Max.   :1.00  
```



70
60
50
40
30
20
age
0
1
CHD
chd
0
1


1.00
0.75
0.50
0.25
0.00
chd
20
30
40
50
60
70
age


1.00
0.75
0.50
0.25
0.00
chd
20
30
40
50
60
70
age
chd
0
1


1.00
0.75
0.50
0.25
0.00
chd
20
30
40
50
60
70
age
chd
0
1

Thinking Inside the Box

LS regression

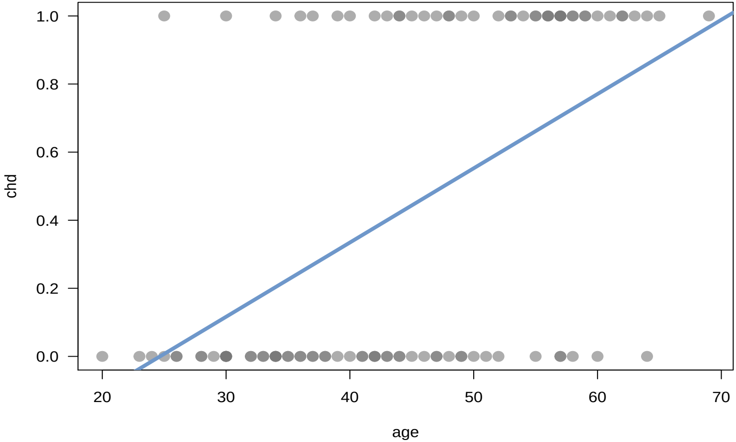
- We would like to predict whether individuals have Coronary Heart Disease or not.
- The Y variable `chd` is categorical: 0 or 1.
- Can we use linear regression when Y is categorical?

Let's try an ordinary LS regression

```r
reg = lm(chd ~ age, data = dat)
summary(reg)
```

```
##
## Call:
## lm(formula = chd ~ age, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85793 -0.33992 -0.07274  0.31656  0.99269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.537960   0.168809  -3.187  0.00193 **
## age          0.021811   0.003679   5.929 4.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 98 degrees of freedom
## Multiple R-squared:  0.264,Adjusted R-squared:  0.2565
## F-statistic: 35.15 on 1 and 98 DF,  p-value: 4.575e-08
```

Regression Line




```r
plot(dat, las = 1, col = "#77777799", pch = 19, cex = 1.5)
abline(reg, col = "#6E97CA", lwd = 4)
```

Regression Line

- At first glance the fit looks a bit awkward
- But the slope of the line kind of makes sense
- Regression line has a positive slope
  (there are more CHD cases in older people than in young poeple)
- When X (age) is small, Y (CHD) tends to be 0
- Likewise when X (age) is large, Y (CHD) tends to be 1

OLS Regression?

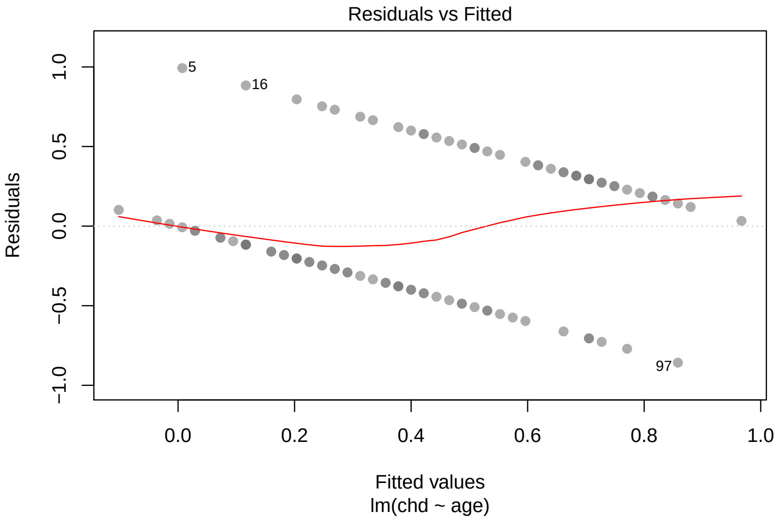
What's the issue with using OLS regression?

What's the issue with using OLS regression?

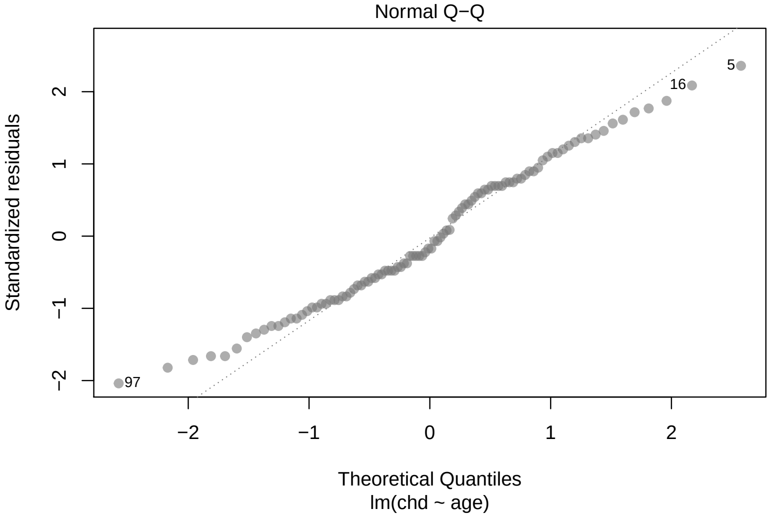
Let's check standard diagnostic tools for regression

```r
# residuals plot
plot(reg, which = 1, col = "#77777799", pch = 19)

# qq-plot
plot(reg, which = 2, col = "#77777799", pch = 19)
```

Residuals Plot



QQ-Plot


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(chd ~ age)
−2
−1
0
1
2
97
16
5

Reminder: Classic Linear Regression Model

Linear Regression Model assumptions:

- $Y = \beta X + \varepsilon$
- Y quantitative response
- X quantitative predictor
- NIID: independent error terms $\varepsilon_i \sim N(0, \sigma^2)$
  - $E(\varepsilon_i) = 0$
  - $Var(\varepsilon_i) = \sigma^2$

Most assumptions don't hold for classification purposes

Regression Framework

In the regression framework, the conditional expectation is typically modeled as:

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

under the assumption that Y is quantitative.

Regression Framework

In the regression framework, the conditional expectation is typically modeled as:

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

under the assumption that Y is quantitative.

But ...

- But what about Y qualitative?
- In particular, what about a binary Y?

Regression Idea

Right now we are considering simple regression (one Y, one X) in which Y is a binary variable:

$$E(Y|X) = \beta_0 + \beta_1 X$$

With the CHD example we have:

$$E(CHD|Age) = \beta_0 + \beta_1 Age$$

Because Y takes two possible values 1 and 0, we can think of it as having a Bernoulli distribution.

Review: Bernoulli Distribution

The **Bernoulli distribution** is the probability distribution of a random variable Y which takes the values of:

- 1 with probability p
- 0 with probability $1 - p$

The mean or expected value of Y is:

$$E(Y) = 1 \times p + 0 \times (1 - p) = p$$

The variance of Y is:

$$Var(Y) = E(Y^2) - E^2(Y) = p(1 - p)$$

Conditional Expectation

- We are actually dealing with $Y|X$
- So we assume that $Y|X$ has a Bernoulli distribution with parameter $p(x) = Prob(Y = 1|X = x)$

$$y_i = \begin{cases} 1 & \text{with} \quad Prob(1|x_i) = p_i \\ 0 & \text{with} \quad Prob(0|x_i) = 1 - p_i \end{cases}$$

- Thus the conditional expectation becomes:

$$E(Y|X) = Prob(Y = 1|X = x) = p(x)$$

Issues with using standard regression

With a binary response Y, we have that

$$E(Y|X) = Prob(Y = 1|X = x) = p(x)$$

In this case, if we use the standard model:

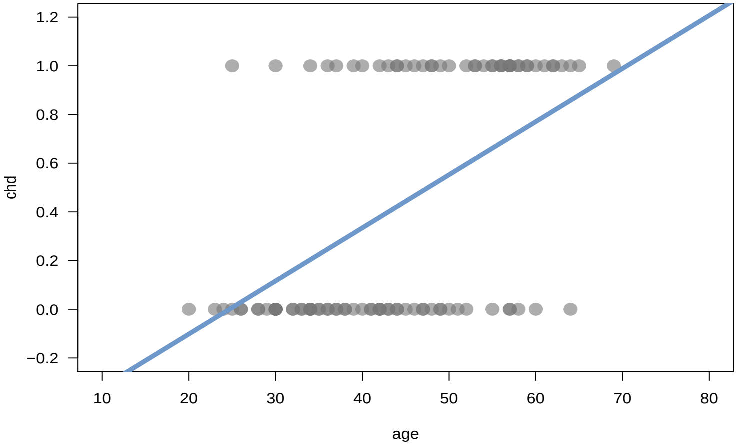
$$E(Y|X) = \beta_0 + \beta_1 X$$

we are actually modeling the probability $p(x)$ as a linear model:

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

Any issues with using this approach?

Issues with a linear model for $p(x)$



Issues with using standard regression

Naively applying OLS regression for binary Y turns out into:

$$E(Y|X) = \hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}$$

This fit will produce output values inside and outside of the range $[0, 1]$. In other words, we would have $-\infty < \hat{\mathbf{y}} < \infty$., because linear functions are unbounded.

Issues with using standard regression

Naively applying OLS regression for binary Y turns out into:

$$E(Y|X) = \hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

This fit will produce output values inside and outside of the range $[0, 1]$. In other words, we would have $-\infty < \hat{\mathbf{y}} < \infty$., because linear functions are unbounded.

However, probability values are only in the range $[0, 1]$.
Conclusion: the standard regression model (which assumes Y quantitative) is not really a good choice for categorical Y.

Other ideas?

Perhaps an obvious idea is to let $log(p(x))$ be a linear function

$$log(p(\mathbf{x})) = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

so that changing the input variable *multiplies* the probability by a fixed amount.

The problem is that logarithms are unbounded in only one direction, but linear models (in egenral) are not bounded.

Thinking Outside the Box

Transforming variables

Given that we don't have many data points for all possible ages, it is more convenient to bin the observations by groups of ages: e.g. 20 to 29, 30 to 34, 35 to 39, 40 to 49, ..., 55 to 59, 60 - 69

```
# regrouping by ages
groups <- c(19, seq(29, 59, by = 5), 69)
group_labels <- paste(c(groups[-9]+1), c(groups[-1]), sep = "-")
age_group <- cut(dat$age, breaks = groups, labels = group_labels,
                 include.lowest = TRUE)
table(age_group)
```

```
## age_group
## 20-29 30-34 35-39 40-44 45-49 50-54 55-59 60-69
##    10    15    12    15    13     8    17    10
```

Transforming the data

Now that we have age by groups, we can get the proportion of coronary heart disease cases in each age group

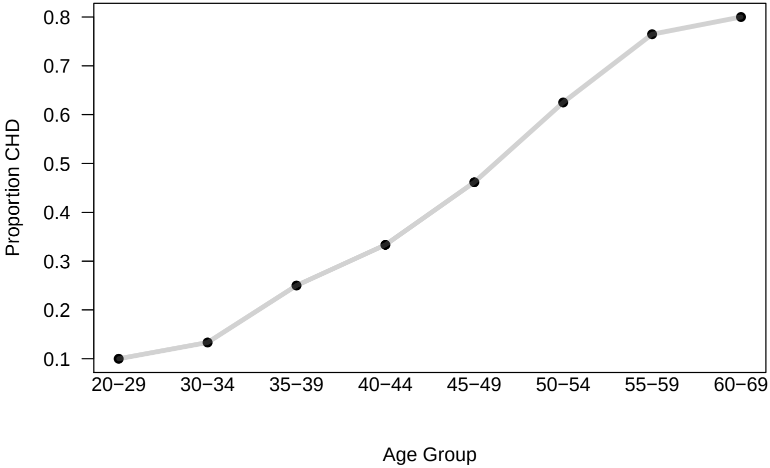
```
dat$age_group <- age_group

tbl <- dat %>%
  group_by(age_group) %>%
  summarize(prop_chd = mean(chd))
```

Transforming the data

```
# A tibble: 8 x 2
  age_group prop_chd
  <fct>        <dbl>
1 20-29        0.100
2 30-34        0.133
3 35-39        0.250
4 40-44        0.333
5 45-49        0.462
6 50-54        0.625
7 55-59        0.765
8 60-69        0.800
```

Now we can treat the proportions of CDH (`prop_chd`) as probabilities! (i.e. values ranging in interval $[0, 1]$).

Replotting the data



Replotting the data

R code to get the previous plot:

```r
plot(1:nrow(tbl), tbl$prop_chd, las = 1, pch = 19, xaxt = 'n',
     xlab = "Age Group", ylab = "Proportion CHD")
# connect points with a line
lines(1:nrow(tbl), tbl$prop_chd, lwd = 4, col = '#77777755')
# add better labels to x-axis
mtext(text = tbl$age_group, side = 1, at = 1:nrow(tbl))
```

I still like to use base graphics in addition to ggplot2

What's going on?

- Plotting the proportions of CHD by age-group produces an interesting plot.
- The shape of the curve roughly follows a typical **sigmoid** curve.
- This curve pattern is better to model probabilities.
- Various mathematical functions produce sigmoid-shape curves.
- One of such functions is the **logistic** function.

Logistic Function

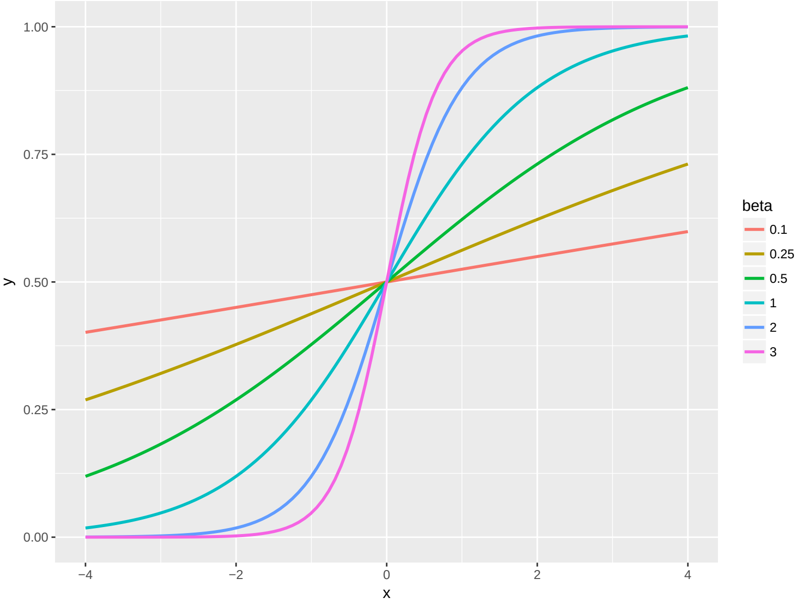
$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

About the Logistic function

Logistic Function

- It behaves like the distribution function of a symmetrical density, with midpoint at zero.
- Its domain moves through the real number axis.
- It rises monotonically between the bounds of 0 and 1.
- Originally developed to describe the course of a *proportion* over time t with $z = a + bt$.
- It is a *growth curve* since $f(t)$ rises monotonically with t.

Examples of Logistic Curves



Logistic Curves

```r
# x-y coordinates for various logistic functions
n = 100
beta_vals <- c(0.1, 0.25, 0.5, 1, 2, 3)
betas <- rep(beta_vals, each = n)
x <- rep(seq(-4, 4, length.out = n), length(beta_vals))
y <- exp(betas*x) / (1 + exp(betas*x))

# assemble data frame for plotting purposes
logistic <- data.frame(
  x = x, y = y, beta = as.factor(betas)
)

# some examples of logistic curves
ggplot(data = logistic, aes(x = x, y = y, group = beta)) +
  geom_line(aes(col = beta), size = 1)
```

Logistic Function

For logisitc regression purposes, we prefer this format:

$$f(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Logistic Function

Sometimes you may also find the logistic equation in an alternative form:

$$
\begin{aligned}
f(x) &= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \\
&= \frac{1}{\frac{1+e^{\beta_0+\beta_1 x}}{e^{\beta_0+\beta_1 x}}} \\
&= \frac{1}{\frac{1}{e^{\beta_0+\beta_1 x}}+1} \\
&= \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}
\end{aligned}
$$

Logistic Approach

Since probability values range inside $[0, 1]$, instead of using a line to try to approximate these values, we should use a more adequate curve.

This is the reason why sigmoid-like curves, such as the logistic function, are preferred for this purpose.

Logistic Function

The logistic function:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

can be used to model the conditional expectation (which we now know that takes the form of a probability)

$$E(Y|X = x_i) = p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

So far ... So good?

References

- **The origins and development of the logit model** by J.S. Cramer (2003)
  http://www.cambridge.org/resources/0521815886/1208_default.pdf
- **Applied Logistic Regression** by Hosmer and Lemeshow (2000).
- **Statistical Regression and Classification** by Norman Matloff (2017)
  Chapter 4:Generalized Linear and Nonlinear Models. CRC Press.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*. Wiley.

References (French Literature)

- **Modeles Statistiques pour Donnees Qualitatives** by Droesbeke et al (2005). *Chapter 6: Modele a reponse dichotomique* by P.L. Gonzalez. Editions Technip, Paris.
- **Statistique Explicative Appliquee** by Nakache and Confais (2003). *Chapter 4: Modele logistique binaire*. Editions Technip, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 11: La regression logistique binaire*. Dunod, Paris.