Data603LogisticModellingPart1

November 7, 2019 5:31 PM

Data603Lo...

Data 603: Statistical Modelling with Data

Logistic Regression

Part I : Introduction to the Logistic Regression Model

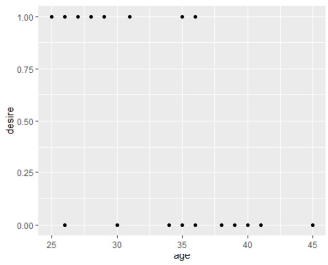
Example: The *desire* data show the distribution of 24 currently married and fecund women interviewed in the Fiji Fertility Survey, according to age, education, desire for more children (wife's perception of husband's desire for additional children). The data are provided in **desire.xlsx** file

X1 = age (year)

X2 = education (0=none, 1=some),

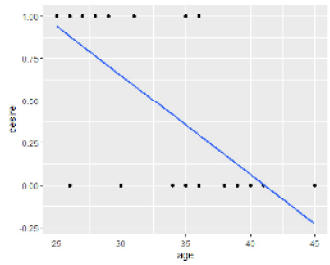
Y = desire for more children (0=no more, 1=more),

```
library("readxl")
desire <- read_excel("c:/Users/thuntida.ngamkham/OneDrive - University of Cal
gary/dataset603/desire.xlsx")
library(ggplot2)
ggplot(data = desire, mapping = aes(x = age, y = desire))+
  geom_point()
```

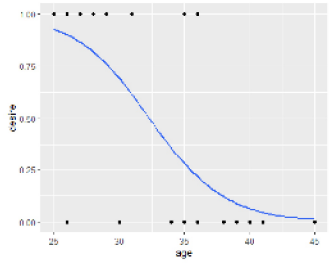


Using linear regression model

```
library("readxl")
library(ggplot2)
desire <- read_excel("c:/Users/thuntida.ngamkham/OneDrive - University of Cal
gary//dataset603/desire.xlsx")
ggplot(data = desire, mapping = aes(x = age, y = desire))+geom_point()+
geom_smooth(method=lm,se=F)
```




```
library('readxl')
library(ggplot2)
desire <- read_excel('C:/Users/Olamilide.Ogunbiyi/OneDrive - University of Calgary/Values.xls/desire.xlsx')
ggplot(data = desire, mapping = aes(x = age, y = desire))+geom_point()+
  stat_smooth(method='glm',method.args=list(family='binomial'),se=FALSE)
```



The linear regression model discussed in Multiple Regression assumes that the response variable Y is **quantitative**. But in many situations, the response variable is instead **qualitative**. For example, eye color is qualitative, taking qualitative values blue, brown, or green. Often qualitative variables are referred to as categorical.

What distinguishes a 'logistic regression model' from the linear regression model is that the outcome variable in logistic regression is **binary** or **dichotomous**.

In this topic, we study approaches for predicting qualitative responses, a process that is known as classification. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this topic we discuss simple Logistic Regression and Multiple Logistic Regression for a qualitative binary response.

What is Logistic Regression ?

Logistic Regression seeks to:

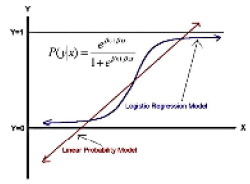
1. **Model** the probability of an event occurring depending on the value of the independent variables, which can be categorical or numerical.
2. **Estimate** the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur.
3. **Predict** the effect of a series of variables on a binary response variable.
4. **Classify** observations by estimating the probability that an observation is in a particular category.

$$y = \beta_0 + \beta_1 X + \beta_2 X + \dots \varepsilon$$

Why Not Linear Regression

1. **Linear regression assumptions** The linear regression model is based on an assumption that the response Y is continuous, with errors are normally distributed. If the response variable is binary, this assumption is clearly violated, and so in general we might expect our inferences to be invalid.
2. **Predicted values may be out of range** For a binary outcome, the mean is the probability of a 1, or success. If we use linear regression to model a binary outcome it is entirely possible to have a fitted regression which gives predicted values for some individuals which are outside of the $[0, 1]$ range of probabilities.

We will also illustrate the concept of classification using the simulated **Default data** set. We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance. In the *Default* data set, the response **default** falls into one of two categories, Yes or No. Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category.



$$P(Y=1 \mid X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Figure 1 Comparing Graphs between simple linear regression and logistic regression

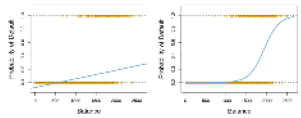


Figure2 Classification using Default data for simple linear regression and logistic regression

Left figure: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default (No or Yes).

Right figure: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

For the Default data, logistic regression models the probability of default. For example, the probability of default given *balance* can be written as

$$Pr(default = Yes|balance).$$

where

$balance$ = the independent variable

$default$ = the response variable which is a binary outcome (YES/NO)

The values of Pr(default = Yes|balance), will range between 0 and 1. Then for any given value of balance, a prediction can be made for *default*.

For example, one might predict default = Yes for any individual for whom Pr(default = Yes|balance) > 0.5. Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as Pr(default = Yes|balance) > 0.1.

The Logistic Model

How should we model the relationship between $p(X) = Pr(Y = 1|X)$ and X? (For convenience we are using the generic 0/1 coding for the response). If we use the approach $p(X) = \beta_0 + \beta_1 X$ to predict default=Yes using balance, then we obtain the model shown in the left-hand panel of Figure 2.1.

Here we see the problem with this approach: for balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1. These predictions are not sensible, since of course the true probability of default, regardless of credit card balance, must fall between 0 and 1. This problem is not unique to the credit default data. Any time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others (unless the range of X is limited). To avoid this problem, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X. Many functions meet this description.

Simple Logistic Regression Model for a Binary Dependent Variable (Quantitative independent variable)

$$p(y) = P(y = 1|X) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

where

y = 1 if category A occurs

= 0 if category B occurs

$P(y)$ = P(Category A occurs) = π

Note that the general logistic model is not a linear function of the β parameters. Obtaining the parameter estimate of a nonlinear regression model, such as the logistic model, is a

dataset	Y	X
	1	25
	0	30
	1	35
	0	15

$$\log \prod \left(P(X=n) \binom{m}{x} p^x q^{m-x} \right)$$

numerically tedious process and often requires sophisticated computer programs. We use a method called **maximum likelihood estimation** to estimate the β_1 parameter.

The right-hand panel of Figure 2.1 illustrates the fit of the logistic regression model to the Default data. Notice that for low balances we now predict the probability of default as close to, but never below, zero. Likewise, for high balances we predict a default probability close to, but never above, one.

The logistic function will always produce an S-shaped curve of this form, and so regardless of the value of X, we will obtain a sensible prediction. We also see that the logistic model is better able to capture the range of probabilities than is the linear regression model.

Success — $\beta_1 > 0$ $\beta_1 < 0$ — Failure

β_0 β_1 β_2

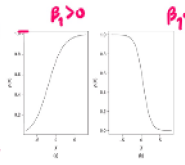


Figure X

Focusing on the single predictor case, if the parameter $\beta_1 > 0$ then the logistic regression model assumes that the probability of success is a monotonically increasing function of X. That is, the probability never decreases as X gets larger; it stays the same or increases. If the parameter $\beta_1 < 0$ the reverse is true.

Figure 3(a) shows the regression line when $\beta_1 = 1$ and $\beta_0 = 0.5$. As is evident, curvature is allowed and predicted probabilities always have a value between 0 and 1. Figure 3(b) shows the regression line when $\beta_1 = -1$ and $\beta_0 = 0.5$. So now, the regression line is monotonically decreasing. (The predicted probability never increases).

P(success) = 0.8

P(failure) = 0.2

the odds = $\frac{P(\text{success})}{P(\text{failure})}$ = 4

The overall odds of wining a lottery prize one 1 in 13

$\therefore$ P(wining) = $\frac{1}{13}$ ✗ = $\frac{1}{14}$

Considering the logistic Regression Model, we find that

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1|x)}{1-P(y=1|x)}$$

$$= \frac{\pi}{1-\pi}$$

$$= e^{\beta_0+\beta_1 x}$$

$$\pi = P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$\frac{P(Y=1|X)}{P(Y=0|X)} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$$

The quantity $\frac{\pi}{1-\pi}$ is called **the odds**, and can take on any value between 0 and ∞.

What is the odds ?

To appreciate the logistic model, it's helpful to have an understanding of odds. Most people regard probability as the "natural" way to quantify the chances that an event will occur. We automatically think in terms of numbers ranging from 0 to 1, with a 0 meaning that the event will certainly not occur and a 1 meaning that the event certainly will occur. But there are other ways of representing the chances of event, one of which -the odds- has a nearly equal claim to being "natural."

For example,

An odds of 4 means we expect 4 times as many occurrences as non-occurrences.

An odds of 1/5 means that we expect only one-fifth as many occurrences as non-occurrences.

In general

$$\text{Odds} = \frac{P(y=1|x)}{1-P(y=1|x)} = \frac{\text{probability of event}}{\text{probability of no event}}$$

The odds (model) = $e^{\beta_0+\beta_1 x}$

Considering the logistic Regression Model, we find that:

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1|x)}{1-P(y=1|x)} = \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x}$$

$$\pi = P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

The quantity $\frac{\pi}{1-\pi}$ is called **the odds**, and can take on any value between 0 and ∞.

What is the odds ?

To appreciate the logistic model, it's helpful to have an understanding of odds. Most people regard probability as the "natural" way to quantify the chances that an event will occur. We automatically think in terms of numbers ranging from 0 to 1, with a 0 meaning that the event will certainly not occur and a 1 meaning that the event certainly will occur. But there are other ways of representing the chances of events, one of which -the odds- has a nearly equal claim to being "natural."

For example,

An odds of 4 means we expect 4 times as many occurrences as non-occurrences.

An odds of 1/5 means that we expect only one-fifth as many occurrences as non-occurrences.

In general:

$$\text{Odds} = \frac{P(y=1|x)}{1-P(y=1|x)} = \frac{\text{probability of event}}{\text{probability of no event}}$$

$$\frac{P(Y=1|X)}{P(Y=0|X)} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$$

$$\text{The odds (model)} = e^{\beta_0+\beta_1 X}$$

Relationship between Odds and Probability

Probability	Odd
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

$p = 0.5, 1-p = 0.5$

odds < 1
p(success) < p(failure)

odds > 1
p(success) > p(failure)

Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.

In general,

If an odds > 1, then the probability of success is higher than failure.

If an odds < 1, then the probability of success is lower than failure.

If an odds = 1, then the probability of success is equal for failure.

For more example about Default data, on average 1 in 5 people with an odds of 1/4 will default, since $p(y=1|X)=0.2$ implies an odds of $\frac{0.2}{1-0.2} = 1/4$.

Likewise on average nine out of every ten people with an odds of 9 will default, since $p(y=1|X)=0.9$ implies an odds of $\frac{0.9}{1-0.9} = 9$.

By taking **the natural logarithm** of both sides of the Logistic model, we arrive at:

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1|x)}{1-P(y=1|x)} = \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x}$$

$$\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0+\beta_1 x \quad \text{------*}$$

The * is called **the log-odds or the logit**.

This transformation is useful, because it creates a variable with a range from $-\infty$ to $+\infty$. Hence, this transformation solves the problem we encountered in fitting a linear model to probabilities. Because probabilities [the dependent variable] only range from 0 to 1, we can get linear predictions that are outside of this range. If we transform our probabilities to logits, then we do not have this problem because the range of the logit is not restricted. In addition, the interpretation of logits is simple-take the exponential of the logit and you have the odds for the two groups in question.

Fitting the Logistic Regression Model

In linear regression, the method used to estimate β_i is least square. In that method, we choose those values of $\beta_0, \beta_1, \ldots, \beta_p$ which minimize the sum of squared residuals. In logistic regression, we estimate β_i by maximizing the log likelihood expression.

In this class, we will use R software to calculate the regression coefficients, so we do not need to concern with the details of the maximum likelihood fitting procedure.

For example, using **Default data** to predict the probability of default using balance.

```
library(ISLR) # for Default data set
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
mylogit <- glm(default ~ balance, data = Default, family = "binomial")
coefficients(mylogit)
```

```
##   (Intercept)       balance
## -10.651330614   0.005498917
```

$\hat{\beta}_0$ $\hat{\beta}_1$

R function coefficients() : perform the regression coefficients

The output shows the coefficient estimates and related information that result from fitting a logistic regression model on the Default data. The maximum likelihood estimates of β_0 and β_1 are $\hat{\beta}_0 = -10.6513$ and $\hat{\beta}_1 = 0.0055$. Therefore,

The estimated logistic regression model is

$$p(\hat{Y}=1 \mid X)_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{-10.6513+0.0055X}}$$ the odds

The Logit is

$$\widehat{logit} = -10.6513 + 0.0055X$$

The odds = $e^{-10.6513+0.0055X}$

A confidence Interval for the Logit

A $100(1-\alpha)\%$ confidence interval for β_1 would be:

```
library(ISLR) # for Default data set
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
mylogit <- glm(default ~ balance, data = Default, family = "binomial")
confint(mylogit)
```

```
##                 2.5 %       97.5 %
## (Intercept) -11.383289955 -9.956559064
## balance       0.005078925  0.005943365
```

R function confint() : perform a 95% confidence interval

$\widehat{logit} = -10.6513 + 0.0055X$

balance

Interpretations of Logistic Regression Coefficients in the Logistic Model (Quantitative independent variable)

In general, the coefficient $\hat{\beta}_1$ in the logistic model estimates the change in the log-odds (same concept as linear regression). For example, from the default output,

$\hat{\beta}_1 = 0.0055$. This indicates that an increase in balance is associated with an increase in the probability of default. To be precise,

a one-unit **increase** in balance is associated with an **increase** in the log odds of default by 0.0055 units or

For every $\$1$ increase in balance, we estimate the log odds of a default increase by 0.0055. What does it really mean??

$$\hat{y} = 2 + 5X_0 = (7)$$

at $X = 1$

on average every 1 unit increase in X, y increas 5 units

at $X = X_0 + 1 = (2)$

$$\hat{y} = 2 + 5(X_0+1) = \underbrace{2+5(X_0)}_{7} + 5 = 12$$

at X_0

The odds = $\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 X_0}$

By computing $e^{\hat{\beta}_1}$ (antilog of the coefficient), so we interpret in terms of the odds.

$$\hat{\beta}_1 = 0.0055$$
$$e^{\hat{\beta}_1} = e^{0.0055} = 1.0055$$

For every 1 dollar increase in balance, we estimate the odds of a default to be multiplied by about 1.0055 i.e. there is an increase of 0.5% [=(1.005-1)*100%] of the odds of a default.

1.005

at X_0+1 ⇒ The odds = $e^{\hat{\beta}_0+\hat{\beta}_1(X_0+1)}$

$= e^{\hat{\beta}_0+\hat{\beta}_1 X_0+\hat{\beta}_1}$

$= e^{\hat{\beta}_0+\hat{\beta}_1 X} \cdot e^{\hat{\beta}_1}$

The odds .

Note! R function to calculate the antilog for $\hat{\beta}_1$ is provided below.

```
library(ISLR) # for Default data Set
mylogit <- glm(default ~ balance, data = Default, family = "binomial")
sum.coef<-summary(mylogit)$coef
est<-exp(sum.coef[,1])
print(est)
```

```
##  (Intercept)      balance
## 2.369833e-05 1.005514e+00
```

Note! R function to calculate the antilog for a confidence interval for β_1 is provided below.

```
library(ISLR) # for Default data Set
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21342
```

Note! R function to calculate the confidence interval for β_1 is provided below.

```
library(ISLR) # for Default data set
summary(Default)

##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554

mylogit <- glm(default ~ balance, data = Default, family = "binomial")
cbt< exp(confint(mylogit))
print(cbt)

##                 2.5 %      97.5 %
## (Intercept) 1.138415e-05 4.584352e-05
## balance     1.982092e+02 1.029951e+02
```

Inclass Practice Problem

Example: The desire data, showing the distribution of 2% currently married and fecund women interviewed in the Fiji Fertility Survey, according to age, education, desire for more children, the data are provided in **desire.xlsx** file

X1= age (year)

X2= education (0=none, 1=some),

Y= desire for more children (0=no more, 1=more),

a) Fit the Logistic Regression Model to predict the probability of desire for more children using age.

b) Construct a 95% confidence interval for the logit model.

For every 1 year increases in age, we estimate the odds of desire to be multiplied by about 0.705 (i.e. a decrease of (1-0.705)*100%=29.5% of the odds).

Testing For The Significance of The Coefficients

After estimating the coefficients, there are several steps involved to assessing the appropriateness, adequacy and usefulness of the model. **Firstly**, the importance of each of the explanatory variables is assessed by carrying out statistical tests of the significance of the coefficients. **Secondly**, the overall goodness of fit of the model is then tested. **Lastly**, the model fit or the ability of the model to discriminate between the two groups defined by the response variable is evaluated.

Significance Of The Coefficients

The Wald Z test or the Wald χ^2 test is the test of significance for individual regression coefficients in logistic regression. (Note that we used t-test for Simple and Multiple Linear Regression). The ratio

1) H_0: $\beta_1 = 0$
The probability of $y = 1$ does not depend on X
H_1: $\beta_1 \neq 0$
The probability of $y = 1$ depends on X

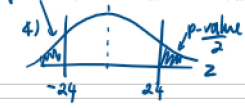
2) $\alpha = 0.05$

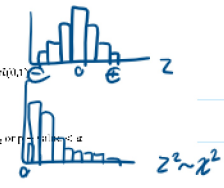
$Z_{cal} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$ follows approx. a $Normal(0,1)$

or $Z^2_{cal} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right)^2$ follows approx. a χ^2_1

3) We reject the null hypothesis when $|Z_{cal}| > Z_{\alpha/2}$ or p-value $< \alpha$

4) p-value






```
library(ISLR) # for Default data set
library(aod) # for Wald test
summary(Default)

##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554

mylogit <- glm(default ~ balance, data = Default, family = "binomial")
# The Wald Z test
summary(mylogit)

##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <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: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8

# Wald chi-square test for full vs reduced model
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 2) # Terms tells
# which terms in the model are to be tested.

## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 622.7, df = 1, P(> X2) = 0.0
```

R function

The odds .

$odd_{s_1} = 1$

$odd_{s_2} = 1 \times 1.0055 = 1.0055$

$odds = e^{5-2X} = 0.7$

Income = 10,000

$odd_{s_1} = 0.7$ $\frac{7}{10}$

Income 10,001

$odd_{s_2} = 0.7 \times e^{\hat{\beta}_1} = 0.7 \times 0.8 = 0.56$

$\left| \frac{0.56 - 0.7}{0.7} \right| = |-0.2|$

decrease

$= 0.2$

$= 20\%$

95% CI for the logit model.

$0.0004 < \beta_1 < 0.00695$

$\ominus < \beta_1 < \oplus$

$Z_{cal} = 24.95$

$\chi^2_{cal} = (24.95)^2 = 622.7$

$y = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$ (term2)

wald.test(b = coef(.....), Sigma = vcov(.....), Terms = ...): *perform full or reduced model for the chi square test.*

Deviance: The deviance is negative two times the maximized log-likelihood; the smaller the deviance, the better the fit (see Part 2 for more details).

Note!

Terms tells R which terms in the model are to be tested, in the example, terms 2 represents testing for $\beta_1 = 0$

From the Default data, for the Wald Z test, Yeal=24.95 with p-value <2.10011, we reject H_0 therefore the probability of default depends on balance. For the Wald χ^2 test, $\chi^2 = 622.7$ with p-value zero. We also conclude that the probability of default depends on *balance*.

Making Predictions

Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance. For example, using the coefficient estimates given in the output, we predict that the default probability for an individual with a balance of $1,000 is

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \times 1000}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \times 1000}} = 0.00576$$

which is below 1%. In contrast, the predicted probability of default for an individual with a balance of $2,000 is much higher, and equals 0.586 or 58.6%.

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \times 2000}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \times 2000}} = 0.586$$

Note! Using R function to calculate $\hat{p}$ when balance =1000 dollars

```
library(ISLR) # for Default data Set
mylogit <- glm(default ~ balance, data = Default, family = "binomial")
newdata = data.frame(balance=1000)
predict(mylogit, newdata, type='response')
```

```
##           1
## 0.005752145
```

A confidence Interval for the odds e^{β_1}

Letting $SE_{\hat{\beta}_1}$ denote the estimated standard error of $\hat{\beta}_1$, the estimate of β_1, a 95-% confidence interval for the change of the odds is

$$e^{\hat{\beta}_1 \pm 2 \cdot SE_{\hat{\beta}_1}}$$

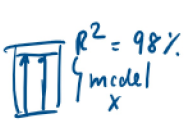
```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <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: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

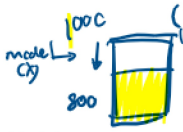
odds = 1.0055
95% CI for the odds
1.00508 <odds < 1.005948

less than 1 <odds < more than 1

```
##                   est     lower.ci     upper.ci
## (Intercept) 2.3069J2e-05 1.1610J7e-05 4.5844J7e-05
## balance     1.0055514e+00 1.0250932e+00 1.0055848e-00
```

From the output, a 95% confidence interval for e^{β_1} is 1.005 to 1.006 implies that there is a positive relationship between *balance* and *default* as this confidence interval doesn't cover 1 and it covers a range greater than 1. This means that an increasing in $1 in balance increases the odds of *default* between 0.5% and 0.6%.





(null deviance) unexplained variation

$\pi = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$

Model Fit in Logistic Regression Model

In linear regression, R^2 is a very useful quantity, describing the fraction of the variability in the response that the explanatory variables can explain. There are a number of ways one can define an analog to R^2 in the logistic regression case, but none of them are as widely useful as R^2 in linear regression. To evaluate the performance of a logistic regression model, you would work on, always look for:

1. Deviance → a measure how much UNEXPLAINED variation in a logistic model

Deviance is a measure of goodness of fit of a generalized linear model [GLM]. Or rather, R software reports two forms of deviance

- the null deviance and the residual deviance. The null deviance shows how well the response variable is predicted by a model that includes only the intercept.
- the residual deviance indicates how well the response is predicted by the model with independent variables.

For our example, we have a value of 2920.6 points on 9999 degrees of freedom. Including the independent variable (balance) decreased the deviance to 1596.5 points on 9998 degrees of freedom, a significant reduction in deviance.

The Residual Deviance has reduced by 1324.1 points with a loss of one degrees of freedom.

2. AIC (Akaike Information Criteria) estimates the relative amount

For our example, we have a value of 2920.6 points on 9999 degrees of freedom. Including the independent variable (balance) decreased the deviance to 1596.5 points on 9998 degrees of freedom, a significant reduction in deviance.

The Residual Deviance has reduced by 1324.1 points with a loss of one degrees of freedom.

2. AIC (Akaike Information Criteria)

estimates the relative amount of information lost by a given model.

The Akaike Information Criterion (AIC) provides a method for assessing the quality of your model through comparison of related models. It's based on the Deviance.

However, unlike adjusted R-squared, the number itself is not meaningful. If you have more than one similar candidate models (where all of the variables of the simpler model occur in the more complex models), then you should select the model that has the smallest AIC. So it's useful for comparing models, but isn't interpretable on its own.

Note!

Fisher's scoring algorithm is a derivative of Newton's method for solving maximum likelihood problems numerically.

3. ROC curve

ROC stands for **Receiver Operating Characteristic**. To explain the ROC curve, we need to understand the important notions of sensitivity and specificity of a test or prediction rule.

Sensitivity and specificity

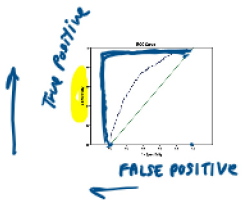
The sensitivity is defined as the probability of the prediction rule or model predicting an observation as 'positive' given that in truth (Y=1). In words, the sensitivity is the proportion of truly positive observations which is classified as such by the model or test. Conversely, the specificity is the probability of the model predicting 'negative' given that the observation is 'negative' (Y=0).

A model needs to not only correctly predict a positive as a positive, but also a negative as a negative.

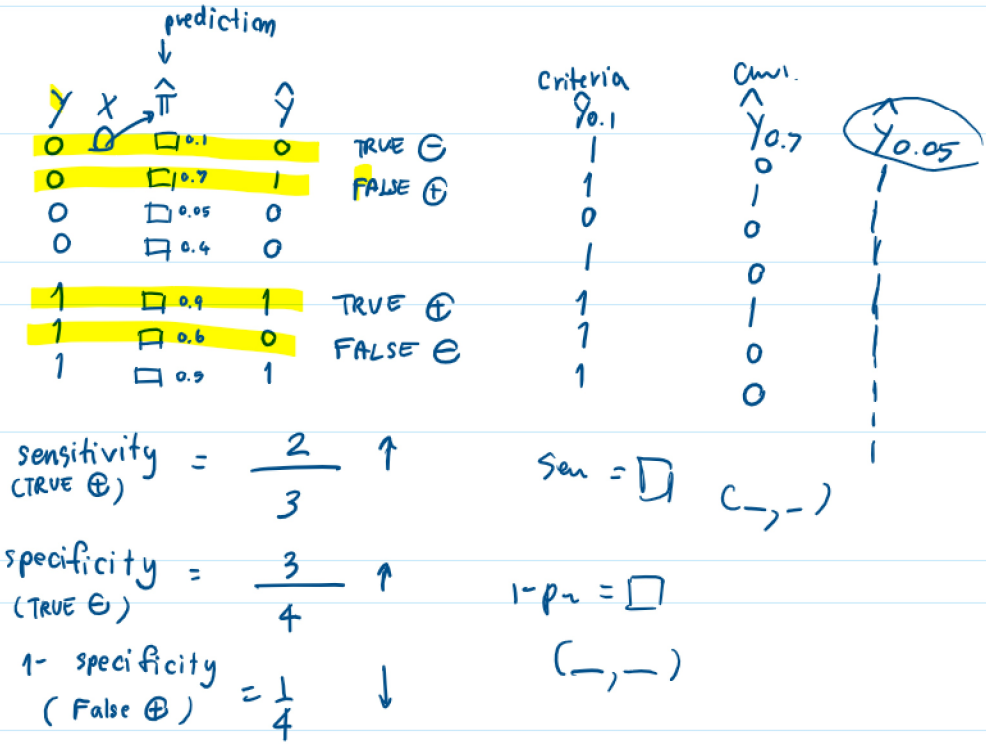
Our model or prediction rule is perfect at classifying observations if it has 100% sensitivity and 100% specificity. Unfortunately, in practice this is (usually) not attainable. So how can we summarize the discrimination ability of our logistic regression model?

The ROC curve does this by plotting **the true positive rate (sensitivity)**, the probability of predicting a real positive will be a positive, against **false positive rate (1-specificity)**, the probability of predicting a real negative will be a positive. **The best decision rule** is high on sensitivity and low on 1-specificity. It's a rule that predicts most true positives will be a positive and few true negatives will be a positive.

How to explore the ROC curve



1. The further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general.



2. There are useful statistics that can be calculated from this curve, the Area Under the Curve (AUC). This tells us how well the model predicts the probability of *Y*.

Inclass Pratice Problem

From the default data,

a) Write both the logistic regression model of Default on Income and the logit transformation of this logistic regression model.

b) Interpret the logistic regression coefficient $e^{\hat{\beta}_1}$ in logistic model

c) Test if The probability of default depends on Income at $alpha = 0.05$

d) Find a 95% Confidence Interval for the logistic regression coefficient $e^{\hat{\beta}_1}$

e) Use the method of Model Fit in Logistic Regression Model to evaluate the performance of a logistic regression model

f) Predict the probability of default when Income= 60,000 dollars. Would you consider a person with $60,000 income defaults on payment?

a) $\hat{\pi} = P(\hat{Y}=1|income)$

$= \frac{e^{-3.094 - 0.00008353X}}{1 + e^{-3.094 - 0.00008353X}}$

$\hat{logit} = -3.094 - 0.00008353X$

the odds model $= e^{-3.094 - 0.00008353X}$

b) $e^{\hat{\beta}_1} = e^{-0.00008353} = 0.99991645$

d) 95% CI for e^{β_1}

$0.99998340 < e^{\beta_1} < 0.99999989$

e) ROC

AIC

AUC = 0.5327

Inclass Practice Problem

Experience in hiring. Suppose you are investigating years of experience the hiring practices of a particular firm. Is there any sufficient evidence to indicate that the years of experience is an important predictor of hiring status? If yes, interpret the logistic regression coefficient $e^{\hat{\beta}_1}$ in logistic model.The data are provided in **DISCRIM.csv file**

a) Write both the logistic regression model of Hire on Experience and the logit transformation of this logistic regression model.

b) Interpret the logistic regression coefficient $e^{\hat{\beta}_1}$ in logistic model

c) Test if The probability of hiring depend on experience at $alpha = 0.05$

d) Find a 95% Confidence Interval for the odds ratio for the logistic regression coefficient $e^{\hat{\beta}_1}$

e) Using the method of Model Fit in Logistic Regression Model to evaluate the performance of a logistic regression model

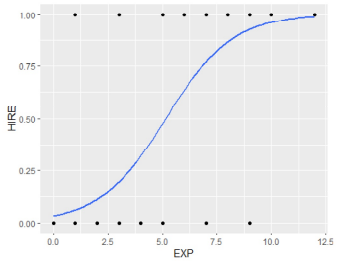
f) Predict the probability of hiring when experience= 1 years. Would you consider a person with a 7 years job experience will be hired?

Visualizing the data and logistic regression model

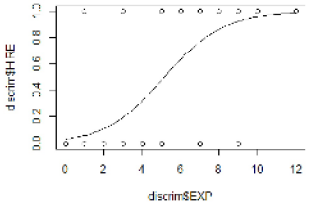
The data and logistic regression model can be plotted with ggplot2 or base graphics:

```
library(ggplot2)
discrim=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary
/dataset603/DISCRIM.csv", header = TRUE)
mylogit <- glm(HIRE ~EXP, data = discrim, family = "binomial")

#option1 using ggplot function
ggplot(discrim, aes(x=EXP, y=HIRE)) + geom_point() +
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```




```
#option2 using plot (base graphic)
plot(discrim$EXP, discrim$HIRE)
curve(predict(mylogit, data.frame(EXP=x), type="response"), add=TRUE)
```



Simple Logistic Regression Model with a Qualitative Independent variable.

One can use qualitative predictors with the logistic regression model using the dummy variable approach as well. For example, the Default data set contains the qualitative variable student. To fit the model we simply create a dummy variable that takes on a value of 1 for students and 0 for non-students. The logistic regression model that results from predicting probability of default from student status can be seen in the output

```
library(ISLR) # for Default data Set
summary(Default)
```

```
##  default    student       balance           income     
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772  
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
##                        Median : 823.6   Median :34553  
##                        Mean   : 835.4   Mean   :33517  
##                        3rd Qu.:1166.3   3rd Qu.:43808  
##                        Max.   :2654.3   Max.   :73554
```

```
mylogit <- glm(default ~ student, data = Default, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = default ~ student, family = "binomial", data = Default)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max  
## -0.2970  -0.2970  -0.2434  -0.2434   2.6585  
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
## studentYes   0.40489    0.11502    3.52 0.000431 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 2908.7  on 9998  degrees of freedom
## AIC: 2912.7
##
## Number of Fisher Scoring iterations: 6
```

From the output, the logit is

$$\widehat{\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right)} = -3.50413 + 0.40489X$$

and the estimated logistic regression model is

$$\hat{p} = \frac{e^{-3.50413 + 0.40489X}}{1 + e^{-3.50413 + 0.40489X}}$$

Interpretations of Logistic Regression Coefficients in the Logistic Model for Qualitative independent variable

By computing $e^{\hat{\beta}_1}$ (antilog of the coefficient), so we interpret in terms of the odds ratio.

Odds review: if π_1 is the probability of default for students, and π_2 is for non students, the odds ratio for student nonstudent is defined as $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

$$\hat{\beta}_1 = 0.40489$$
$$e^{\hat{\beta}_1} = e^{0.40489} = 1.4991288$$

Note! R function to calculate the antilog for $\hat{\beta}_1$

```
library(ISLR) # for Default data Set
mylogit <- glm(default ~ factor(student), data = Default, family = "binomial"
)
sum.coef<-summary(mylogit)$coef
coef<-exp(sum.coef[,1])
print(coef)
```

```
##         (Intercept) factor(student)Yes 
##          0.03002999          1.49912885
```

The odds ratio of $e^{\hat{\beta}_1} = 1.4991288$ tells us that the predicted odds of a default for students are 1.4991288 times the odds for non students. In other words, the odds of a default for students are 49.91288% higher than the odds for non students.

Making Predictions

Once the coefficients have been estimated, it is a simple matter to compute the probability of default for each dummy variable (0 or 1). For example, using the coefficient estimates given in the output, we predict that the default probability for a student and non student

$$\hat{p} = \frac{e^{-3.50413 + 0.40489X}}{1 + e^{-3.50413 + 0.40489X}}$$

which:

$$P(default|student = YES) = \frac{e^{-3.50413 + 0.40489(1)}}{1 + e^{-3.50413 + 0.40489(1)}} = 0.0431$$

$$P(default|student = No) = \frac{e^{-3.50413 + 0.40489(0)}}{1 + e^{-3.50413 + 0.40489(0)}} = 0.0292$$

The predicted probability of default for a student with a balance is 4.31% while 2.9% for non-student. This indicates that students tend to have higher default probabilities than non-students.

Inclass Practice Problem

Gender discrimination in hiring. Suppose you are investigating allegations of gender discrimination in the hiring practices of a particular firm. An equal-rights group claims that females are less likely to be hired than males. Is there any sufficient evidence to indicate this gender bias in your hiring studies?

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

The odds $= e^{\hat{\beta}_0 + \hat{\beta}_1 X}$

The odds of student $= e^{\hat{\beta}_0 + \hat{\beta}_1} = \frac{\hat{\pi}_{st}}{1 - \hat{\pi}_{st}}$

The odds of nonstudent $= e^{\hat{\beta}_0} = \frac{\hat{\pi}_{nonst}}{1 - \hat{\pi}_{nonst}}$

$$\frac{\text{The odds of St}}{\text{The odds of non St}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{e^{\hat{\beta}_0}} = \frac{e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1}}{e^{\hat{\beta}_0}}$$

The odds ratio $= e^{\hat{\beta}_1}$

non-students.

Inclass Practice Problem

Gender discrimination in hiring. Suppose you are investigating allegations of gender discrimination in the hiring practices of a particular firm. An equal-rights group claims that females are less likely to be hired than males. Is there any sufficient evidence to indicate that gender is an important predictor of hiring status?

DISCRIM.CSV

Male = 1
Female = 0

$e^{\beta_1} = e^{1.7918}$

$$\text{The odds ratio} = \frac{\frac{\hat{\pi}_{male}}{1-\hat{\pi}_{male}}}{\frac{\hat{\pi}_{female}}{1-\hat{\pi}_{female}}} = e^{\hat{\beta}_1} = e^{1.7918} = 6$$

Example: The desire data, showing the distribution of 24 currently married and fecund women interviewed in the Fiji Fertility Survey, according to age, education, desire for more children. the data are provided in **desirealtx.txt** file

X1= age (year)

X2= education [0=none, 1=some],

Y= desire for more children: [0=no more, 1=more],

Is there any sufficient evidence to indicate that education is an important predictor of desire for children?

a) Fit the logistic regression model

```
library('readxl')
desire <- read_excel("c:/Users/chantida.ngamkham/OneDrive - University of Fal
gary/Desktop/685/desire.xlsx")
mylogit<-glm(desire~factor(education),data=desire, family = "binomial")
summary(mylogit)

##
## Call:
## glm(formula = desire ~ factor(education), family = "binomial",
##     data = desire)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
##  1.73448  0.98652  0.69573  0.78898  1.48258
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.0321     0.9477  -1.286   0.2067
## factor(education)1  1.9459     0.9710   2.004   0.0451 *
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.271  on 23  degrees of freedom
## Residual deviance: 28.650  on 22  degrees of freedom
## AIC: 32.65
##
## Number of Fisher Scoring iterations: 4

sum.coef<-summary(mylogit)$coef
est<-exp(sum.coef[,1])
print(est)

##       (Intercept) factor(education)1
##         0.1310345         7.0000000
```

The odds ratio of $e^{1.9459}=7$ tells us that the predicted odds of desire for more children for educated person are 7 times the odds for uneducated person. In other words, the odds of desire for more children for educated person are 600% higher than the odds for uneducated person.