CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 58
Generalized Linear Modeling

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Generalized Linear Model (GLM)

- Let y_i have any probability distribution so long as it is from the "exponential" family
  - e.g., normal, log-normal, exponential, gamma, chi-squared, beta, Bernoulli, Poisson, etc.
  - Not included are Student's t, mixed distributions
- Allow for any transformation of y (the link function)
  - Must be monotonic and differentiable
- Linear in the model parameters

Link function ⟶ $g(\hat{y}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$

Non-Normal Distributions

- For a non-normal distribution, Least Squares ≠ MLE
- Combine any link function with any (exponential family) probability distribution for y, then find the maximum likelihood estimates for the parameters
  - Solve with iteratively reweighted least squares
  - Many software packages can do this regression

Typical Distribution/Link Pairings

Distribution	Typical uses	Link Name	Link function
Normal	Linear-response data	Identity	$g(y) = y$
Exponential or Gamma	Exponential-response data, scale parameters	Inverse	$g(y) = 1/y$
Poisson	count of occurrences in fixed amount of time/space	Log	$g(y) = \ln(y)$
Bernoulli or Binomial	outcome of single yes/no occurrence	Logit (logistic model)	$g(y) = \ln\left(\frac{y}{1-y}\right)$

What if our Response is Binary?

- Sometimes the response is binary
  - The patient lives or dies
  - The part passes or fails
  - The customer buys or doesn't buy
- The response y will follow a Bernoulli distribution
  - $E[y] = \pi$, the probability of "success" (y = 1)
  - $var[y] = \pi(1-\pi)$ (a function of the mean)
- We want to model the probability of success
  - $\hat{y} = E[y] = \pi$

Predicting Proportions

- We want to predict a proportion (or probability), π, for a categorical variable
  - Fraction of people that die of a heart attack
  - Fraction of molecules that decompose
- Consider the following linear model

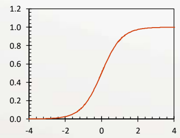
Suppose $\hat{\pi}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$

- Problem: π is constrained to between 0 and 1
  - This model does not force a constraint on $\hat{\pi}$
  - Additionally, the variance is not constant

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Probit Regression

- Instead of a linear model, assume π is sigmoidally shaped
  - Note that virtually every cdf (cumulative distribution function) has a sigmoidal shape



- Example: Probit model
  - Assume π is Bernoulli distributed, then

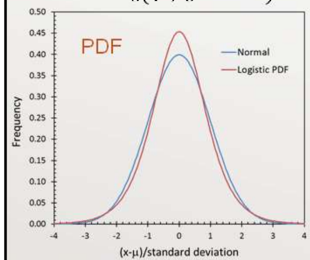
$$probit(\pi) = NormInv(\pi) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$$

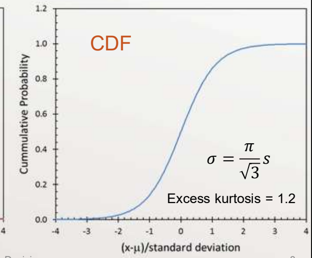
Link function

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Logistic Distribution

$$pdf \quad \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-(x-\mu)/s}\right)^2} \qquad cdf = \frac{1}{1+e^{-(x-\mu)/s}}$$





THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Logistic Distribution

- The logistic cdf can be analytically inverted

$$cdf = \frac{1}{1+e^{-(x-\mu)/s}} \qquad \ln\left(\frac{cdf}{1-cdf}\right) = \frac{x-\mu}{s}$$

- We'll model π, our output fraction of a binary variable, with this cdf
- This is called logistic regression

$$\mathrm{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = link\ function$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Logistic Regression

- Instead of the probit model's inverse normal function, use the inverse logistic cdf (and assume π is Bernoulli distributed)

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$$

- Note: $odds = \dfrac{probability\ of\ success}{probability\ of\ failure} = \dfrac{\pi}{1-\pi}$

Inverting, $\pi(x_1, \ldots) = \dfrac{e^{\beta_0+\beta_1 x_{1i}+\cdots}}{1+e^{\beta_0+\beta_1 x_{1i}+\cdots}}$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Uses of Logistic Regression

- Predict binary outcome events
  - Patient mortality after surgery (as a function of age, prior health indicators, sex)
  - Probability of contracting a certain disease (as a function of age, ethnicity, fitness, sex)
  - Probability customer will make a purchase (as a function of income, age, sex, neighborhood)
  - Probability of defaulting on a mortgage (as a function of price, income, interest rate, mortgage type)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lecture 58: What have we learned?

- What are the three requirements of a generalized linear model?
- What are some common distribution/link function pairs?
- What is the distribution/link function pair for logistic regression?
- Name three examples of where you might want to use a logistic regression