RUG1

university of groningen

Binomial Logistic Regression
Applied to Gradability in Dutch

› Nick Ruiz (with Edgar Weiffenbach)

› Statistics Seminar, Spring 2010

Slide 1

RUG1 To set the date:
* >Insert >Date and Time
* At Fixed: fill the date in format mm-dd-yy
* >Apply to All

RUG; 30-8-2007

Contents

Why Regression?

- Regression is used to understand how a dependent variable relies on a series of independent factors.
- Typically used in prediction and forecasting.

Review: Linear Regression

› Formula:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon_i$$

› Assumptions:

- Linear relationships between independent variables and dependent variable
- Normality of residuals
- Dependent variable is unbound

Logistic Regression

› Used to predict a categorical variable based on one or more independent factors
› Analyze the dependency of categorical variables on other factors
› Predicts the likelihood (probability) of a categorical value's occurrence within data.

university of groningen

Why not use normal regression?

- In normal regression models, the dependent variable is unbound.

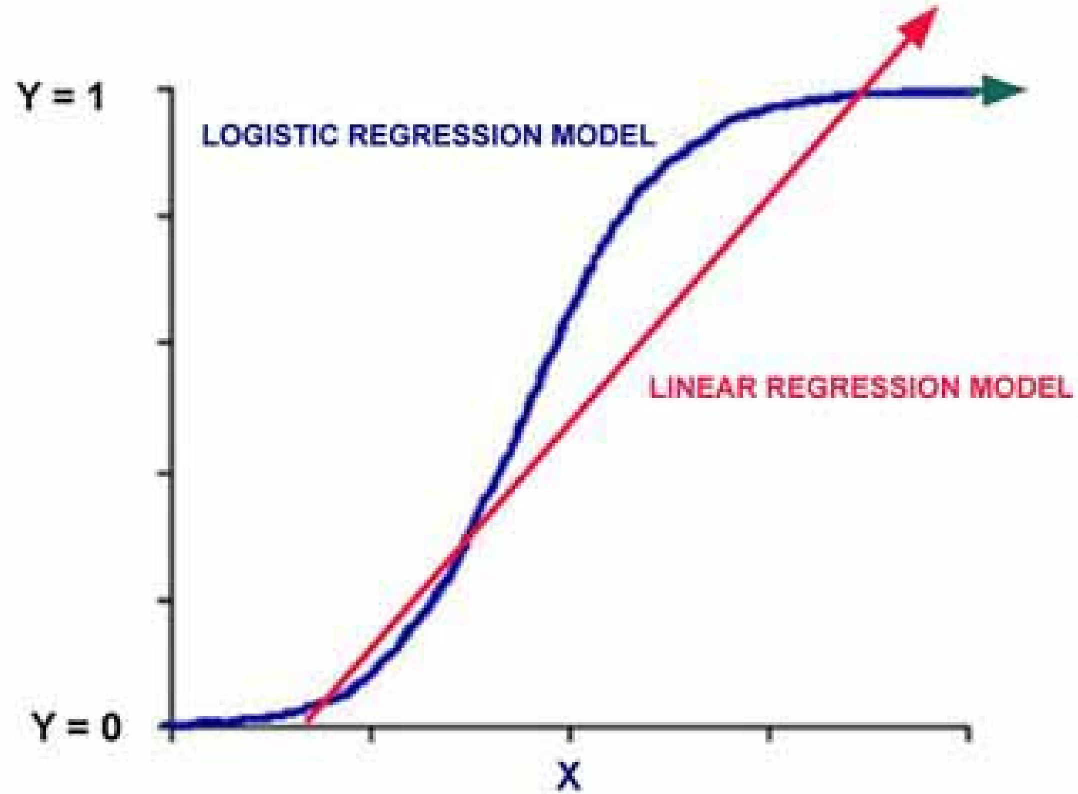
$$-\infty \leq v \leq \infty$$

- In our analysis, we wish to develop a regression model to predict the occurrence of a categorical variable.

$$0 \leq \mathrm{P}(event) \leq 1$$

- How do we model a constrained variable with regression?
  - Logistic regression expresses the equation in logarithmic terms, overcoming the linearity constraint.

Logistic vs. Linear Regression



› Source: http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

Logit function

- Odds

$$odds_{event} = \frac{P(event)}{1 - P(event)}$$

- Log odds

$$\mathrm{logit}(p) = \ln odds_{event} = \ln \frac{P(event)}{1 - P(event)}$$

- Bounds of logit function:

$$-\infty \leq \mathrm{logit}(p) \leq \infty$$

Predicting logit values

- Use regression to find the optimal coefficients.

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- The logit function is a similar equation to linear regression.

university of groningen

Applying logit to regression

- Recall that we wish to predict the likelihood (probability) of a dependent variable.
- Logistic regression seeks to predict the likelihood of a logit value.
- Logarithmic function:

$$\mathrm{P}(v) = \frac{1}{1 + e^{-v}}$$

Applying logit to regression

› Plugging logit into the logistic function:

$$\mathrm{P}(\mathrm{logit}(v)) = \frac{1}{1 + e^{-\mathrm{logit}(v)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Maximal Likelihood Estimation

- Recall that Ordinary Least Squares (OLS) regression seeks to minimize the squared differences of the data points to the regression line.
  - Also known as minimizing the least squared error
- Logistic regression uses Maximal Likelihood Estimation
  - Maximizing the odds that the observed values of the dependent variable are predicted from the independent variables

Maximum Likelihood Estimation

› Dependent variables in binomial regression have 2 possible values.
  - Follows a binomial distribution

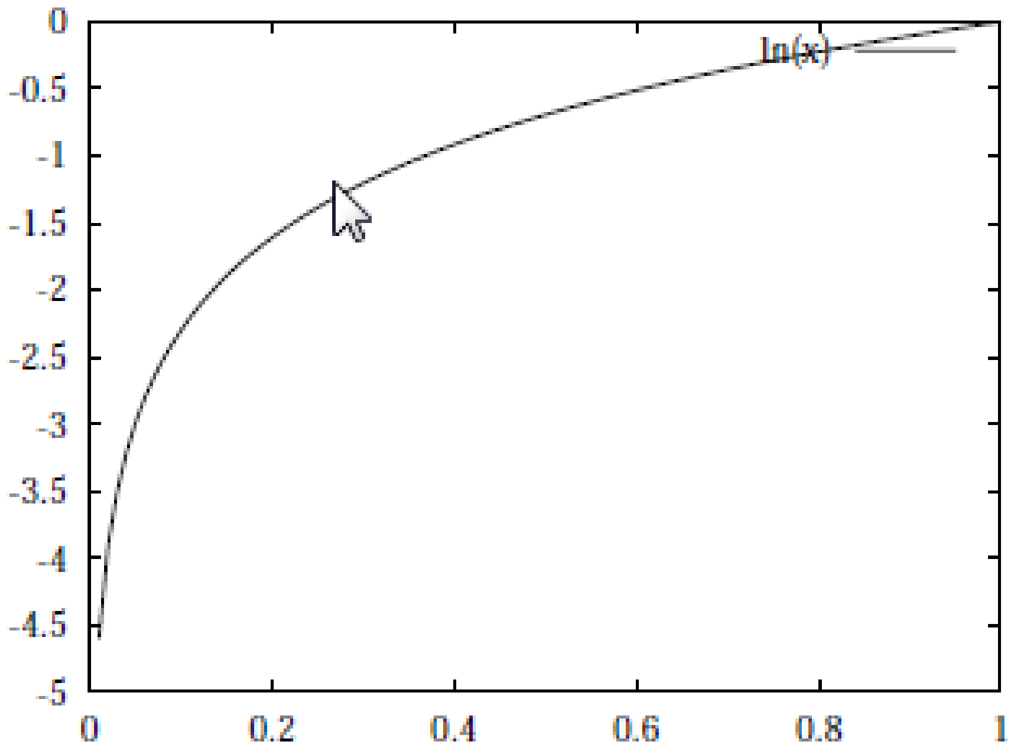
$$p^k(1-p)^{n-k}$$

› Goal: Maximize the log likelihood (L) of an event by estimating the parameters in the model.

$$L = \ln p^k(1-p)^{n-k} = k\ln p + (n-k)\ln(1-p)$$

university of
groningen

Log Probabilities



Significance tests

- To determine the “goodness of fit”, several methods are possible:
  - e.g. Hosmer and Lemeshow’s Chi-Square test
  - *-2L* has approximately a Chi-Square distribution with *n-1* degrees of freedom

Application to Gradability Analysis in Dutch

› Gradability examples in English:
  - “Some day I’m going to be a big star.”

› Non-gradable examples in English:
  - “The big basketball player”

Analysis

- Extract adjective-noun pairs for the following Dutch adjectives:
  - Gigantisch
  - Reusachtig
  - Kolossaal
- Goal: Try to predict the occurrence of a gradable reading of these adjectives, given the sentence context.
- Statistical technique: Binomal logistic regression

Variables in the analysis

- Dependent Variable:
  - Gradable Reading (binary)
- Independent Variables:
  - Semantic Role {Subject, Object, Predicative}
  - Adjective
  - Article of Preceding Determiner {Definite, Indefinite}
    - e.g. "*die* gigantisch boom"
  - Less important features:
    - Following Preposition (binary)

Analysis

- Extracted adjective-noun pairs from the Lassy Dutch Corpus
  - http://www.let.rug.nl/~vannoord/Lassy/
  - Newspaper
- Over 8,000 examples with gigantisch, reusachtig, and kolossaal.
  - Edgar manually annotated each reading with a gradable/non-gradable interpretation.

Sample analysis in R (using Design package)

```
> adjfreq3.lrm = lrm(GradableReading ~ SemanticRole + Adj + DefiniteArticle +
   FollowingPP, data = adjfreq3, x = T, y = T)

Frequencies of Responses
  n   y
381 197

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy
       578      2e-11      92.24          5          0      0.718      0.437
     Gamma      Tau-a         R2      Brier
     0.484      0.197      0.204      0.192
```

	Coef	S.E.	Wald Z	P
Intercept	-0.55674	0.2021	-2.76	0.0059
SemanticRole=predc	1.10901	0.3199	3.47	0.0005
SemanticRole=su	-0.07674	0.2266	-0.34	0.7349
Adj=reusachtig	-1.88641	0.2710	-6.96	0.0000
DefiniteArticle=indefinite	0.24457	0.2087	1.17	0.2411
FollowingPP=yes	0.25775	0.2235	1.15	0.2487

Evaluation

› R^2 measures the accuracy of the predictions

› C: Index of concordance between the predicted probability and the observed response

- $C = 0.5$ implies that the predictions are random
- $C = 1.0$ implies that the predictions are perfect
- Our C: 0.718

Evaluation

- Somers' D_{xy} provides a rank correlation between the predicted probability and the observed responses

$$0 \leq D_{xy} \leq 1$$

- In our analysis:

$$D_{xy} = 2(C - 0.5) = 0.437$$

Conclusion

- We have attempted a simple logistic regression model with a subset of data; the results are not yet promising.
- After analyzing all of the data, we will determine if there are any interactions between the independent variables.

university of groningen

Thank you