WHY LOGISTIC REGRESSION?

- There are many important research topics for which the dependent variable is "limited" i.e. it has known but limited states (2 or more) and independent variables are continuous or discrete and may not have any special distribution requirement.

- For example:
  - A researcher wants to evaluate the influence of grade point average, test scores and curriculum difficulty on the outcome variable of admission (or no admission) to a particular university.
  - Here, outcome is Dichotomous.

WHY LOGISTIC REGRESSION?

- For such problems, researcher can employ a technique known as **Logistic Regression or Logit transformation.**
  - Name is actually misleading as it is a technique for classification and not regression.
    - Word 'Regression' is utilized because regression is used to fit a curve to data in which the dependent variable is binary, or dichotomous.
- It is a statistical technique that is most useful for understanding the influence of **several independent variables on a single dichotomous outcome variable** (can also be extended to polytomous outcome variables).

LOGISTIC REGRESSION

- When this technique is employed for Dichotomous or two class classification/prediction it is known as Binary Logistic Regression or just Logistic Regression, by default.
- Its prediction of outcome based on several predictor variables that may be mix of categorical and continuous natures.

LOGISTIC REGRESSION vs LINEAR REGRESSION

- In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values.
  - Linear regression is line fitting.

- In logistic regression, the outcome (dependent variable) has only a limited number of possible values.
  - Logistic regression is curve fitting.

EXAMPLE

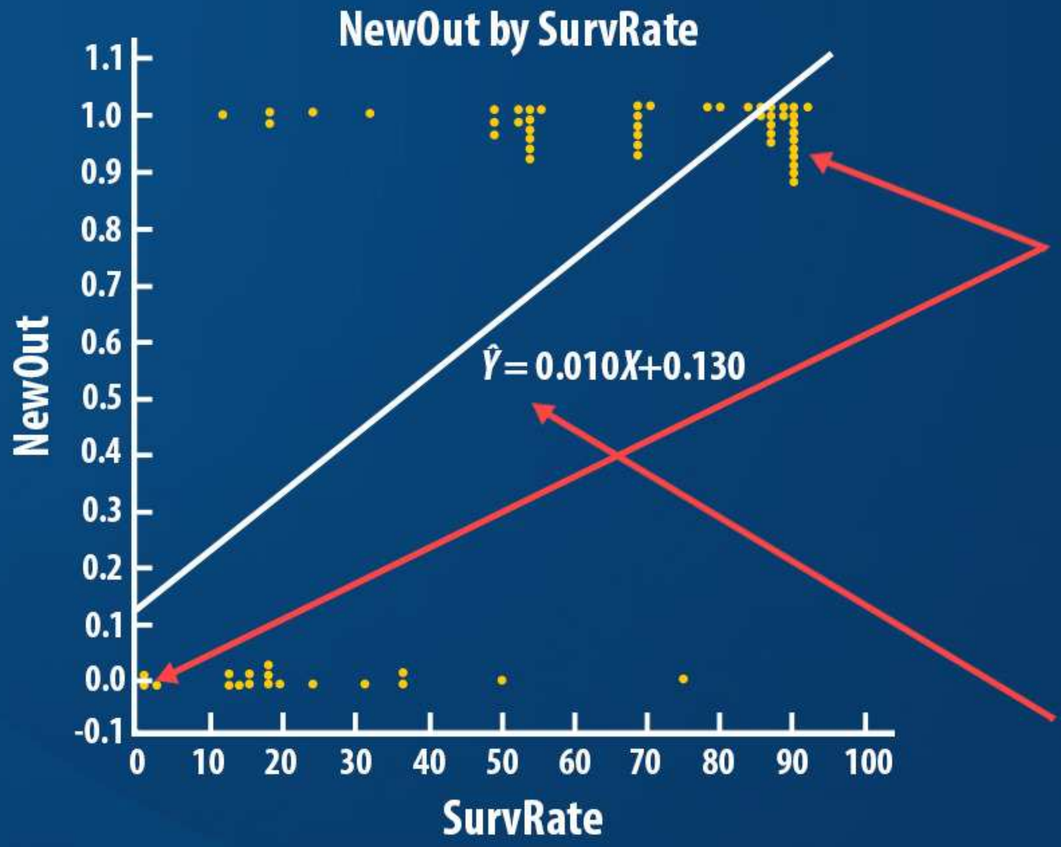


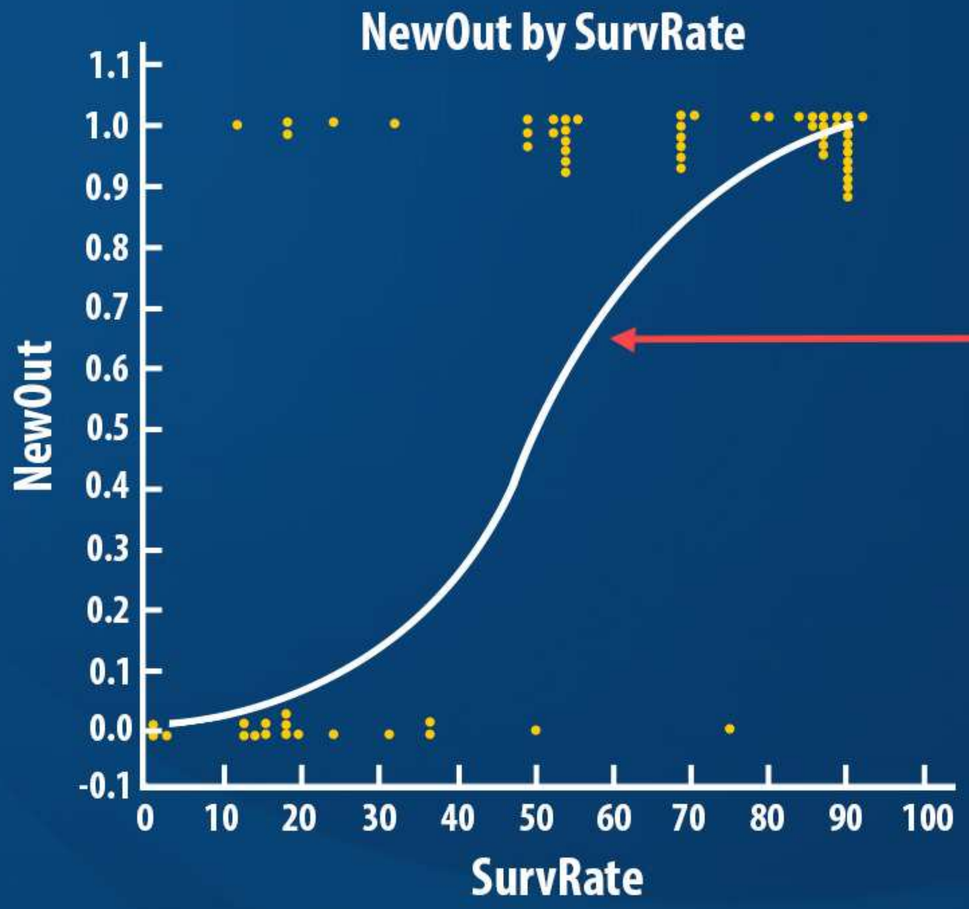
Figure I 5.7 *Outcome as a function of SurvRate*

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

Regression:
Standard Linear Regression

Problem: Extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0,1].

A BETTER SOLUTION



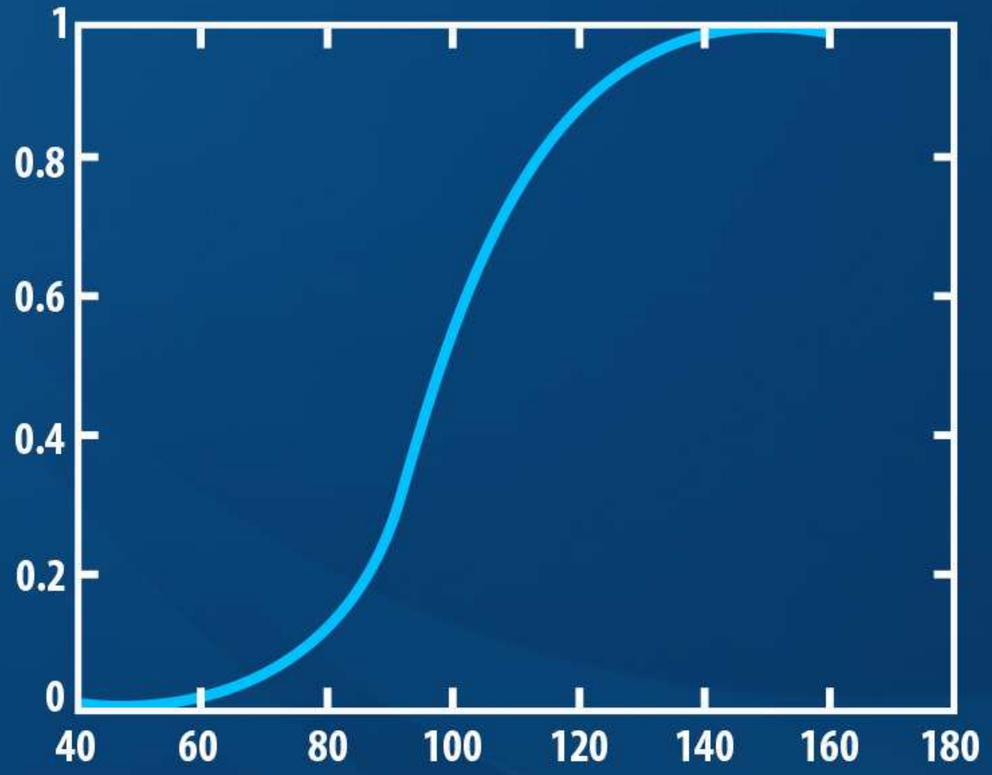
Regression Curve:
Sigmoid function

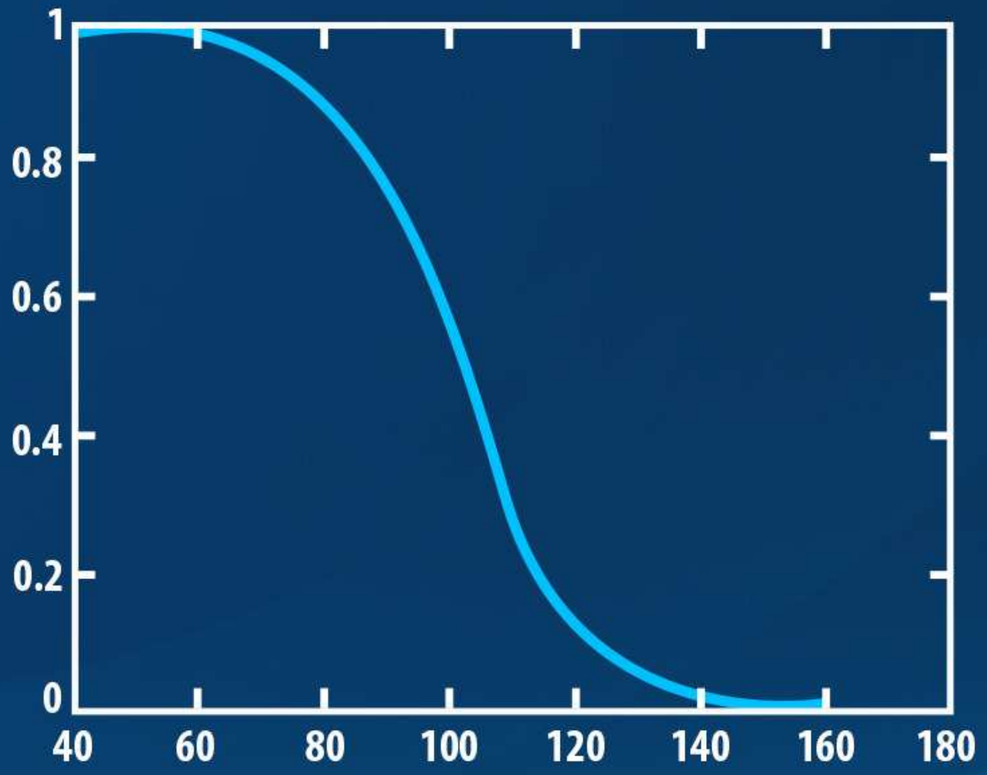
(bounded by asymptotes y=0 and y=1)

Figure I 5.8 ***More appropriate regression line for predicting outcome***

LOGISTIC RESPONSE FUNCTION

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





UNDERSTANDING ODDS

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

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

- Consider the following data:

Testosterone	Delinquent: Yes	Delinquent: No	Total
Normal	402	3614	4016
High	101	345	446
	503	3959	4462

- The odds of being delinquent if you are in the Normal group are:

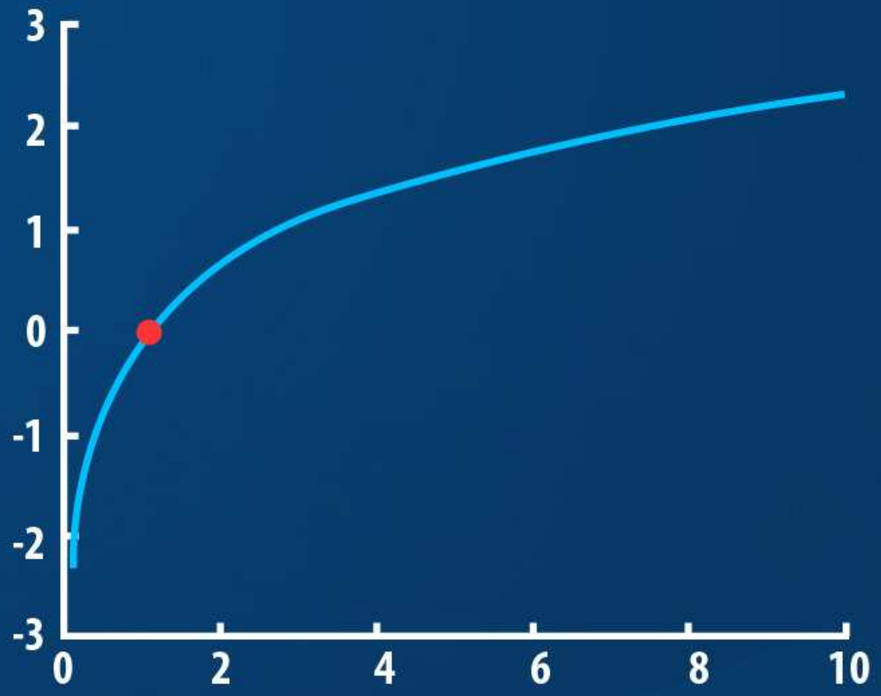
$$p(\text{delinquent})/(1-p(\text{delinquent})) = (402/4016) / (1 - (402/4016))$$
$$= 0.1001 / 0.8889$$
$$= 0.111$$

ODDS RATIO

- The odds of being not delinquent in the Normal group is the reciprocal of this:
  - 0.8999/0.1001 = 8.99
- Now, for the High testosterone group:
  - odds(delinquent) = 101/345 = 0.293
  - odds(not delinquent) = 345/101 = 3.416
- When we go from Normal to High, the odds of being delinquent nearly triple:
  - Odds ratio: 0.293/0.111 = 2.64
  - 2.64 times more likely to be delinquent with high testosterone levels.

LOGIT TRANSFORM

- The logit is the *natural* log of the odds-ratio



- logit(p) = ln(odds-ratio) = ln (p/(1-p))

LOGISTIC REGRESSION

- In logistic regression, we seek a model:

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

- That is, the log odds (logit) is assumed to be linearly related to the independent variable X.
- So, now we can focus on solving an ordinary (linear) regression!

RECOVERING PROBABILITIES

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

$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

which gives *p* as a sigmoid function.

INTERPRETATION OF β1

- Let:
  - odds1 = odds for value X (p/(1–p))
  - odds2 = odds for value X + 1 unit
- Then,

$$\frac{odds2}{odds1} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}}$$

$$= \frac{e^{(\beta_0+\beta_1 X)+\beta_1}}{e^{\beta_0+\beta_1 X}} = \frac{e^{(\beta_0+\beta_1 X)}e^{\beta_1}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$$

- Hence, the exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X.

EXAMPLE

- Suppose a cancer study yields:
  - log odds = −2.6837 + 0.0812 SurvRate
- Consider a patient with SurvRate = 40
  - log odds = −2.6837 + 0.0812(40) = 0.5643
  - odds = $e^{0.5643}$ = 1.758
  - patient is 1.758 times more likely to be improved than not
- Consider another patient with SurvRate = 41
  - log odds = −2.6837 + 0.0812(41) = 0.6455
  - odds = $e^{0.6455}$ = 1.907
  - patient's odds are 1.907/1.758 = 1.0846 times (or 8.5%) better than those of the previous
- Using probabilities,
  - p40 = 0.6374 and p41 = 0.6560
  - Improvements appear different with odds and with *p*