



Logistic Regression

DR. HEMAL PANDYA

LOGISTIC REGRESSION

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy).

The prediction is based on the use of one or several predictors (numerical and categorical).

A logistic regression produces a logistic curve, which is limited to values between 0 and 1.

Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability.

Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

WHY USE LOGISTIC REGRESSION?

- There are many important research topics for which the dependent variable is "limited."
- For example: whether or not a person smokes, or drinks, or skips class, or takes advanced mathematics. For these the outcome is not continuous or distributed normally.
 - Example: Are mother's who have high school education less likely to have children with IEP's (individualized plans, indicating cognitive or emotional disabilities
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not smoke) or 1(did smoke)

ASSUMPTIONS UNDERLYING LOGISTIC REGRESSION

- Assumptions Not Required:
- First, logistic regression **does not require a linear relationship** between the dependent and independent variables.
- Second, **the error terms (residuals) do not need to be normally distributed.**
- Third, **homoscedasticity is not required.**
- Finally, the **dependent variable in logistic regression is not measured on an interval or ratio scale.** However, some other assumptions still apply.

Assumptions Required:

- First, Binary Logistic Regression **requires the dependent variable to be binary and ordinal** **logistic regression requires the dependent variable to be ordinal** i.e.; the dependent variable to be ordered, for example *Too Little, About Right, Too Much..*
- Second, logistic regression **requires the observations to be independent of each other.** In other words, the observations should not come from repeated measurements or matched data.
- Third, logistic regression **requires there to be little or no multicollinearity** among the independent variables. This means that the independent variables should not be too highly correlated with each other.
- Fourth, logistic regression assumes linearity of independent variables and log odds. although this analysis **does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.**
- Finally, logistic regression typically **requires a large sample size.** A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of $500 = (10 * 5 / .10)$.

Logistic Regression vs. Linear Discriminant Analysis

Like Discriminant Analysis , Logistic Regression is used to distinguish between two or more group while either technique can be used for such applications.

Logistic Regression is generally preferred when there are only two categories of the dependent variable, while Discriminant Analysis is the method of choice when there are more than two. Typical application areas are cases where one wishes **to predict the likelihood of an entity belonging to one group or another** such as in response to a marketing effort (likelihood of purchase /non-purchase), creditworthiness (high/low risk of default) insurance (high/low risk of accident claim) , medicine (high/low risk of some ailment like heart attack), sports (likelihood of victory /loss)

Logistic Regression vs. Linear Discriminant Analysis

In terms of predicting power, there is a debate over which technique perform better and there is no clear winner. As stated before the general view is that **Logistic Regression is preferred for binormal dependent variables, while discriminant is better when there are more than 2 values of the dependent.**

However, there is a practical advantage to Logistic Regression.

An organization such as a credit card lender typically has several models of this type (with binary departments) to predict various aspects of a customers' behaviour. For instance, each month it may evaluate all its customers to see how likely they are to become delinquent in the near future.

They may have another model to predict the likelihood of a customer leaving them to switch to a competitor, and yet another to predict the ones most likely to purchase additional financial products.

Logistic Regression vs. Linear Discriminant Analysis

If each of these models is built using Logistic regression, then all the different scores that a customer receives on these models can be interpreted the same way. For instance, a score of 0.75 on any of the models means the same thing – it implies a 75% chance that the customer will belong to the group with value 1 (as opposed to 0) for that model. From a manager's both yield the same score can mean different things. A score of 0.75 may be a “good” score on one model and “bad” one on another.

Logistic Regression vs. Linear Discriminant Analysis

Discriminant function analysis is very similar to logistic regression, and both can be used to answer the same research questions.

Logistic regression does not have as many assumptions and restrictions as discriminant analysis. However, when Discriminant Analysis' assumptions are met, it is more powerful than Logistic Regression.

Unlike logistic regression, Discriminant Analysis can be used with small sample sizes. It has been shown that when sample sizes are equal, and homogeneity of variance/covariance holds, discriminant analysis is more accurate.

With all this being considered, Logistic Regression has become the common choice, since the assumptions of Discriminant Analysis are rarely met.

Logistic Regression vs. Linear Discriminant Analysis

Mathematically too, Logistic Regression is less encumbered by the assumptions of Discriminant Analysis. The independent variables in Logistic Regression may be anything from Normal to Ratio scaled and there are no distribution assumptions.

Logistic regression answers the same questions as discriminant analysis. It is often preferred to discriminate analysis as it is more flexible in its assumptions and types of data that can be analyzed.

“Logistic Regression can handle both categorical and continuous variables, and the predictors do not have to be normally distributed, linearly related, or of equal variance within each group (Tabachnick and Fidell 1996)”.

TERMS

Logit: The **logit** function or the **log-odds** is the logarithm of the odds $p/(1 - p)$ where p is the probability of success.

MLE: Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a **maximum likelihood estimate**, which is also abbreviated as MLE.

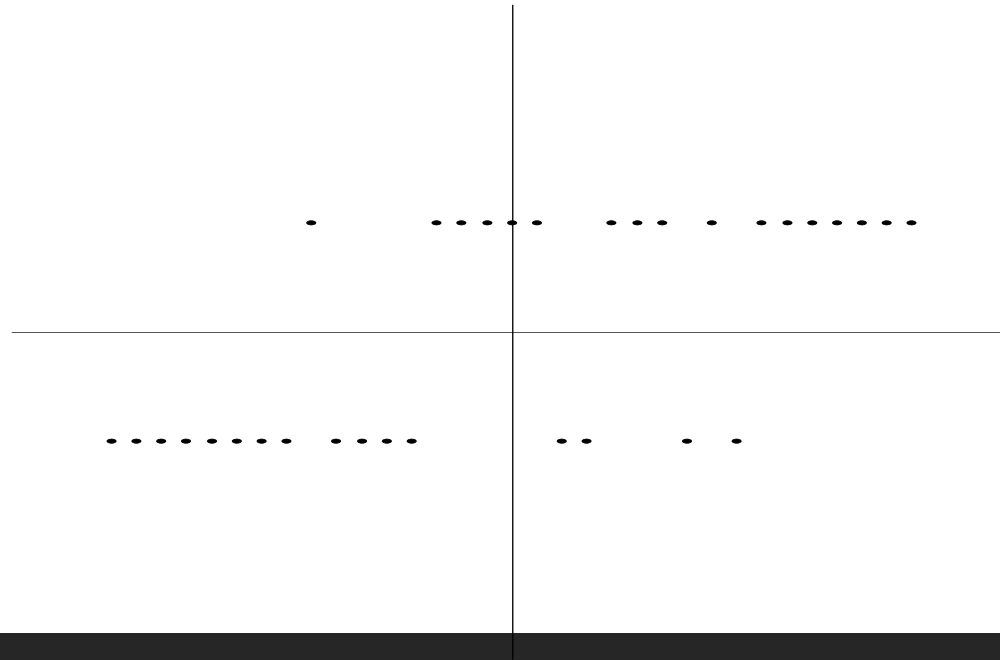
Types of Logistic Regression

- Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.
- Multinomial Logistic Regression: The target variable has three or more nominal categories such as predicting the type of Wine.
- Ordinal Logistic Regression: the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

A PROBLEM WITH LINEAR REGRESSION

However, transforming the independent variables does not remedy all of the potential problems. What if we have a non-normally distributed dependent variable? The following example depicts the problem of fitting a regular regression line to a non-normal dependent variable).

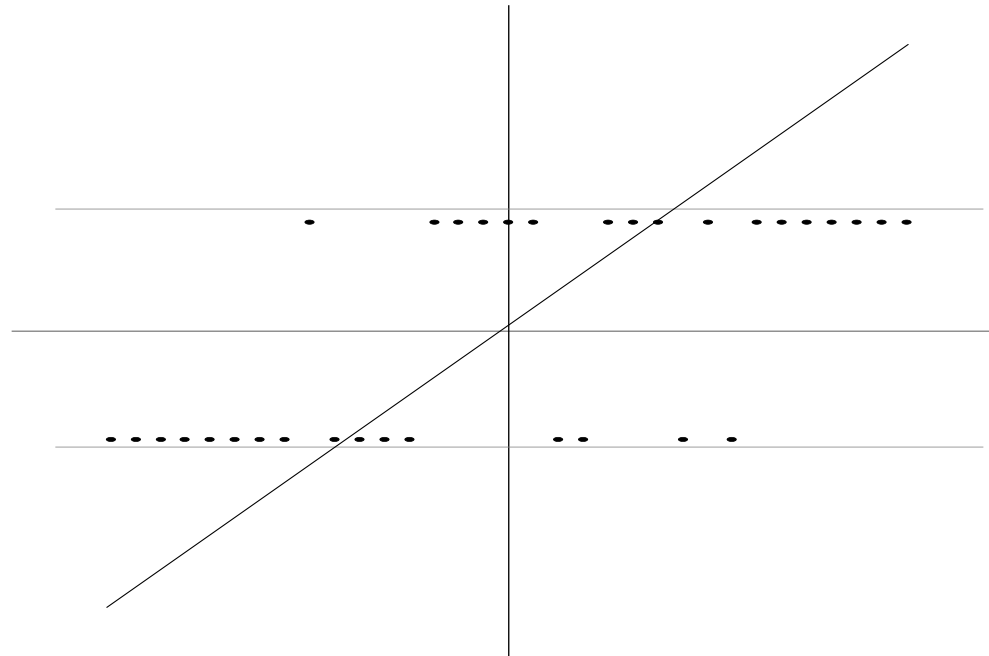
Suppose you have a binary outcome variable. The problem of having a non-continuous dependent variable becomes apparent when you create a scatterplot of the relationship. Here, we see that it is very difficult to decipher a relationship among these variables.



A PROBLEM WITH LINEAR REGRESSION

We could severely simplify the plot by drawing a line between the means for the two dependent variable levels, but this is problematic in two ways: (a) the line seems to oversimplify the relationship and (b) it gives predictions that cannot be observable values of Y for extreme values of X .

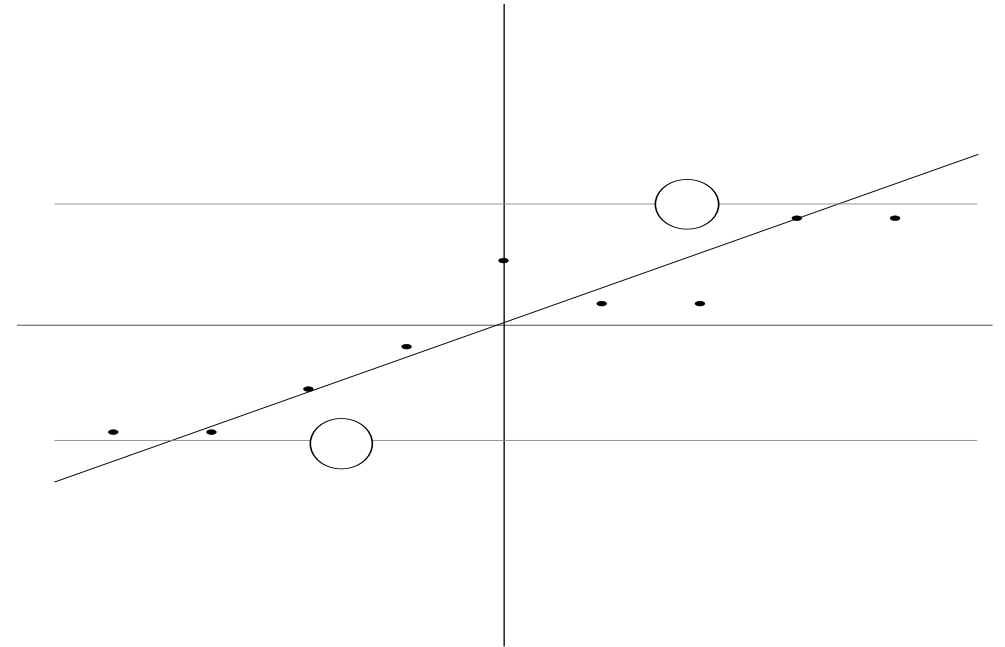
The reason this doesn't work is because the approach is analogous to fitting a linear model to the probability of the event. As you know, probabilities can only take values between 0 and 1. Hence, we need a different approach to ensure that our model is appropriate for the data.



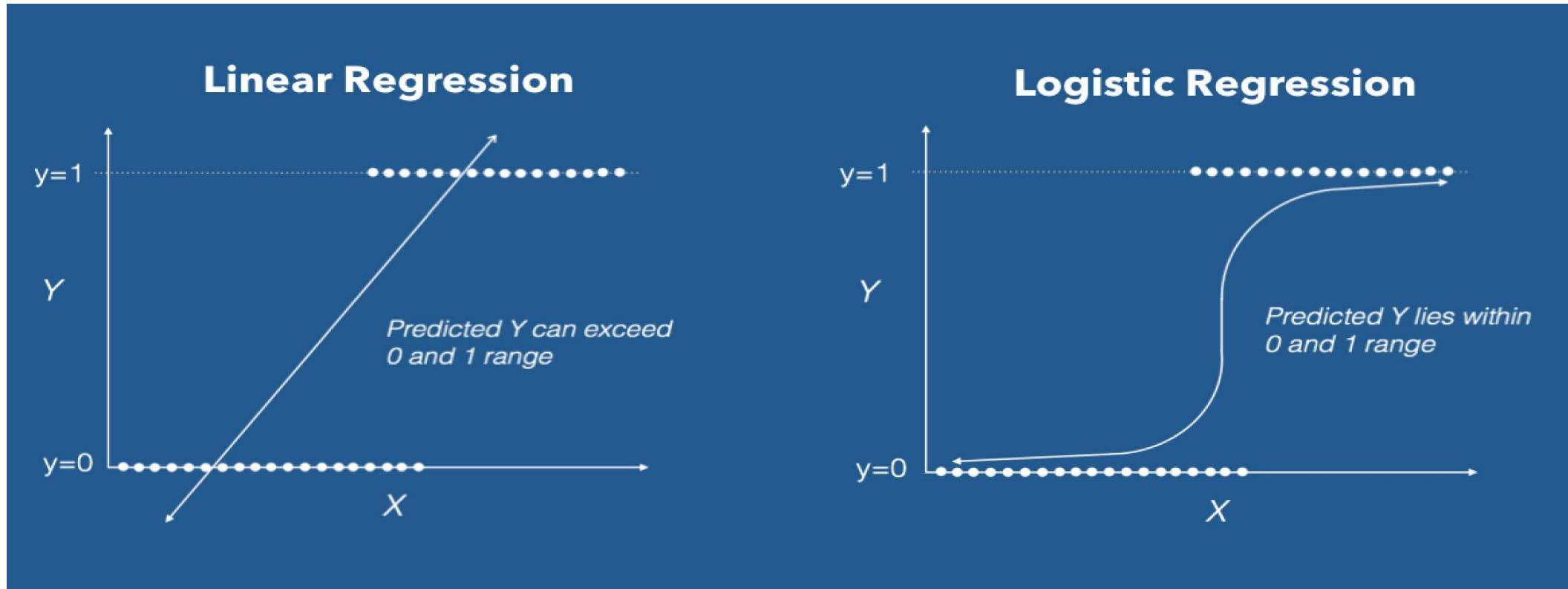
A PROBLEM WITH LINEAR REGRESSION

The mean of a binomial variable coded as (1,0) is a proportion. We could plot conditional probabilities as Y for each level of X. Of course, we could fit a linear model to these conditional probabilities, but (as shown) the linear model does not predict the maximum likelihood estimates for each group (the mean—shown by the circles) and it still produces unobservable predictions for extreme values of the dependent variable.

This plot gives us a better picture of the relationship between X and Y. It is clear that the relationship is non-linear. In fact, the shape of the curve is sigmoid.



LINEAR REGRESSION V/S LOGISTIC REGRESSION



Linear Regression Vs. Logistic Regression

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example of the continuous output is house price and stock price. Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn.

Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

Linear Regression Vs. Logistic Regression

The **MLE** is a "likelihood" maximization method, while **OLS** is a distance-minimizing approximation method.

Maximizing the likelihood function determines the parameters that are most likely to produce the observed data.

From a statistical point of view, MLE sets the mean and variance as parameters in determining the specific parametric values for a given model. This set of parameters can be used for predicting the data needed in a normal distribution.

Ordinary Least squares estimates are computed by fitting a regression line on given data points that has the minimum sum of the squared deviations (least square error). Both are used to estimate the parameters of a linear regression model.

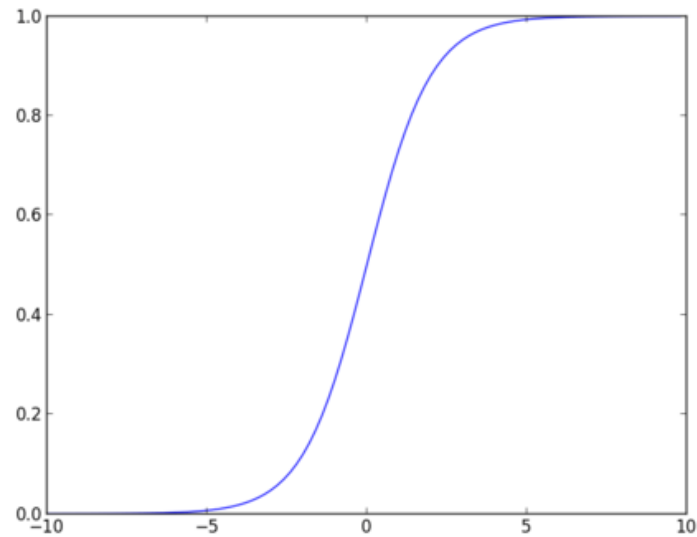
MLE assumes a joint probability mass function, while OLS doesn't require any stochastic assumptions for minimizing distance.

Logistic Regression

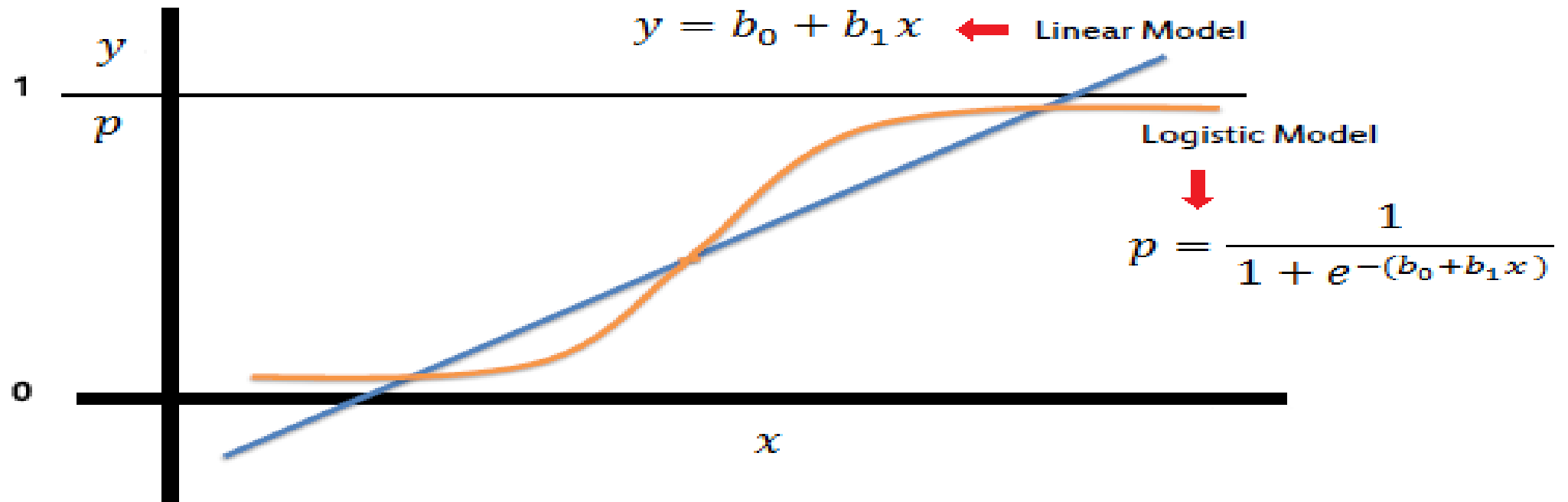
- One commonly used algorithm is Logistic Regression
- Assumes that the dependent (output) variable is binary which is often the case in medical and other studies. (Does person have disease or not, survive or not, accepted or not, etc.)
- Like Quadric, Logistic Regression does a particular non-linear transform on the data after which it just does linear regression on the transformed data
- Logistic regression fits the data with a sigmoidal/logistic curve rather than a line and outputs an approximation of the probability of the output given the input

SIGMOID FUNCTION

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



SIMPLE LOGISTIC REGRESSION



LOGISTIC REGRESSION

In the logistic regression the constant (b_0) moves the curve left and right and the slope (b_1) defines the steepness of the curve.

$$\frac{p}{1-p} = \exp(b_0 + b_1 x)$$

By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

Finally, taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient (b_1) is the amount the logit (log-odds) changes with a one unit change in x .

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

The outcome in logistic regression analysis is often coded as 0 or 1, where 1 indicates that the outcome of interest is present, and 0 indicates that the outcome of interest is absent. If we define p as the probability that the outcome is 1, the multiple logistic regression model can be written as follows:

Multiple Logistic Model

The outcome in logistic regression analysis is often coded as 0 or 1, where 1 indicates that the outcome of interest is present, and 0 indicates that the outcome of interest is absent. Logistic regression can handle any number of numerical and/or categorical variables.

If we define p as the probability that the outcome is 1, the multiple logistic regression model can be written as follows:

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p)}$$

Multiple Logistic Model

$\hat{p}$ is the expected probability that the outcome is present; X_1 through X_p are distinct independent variables; and b_0 through b_p are the regression coefficients. The multiple logistic regression model is sometimes written differently. In the following form, the outcome is the expected log of the odds that the outcome is present,

The Logit is defined as: $\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right)$

The Multiple Logistic Model is defined as under:

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

Sigmoid Function

Sigmoid Function

- The sigmoid function, also called logistic function gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1.
- If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0.
- If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO.
- For example: If the output is 0.75, we can say in terms of probability as: There is a 75 percent chance that patient will suffer from cancer.

Controlling for the Number of Predictor Variables

- The decision about whether there are too many predictor variables is related to the number of cases
- A high number of predictor variables can lead to a model that over fits the data and the model becomes too complex in relation to the number of candidate predictors and the number of cases
- One technique for controlling for the number of candidate predictors is to test for multicollinearity

Formula to Control for Overfitting the Model

□ $p < m/10$

□ Where p is the number of candidate predictor variables and m is the number of cases in each group of the dependent variable

□ If m was 981 then p equals 98

Need to Consider the Following when Conducting LR Analysis

- ❖ Setting Up Logistic Regression
- ❖ Setting the Cut-off Value
- ❖ Selecting the best model
- ❖ Interpreting the individual predictors
- ❖ Interpreting the odds ratios when they are negative

Setting the Cut-off Value

- ❖ The cutoff value is the probability of obtaining a 1 (e.g.: course success)
- ❖ The cutoff value directly impacts the results generated for the classification tables
- ❖ The default is set at 0.50
- ❖ Set the cutoff value to match the current probability of success
- ❖ Example: If trying to increase success in an English course and the success rate is 61%, set the cutoff value as 0.61

Selecting the Best Model

- ❖ An acceptable model would have an overall percentage correct greater than cutoff value
- ❖ None of the models reached this threshold
- ❖ Any Model which has the highest overall percentage correct of greater than the cutoff value, say, 66.2%, will be acceptable.

Case Study

A Pharmaceutical Firm that developed a particular drug for women wants to understand the characteristics that cause some of them to have an adverse reaction to a particular drug. They collect data on 15 women who had such a reaction and 15 who did not. The variables measured are:

1. Systolic Blood Pressure
2. Cholesterol Level
3. Age of the person
4. Whether or not the woman was pregnant (1=Yes)

The Dependent Variable indicates if there was an adverse reaction (1=Yes)

The Data for this case is provided in Logistic Regression 2.xls

SPSS Commands for Logistic Regression

- Click on **Analyze > Regression> Binary Logistic**.
- Select the dependent variable, independent variable (covariates) and method (The **Enter** Method is preferred, where you choose to enter all the independent variable in the model. For small sample sizes, the stepwise procedure will tend to throw out most variables, since the statistical test like Wald's will find them insignificant).
- Click on **SAVE** and select Probabilities and Group Membership, then Choose Continue.
- Click on **OPTIONS** and select the statistics (like the Hosmer Lemeshow test) and plots you wish to see.
- Click **OK** to Get the output.

Testing the Statistical Significance of the Estimated Model

Statistical Significance can be tested using the $-2 \log \text{likelihood}$ measure, which essentially measures how likely it is that one can predict the observed values of Y based on the model. Since likelihood varies between 0 and 1, the log value varies from negative infinity to 0. The negative log, therefore, goes from 0 to infinity. A prediction likelihood of 1, or perfect model, has a corresponding $-2 \log$ Likelihood value of 0. This is the case in this example as shown below.

TESTING STATISTICAL SIGNIFICANCE

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	21.841 ^a	.482	.643

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Pseudo R- Square Measures

In the linear regression model, the coefficient of determination, R^2 , summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R^2 values indicating that more of the variation is explained by the model, to a maximum of 1.

For regression models with a categorical dependent variable, it is not possible to compute a single R^2 -Statistic that has all of the characteristics of R^2 in the linear regression model, so these approximations are computed instead.

Pseudo R-Square Measures

The following methods are used to estimate the coefficient of determination.

- Cox and Snell's R^2 is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model.
- Nagelkerke's R^2 is an adjusted version of the Cox & Snell R -square that adjusts the scale of the statistic to cover the full range from 0 to 1.
- McFadden's R^2 is another version, based on the log-likelihood kernels for the intercept-only model and the full estimated model.

What constitutes a “good” R^2 value varies between different areas of application. While these statistics can be suggestive on their own, they are most useful when comparing competing models for the same data. The model with the largest R^2 statistic is “best” according to this measure.

Pseudo R-Square Measures

- All of the Pseudo R-squares reported as less than 0.5 generally agree that the estimated model better fits the outcome data than the previous model.
- While Pseudo R-squares cannot be interpreted independently or compared across datasets, they are valid and useful in evaluating multiple models predicting the same outcome on the same dataset. In other words, a pseudo R-squared statistic without context has little meaning.
- A Pseudo R-square only has meaning when compared to another pseudo R-square of the same type, on the same data, predicting the same outcome. In this situation, the higher pseudo R-squared indicates which model better predicts the outcome.

Pseudo R-Square Measures

For Further information one can refer to:

Cox, D. R., and E. J. Snell. 1989. *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

² Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. *Biometrika*, 78:3, 691-692.

³ McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In: *Frontiers in Economics*, P. Zarembka, eds. New York: Academic Press.

TESTING STATISTICAL SIGNIFICANCE OF THE ESTIMATED Model: Hosmer-Lemeshow Goodness-of-Fit Test

Another Statistical Test that is used is the Hosmer and Lemeshow Goodness-of-Fit test (Chi-square test), which test how well the model fits. Lack of significance on this test is a good thing- in this case with a significance value of 1.000, the test is completely insignificant, meaning that there is no significant difference between the model predictions observed values.

Hosmer-Lemeshow Goodness-of-Fit Test

The Hosmer-Lemeshow test is a statistical test for goodness of fit for the logistic regression model. The data is divided into a number of groups (ten groups is a good way to start). The observed and expected number of cases in each group is calculated and a Chi-squared statistic is calculated as follows:

$$\chi^2_{HL} = \sum_{g=1}^G \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)}$$

where O_g signifies the observed events, E_g signifies the expected events and n_g signifies the number of observations for the g th group, and G is the number of groups. The test statistic follows a Chi-squared distribution with $(G-2)$ degrees of freedom. A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.

HOSMER and LEMESHOW Test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	4.412	8	.818

Contingency Table for Hosmer and Lemeshow Test						
		Reaction = .00		Reaction = 1.00		Total
		Observed	Expected	Observed	Expected	
Step 1	1	3	3.000	0	.000	3
	2	3	2.976	0	.024	3
	3	2	2.711	1	.289	3
	4	2	2.128	1	.872	3
	5	2	1.425	1	1.575	3
	6	2	1.124	1	1.876	3
	7	1	.886	2	2.114	3
	8	0	.513	3	2.487	3
	9	0	.200	3	2.800	3
	10	0	.037	3	2.963	3

Lack of Significance of Chi-Square test indicates that the model is a Good Fit

CLASSIFICATION TABLE

Classification Table ^a					
	Observed		Predicted		
			Reaction		Percentage Correct
			0	1	
Step 1	Reaction	0	11	4	73.3
		1	2	13	86.7
	Overall Percentage				80.0

The cut value is 0.50

CLASSIFICATION TABLE

The Classification Table above shows that the model makes a correct classification of 80% of the time overall. Out of the 15 women with no reaction, the model identified 11 of them as not likely to have one. Similarly, out of 15 women who had the reaction, the model identified correctly 13 as likely to have one.

Estimated Coefficients

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	BP	-.018	.027	.463	1	.496	.982	.932	1.035
	Cholesterol	.027	.025	1.182	1	.277	1.027	.979	1.078
	Age	.265	.114	5.404	1	.020	1.304	1.043	1.631
	Pregnancy(1)	-8.501	3.884	4.790	1	.029	.000	.000	.412
	Constant	-9.373	7.030	1.778	1	.182	.000		
a. Variable(s) entered on step 1: BP, Cholesterol, Age, Pregnancy.									

Wald's Test

Wald and Sig. – This is the Wald Chi-Square test that tests the null hypothesis that the constant equals 0. This hypothesis is rejected because the p-value (listed in the column called “Sig.”) is smaller than the critical p-value of .05 (or .01). Hence, we conclude that the constant is not 0.

Here, Since BP and Cholesterol show up as not significant, one can try to run the Regression again without those variables to see how it impacts the prediction accuracy. Since the sample size here is low, one cannot assume that they are insignificant. Wald's test is suited to large sample sizes.

Predicting Probability and Predicted Group Membership

The Prediction equation will be:

$\text{Log}(\text{Odds of Reaction to the Drug}) = -17.874 - 0.018(\text{BP}) + 0.027(\text{Cholesterol}) + 0.265(\text{Age}) + 8.501(\text{Pregnant})$

For the First Women Respondent,

$\text{Log}(\text{Odds of Reaction to the Drug}) = -17.874 - 0.018 \times (100) + 0.027 \times (150) + 0.265 \times (20) + 8.501 \times (0) = -10.324$

Hence $\log(p/(1-p)) = -10.324$,

Hence $p = (\text{Exp}((-10.324)) / (1 + \text{Exp}(-10.324))) = 0.0000328355$ is the predicted probability. Which is less than the cut-off of 0.5, which classifies her in Group 1 with no chance of getting Drug Reaction.

Predicted Probabilities and Classification of Each Respondent

SEE THE SPSS DATA SHEET. NEW COLUMNS WILL BE CREATED IN THE DATA SHEET.

Advantages and Disadvantages of Logistic Regression

Advantages

Because of its efficient and straightforward nature, doesn't require high computation power, easy to implement, easily interpretable, used widely by data analyst and scientist. Also, it doesn't require scaling of features. Logistic regression provides a probability score for observations.

Disadvantages

Logistic regression is not able to handle a large number of categorical features/variables. It is vulnerable to overfitting. Also, can't solve the non-linear problem with the logistic regression that is why it requires a transformation of non-linear features. Logistic regression will not perform well with independent variables that are not correlated to the target variable and are very similar or correlated to each other.

Thank You
Have a Nice Day