CIS9660: Data Mining for Business Analytics

4. Introduction to Logistic Regression

Imp 3 4 5 6 8 9 10 13 14

(Gujarati, D.N., 2011. Econometrics by example. New York: Palgrave Macmillan. Chapter 8)

An Illustrative Example:
To smoke or Not to smoke

There is a random sample of 1,196 US males. The variables are as follows:

- Smoker = 1 for smokers and 0 for nonsmokers
- Age = age in years
- Education = number of years of schooling
- Income = family income
- Pcigs = price of cigarettes in individual states in 1979

Build a regression model to determine smoking behavior in relation to age, education, family income, and price of cigarettes.

Question: Is linear regression the most appropriate model here?

Dv is not continuous its 0 or 1 so we use logistic regression

What is logistic regression?

Although binary dependent variable models can be estimated by OLS, in which case they are known as linear probability models (LPM), OLS (a method used to estimate regression) is not the preferred method of estimation for such models because of two limitations：

1) The estimated probabilities from LPM do not necessarily lie in the bounds of 0 and 1
2) LPM assumes that the probability of a positive response increases linearly with the level of the explanatory variable, which is counterintuitive.

To address this problem, we introduce ***Logistic regression*** which is a nonlinear regression model specifically designed for binary dependent variables.

What is logistic regression?

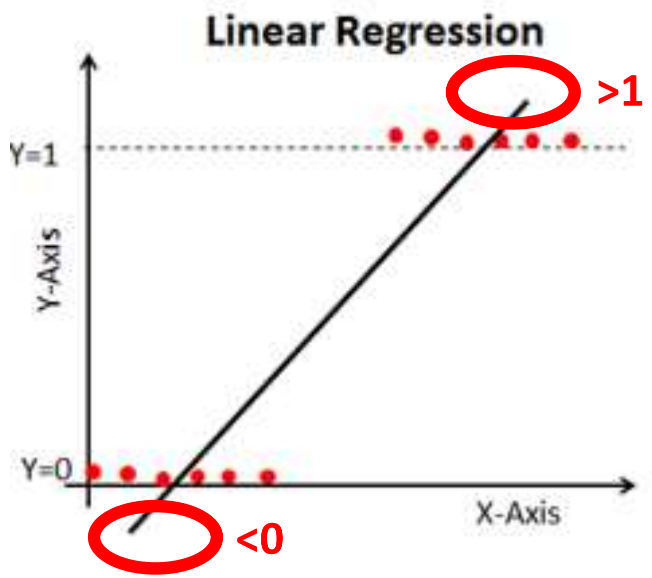
Logistic Regression

- Basic form (this is our focus in this course):
  - Binary logistic regression: the dependent variable is a dummy variable: coded 0 or 1
    - E.g., 1 (leave) or 0 (not leave), 1 (die) or 0 (not die), 1 (pass) or 0 (not pass)
- Extended forms:
  - Multinomial logistic regression: the dependent variable has categorical outputs
    - E.g., determining the probability an image contains a car, a motorcycle, or a bicycle, etc...
  - Ordinal logistic regression: the dependent variable has ordered categorical outputs
    - E.g., determining the probability a student gets an A, B, C or a lower grade in one course etc...

Why logistic regression?

The major problems with using linear model to estimate class probability:

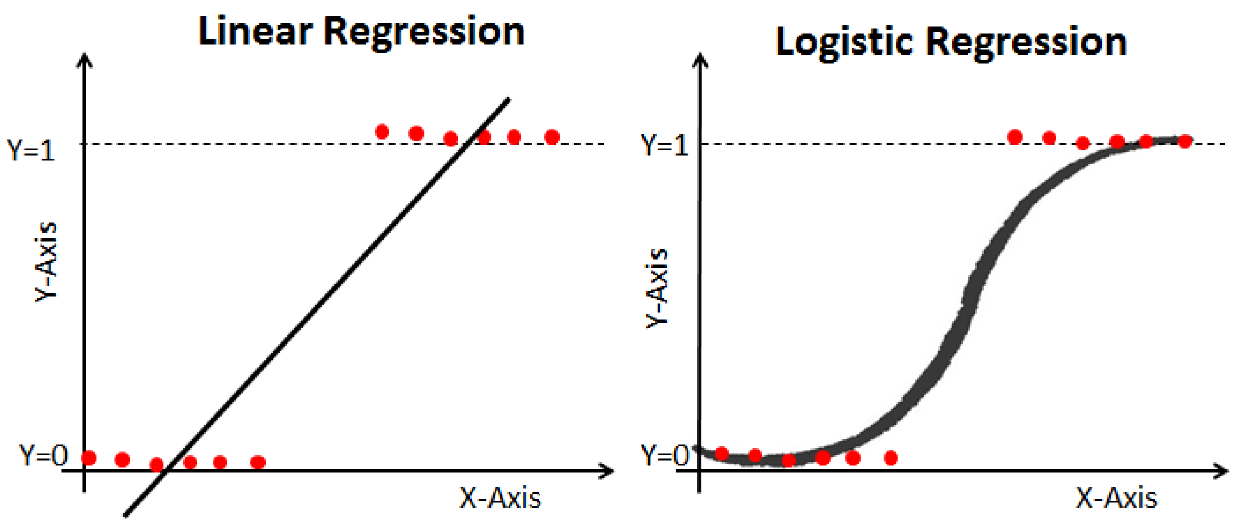
- Violate some assumptions of linear regression (e.g., linear correlation)
- Predicted value of y ranges from $-\infty$ to ∞, but a probability should range from zero to one.



Why logistic regression?

The benefit of using logistic models to estimate class probability:

- Based on quite different assumptions from those of linear regression
- Logistic function is mathematically correct. Linear regression can only give an approximation of the truth
- Predicted value of y ranges from zero to one



Transformation: Logistic to Linear

- To better understand logistic regression, we will discuss another representation of the likelihood of an event: **Odds ratio**

$$\textbf{Odds=}\frac{P_i}{1-P_i}$$

  – The odds of an event is the ratio of the probability of the event occurring to the probability of the event not occurring

Probability	Corresponding odds
0.5	50:50 or 1
0.9	90:10 or 9
0.999	999:1 or 999

Transformation: Logistic to Linear

- Then we take the logarithm of the odds (called the "log-odds")

$$\text{Log-odds} = \ln\left(\frac{P_i}{1-P_i}\right)$$

  - For any number in the range 0 to ∞ its log will be between –∞ to ∞

Probability	**Odds**	**Log-odds**
0.5	50:50 or 1	0
0.9	90:10 or 9	2.19
0.999	999:1 or 999	6.9

Transformation: Logistic to Linear

The "logistic" model:

$$\ln[p/(1-p)] = \beta_0 + \beta_1 X$$

- It models the logit-transformed probability (log odds) as a linear relationship with the predictor variables.
- In another word, the logit model assumes that the log of the odds ratio is linearly related to X.
- p is the probability that the event y occurs: [range=0 to 1]
- p/(1-p) is the "odds" that the event y occurs: [range=0 to ∞]
- ln[p/(1-p)]: logit (log of the odds ratio): [range=-∞ to +∞]
- Interpretation of coefficient β_0: The estimated log odds of the event when all x=0
- Interpretation of coefficient β_1: The estimated change in log of odds when there is a one-unit increase in x
- How to translate log-odds into probability?

$$odds = e^{\beta_0 + \beta_1 X} \Longrightarrow p(x) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}} \Longrightarrow p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Parameter Estimation in R

Consider the following dataset from a study of risk factors associated with low birthweight described in Hosmer, Lemeshow, and Sturdivant (2013, 24).

variable name	storage type	display format	value label	variable label
id	int	%8.0g		identification code
low	byte	%8.0g		birthweight<2500g
age	byte	%8.0g		age of mother
lwt	int	%8.0g		weight at last menstrual period
race	byte	%8.0g	race	race
smoke	byte	%9.0g	smoke	smoked during pregnancy
ptl	byte	%8.0g		premature labor history (count)
ht	byte	%8.0g		has history of hypertension
ui	byte	%8.0g		presence, uterine irritability
ftv	byte	%8.0g		number of visits to physician during 1st trimester
bwt	int	%8.0g		birthweight (grams)

Parameter Estimation in R

Dependent variable **Independent variables**

```
> logitMod <- glm(low ~ age + lwt + smoke + ptl + ht + ui, data=lbw, family=binomial(link="logit"))
> summary(logitMod)

Call:
glm(formula = low ~ age + lwt + smoke + ptl + ht + ui, family = binomial(link = "logit"),
    data = lbw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0763  -0.8085  -0.6247   1.0281   2.0218

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.378960   1.088892   1.266  0.20537
age         -0.042254   0.034584  -1.222  0.22178
lwt         -0.014288   0.006652  -2.148  0.03172 *
smoke        0.550631   0.343629   1.602  0.10907
ptl          0.593255   0.348419   1.703  0.08862 .
ht           1.862491   0.686229   2.714  0.00665 **
ui           0.736790   0.456488   1.614  0.10652
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 208.79  on 182  degrees of freedom
AIC: 222.79

Number of Fisher Scoring iterations: 4
```

Name of the data file

Logistic regression model

β_0

β_1

P value for significance test

How to interpret the parameters?

-Holding other variables constant, if lwt increases by one unit, the average log odds in favor of having low birthweight goes down by 0.014.

An Example

Attitude towards women's roles

In 1979, women's and men's attitudes toward women's familial roles were examined using the questionnaire \Do women belong in the home?" in a cross-sectional survey of US adults, cross-classified by respondent's gender and their formal education measured in years. What model should we use to examine the determinants of people’s attitudes toward women's familial roles?

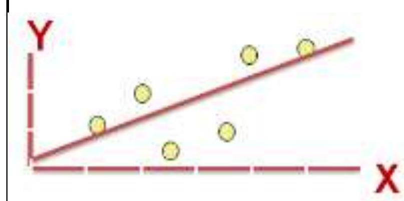
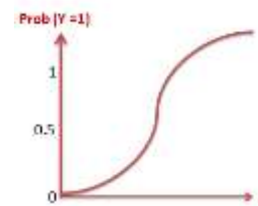
Key variables:

- yes: the number of people who responds yes
- no: the number of people who responds no
- gender: the gender, treated as categorical variable
- educ: years of education, treated as continuous

A sample of the data set:

```
> women[1:4,]
  yes  no gender educ
1   4   2      M    0
2   4   2      F    0
3   2   0      M    1
4   1   0      F    1
```

Logistic Regression versus Linear Regression

	Linear Regression	**Logistic Regression**
Dependent variable	Continuous	Binary or category
Parameter estimation	Ordinary least squares (OLS)	Maximum Likelihood Estimation (MLE)
Equation	y = $\beta_0 + \beta_1 X + \varepsilon$	$p(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$
Curve		
Parameter interpretation	Change in y caused by a one-unit change in x	Change in log odds caused by a one-unit change in x
Interpretation of interaction terms	Straight forward	Complicated

Logistic Regression versus Linear Regression

- Examine whether online reviews influence product sales
- Examine factors associated with the valance of ratings
- Examine risk factors associated with low ratings (<3stars)
- Examine how monetary incentives influence review numbers
- Build an algorithm to improve online ratings of a restaurant

Summary

- Why logistic regression?
- What is odds ratio?
- What is the basic form of logistic regression?
- How to estimate the parameters of logistic regression?
- How to interpret the parameters of logistic regression?
- How to choose between logistic and linear regression?