

Logistic Regression

BUS 735: Business Decision Making and Research

1 Goals and Agenda

Goals of this section

- Specific goals:
 - Learn how to conduct regression analysis with a dummy independent variable.
- Learning objectives:
 - LO2: Be able to construct and use multiple regression models (including some limited dependent variable models) to construct and test hypotheses considering complex relationships among multiple variables.
 - LO6: Be able to use standard computer packages such as SPSS and Excel to conduct the quantitative analyses described in the learning objectives above.
 - LO7: Have a sound familiarity of various statistical and quantitative methods in order to be able to approach a business decision problem and be able to select appropriate methods to answer the question.

Agenda

Learning Objective	Active Learning Activity
Understand purpose and background of Logistic Regression	Lecture
Understand how to predict the impact explanatory variables have on the outcome variable	Lecture
Estimate a logistic regression model	In-class Exercise with SPSS
More practice!	Homework assignment, due Monday, Oct 6.

2 Regression Framework

2.1 Binary Dependent Variable

Logistic Regression

- **Logistic Regression:** method for estimating a regression with a *dummy (binary) dependent variable*.
- Will a potential customer purchase a product?(YES=1, NO=0).
 - Might use explanatory variables: age, gender, income, etc.
- Will a potential employee be retained after one year?(YES=1, NO=0).
 - Might use explanatory variables: age, gender, years experience, past income, education dummy (4-year = 1, otherwise = 0).
- Why not run a regression? Which assumption is violated?

2.2 Structural Form

Structural Form

- Normal regression:

$$y_i = b_0 + b_1X_{1,i} + b_2X_{2,i} + \dots + b_{k-1}X_{k-1,i} + e_i$$

- Logistic regression:

$$\log(Odds) = b_0 + b_1X_{1,i} + b_2X_{2,i} + \dots + b_{k-1}X_{k-1,i} + e_i$$

$$Odds = \frac{P(y_i = 1)}{1 - P(y_i = 1)} = \frac{P(y_i = 1)}{P(y_i = 0)}$$

- Remember, $y_i \in 0, 1$ indicates YES=1 event did occur, or NO=0 event did not occur.

2.3 Predicted Value

Predicted Value

- Predicted value from a regular regression:

$$\hat{y}_i = b_0 + b_1X_{1,i} + b_2X_{2,i} + \dots + b_{k-1}X_{k-1,i}$$

- For a logistic regression, you can get predicted logit (not too interesting yet):

$$\hat{l}_i = \ln(\widehat{Odds}) = b_0 + b_1X_{1,i} + b_2X_{2,i} + \dots + b_{k-1}X_{k-1,i}$$

- To uncover the **predicted probability of the event occurring**:

$$P(\widehat{y_i = 1}) = \frac{1}{1 + e^{(-\hat{l}_i)}}$$

3 Marginal Effect

3.1 Regression versus Logistic Regression

Marginal Effects

- **Marginal effect for regression:** measure of how much y changes when x increases by 1.
 - Example: How much does public expenditure per capita increase (or decrease) when economic ability increases by one unit?
- **Marginal effect for logistic regression:** measure of how much $P(y_i = 1)$ changes when x increases by 1.
 - How much more (or less) likely will an interview candidate be working here after one year if she/he has a four year college education?

3.2 Coefficients versus Marginal Effects

Coefficients versus Marginal Effects

- For a regular regression, for coefficient b_2 :
 - The sign (positive/negative) indicates whether x_2 causes y to increase or decrease.
 - The magnitude tells *how much* y increases when increasing x_2 by 1.

- Coefficient = Marginal Effect.
- For a logistic regression, the coefficient b_2 :
 - The sign (positive/negative) indicates whether x_2 causes y to increase or decrease.
 - The magnitude is pretty meaningless.
 - Need to do more to figure out marginal effect.
- Because SPSS is stupid, it cannot figure out marginal effects.
- For a specific individual (specific values for X's), compute predicted probabilities by hand for:
 - The individual's actual set of values for X's.
 - Same set of X's, except increase one of them by 1.
 - Take the difference = Marginal effect.

4 Example

Example

- Well Being Dataset (Described on pages 356-357 of SAB).
- 182 college students in Washington, D.C. area.
- Variables of interest:
 - Weight (4 categories: UNDERWEIGHT=1, NORMAL=2, OVERWEIGHT=3, OBESE=4)
 - Gender (2 categories: MALE=0, FEMALE=1)
 - Age
 - Race (AFRICAN AMERICAN=1, WHITE=2, OTHER=3)
 - Marital Status (SINGLE=1, MARRIED=2, OTHER=3)
 - Financial Status (OVEREXTENDED=1, MAKING ENDS MEET=2, COMFORTABLE=3)
 - Physical Health (5 point Likert scale)
 - Depression Scale (0-60, values > 15 indicate depression)
- Find preliminary evidence whether each variable is related to weight.
- Logistic Regression Analysis.