

Victoria BC SASUG



"A Choice of Prediction
Rules in Logistic Regression
Models"

Melvin Ott, PhD

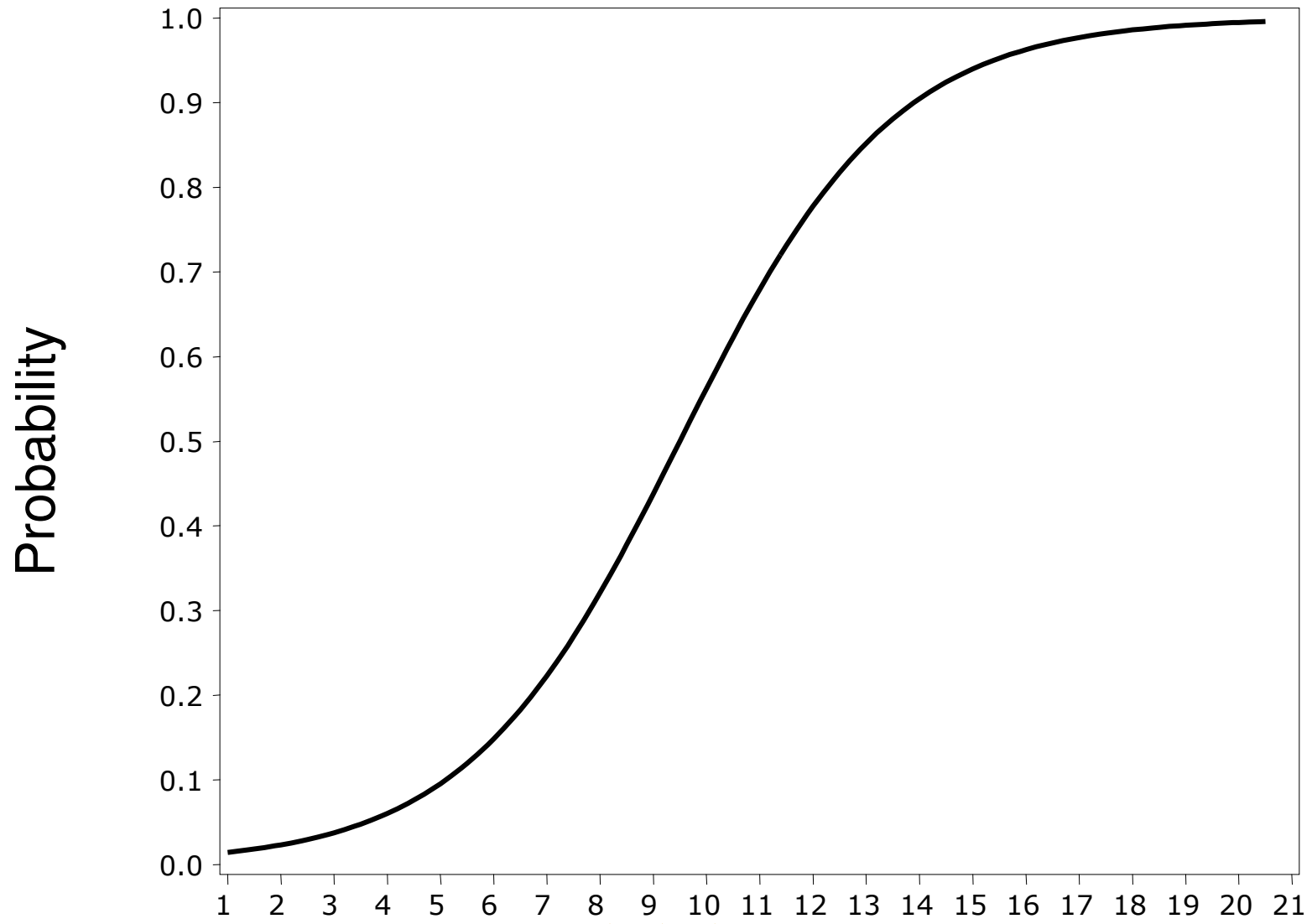
www.melvinott.com

May 31, 2006

What Does Logistic Regression Do?

- ◆ The logistic regression model uses the predictor variables, which can be categorical or continuous, to **predict the probability** of specific outcomes.
- ◆ Logistic regression is designed to describe probabilities associated with the values of the response variable.
- ◆ Because you are modeling probabilities, a continuous linear regression model would not be appropriate. One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1. **Logistic regression models ensure that the estimated probabilities are between 0 and 1.**
- ◆ And, the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an **S-shaped curve**.

Logistic Regression Curve



X

Logit Transformation

Logistic regression models transform probabilities to values called *logits*.

$$\text{logit}(p_i) = \log \left(\frac{p_i}{1 - p_i} \right)$$

where

i indexes all cases (observations)

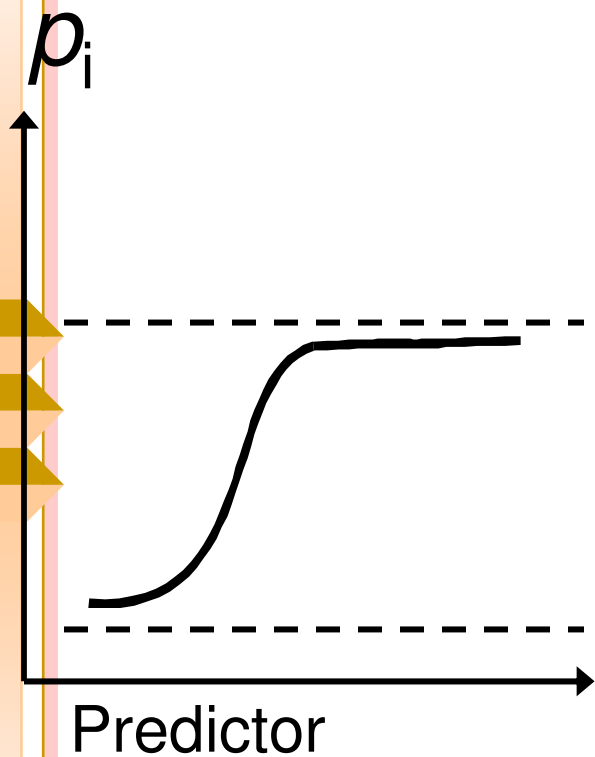
p_i is the probability the event (a sale, for example) occurs in the i^{th} case

\log is the natural log (to the base e).

A logistic regression model applies a transformation to the probabilities. The probabilities are transformed because the relationship between the probabilities and the predictor variable is nonlinear.

The logit transformation ensures the model generates estimated probabilities between 0 and 1.

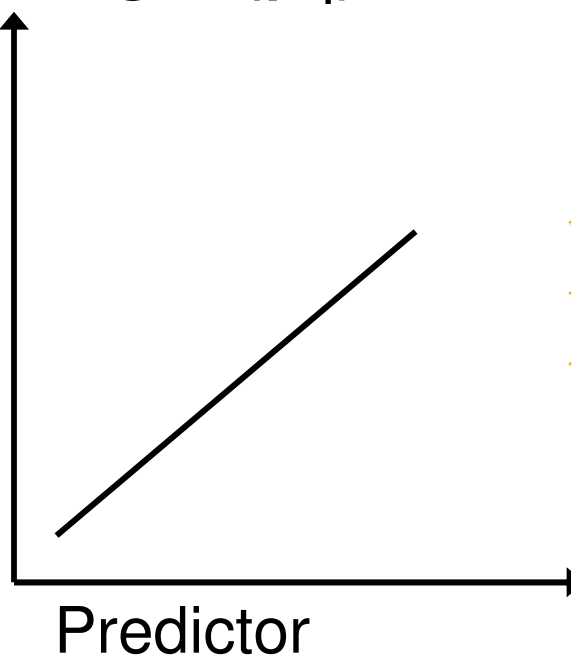
Assumption



Logit
Transform



Logit (p_i)



Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

Where $\text{logit}(p_i)$ is the logit transformation of the probability of the event.

β_0 is the intercept of the regression line.

β_1 is the slope of the regression line.

ε_i error (residual) associated with each observation.

For a binary outcome variable, the linear logistic model with one predictor variable has the form above.

Unlike linear regression, the categorical response is not normally distributed and the variances are not the same. Also, logistic regression usually requires a more complex iterative estimation method called maximum likelihood to estimate the parameters than linear regression does. This method finds the parameter estimates that are most likely to occur given the data. It accomplishes this by maximizing the likelihood function that expresses the probability of the observed data as a function of the unknown parameters.

Some Examples

- ◆ Banking-Estimate the probability of loan failure based on financial and other data
- ◆ Marketing-Estimate the probability that someone will buy my product based on price and other demographic data
- ◆ Health Care-Estimate the probability that a patient has a medical condition based on some clinical test
- ◆ And on and on...

Issues of Concern

- ◆ “Hits” and “False Alarms” per Swets 1988, also known as true-positive’s and false-positive’s
- ◆ Measure of accuracy for diagnostic tools is the proportion of area beneath the ROC curve, per Hanley and McNeil 1982, this measures the probability that in randomly paired individuals, the logistic model will allow them to be correctly identified. SAS will count concordant, discordant and tied pairs for the model.
- ◆ However, in the examples of the previous slide, I really want to know, the probability for a single individual that they will be correctly identified, and what to do with this information...so, how do I use my logistic regression model?

Prediction Rules

- ◆ Neter, et al in his text Applied Linear Regression Models, suggests three rules:
 1. “Use .5 as the cutoff.” If $p > .5$ then predict Yes; otherwise predict No.
 2. “Find the best cutoff for the data set on which the multiple logistic regression model is based.” Requires evaluating different cutoffs. Then, apply the rule on the n cases to determine the proportion of incorrect predictions. The cutoff for which the proportion of incorrect predictions is lowest is the one then selected.
 3. Use prior probabilities and costs of incorrect predictions in determining the cutoff.

ROC Curve

◆ View ROC



Demo SAS EG for Health Data

"A Choice of Prediction Rules in Logistic Regression Models"

- ◆ Follow Neter?
- ◆ Look at the application and the consequences of the decision?
- ◆ Find optimal operating point (OOP) from ROC curve per Halpern, Gallop and others
- ◆ **Suggest, look at the SAS EG predicted file and talk to a content expert/decision maker to arrive at a prediction rule.**

Prediction Rules per Mel

- ◆ Market Research...0.5
- ◆ Health Care...

Suggest high probability cutoff for **immediate treatment** ($p_2 \geq 0.90$), low probability for **rule out** ($p_1 \leq 0.10$), with action for range between of **wait and retest** ($p_1 < p_3 < p_2$).

Establish p_1 and p_2 through the SAS EG predicted file and discussion with content expert.

Summary

- ◆ Estimate probability using logistic regression and SAS EG
- ◆ Examine ROC & area under the curve
- ◆ Establish p_1 and p_2 using predicted file and content expert
- ◆ Other models may predict catastrophic outcome if medical condition is true