NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W6: 9/29/15-10/1/15

NC STATE UNIVERSITY

Administrative

- Updated Weekly Schedule (on Moodle)
- 1st Midterm: 10/6/15
  - Review: 10/1/15 (important, don't miss)
- Additional Reading Materials
  - Logistic Regression, by Kleinbaum (Springer, Through NCSU Library)
  - Computing Primer for Applied Linear Regression, 4ed, Using R. http://z.umn.edu/alrprimer

Regression, So far

- Linear Regression
  - Least Squares
- Correlation
- Regression Parameters
- Properties
- Significance of "r"
- Total Variation (explained + unexplained)
- Coefficient of Determination
- Standard Error of estimate, Prediction Interval
- Multiple Linear Regression
- Multiple Correlation Coefficient (R)
- Testing for significance of R

Regression As Classification

- So far, regression as prediction
- Today, regression as classification

Binary Classification Problems

- Problems that have two outcomes
- True/False categorical outcomes
  - Had Fever/no fever
  - Had a disease/no disease
  - (mechanical) failed/not failed
  - Win/loss
- Dichotomized categorical outcomes
  - Yes/no
  - Agree/Disagree

Recall That

- 0/1 Outcomes
  - Bernoulli outcomes
- Collection of exchangeable outcomes for same attribute (or covariate) data
  - Binomial outcome

NC STATE UNIVERSITY

Check if Regression Works

- Problem
  - Predicting NCSU Wolfpack (WF) game based on points (P)

$$WF_i = b_0 + b_1 P_i + \varepsilon_i$$

- $WF_i = \{0,1\}$

NC STATE UNIVERSITY

Observations

- Errors
  - Can't be normally distributed
- Error variance is not constant
  - It depends on the level of X_i
- Response variable is bounded

  $0 \le Pr \le 1$

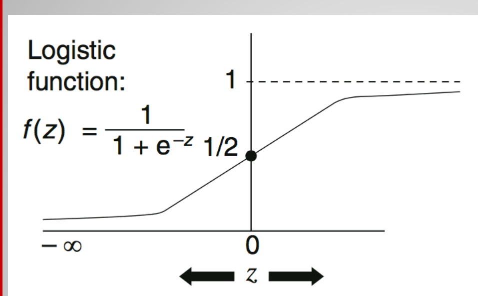
NC STATE UNIVERSITY

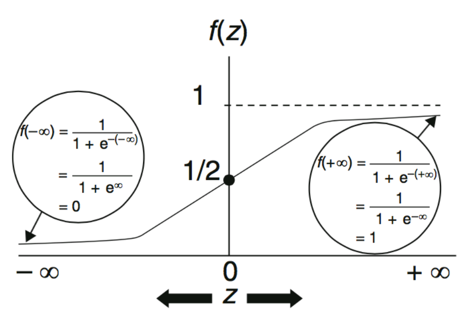
Response Function

- Logit
- Probit
- Log-Log

NC STATE UNIVERSITY

Logistic Function





Range: $0 \leq f(z) \leq 1$

NC STATE UNIVERSITY

Few Definitions

- Outcome (Binary; 0/1): $WF_i\ (Y_i)$
- Probability (0, 1): $Pr(Y_i \mid X_i, b_0, b_1)$
- Odd (0, ∞): (1/(1- Pr())): $\frac{\Pr(Y_i \mid X_i, b_0, b_1)}{1-\Pr(Y_i \mid X_i, b_0, b_1)}$
- Log odds (-∞, ∞): (Logit) $\log \frac{\Pr(Y_i \mid X_i, b_0, b_1)}{1-\Pr(Y_i \mid X_i, b_0, b_1)}$

11/3/15 © Raju Vatsavai CSC-591. 11

NC STATE UNIVERSITY

Linear vs. Logistic Regression

- Linear: $Y_i = b_0 + b_1X_i + \varepsilon_i$

$$E[Y_i \mid X_i, b_0, b_1] = b_0 + b_1 X_i$$

- Logistic: $\Pr(Y_i \mid X_i, b_0, b_1) = \hat{\pi}_i = \frac{e^{(b_0 + b_1 X_i)}}{1 + e^{(b_0 + b_1 X_i)}}$

11/3/15 © Raju Vatsavai CSC-591. 12

NC STATE UNIVERSITY

Fitted Logit Response Function

$$\hat{\pi}_i = \frac{e^{(b_0 + b_1 X_i)}}{1 + e^{(b_0 + b_1 X_i)}}$$

- Log Odds: $\hat{\pi}_i^{'} = \log_e \frac{\hat{\pi}_i}{1 - \hat{\pi}_i}$

$$\hat{\pi}_i^{'} = b_0 + b_1 X_i$$

NC STATE UNIVERSITY

In Summary

- The Logistic model

$$z = \underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

- In essence, *z* is an index that combines Xs

Epidemiology Example

- Let $X_1, X_2, \ldots, X_k$ are observations on a group of subject at Time T_0
- For each of those observation, we also determined the disease status, as either 1 if "with disease" or 0 if "without disease".
- Objective
  - We wish to use this information to describe the probability that the disease will develop during a defined study period, say T_0 to T_1, in a disease- free individual with independent variable values X_1, X_2, up to X_k, which are measured at T_0

11/3/15 © Raju Vatsavai CSC-591. 15

Epidemiology Example

$$P(D = 1 | X_1, X_2, \ldots, X_k)$$

$$= P(\mathbf{X})$$

Model formula:

$$\boxed{P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}}$$

- Then, from observed data, we can estimate the parameters; $b_0, b_1, b_2, \ldots b_k$
- For a given patient, we can estimate the risk by simply plug-in the observations (X_i) into to model

11/3/15 © Raju Vatsavai CSC-591. 16

NC STATE UNIVERSITY

Acknowledgements

- G. James, et. al., Moore, et. al.
- Kleinbaum, et. al.