LOGISTIC REGRESSION ANALYSIS

Seth Howells
Concepts of Statistics II
Week# 6 Assignment – Logistic Regression
08/16/20

TABLE OF CONTENTS

OVERVIEW

Logistics regression is a multivariate statistical technique under the multivariate dependence methods umbrella, *see FIG 6-1*, that contains one nonmetric dependent variable. Because the level of measurement helps determine similar multivariate techniques that share the same data type, researchers would likely compare logistic regression analysis with discriminant analysis since both require a nonmetric (categorical) dependent variable. Despite the similarities and *closeness* in the FIG 6-1, logistic regression is far different in its approach incorporating categorical dependent variables.

EQUATION

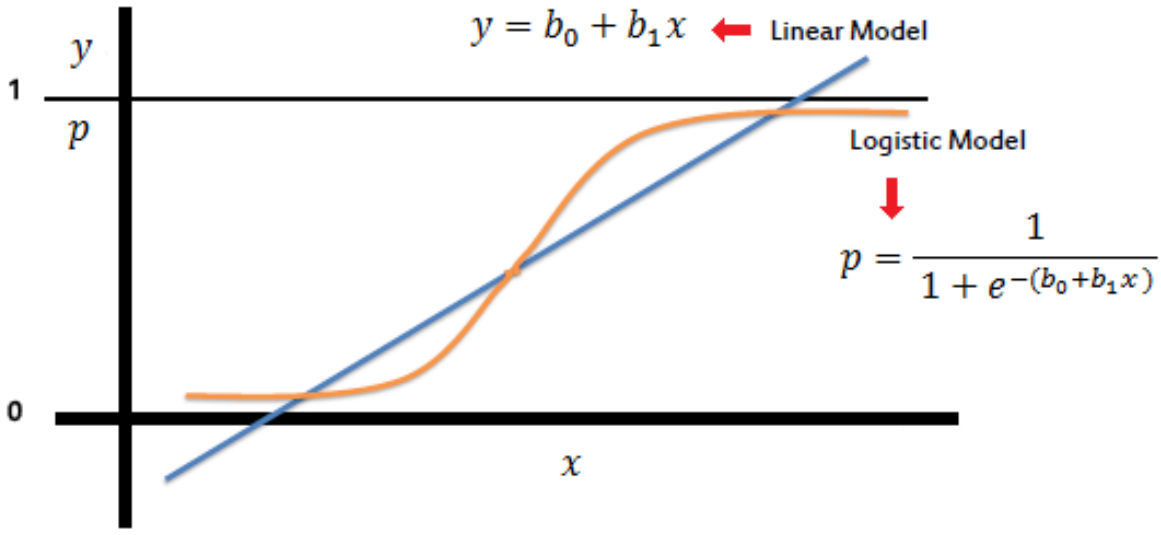
Fundamental takeaways are often found in the equation because the researcher can conceptually understand the relationship among the multiple variables in the multivariate model.

Linear Model: $y = b_0 + b_1 x$

Logistic Model: $p = \frac{1}{1+e^{-b_0+b_1 x}}$

It is apparent that logistic regression refers to a probability while linear refers to a slope weighted by the variables' coefficients. This idea can be graphically explained in FIG 6-3, where the linear model measures in a different scope than the logistic model.

FIG 6-2: Linear Model vs Logistic Model



LOGISTICS REGRESSION VS DISCRIMINANT ANALYSIS VS MULTIPLE REGRESSION

Discriminant analysis assumes on multivariate normality and equal variance-covariance matrices while in contradistinction, logistic regression is not restricted to that assumption and is more robust when multivariate normality is not met. Thus, many researchers will find logistic regression to be applicable to a wider range of analysis.

Multiple Regression	**Logistic Regression**
Total Sum of Squares	-2LL of Base Model
Error Sum of Squares	-2LL of Proposed Model
Regression Sum of Squares and Proposed Models	Difference of -LL for Base
F test of model fit	Chi-square Test of -2LL Difference
Coefficient of determination	"Pseudo" R^2 measures

6-STAGE REGRESSION PROCESS

Stage 1: Objectives of Logistic Regression

- Identify independent variables that impact group membership in the dependent variable
- Establishing binary classification for determining group membership

Stage 2: Research Design

- Due to binary nature of logistic regression, error term has a binomial distribution
- No assumption of normality
-

Stage 3: Assumptions

- Less assumptions (in comparison to, for example, discriminant analysis) is an advantage as it allows flexibility for the researcher
- Does not require any specific distributional form for the independent variables
- Heteroscedasticity is not required for the independent variables
- Linear relationships between the independent variables and dependent variables are not required.

Stage 4: Estimation & Assessing Overall Fit

Steps: Estimating the Coefficients

1. *Transforming a probability into odds and logit values*
    a. ***Odds:*** *probability*
    b. ***Logit:*** *used to model the probability of a certain class: win/lose, good credit/poor credit, male/female etc.*

    c. *Logistic transformation steps: restating a probability as odds and calculating the logit values*
2. *Model does not use least squares method, but rather an estimation using a maximum likelihood approach.*
    a. ***Maximum likelihood approach:** maximizes the likelihood that an event would occur*
3. *Null model vs proposed model:*
    a. ***Null Model:** acts as baseline for making comparison of improvement in the model fit*
    b. ***Estimate Proposed Model:** model containing the independent variables to be included in the logistic regression*
    c. ***2LL Difference:** a measure of how well the estimated model fits the likelihood. The value represents how much unexplained information there is after the model has been fitted, and thus a lower -2LL represents an overall fit.*

Stage 5: Interpretation

- Coefficients:
    - In multiple regression, a coefficient of 0 indicates that the independent variables has no impact on the dependent variable while in contradistinction, logistic regression coefficient of 0.50 indicates that there is no impact on the dependent variable since it is based on a 0.00 to 1.00 probability.
- Wald statistic
- Directionality of relationship:
    - Positive coefficient: increase in independent variables increases the predicted probability of the dependent variable
    - Negative coefficient: decrease in independent variables coefficient decreases the predicted probability of the dependent variable
- Magnitude of the relationship of metric independent variables
    - Interpreted differently than multiple regression since multiple regression is a linear function while logistic regression is nonlinear (0 to 1).
    - Exponentiated logistic coefficients is used to determine the magnitude of the relationship.
        - 1.0 coefficient does not impact the dependent variable
        - 0.20, for example, would be interpreted as a one-unit change in the independent variable will reduce the odds by 80%
- Nonmetric independent variables

Stage 6: Validation of Results

- Internal and External validity of the results
    - Estimating external validity: creation of a holdout or validation sample and calculating the hit ratio (most common).

        - **Hit ratio:** percentage of the observations that is correctly predicted by the model
    - Cross validation achieved with a jackknife (leave-one-out) process of calculating the hit ratio

FIG 6-1: Multivariate Methods Diagram

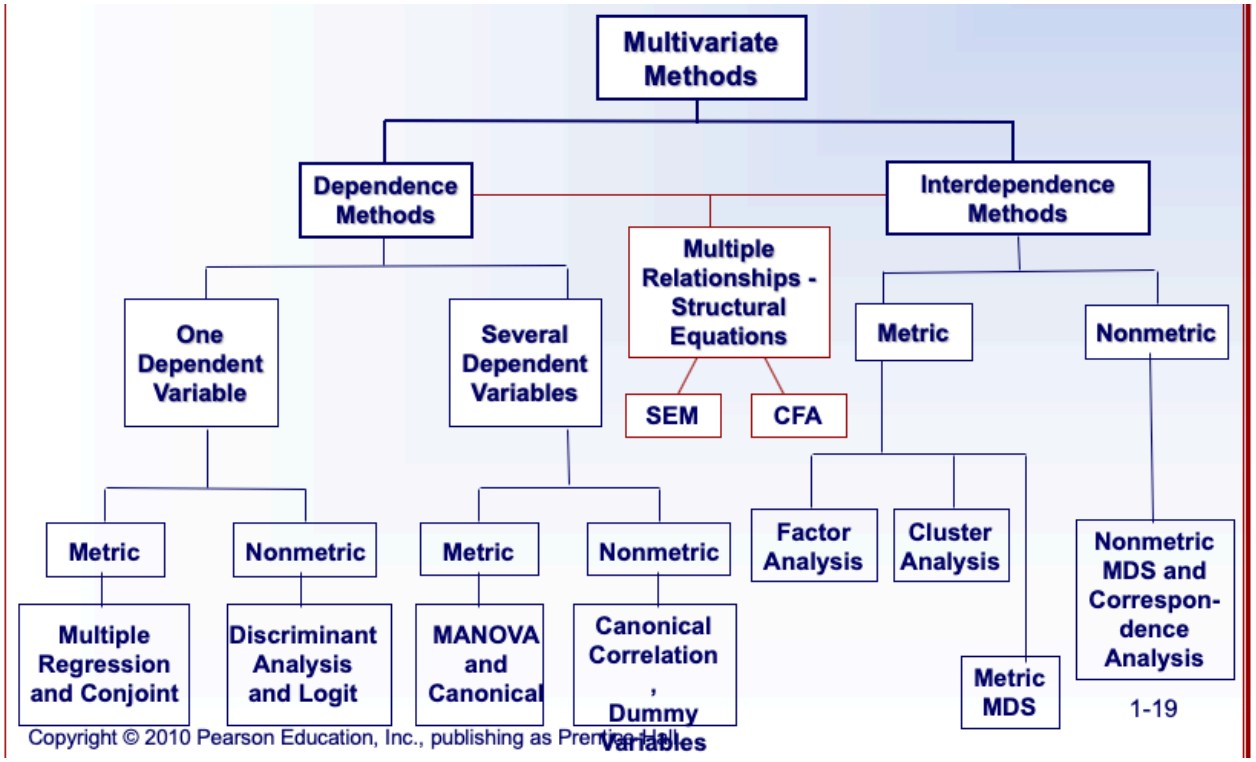


FIG 6-3: Linear Regression vs Logistic Regression