IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Linear Regression Model Vs Logistic Regression Model

Dr. A. Ramesh
DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Comparison of Linear Regression model and Logistic regression model

Estimating the relationship

Linear regression model

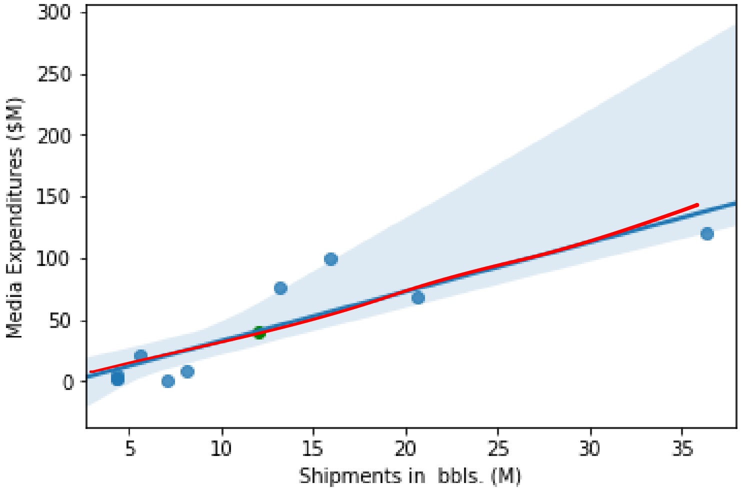
- $Y_1 = X_1+X_2+\ldots+X_n$
- Where ,
  - Y_1 = continuous data
  - Independent variables = nonmetric and metric

Logistic regression model

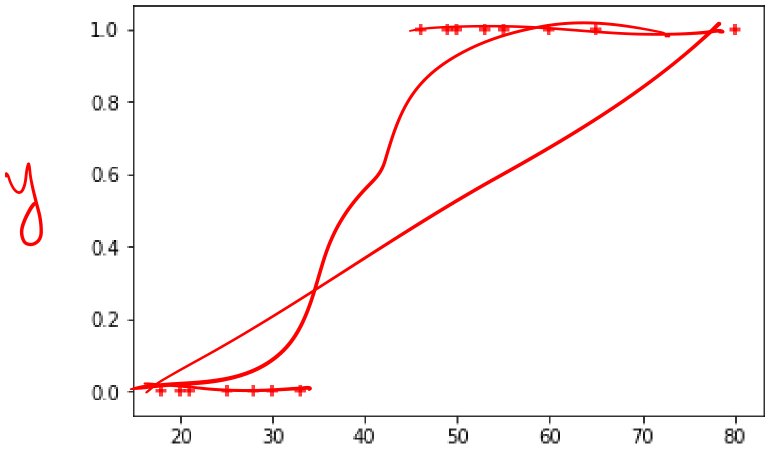
- $Y_1 = X_1+X_2+\ldots+X_n$
- Where ,
  - Y_1 = Binary nonmetric
  - Independent variables = nonmetric and metric

Graphical representation

- **Linear regression**



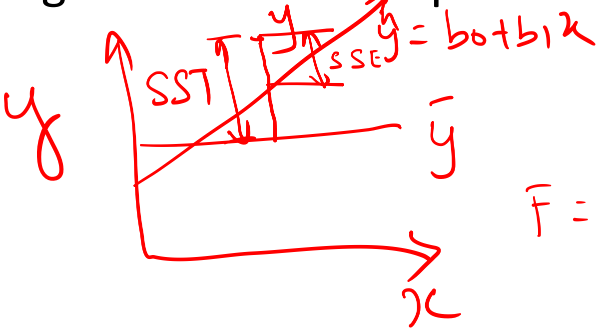
- **Logistic regression**



Correspondence of Primary Elements of Model Fit

Linear Regression

- Total sum of squares SST
- Error sum of squares SSE
- F test of model fit
- Coefficient of determination (R^2)
- Regression sum of squares SSR



$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \qquad R^2 = \frac{SSR}{SST}$$

Logistic Regression

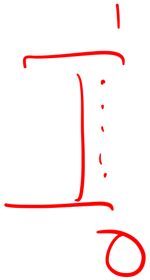
- -2LL of base model
- -2LL of proposed model
- Chi-square test of -2LL difference (G)
- Pseudo R^2 measures
- Difference of -2LL for base and proposed models

Objective of logistic regression

- Logistic regression is identical to discriminant analysis in terms of the basic objectives it can address
- Logistic regression is best suited to address two research objectives:
  - Identifying the independent variables that impact group membership in the dependent variable
  - Establishing a classification system based on the logistic model for determining group membership

The fundamental difference

- Logistic regression differs from linear regression, in being specifically designed to predict the probability of an event occurring (ie., the probability of an observation being in the group coded 1)
- Although probability values are metric measures, there are fundamental differences between linear regression and logistic regression



Log likelihood

- Measure used in logistic regression to represent the lack of predictive fit
- Even though this method does not use the least squares procedure in model estimation, as is done in linear regression, the likelihood value is similar to the sum of squared error in regression analysis

SSE

Logistic vs discriminant

- Logistic regression may be preferred for two reasons
- First, discriminant analysis relies on strictly meeting the assumptions of
  - Multivariate normality and equal variance
  - Covariance matrices across groups
  - Assumptions that are not met in many situations
- Logistic regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations

Logistic vs discriminant

- Second, even if the assumptions are met, many researchers prefer logistic regression because it is similar to multiple regression
- It has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics
- Logistic regression is equivalent to two-group discriminant analysis and may be more suitable in many situations

Logistic vs discriminant : Sample size

- One factor that distinguishes logistic regression from the other techniques is its use of maximum likelihood (MLE) as the estimation technique
- MLE requires larger samples such that, all things being equal, logistic regression will require a larger sample size than multiple regression
- As for discriminant analysis, there are considerations on the minimum group size as well

Logistic vs discriminant : Sample size

- The recommended sample size for each group is at least 10 observations per estimated parameter

3 –

- This is much greater than multiple regression, which had a minimum of five observations per parameter, and that was for the overall sample, not the sample size for each group, as seen with logistic regression

Regression : 1 – 5

L. R : 1:10

Determination of coefficients

Linear regression

- R^2
  - r^2 = SSR/SST

  where:

  SSR = sum of squares due to regression

  SST = total sum of squares

Logistic regression

$$R^2 Logit = \frac{-2LLnull - (-2LLmodel)}{-2LLnull}$$

Where:

LL = Loglikelihood

$-2LL_{null}$ = -2LL of base model

$-2LL_{model}$ = -2LL of proposed model

Determination of coefficients

- **Linear regression**

OLS Regression Results

Dep. Variable:	Media Expenditures ($M)	**R-squared:**	0.783
Model:	OLS	**Adj. R-squared:**	0.756
Method:	Least Squares	**F-statistic:**	28.93
Date:	Wed, 10 Oct 2018	**Prob (F-statistic):**	0.000663
Time:	18:16:26	**Log-Likelihood:**	-44.355
No. Observations:	10	**AIC:**	92.71
Df Residuals:	8	**BIC:**	93.32
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.6277	11.485	-0.664	0.525	-34.112	18.857
Shipments in bbls. (M)	4.0065	0.745	5.378	0.001	2.289	5.724

Omnibus:	4.361	**Durbin-Watson:**	1.473
Prob(Omnibus):	0.113	**Jarque-Bera (JB):**	2.129
Skew:	1.129	**Prob(JB):**	0.345
Kurtosis:	2.925	**Cond. No.**	24.6

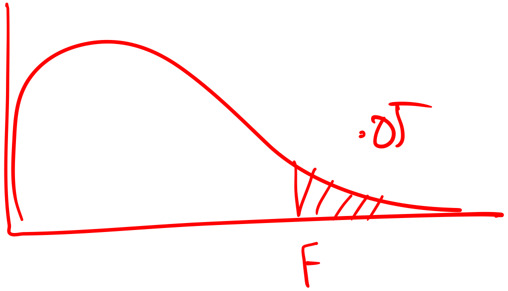
- **Logistic regression**

```
                           Results: Logit
==============================================================
Model:               Logit              Pseudo R-squared: 0.192
Dependent Variable:  Coupon             AIC:              12.0864
Date:                2019-09-08 11:07   BIC:              12.6916
No. Observations:    10                 Log-Likelihood:   -4.0432
Df Model:            1                  LL-Null:          -5.0040
Df Residuals:        8                  LLR p-value:      0.16568
Converged:           1.0000             Scale:            1.0000
No. Iterations:      7.0000
--------------------------------------------------------------
            Coef.   Std.Err.     z     P>|z|   [0.025   0.975]
--------------------------------------------------------------
Spending   -0.6318   0.4566  -1.3838  0.1664  -1.5267  0.2630
Card       -0.0029   1.4887  -0.0020  0.9984  -2.9207  2.9149
==============================================================
```

Testing for overall significance

Linear regression

- F-test of model fit
- F = MSR/MSE

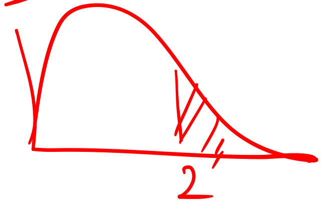


- **Logistic Regression**
- G-test of model fit

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

$$= -2[-5 - (-4)]$$

$$= -2[-1] = 2$$



Testing for overall significance

- **Linear regression**

OLS Regression Results

Dep. Variable:	Media Expenditures ($M)	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	28.93
Date:	Wed, 10 Oct 2018	Prob (F-statistic):	0.000663
Time:	18:16:26	Log-Likelihood:	-44.355
No. Observations:	10	AIC:	92.71
Df Residuals:	8	BIC:	93.32
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.6277	11.485	-0.664	0.525	-34.112	18.857
Shipments in bbls. (M)	4.0065	0.745	5.378	0.001	2.289	5.724

Omnibus:	4.361	Durbin-Watson:	1.473
Prob(Omnibus):	0.113	Jarque-Bera (JB):	2.129
Skew:	1.129	Prob(JB):	0.345
Kurtosis:	2.925	Cond. No.	24.6

- **Logistic regression**

```
                        Results: Logit
===============================================================
Model:              Logit            Pseudo R-squared: 0.192
Dependent Variable: Coupon           AIC:              12.0864
Date:               2019-09-08 11:07 BIC:              12.6916
No. Observations:   10               Log-Likelihood:   -4.0432
Df Model:           1                LL-Null:          -5.0040
Df Residuals:       8                LLR p-value:      0.16568
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
---------------------------------------------------------------
             Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
---------------------------------------------------------------
Spending    -0.6318   0.4566 -1.3838 0.1664 -1.5267 0.2630
Card        -0.0029   1.4887 -0.0020 0.9984 -2.9207 2.9149
===============================================================
```

Testing for significance

Linear regression

- t-test

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$t = \frac{b_1 - \beta_1}{S_b}$$

where: $S_b = \frac{S_e}{\sqrt{SS_{XX}}}$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{XX} = \sum X^2 - \frac{(\sum X)^2}{n}$$

β_1 = the hypothesized slope

$df = n - 2$

Logistic regression

- Wald-test

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} = \frac{-0.0029}{1.4882}$$

$=$

Testing for significance

- **Linear regression**

OLS Regression Results

Dep. Variable:	Media Expenditures ($M)	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	28.93
Date:	Wed, 10 Oct 2018	Prob (F-statistic):	0.000663
Time:	18:16:26	Log-Likelihood:	-44.355
No. Observations:	10	AIC:	92.71
Df Residuals:	8	BIC:	93.32
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.6277	11.485	-0.664	0.525	-34.112	18.857
Shipments in bbls. (M)	4.0065	0.745	5.378	0.001	2.289	5.724

Omnibus:	4.361	Durbin-Watson:	1.473
Prob(Omnibus):	0.113	Jarque-Bera (JB):	2.129
Skew:	1.129	Prob(JB):	0.345
Kurtosis:	2.925	Cond. No.	24.6

- **Logistic regression**

```
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.192
Dependent Variable: Coupon           AIC:              12.0864
Date:               2019-09-08 11:07 BIC:              12.6916
No. Observations:   10               Log-Likelihood:   -4.0432
Df Model:           1                LL-Null:          -5.0040
Df Residuals:       8                LLR p-value:      0.16568
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
              Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
-----------------------------------------------------------------
Spending     -0.6318    0.4566  -1.3838  0.1664  -1.5267  0.2630
Card         -0.0029    1.4887  -0.0020  0.9984  -2.9207  2.9149
=================================================================
```

Model Estimation fit

- The basic measure of how well the maximum likelihood estimation procedure fits is the likelihood value, similar to the sums of squares values used in multiple regression SSE
- Logistic regression measures model estimation fit with the value of -2 times the log of the likelihood value, referred to as -2LL or -2 log likelihood
- The minimum value for -2LL is 0, which corresponds to a perfect fit (likelihood = 1 and -2LL is then 0)

Model Estimation fit

- The lower the -2LL value, the better the fit of the model
- The -2LL value can be used to compare equations for the change in fit

Between Model Comparison

- The likelihood value can be compared between equations to assess the difference in predictive fit from one equation to another, with statistical tests for the significance of these differences
- The basic approach follows three steps:

Step 1 : Estimate a null model

- The first step is to calculate a null model, which acts as the baseline for making comparisons of improvement in model fit.
- The most common null model is one without any independent variables, which is similar to calculating the total sum of squares using only the mean in linear regression. $\bar{y}$
- The logic behind this form of null model is that it can act as a baseline against which any model containing independent variables can be compared.

Step 2: Estimate the proposed model

- This model contains the independent variables to be included in the logistic regression model.
- This model fit will improve from the null model and result in a lower -2lL value.
- Any number of proposed models can be estimated

Step 3: **Assess -2LL difference**:

- The final step is to assess the statistical significance of the -2LL value between the two models (null model versus proposed model).
- If the statistical tests support significant differences, then we can state that the set of independent variable(s) in the proposed model is significant in improving model estimation fit.

Between model comparison

Linear regression

- SSE
- $= \sum(y_i - \hat{y}_i)^2$

Logistic Regression

- -2LL of proposed model

Between model comparison

Linear Regression

- SSR = = $\sum(y_i \_ \bar{y}_i)^2$
- SST-SSE

Logistic regression

- Difference between log likelihood
- = $2LL_{null} - (2LL_{model})$

Normality of Residual (Error)

Linear regression

- Normally distributed
- Linear regression assumes that residuals are approximately equal for all predicted dependent variable values

Logistic regression

- Binomially distributed
- Logistic regression does not need residuals to be equal for each level of the predicted dependent variable values

Estimation Methods

- Linear regression is based on least square estimation
- Regression coefficients should be chosen in such a way that it minimizes the sum of the squared distances of each observed response to its fitted value

S

- logistic regression is based on Maximum Likelihood Estimation
- Coefficients should be chosen in such a way that it maximizes the Probability of Y given X (likelihood)
- With MLE, the computer uses different "iterations" in which it tries different solutions until it gets the maximum likelihood estimates

Interpretation

Coefficients of linear regression is interpreted as:

- Keeping all other independent variables constant, how much the dependent variable is expected to increase/decrease with an unit increase in the independent variable

In logistic regression, we interpret odd ratios as:

- The effect of a one unit of change in X in the predicted odds ratio with the other variables in the model held constant

3

THANK YOU