Logistic Regression

semicolon

Learning Outcomes

- Participants should understand how logistic regression works

Introduction

- Logistic Regression is used to fit a curve to data in which the dependent variable is binary, or dichotomous
  - We might want to predict response to treatment, where we might code survivors as 1 and those who don’t survive as 0
- A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables

Example

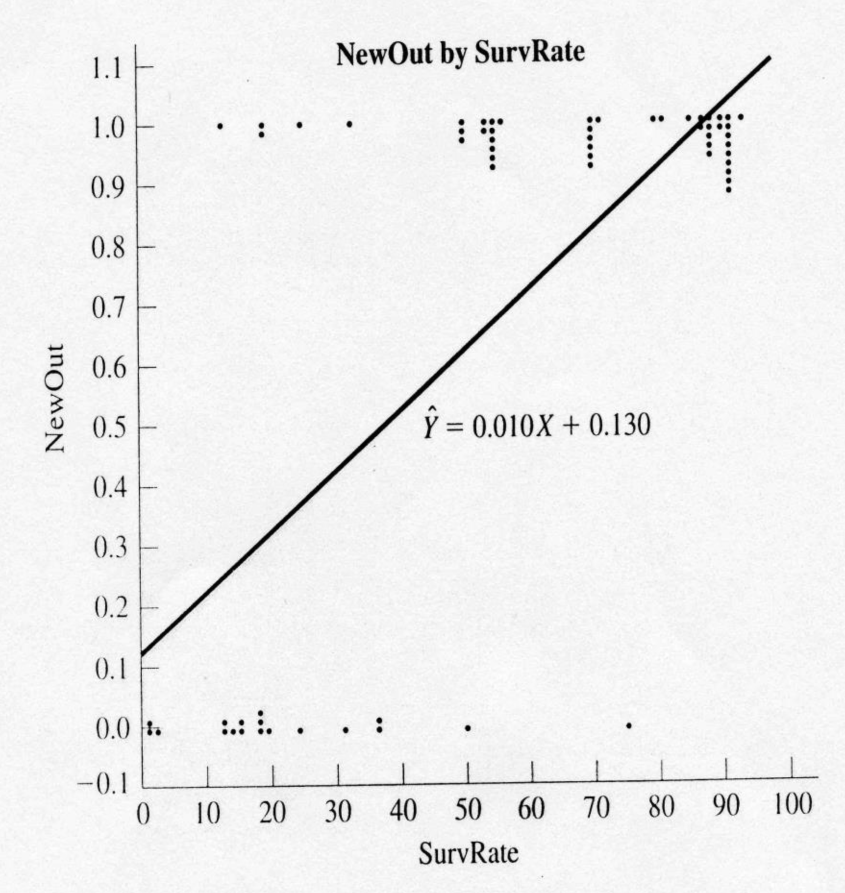


FIGURE 15·7 *Outcome as a function of SurvRate*

- Observations:
- For each value of SurvRate, the number of dots is the number of patients with that value of NewOut

Regression:
Standard linear regression

A Better Solution

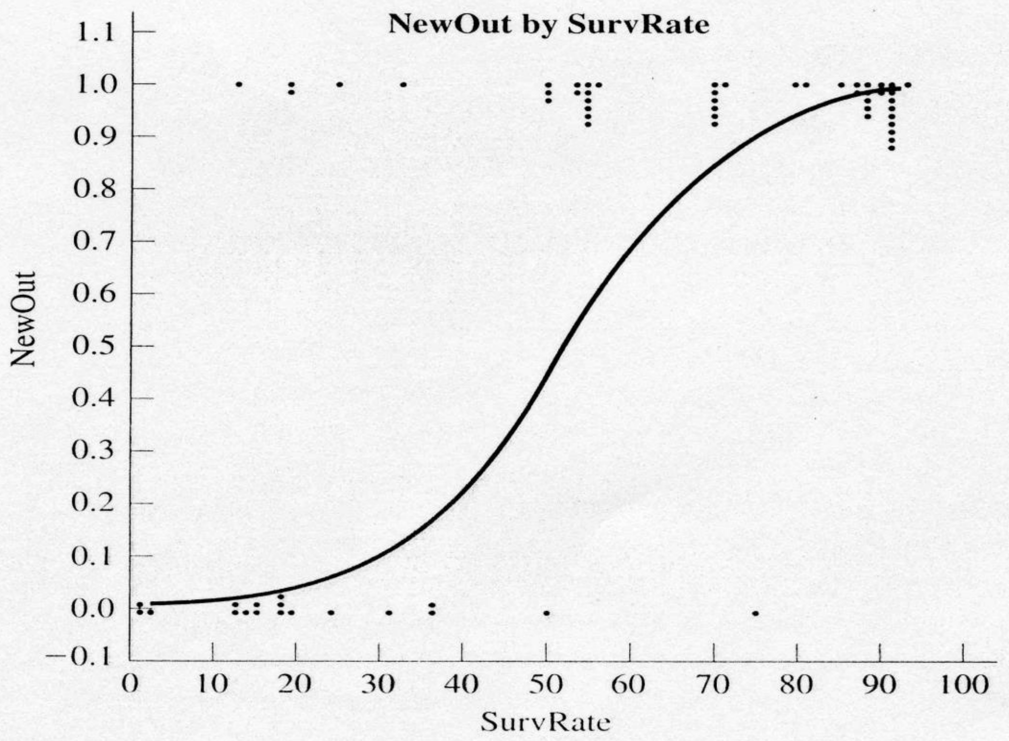


FIGURE 15.8 *More appropriate regression line for predicting outcome*

- Regression Curve:
- Sigmoid function!
- (bounded by asymptotes y=0 and y=1)

Odds

- Given some event with probability p of being 1, the odds of that event are given by:

$$\boxed{\text{odds} = p \,/\, (1 - p)}$$

- Consider the following Financial data

		Loan Default		
		Yes	No	Total
Gender	Male	402	3614	4016
	Female	101	345	446
		503	3959	

- The odds of defaulting if you are a Male is:

  Prob.Yes/(1–Prob.Yes) = (402/4016) / (1 - (402/4016)) = 0.1001 / 0.8889 = 0.111

Odds Ratio

- The odds of not defaulting on Loan if you are a male is the reciprocal of this:
  - 0.8999/0.1001 = 8.99
- Now, for the Female group
  - odds(default) = 101/345 = 0.293
  - odds(not default) = 345/101 = 3.416
- When we go from Male to Female, the odds of defaulting nearly triple:
  - Odds ratio: 0.293/0.111 = 2.64
  - 2.64 times more likely to default on loan if you are a female.

Logit Transform

- Equivalent forms of the logistic regression model:
- Logit form

Probability form

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 X$$

$$p=\frac{e^{b_0+b_1 X}}{1+e^{b_0+b_1 X}}$$

Logit to Prob. Transform

- Let x1,⋯,xk be a set of predictor variables. Then the logistic regression of Y on x1,⋯,xk estimates parameter values for β0,β1,⋯,βk via maximum likelihood method of the following equation

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- Exponentiate and take the multiplicative inverse of both sides,

$$\frac{1-p}{p} = \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Logit to Prob. Transform

- Simplifying further

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

- Change 1 to a common denominator,

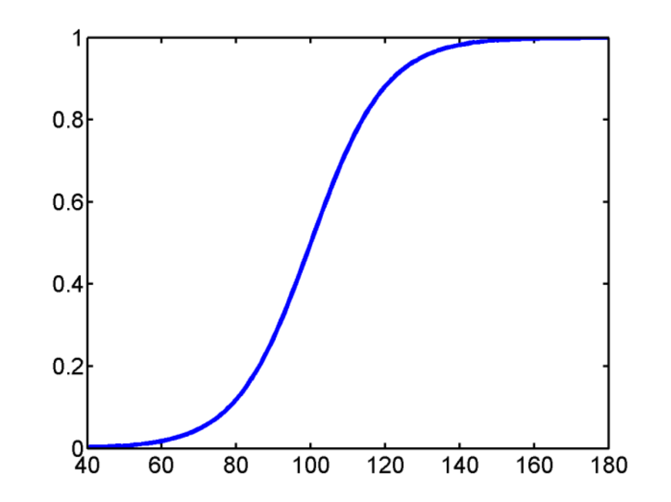
$$\frac{1}{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + 1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

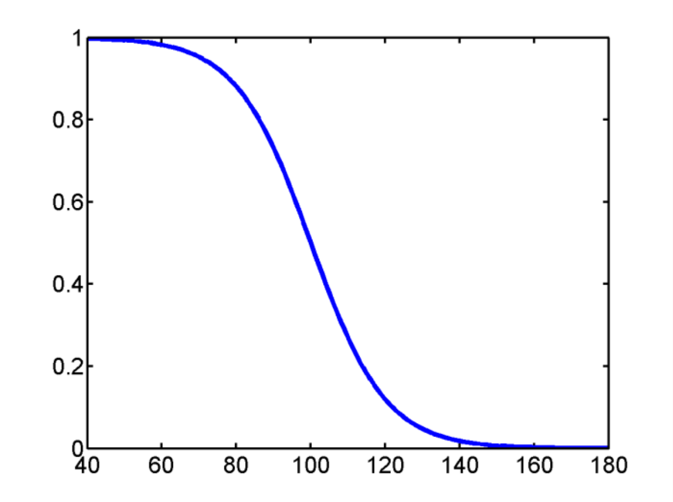
- Finally, take the multiplicative inverse

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Logistic Response Function

- When the response variable is binary, the shape of the response function is often sigmoidal:





Example

- Suppose a Loan Default model yields:
  - log odds = –2.6837 + 0.0812 Age
- Consider a customer with Age = 40
  - log odds = –2.6837 + 0.0812(40) = 0.5643
  - odds = $e^{0.5643}$ = 1.758
  - Customer is 1.758 times more likely to default
- Consider another Customer with Age = 41
  - log odds = –2.6837 + 0.0812(41) = 0.6455
  - odds = $e^{0.6455}$ = 1.907
  - Customers' s odds are 1.907/1.758 = 1.0846 times (or 8.5%) higher than those of the previous customer

Logistic Regression in R

- Example 2: Using the “Student” data, we want determine the factors that affect “hon” (indicating if a student is in an honors class or not.)

- dnt<-read.csv(file.chose(), header=T)
- mod1<-glm(hon~female+read+math,data = dnt,family = binomial)
- summary(mod1)$coeff

Logistic Regression in R

- Estimate Std. Error z value Pr(>|z|)
- (Intercept) -11.77024556 1.71067745 -6.880459 5.966014e-12
- female 0.97994800 0.42162622 2.324210 2.011422e-02
- read 0.05906323 0.02655280 2.224369 2.612361e-02
- math 0.12295888 0.03127553 3.931472 8.442732e-05

We interpret logistic regression with respect to the odds, to obtain the odds for each independent variable, we take the exponent of the Estimate

Logistic Regression in R

- **The** odds of getting into an honors class for females (**female** = 1)over the odds of getting into an honors class for males (**female** = 0) is exp(.979948) = 2.66,   In terms of percent change, we can say that the odds for females are 166% higher than the odds for males.

- For maths, we will see 13% increase in the odds of getting into an honors class for a one-unit increase in math score since exp(.1229589) = 1.13

Questions ??

semicolon

Reference

- All IT eBooks. “Learn R for Applied Statistics - PDF EBook Free Download.” *Allitebooks.In*, 13 Feb. 2019, www.allitebooks.in/learn-r-for-applied-statistics/.

semicolon