MoreSteam.com

The Transactional Dilemma: Understanding Regression with Attribute Data

Smita Skrivanek

August 26, 2010

Agenda

- Welcome
- Introduction of MBB Webcast Series
  - Larry Goldman, MoreSteam.com
- The Transactional Dilemma: Understanding Regression with Attribute Data
  - Smita Skrivanek, MoreSteam.com
- Open Discussion and Questions

MoreSteam.com®

MoreSteam.com – Company Background

- Founded 2000
- Over 250,000 Lean Six Sigma professionals trained
- Serving 45% of the Fortune 500
- First firm to offer the complete Black Belt curriculum online
- Courses reviewed and approved by ASQ
- Registered education provider of Project Management Institute (PMI)

Select Customers:

CardinalHealth

Nationwide

Honeywell

ROLLS ROYCE

IBM

Quest Diagnostics

tyco

XEROX

lenovo
NEW WORLD. NEW THINKING.

TEXTRON

Master Black Belt Program

- Offered in partnership with Fisher College of Business at The Ohio State University
- Employs a Blended Learning model with world-class instruction delivered in both the classroom and online
- Covers the MBB Body of Knowledge with topics ranging from advanced *DOE* to *Leading Change* to *Finance for MBBs*
- Go to http://www.moresteam.com/master-black-belt.cfm for more information about curriculum, prerequisites, and schedule

Today's Presenter

Smita Skrivanek

Principal Statistician, MoreSteam LLC

- Develops content, software functions, exam question banks and simulation games for MoreSteam's diverse client base
- EngineRoom® Product Manager
- Masters in Applied Statistics from The Ohio State University and a MS from Mumbai University, India

The ‘Dilemma’

Examples of categorical responses:

Delinquent payments

Return purchases

Billing errors

Brand preferences

Delayed shipments

Customer satisfaction ratings

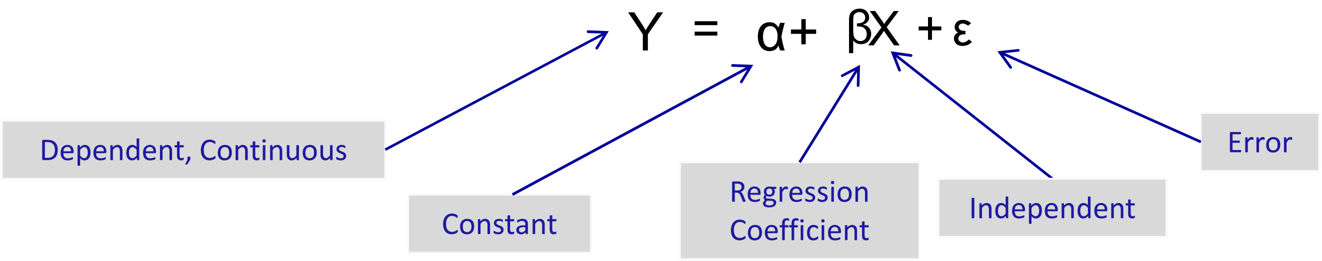
☆ It is unnecessary *(and often inappropriate)* to use continuous data methods on categorical responses. Logistic regression is a more intuitive and powerful method in such cases.

Objectives

- What is binary logistic regression (BLR)
- When is a logistic approach appropriate (and why)
- Probabilities, Odds and Odds Ratios
- Logistic model interpretation
- Methods used to estimate model coefficients, evaluate model fit and compare alternative models
- How to approach the teaching of logistic regression to students

The Regression Model

Ordinary Least Squares (OLS) Regression:



$$Y = \alpha + \beta X + \varepsilon$$

Dependent, Continuous

Constant

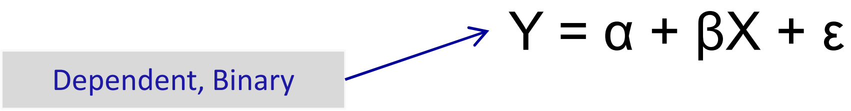
Regression Coefficient

Independent

Error

$$-\text{ infinity} < E(Y|x) = \alpha + \beta x < \textit{infinity}$$

Logistic/Logit Regression:



$$Y = \alpha + \beta X + \varepsilon$$

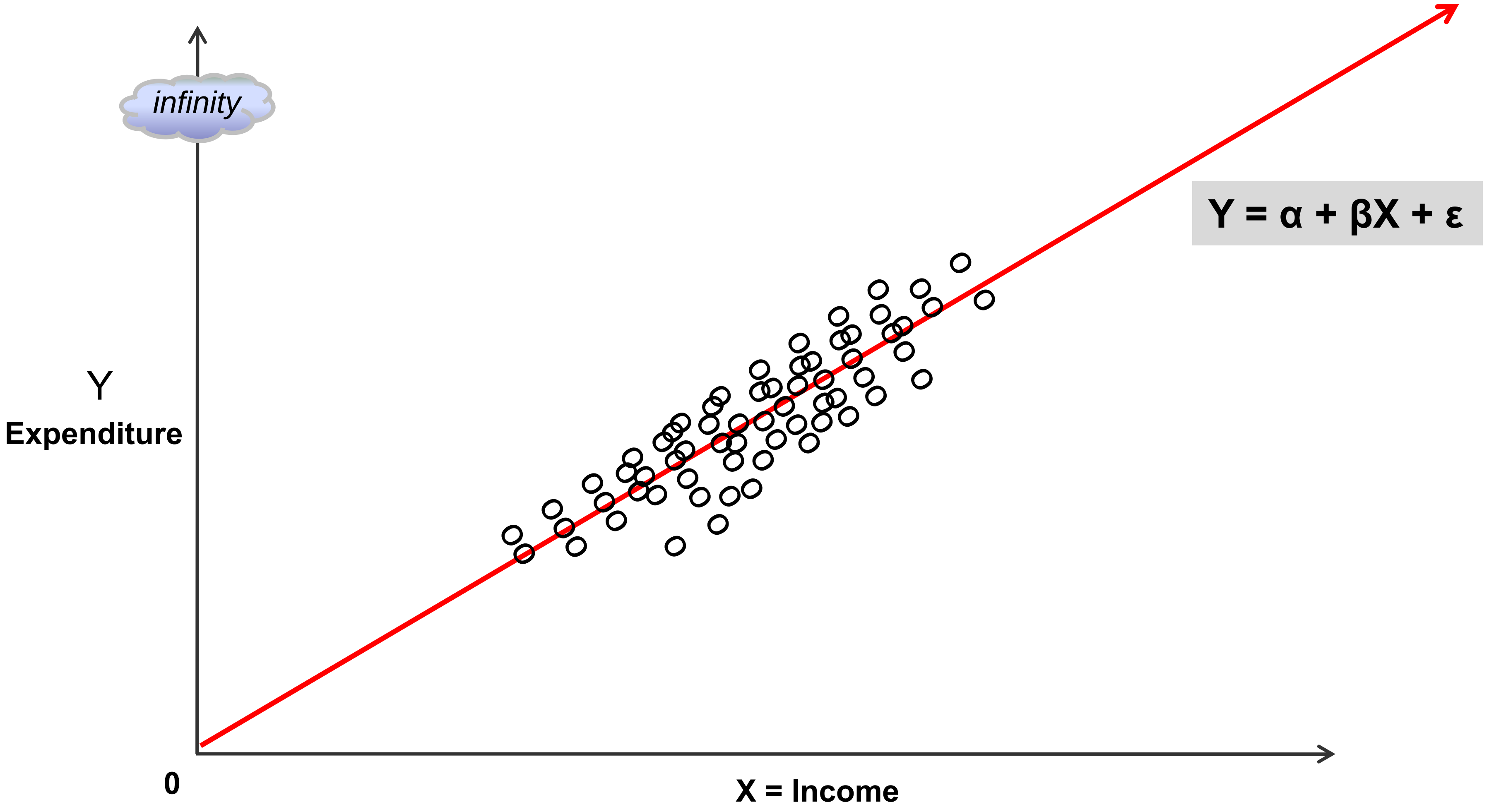
Dependent, Binary

$$0 < E(Y|x) = P(Y|x) =<\alpha 1+ \beta x$$

$$-\text{infinity} < \alpha + \beta x < \text{infinity}$$

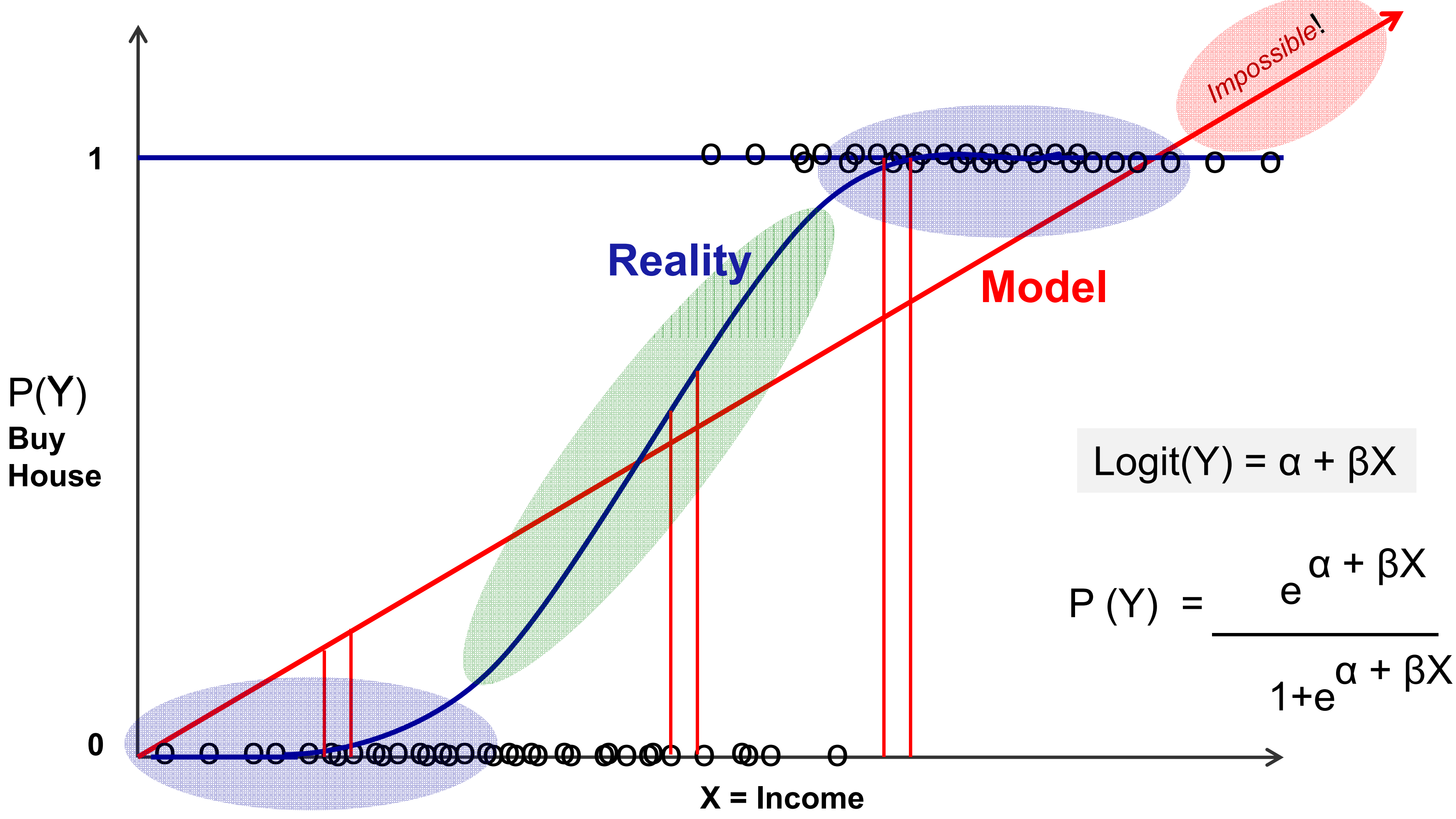
MoreSteam.com®

OLS vs. BLR – the OLS Model



$$Y = \alpha + \beta X + \varepsilon$$

OLS vs. BLR – Where We Go Wrong



Logit(Y) = α + βX

$$P(Y) = \frac{e^{\alpha + \beta X}}{1+e^{\alpha + \beta X}}$$

OLS vs. BLR – Initial Comparison

Ordinary Least Squares (OLS)	Binary Logistic Regression (BLR)
• Independent data	• Independent data
• Errors are normal, *with*	• Errors are bernoulli, *with*
• Constant variance (σ^2)	• Non-constant variance [$p_i(1-p_i)$]
• Y is linear in the predictors	• Logit(Y) is linear in the predictors

MoreSteam.com®

Probabilities and Odds: $\text{Logit}(Y) = \alpha + \beta X + \varepsilon$

$$P(\text{Event}) = \frac{\#\ \text{Events}}{\#\ \text{Total}}$$

$$\text{Odds}(\text{Event}) = \frac{\#\ \text{Events}}{\#\ \text{Non-Events}}$$

$$P(\text{Non-Event}) = \frac{\#\ \text{Non-Events}}{\#\ \text{Total}}$$

$$\text{Odds}(\text{Non-Event}) = \frac{\#\ \text{Non-Events}}{\#\ \text{Events}}$$

$$P(\text{Event}) + P(\text{Non-Event}) = 1$$

$$\text{Odds}(\text{Event}) * \text{Odds}(\text{Non-Event}) = 1$$

$$\boxed{\text{Odds}(\text{Event}) = \frac{\#\ \text{Events}}{\#\ \text{Non-Events}} = \frac{\#\text{Events}|\ \#\text{Totals}}{\#\text{Non-Events}|\ \#\text{Totals}} = \frac{P(\text{Event})}{P(\text{Non-Event})}}$$

Probabilities and Odds: $\text{Logit}(Y) = \alpha + \beta X + \varepsilon$

	Own Car?		
Gender	**Yes**	**No**	**Total**
Male	62	157	219
Female	48	185	233
Total	110	342	452

$$P(\text{Own}) = \frac{110}{452} = 0.24$$

$$\text{Odds}(\text{Own}) = \frac{110}{342} = \frac{0.24}{0.76} = 0.32$$

$$P(\text{Don't own}) = \frac{342}{452} = 0.76$$

$$\text{Odds}(\text{Don't own}) = \frac{342}{110} = \frac{0.76}{0.24} = 3.1$$

$$0.24 + 0.76 = 1$$

$$0.32 * 3.1 = 1$$

The Odds Ratio: A Measure of Association

$$\text{Odds(Event |Group 1)} = \frac{\text{P(Event in Group 1)}}{\text{P(Non-Event in Group 1)}}$$

$$\text{Odds(Event |Group 2)} = \frac{\text{P(Event in Group 2)}}{\text{P(Non-Event in Group 2)}}$$

$$\text{Odds Ratio (Event)} = \frac{\text{Odds(Event/Group 1)}}{\text{Odds(Event/Group 2)}}$$

X = Categorical:

Odds ratio = the increase/decrease in the odds of the event in group 1 relative to group 2

X = Continuous:

Odds ratio = the increase/decrease in the odds of the event for a unit increase in X

Odds Ratio of Owning: Logit(Y) = α + βX + ε

	Own Car?		
Gender	**Yes**	**No**	**Total**
Male	62	157	219
Female	48	185	233
Total	110	342	452

$$\text{Odds(Own/Male)} = \frac{62}{157} = 0.39$$

$$\text{Odds(Own/Female)} = \frac{48}{185} = 0.26$$

$$\text{Odds Ratio(Own)} = \frac{0.39}{0.26} = 1.52$$

Males have 1.52 times greater odds of owning a car than females.

Note:

Log(1.52) = 0.418

Odds and Odds Ratios: Logit(Y) = α + βX + ε

Event |Success: Y = 1 Non-Event | Failure: Y = 0

X = x:

Logit(Y =1) = Log-odds(Y = 1) = α + βx

$$Odds\ (Y = 1) = e^{(\alpha + \beta x)}$$

X = Binary (0, 1):

X = 1: $Odds\ (Y = 1 \mid X = 1) = e^{(\alpha + \beta * 1)} = e^{\alpha + \beta}$

X = 0: $Odds\ (Y = 1 \mid X = 0) = e^{(\alpha + \beta * 0)} = e^{\alpha}$

Odds Ratio (Y=1|X) = $\dfrac{\text{Odds(Y=1 | X=1)}}{\text{Odds(Y=1| X=0)}} = \dfrac{e^{\alpha+\beta}}{e^{\alpha}} = \dfrac{\cancel{e^{\alpha}} e^{\beta}}{\cancel{e^{\alpha}}} = e^{\beta}$

OLS vs. BLR : $\text{Logit}(Y) = \alpha + \beta X + \varepsilon$

Ordinary Least Squares (OLS)

- -infinity < β < infinity
- β < 0 → negative association
- β > 0 → positive association

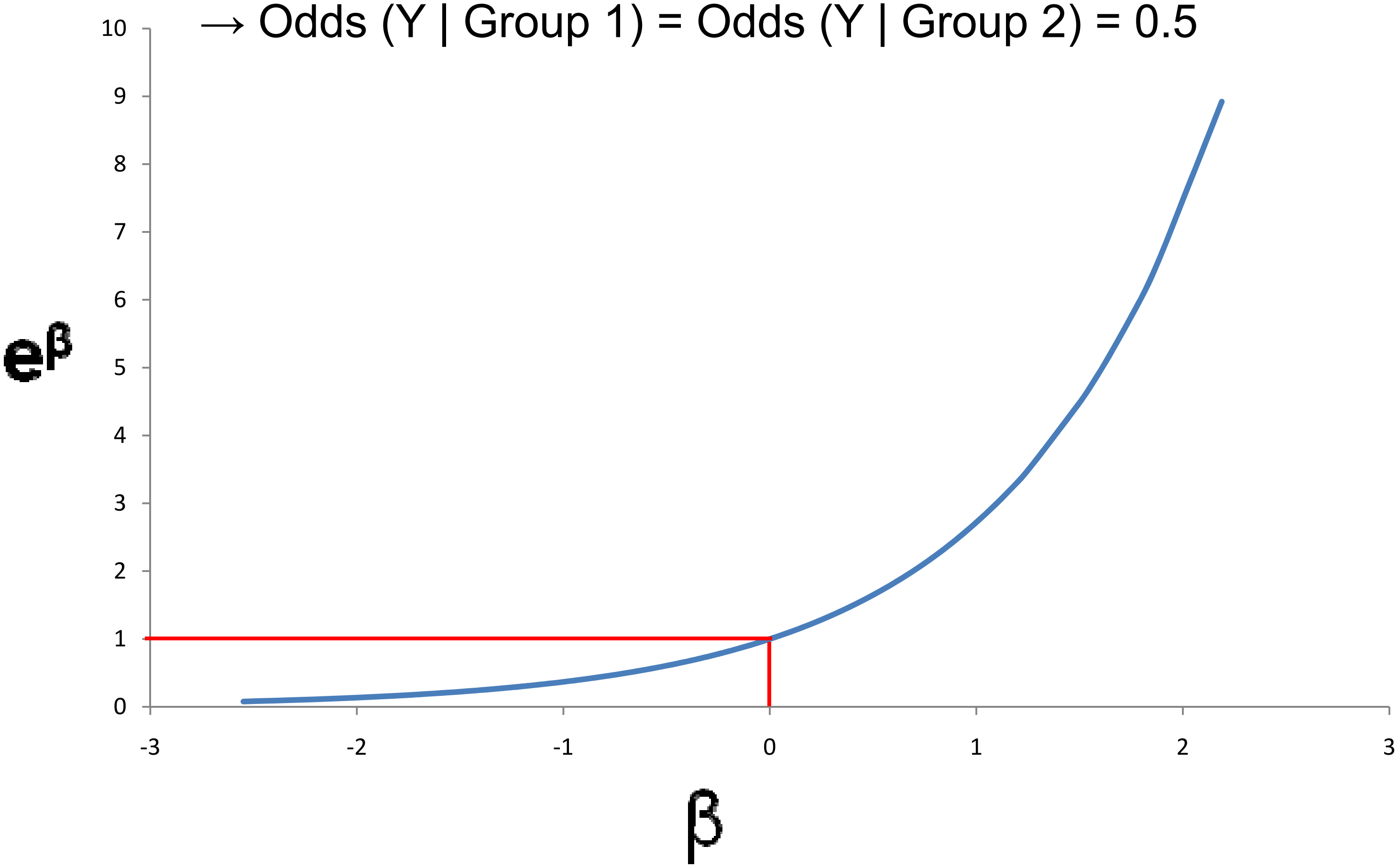
Binary Logistic Regression (BLR)

- 0 < Odds ratio = $\mathbf{e^{\beta}}$ < infinity
- Odds ratio = $\mathbf{e^{\beta}}$ < 1 → decreasing odds
- Odds ratio = $\mathbf{e^{\beta}}$ > 1 → increasing odds

Beta vs. Exp(Beta): Logit(Y) = α + βX + ε

Odds Ratio(Y) = 1

→ Odds (Y | Group 1) = Odds (Y | Group 2) = 0.5



Odds Ratio of Owning: Multiple Predictors

Own car	Coeff (β)	z	P(Z>|z|)
constant	-4.683	-3.18	0.001
income	-0.0102	-0.02	0.986
age	0.246	3.55	0.000
male	0.418	2.02	0.044

Odds Ratio = e^{β}

Income	0.99	A unit increase in income does not change the odds of owning a car.
Age	1.28	A unit increase in age increases the odds of owning a car by 28%.
Male	1.52	Males have a 52% higher odds of owning a car than females

MoreSteam.com®

Estimating the Parameters

OLS Regression uses **Minimum Least Squares** method

- *When applied to a logistic regression model, the estimators lose their desirable statistical properties.*

Logistic regression uses the **Maximum Likelihood** method

- Find values of the parameters α and β which make the probability of observing Y, i.e., P(Y = y) as large as possible.
- "Best" parameters to explain the observed data.

Assessing Fit and Comparing Models

Comparing alternative models

- Does the model which includes the selected variables tell us more about the response variable than a model that does not include those variables?

Assessing Goodness of Fit

- How well does our model 'fit' the observed data (describe the response variable Y)?

Another Example: Late Debt Payments

Do Age Category and/or Home Ownership affect P(Default) and if so, how?

Default	Coeff (β)	Odds ratio (e^{β})
constant	0.4214	
homeowner	-0.2672	0.76
age (<35)	0.1512	1.16
age (35-64)	0.2704	1.31

Qstn: *What is the estimated probability that a renter aged 30 years will default on a loan payment?*

Log-Odds (Default) = 0.4214 – 0.2672*(0) + 0.1512*(1) + 0.2704*(0) = 0.5726

Odds (Default) = $e^{0.5726}$ = 1.773

$$P(Default) = \frac{e^{0.5726}}{1+e^{0.5726}} = 0.64$$

How to Teach Logistic Regression

- Keep it **Simple**.
- Use **analogies** between ordinary least squares (OLS) regression and binary logistic regression (BLR).
- Introduce BLR with a **single independent variable**, as is used to teach OLS.
- Illustrate concepts with **contingency tables**.
- Link logistic regression concepts to the **interpretation** of statistical computer outputs.

References

- *Logistic Regression Models*: Joseph M. Hilbe
- *Applied Logistic Regression*: David W. Hosmer, Stanley Lemeshow
- *Teaching, Understanding and Interpretation of Logit Regression*: Anthony Walsh (Teaching Sociology, Vol. 5, No. 2)
- *Using and Interpreting Logistic Regression*: Ilsa L. Lottes, Alfred DeMaris, Marina A. Adler (Teaching Sociology, Vol. 24, No. 3)

Thank you for joining us

Resource Links and Contacts

Questions? Comments? We'd love to hear from you.

Smita Skrivanek, Principal Statistician - MoreSteam.com
sskrivanek@moresteam.com

Larry Goldman, Vice President Marketing - MoreSteam.com
lgoldman@moresteam.com

Additional Resources:

Archived presentation, slides and other materials:
http://www.moresteam.com/presentations/webcast-regression-analysis-attribute-data.cfm

Master Black Belt Program: http://www.moresteam.com/master-black-belt.cfm