SUPERCAL
IFRAGILIS
TICEXPIAL
IDOCIOUS

Logistic Regression

for classification

Overview

Why not ordinary linear regression?

The logistic regression model: modeling odds of events

Uses for predictive task

Uses for explanatory task

Baby Example: Beer Preference

Beer manufacturer wants to understand what demographics separate **light beer** drinkers from **regular beer** drinkers

Task and Data

Task: Profile beer drinkers in terms of demographics

Beer data & analysis.xls

Two classes
4 explanatory variables

100 records

Demographics (predictors)				output
Gender	Married	Income	Age	Preference
0	1	$39,942	21	Light
0	0	$33,088	22	Light
0	0	$30,841	24	Light
0	1	$33,700	25	Light
1	1	$42,108	26	Light
1	0	$42,775	27	Light
0	0	$43,593	27	Light
0	0	$39,370	28	Light
0	0	$26,598	29	Light
0	0	$35,406	29	Light
1	1	$58,164	30	Light
1	1	$42,404	30	Light
1	0	$23,234	31	Regular
0	1	$44,558	31	Light
1	1	$40,261	31	Light
0	0	$36,821	32	Light
0	1	$48,259	32	Light
1	0	$37,926	33	Light
1	1	$48,957	33	Light
1	0	$28,513	34	Regular

Why not linear regression?

Code response as

$$Y = \begin{cases} 1 & \text{if Light} \\ 0 & \text{if Regular} \end{cases}$$

Fit the model

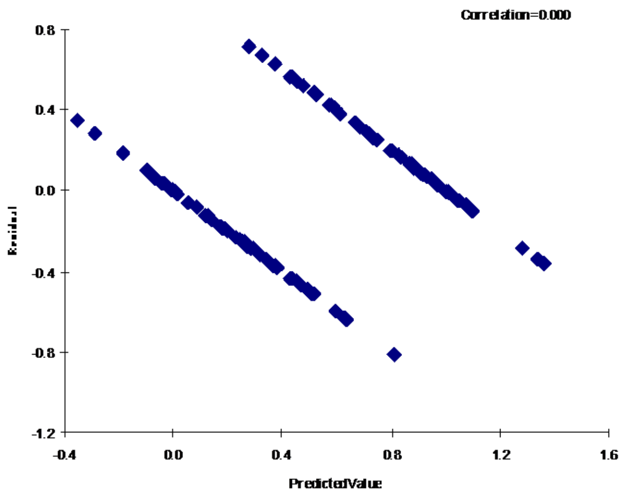
$$Y = \alpha + \beta_1 Gender + \beta_2 Married + \beta_3 Income + \beta_4 Age + \varepsilon$$

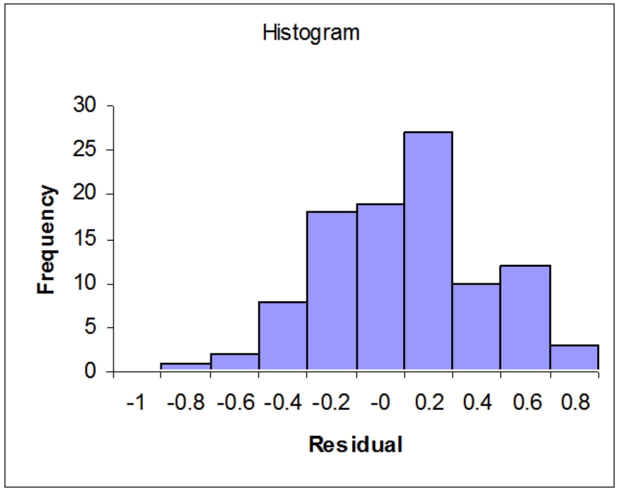
Business Analytics Using Data Mining

ISB

Partial Output

Row Id	Predicted Value	Actual Value	Residual	Gender	Married	Income	Age
1	0.2333295	0	-0.2333295	0	0	$31,779.00	46
2	0.14347264	0	-0.14347264	1	1	$32,739.00	50
3	-0.00633473	0	0.00633473	1	1	$24,302.00	46
4	0.59862394	0	-0.59862394	1	1	$64,709.00	70
5	0.31359163	0	-0.31359163	1	1	$41,882.00	54
6	0.62723779	0	-0.62723779	1	0	$38,990.00	36





Different Formulation

?

Categorical Y $\rightarrow$ continuous Y

How about p = Prob(Y=1)?

p = **probability** that customer prefers light beer

$$p = \alpha + \beta_1 Gender + \beta_2 Married + \beta_3 Income + \beta_4 Age + \varepsilon$$

How about a function of p?

- Range $(-\infty, \infty)$
- Meaningful

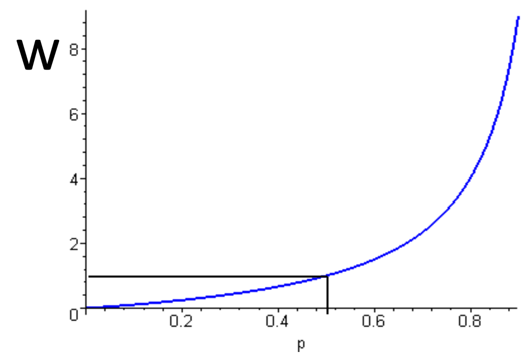
Meaningful functions

Probability of the event Y=1

$$\boxed{p = \mathrm{Prob}(Y=1)}$$

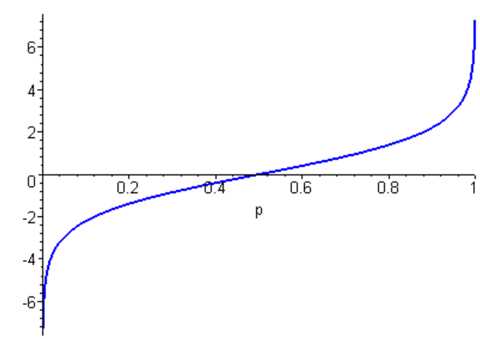
Better: **odds** of the event Y=1

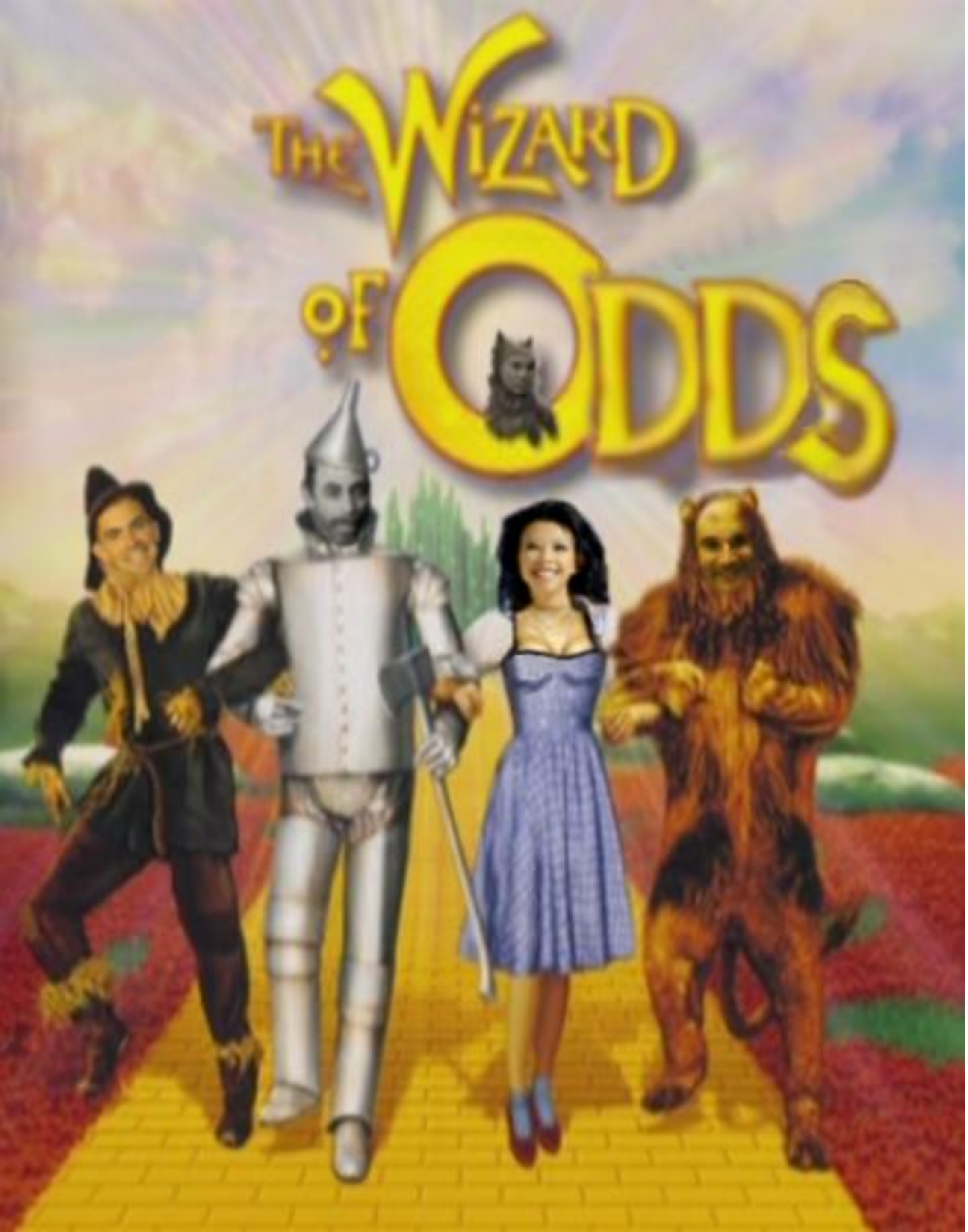
$$\boxed{w = \frac{p}{1-p}}$$



Best: **logit** of the event Y=1

$$\boxed{\mathrm{logit} = \ln(w) = \ln \frac{p}{1-p}}$$




The Wizard of Odds

Probability, odds, logit

Given the *odds* of an event, its probability is:

$$p = \frac{w}{1+w}$$

Given the *logit* of an event, its odds are:

$$w = e^{\text{logit}}$$

and its probability is:

$$p = \frac{e^{\text{logit}}}{1+e^{\text{logit}}} = \frac{1}{1+e^{\text{-logit}}}$$

The logistic regression model

is a **nonlinear** model between Y and predictors

Linear relationship between logit and predictors

$$logit = \alpha + \beta_1 Gender + \beta_2 Married + \beta_3 Income + \beta_4 Age$$

Multiplicative relationship between odds and predictors

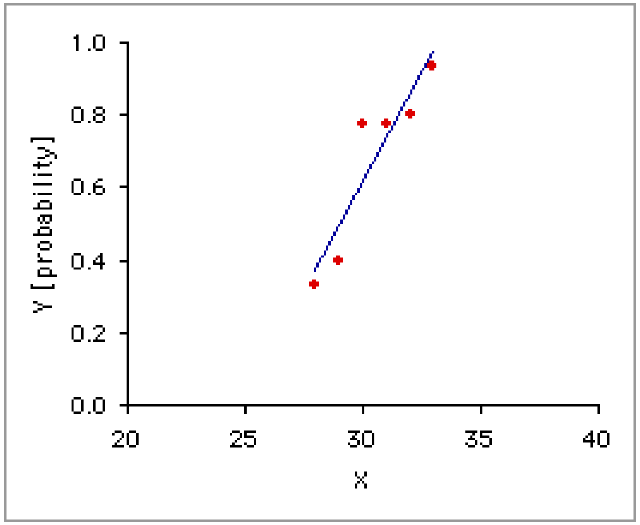
$$odds = exp\{\alpha + \beta_1 Gender + \beta_2 Married + \beta_3 Income + \beta_4 Age\}$$

And…

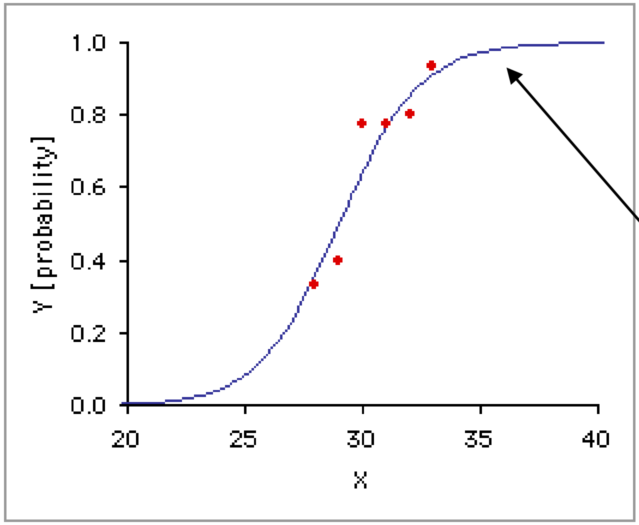
Non-linear relationship between *p* (probability of *Y*=1) and predictors

$$p = \frac{1}{1 + e^{-(\alpha + \beta_1 \mathrm{GENDER} + \beta_2 \mathrm{MARRIED} + \beta_3 \mathrm{INCOME} + \beta_4 \mathrm{AGE})}}$$

Plotting the logistic relationship (single predictor)



Linear



Logistic

S-shaped / sigmoidal function

Estimating the model

Estimate α, β_1, β_2, β_3, β_4

How?

What to use for Y column?

Cannot use least squares (like linear regression)

Instead: Maximum Likelihood Estimation
(find estimates that maximize the chance of obtaining the data that we see); done iteratively

Personal Loan Offer

(UniversalBank.csv)

Outcome variable: accept bank loan (0/1)

Predictors: Demographic info, and info about their bank relationship

Classifying

To classify an observation:

1. Use estimated model to obtain *logit*

2. Estimate p = probability that Y=1

$$p = \frac{e^{\text{logit}}}{1+e^{\text{logit}}} = \frac{1}{1+e^{\text{-logit}}}$$

3. Use cutoff value to determine class membership

Variable Selection

Like in linear regression: stepwise, forward selection, backward elimination, best subsets

- *XLMiner:* "best subsets" button

Metrics for comparing models:

RSS = residual sum of squares (smaller=better)

Cp (should be ≅ # predictors)

Perfectly separable data

Remember perfect multicollinearity in linear regression?

If all records in class Y=0 have X_2<3.5, and all records in class Y=1 have X_2 > 3.5,

The dataset is said to be perfectly separable using X_2.

Trivial classification?

Is X_2 available at time of prediction?

Software: estimation procedure for logistic regression cannot proceed (error message)

Advantages and weaknesses

The Good

- Model-based (little data needed)
- Useful for explaining and predicting
- Interpretable
- Variable selection
- (Similar to linear regression)

The Bad

- Model-based (specify exact relationship)
- Global relationship
- (Similar to linear regression)