

Categorical Dependent Variables

So far, we have considered only quantitative response variables. What if the response variable is a categorical variable?

We use dummy variables to represent categorical explanatory variables (gender); we do the same with categorical response variables. (Here we consider only dichotomous response variables.)

A linear fit won't work because it gives predictions less than 0 and greater than 1. Two approaches:

- logistic regression
- discriminant analysis

3/28/2003

PUAF 610: Logistic Regression

1

Logistic Regression

We can model a dichotomous response variable as a probability: $\hat{y} = P(y|x)$

In the sample data, y either occurs (1) or doesn't occur (0). The model uses this data to predict the probability of y occurring as a function of the values of the explanatory variables.

- probability, p , varies between 0 and 1
- odds, $d = p/(1-p)$, varies between 0 and ∞
- log of odds, $L = \log(d)$, varies between $-\infty$ and ∞ , making it a well-behaved response variable

8/28/2003

PUAF 610: Logistic Regression

2

Logistic Regression

The logistic regression model is:

$$L = \log_e \left(\frac{p}{1-p} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

$$d = \text{odds} = \left(\frac{p}{1-p} \right) = e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

$$p = P(y|x_1, x_2, \dots, x_k) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

The logarithm of the odds ratio is called the "log odds" or the "logit".

3/28/2003

PUAF 610: Logistic Regression

3

Logistic Regression

- Sample data usually contains $y = 0$ or 1 ; because $\log(0)$ is undefined, least squares cannot be used.
- "Maximum likelihood estimation" is used to estimate best-fit values of α , β_1 , β_2, \dots, β_k .
- Interpreting the regression coefficients:
 - when x increases 1 unit, log odds increases β units and the odds increases by a factor of e^β , or, if β is small, by β percent (all else equal)
 - if $\beta < 0$, the odds decrease as x increases;
 - if $\beta > 0$, the odds increase; if $\beta = 0$, no effect

8/28/2003

PUAF 610: Logistic Regression

4

Examples

- Probability that an inmate will violate parole if released, based on type of offense, prior history, and behavior while incarcerated
- Probability that a seismic event is a nuclear explosion, based on ratio of surface wave to body wave magnitudes
- Probability that death sentence will be imposed, based on race of defendant and victim and factors related to nature of crime
- Probability of getting an A in 610, based on GPA, GRE, undergraduate major

3/28/2003

PUAF 610: Logistic Regression

5

Logistic Regression

- No R^2 ; use p-value of "improvement" to judge overall value of model
- As with multiple regression, use p-value of coefficients to make include/exclude decisions
- Validate model using sample splitting; estimate SE for predictions with cross-validation technique
- No analysis of residuals
- If perfect discrimination possible (no overlap), technique fails

8/28/2003

PUAF 610: Logistic Regression

6

Example: Heights of Students

- Previously, we used gender and parent's heights to predict a student's height
- We could also use a student's height, together with parent's heights, to predict gender
- Based on student's height alone, we can correctly classify 78% of students
- Based on student's and parent's heights, we can correctly classify 90% of students

3/28/2003

PUAF 610: Logistic Regression

7

Discriminant Analysis

- Logistic regression finds a "best fit" equation for the probability of occurrence of an event
 - we specify a threshold value of the predicted probability for classifying events (e.g., 0.5)
- Discriminant analysis finds the surface that best divides the data set into two groups
 - we specify the relative costs of misclassification and
 - the prior probability of the events

8/28/2003

PUAF 610: Logistic Regression

8

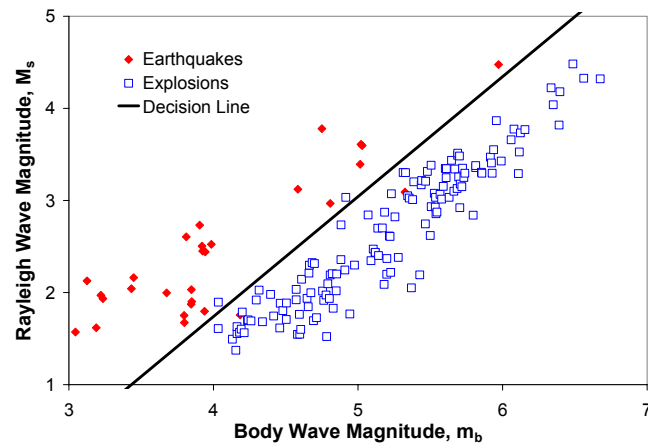
Example: Seismic Verification

- All five nuclear weapon states have signed the CTBT, which prohibits all nuclear explosions
- Verifying the absence of nuclear tests in the atmosphere, oceans, and space is easy
- Verifying the absence of underground nuclear tests is difficult:
 - explosions generate seismic waves, but so do earthquakes
 - explosions generate smaller surface waves than do earthquakes, for a body-wave magnitude

3/28/2003

PUAF 610: Logistic Regression

9



8/28/2003

PUAF 610: Logistic Regression

10

Example: Seismic Verification

- Assuming equal costs of misclassification and equal prior probabilities, we derive a decision line that correctly classifies 93% of earthquakes and 98% of explosions in the sample
- With other assumptions, we can increase one of these probabilities at the expense of the other
- By adding other seismic measures, we can create a multi-dimensional discriminant function with lower false positive and negative rates
- Discriminant analysis is valuable as an aid to decision making, but not for determining the effect of a variable holding all others constant

3/28/2003

PUAF 610: Logistic Regression

11