STAT-411
CATEGORICAL DATA ANALYSIS

Class: BS Statistics 8th regular

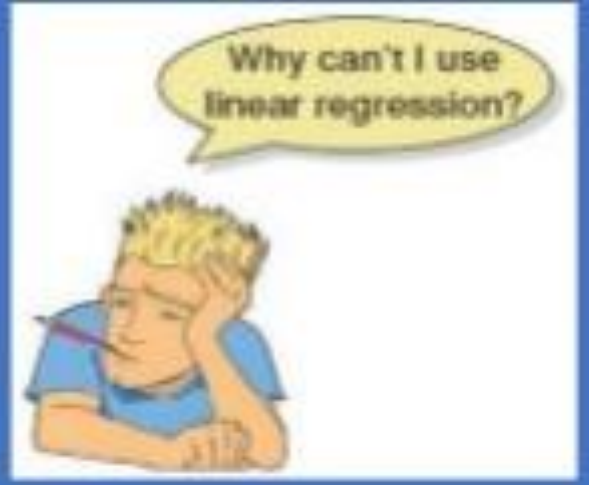
Instructor: Miss Summera Kinat

Aims

- When and Why do we Use Logistic Regression?
  - Binary
  - Multinomial
- Theory Behind Logistic Regression
  - Assessing the Model
  - Assessing predictors
  - Things that can go Wrong
- Interpreting Logistic Regression

When And Why

- To predict an outcome variable that is categorical from one or more categorical or continuous predictor variables.
- Used because having a categorical outcome variable violates the assumption of linearity in normal regression.
- Does not assume a linear relationship between DV and IV



When And Why

- No assumptions about the distributions of the predictor variables.
- Predictors do not have to be normally distributed
- Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.
- Because it does not impose these requirements, it is preferred to discriminant analysis when the data does not satisfy these assumptions.

Logistic regression

- Logistic regression is used to analyze relationships between a dichotomous dependent variable and continue or dichotomous independent variables. (SPSS now supports Multinomial Logistic Regression that can be used with more than two groups, but our focus here is on binary logistic regression for two groups.)

- Logistic regression combines the independent variables to estimate the probability that a particular event will occur, i.e. a subject will be a member of one of the groups defined by the dichotomous dependent variable. In SPSS, the model is always constructed to predict the group with higher numeric code. If responses are coded 1 for Yes and 2 for No, SPSS will predict membership in the No category. If responses are coded 1 for No and 2 for Yes, SPSS will predict membership in the Yes category. We will refer to the predicted event for a particular analysis as the modeled event.

With One Predictor

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+\varepsilon_i)}}$$

- Outcome
  - We predict the *probability* of the outcome occurring
- b_0 *and* b_0
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

With Several Predictor

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n+\varepsilon_i)}}$$

- Outcome
  - We still predict the *probability* of the outcome occurring
- Differences
  - Note the multiple regression equation forms part of the logistic regression equation
  - This part of the equation expands to accommodate additional predictors

The Logistic Regression

Logit $p = \alpha + \beta 1X1 + \beta 2X2 + .. + \beta pXp$

α represents the overall disease risk

$\beta 1$ represents the fraction by which the disease risk is altered by a unit change in X1

$\beta 2$ is the fraction by which the disease risk is altered by a unit change in X2

..... and so on.

What changes is the log odds. The odds themselves are changed by e^{β}

If $\beta = 1.6$ the odds are $e^{1.6} = 4.95$

Measuring the Probability of Outcome

The probability of the outcome is measured by the odds of occurrence of an event.

If P is the probability of an event, then (1-P) is the probability of it not occurring.

Odds of success = P / 1-P

$$\frac{P}{1-P}$$

Methods of Regression

- Forced Entry: All variables entered simultaneously.
- Hierarchical: Variables entered in blocks.
  - Blocks should be based on past research, or theory being tested. Good Method.
- Stepwise: Variables entered on the basis of statistical criteria (i.e. relative contribution to predicting outcome).
  - Should be used only for exploratory analysis.

DECISION PROCESS

Stage 1:

Objectives Of logistic regression

- Identify the independent variable that impact in the dependent variable

- Establishing classification system based on the logistic model for determining the group membership

Types of logistic regression

- ***BINARY LOGISTIC REGRESSION***

  It is used when the dependent variable is dichotomous.

MULTINOMIAL LOGISTIC REGRESSION

_It is used when the dependent or outcomes variable has more than two categories.

Linear Regression

Independent Variable

Dependent Variable

Logistic Regression

Independent Variable

Dependent Variable

Binary logistic regression expression

$$Y = b_o + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots b_kX_k + E$$

BINARY

Y = Dependent Variables

ß. = Constant

$ß_1$ = Coefficient of variable X_1

X_1 = Independent Variables

E = Error Term

SAMPLE SIZE

- Very small samples have so much sampling errors.
- Very large sample size decreases the chances of errors.
- Logistic requires larger sample size than multiple regression.
- Hosmer and Lamshow recommended sample size greater than 400.

SAMPLE SIZE PER CATEGORY OF THE INDEPENDENT VARIABLE

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

Estimation of logistic regression model assessing overall fit

- Logistic relationship describe earlier in both estimating the logistic model and establishing the relationship between the dependent and independent variables.
- Result is a unique transformation of dependent variables which impacts not only the estimation process but also the resulting coefficients of independent variables.

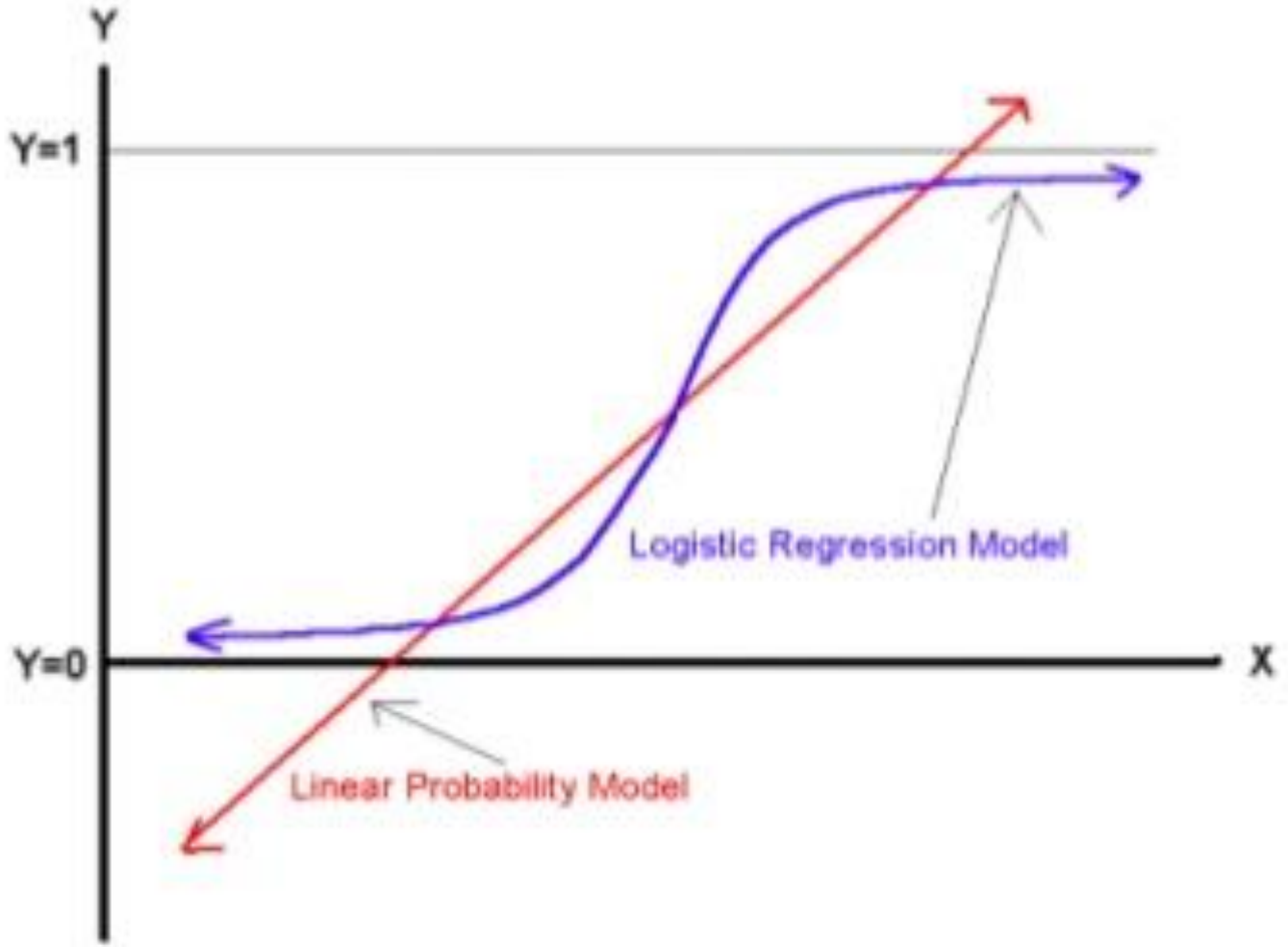
Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample:

  $$L = \text{Prob}\ (p_1 * p_2 * * * p_n)$$

- The higher the L, the higher the probability of observing the ps in the sample.

Comparing the LP and Logit Models



Description of the data

- The data used to conduct logistic regression is from a survey of 30 homeowners conducted by an electricity company about an offer of roof solar panels with a 50% subsidy from the state government as part of the state's environmental policy.

- The variables are:

•IVs: household income measured in units of a thousand dollars

age of householder

monthly mortgage

size of family household

•DV: whether the householder would take or decline the offer.

•Take the offer was coded as 1 and decline the offer was coded as 0.

WHAT IS THE RESEARCH QUESTION?

- To determine whether household income and monthly mortgage will predict taking or declining the solar panel offer

- Independent Variables: household income and monthly mortgage

- Dependent Variables: Take the offer or decline the offer

Two hypotheses to be tested

- There are two hypotheses to test in relation to the overall fit of the model:

◆H0: The model is a good fitting model

◆H1: The model is not a good fitting model (i.e. the predictors have a significant effect)

How to perform logistic regression in spss

1) Click *Analyze*
2) Select *Regression*
3) Select *Binary Logistic*
4) Select the dependent variable, the one which is a grouping variable (0 and 1) and place it into the Dependent Box, in this case, take or decline offer
5) Enter the predictors (IVs) that you want to test into the Covariates Box. In this case, Household Income and Monthly Mortgage
6) Leave *Enter* as the default method

Continuation of SPSS Steps

•7) If there is any categorical IV, click on Categorical button and enter it. There is none in this case.

•8) In the *Options* button, select *Classification Plots, Hosmer-Lemeshow goodness-of-fit, Casewise Listing of residuals*. Retain default entries for *probability of stepwise, classification cutoff,* and *maximum iterations*

•9) *Continue*, then, *OK*