IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION - I

Dr A. RAMESH

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Building Logistic regression Model
- Python Demo on Logistic Regression

Application

$y = a + b_{p0} x_1 + b_p x_2$

$(0,1)$

- In many regression applications the dependent variable may only assume two discrete values.
- For instance, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card or not
- The dependent variable can be coded as *y* =1 if the bank approves the request for a credit card and *y* = 0 if the bank rejects the request for a credit card.
- Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

Example

- Let us consider an application of logistic regression involving a direct mail promotion being used by Simmons Stores.
- Simmons owns and operates a national chain of women's apparel stores.
- Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a $50 discount on purchases of $200 or more.
- The catalogs are expensive and Simmons would like to send them to only those customers who have the highest probability of using the coupon.

Sources: Statistics for Business and Economics,11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

Variables

- Management thinks that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

Data (10 customer out of 100)

x_1 x_2 Y

Customer	Spending	Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

Explanation of Variables

- The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not.
- In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

Logistic Regression Equation

- If the two values of the dependent variable *y* are coded as 0 or 1, the value of $E(y)$ in equation given below provides the *probability* that $y = 1$ given a particular set of values for the independent variables $x_1, x_2, \ldots, x_p$.

LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

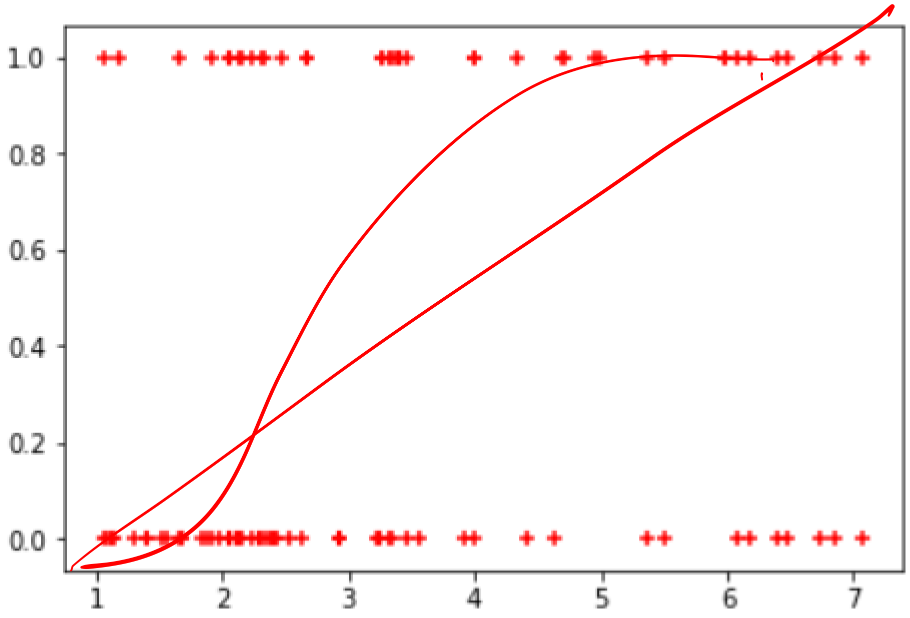
Logistic Regression Equation

- Because of the interpretation of *E*(*y*) as a probability, the **logistic regression equation** is often written as follows

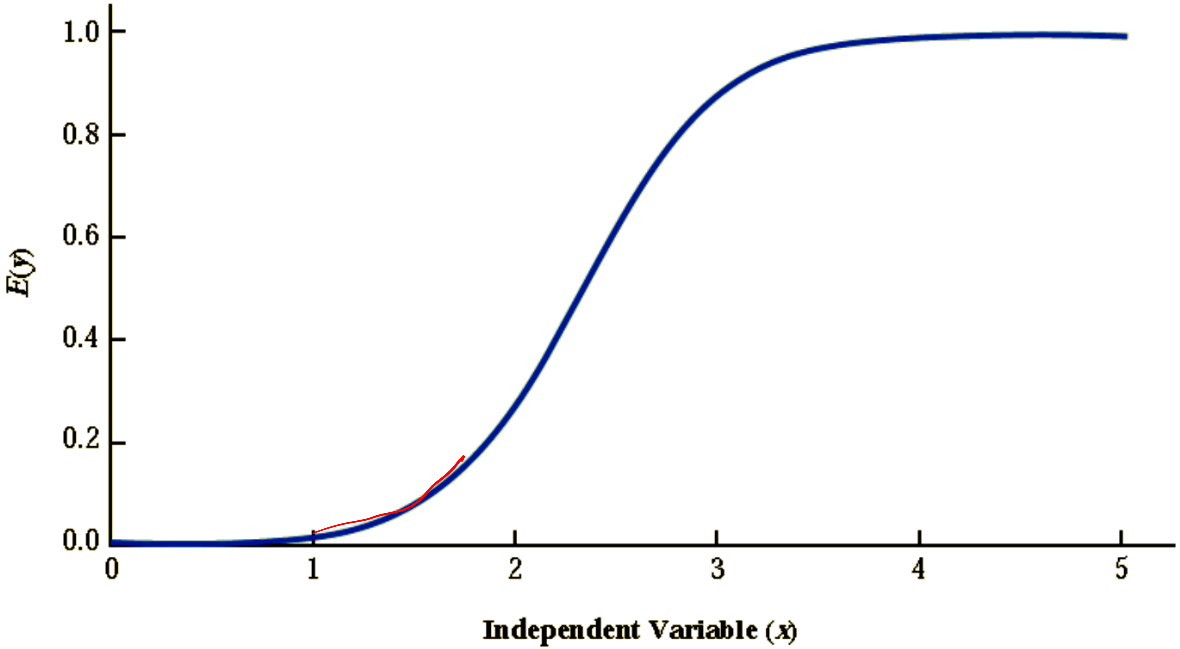
$$E(y) = P(y = 1 | x_1, x_2, \ldots, x_p)$$

```
In [15]: plt.scatter(df.Spending,df.Coupon,marker='+',color= 'red')

Out[15]: <matplotlib.collections.PathCollection at 0x2a1b5b73c50>
```



Logistic regression equation for β_0 and β_1



Logistic regression equation for β_0 and β_1

- Note that the graph is S-shaped.
- The value of $E(y)$ ranges from 0 to 1, with the value of $E(y)$ gradually approaching 1 as the value of x becomes larger and the value of $E(y)$ approaching 0 as the value of x becomes smaller.
- Note also that the values of $E(y)$, representing probability, increase fairly rapidly as x increases from 2 to 3.
- The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (slide no.11) ideally suited to model the probability the dependent variable is equal to 1.

Estimating the Logistic Regression Equation

- In simple linear and multiple regression the least squares method is used to compute $b_0, b_1, \ldots, b_p$ as estimates of the model parameters ($\beta_0, \beta_1, \ldots, \beta_p$).
- The nonlinear form of the logistic regression equation makes the method of computing estimates more complex MLE
- We will use computer software to provide the estimates.
- The **estimated logistic regression equation** is

ESTIMATED LOGISTIC REGRESSION EQUATION

$$\hat{y} = \text{estimate of } P(y = 1 \mid x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}$$

Here, y hat provides an estimate of the probability that $y = 1$, given a particular set of values for the independent variables.

Python Code for Logistic Regression

```
In [12]: x = df[['Card','Spending']]
         y = df['Coupon']

         import statsmodels.api as sm
         x1= sm.add_constant(x)
         logit_model=sm.Logit(y,x1)
         result=logit_model.fit()
         print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604869
         Iterations 5
                        Results: Logit
================================================================
Model:              Logit            No. Iterations:   5.0000
Dependent Variable: Coupon           Pseudo R-squared: 0.101
Date:               2019-09-11 12:54 AIC:              126.9739
No. Observations:   100              BIC:              134.7894
Df Model:           2                Log-Likelihood:   -60.487
Df Residuals:       97               LL-Null:          -67.301
Converged:          1.0000           Scale:            1.0000
-----------------------------------------------------------------
            Coef.   Std.Err.    z     P>|z|    [0.025   0.975]
-----------------------------------------------------------------
const     -2.1464    0.5772 -3.7183  0.0002  -3.2778  -1.0150
Card       1.0987    0.4447  2.4707  0.0135   0.2271   1.9703
Spending   0.3416    0.1287  2.6551  0.0079   0.0894   0.5938
================================================================
```

wald

$x_2 \to$

x_1

G

Variables

$$y = \begin{cases} 0 \text{ if the customer did not use the coupon} \\ 1 \text{ if the customer used the coupon} \end{cases}$$

$$x_1 = \text{annual spending at Simmons Stores (\$1000s)}$$

$$x_2 = \begin{cases} 0 \text{ if the customer does not have a Simmons credit card} \\ 1 \text{ if the customer has a Simmons credit card} \end{cases}$$

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2}} = \frac{e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}{1 + e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}$$

Managerial Use

- $P(\ y = 1/x_1 = 2,\ x_2 = 0) = .1880$

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(0)}}{1+e^{-2.14637+0.341643(2)+1.09873(0)}} = \frac{e^{-1.4631}}{1+e^{-1.4631}} = \frac{.2315}{1.2315} = 0.1880$$

- $P(\ y = 1/x_1 = 2,/\ x_2 = 1) = .4099$

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(1)}}{1+e^{-2.14637+0.341643(2)+1.09873(1)}} = \frac{e^{-0.3644}}{1+e^{-0.3644}} = \frac{.6946}{1.6946} = 0.4099$$

- Probabilities indicate that for customers with annual spending of $2000 the presence of a Simmons credit card increases the probability of using the coupon

Managerial Use

- It appears that the probability of using the coupon is much higher for customers with a Simmons credit card.

		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Testing for Significance

t, F

$$H_0: \beta_1 = \beta_2 = 0$$

H_a: One or both of the parameters is not equal to zero

G Statistics

- The test for overall significance is based upon the value of a *G* test statistic.

  F

- If the null hypothesis is true, the sampling distribution of *G* follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model.

```
In [12]: x = df[['Card','Spending']]
         y = df['Coupon']

         import statsmodels.api as sm
         x1= sm.add_constant(x)
         logit_model=sm.Logit(y,x1)
         result=logit_model.fit()
         print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604869
         Iterations 5
                        Results: Logit
==============================================================
Model:              Logit            No. Iterations:    5.0000
Dependent Variable: Coupon           Pseudo R-squared:  0.101
Date:               2019-09-11 12:54 AIC:               126.9739
No. Observations:   100              BIC:               134.7894
Df Model:           2                Log-Likelihood:    -60.487
Df Residuals:       97               LL-Null:           -67.301
Converged:          1.0000           Scale:             1.0000
---------------------------------------------------------------
            Coef.   Std.Err.    z     P>|z|   [0.025   0.975]
---------------------------------------------------------------
const      -2.1464   0.5772  -3.7183  0.0002  -3.2778  -1.0150
Card        1.0987   0.4447   2.4707  0.0135   0.2271   1.9703
Spending    0.3416   0.1287   2.6551  0.0079   0.0894   0.5938
==============================================================
```

G Statistics

$$G = -2\ln\left[\frac{\text{(likelihood without the variable)}}{\text{(likelihood with the variable)}}\right].$$

$$G = 2(-60.487 - (-67.301) = \boxed{13.628}$$

- The value of G is 13.628, its degrees of freedom are 2,and its p-value is 0.001.
- Thus, at any level of significance α >= .001, we would reject the null hypothesis and conclude that the overall model is significant.

Thank You