Logistic regression

Logistic Regression

- Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous
- Typical application: Medicine
  - We might want to predict response to treatment, where we might code survivors as 1 and those who don't survive as 0

Example

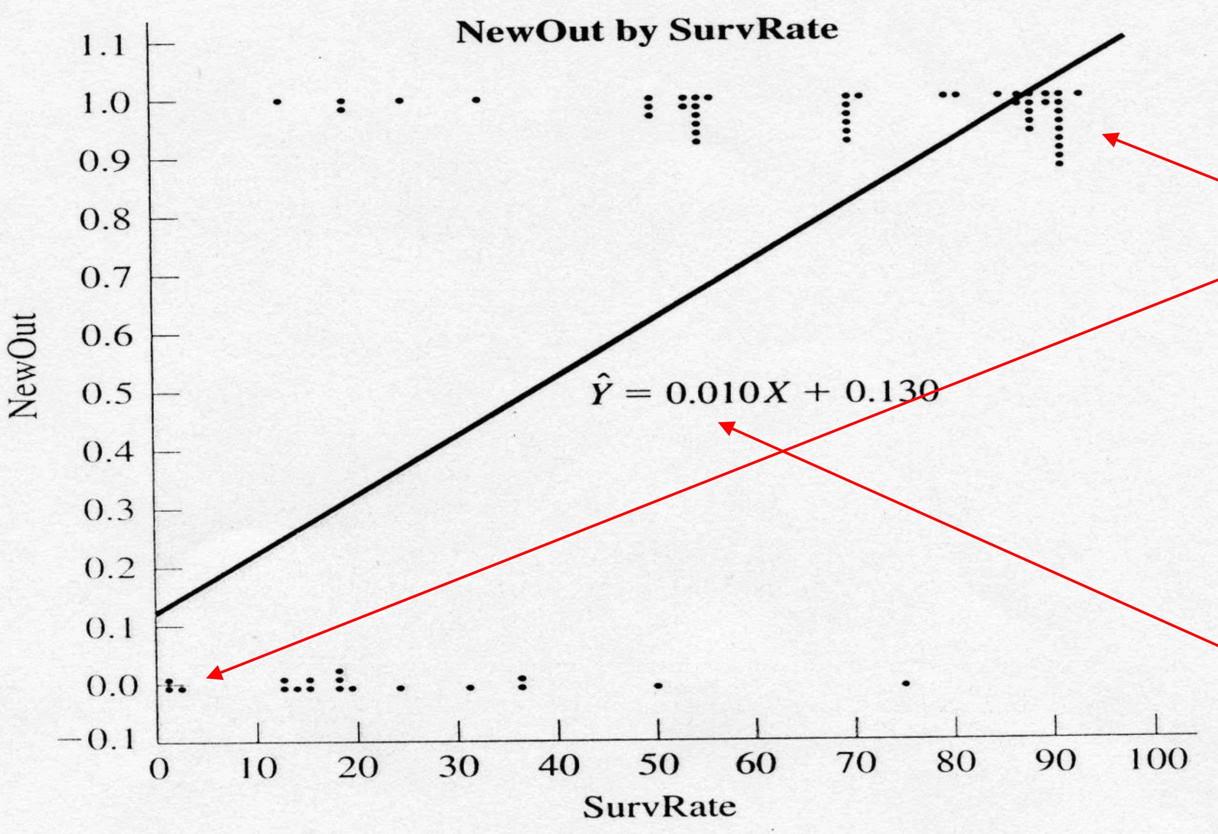


FIGURE 15·7 *Outcome as a function of SurvRate*

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

Regression: Standard linear regression

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

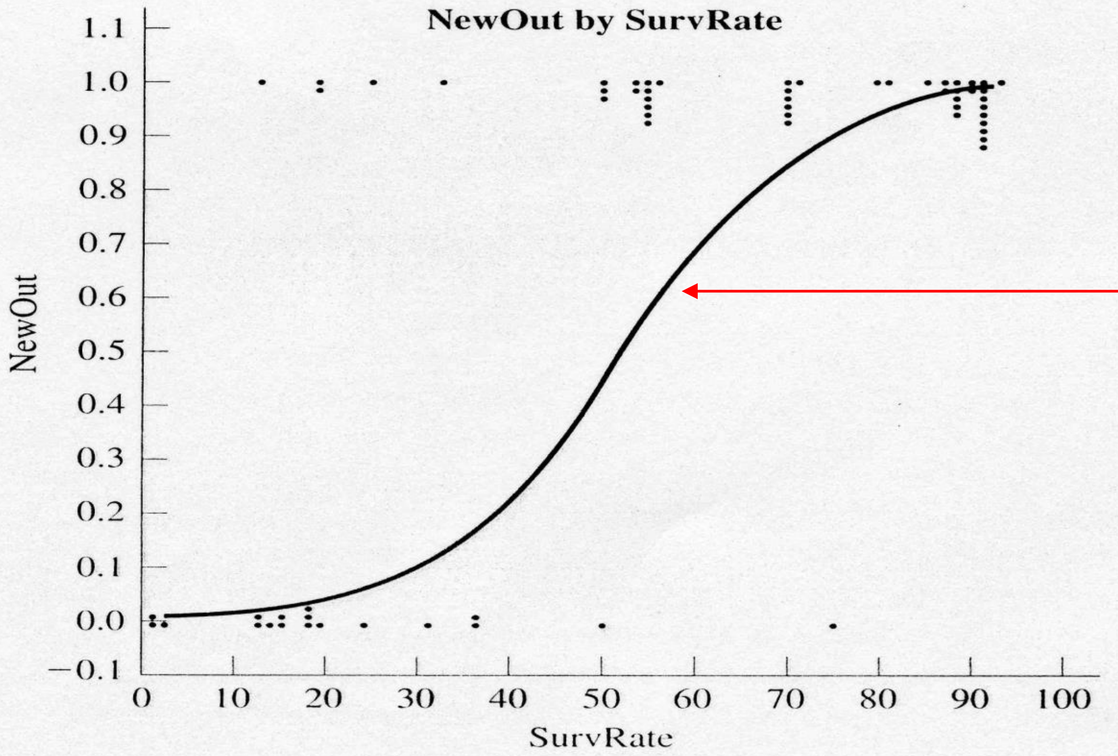


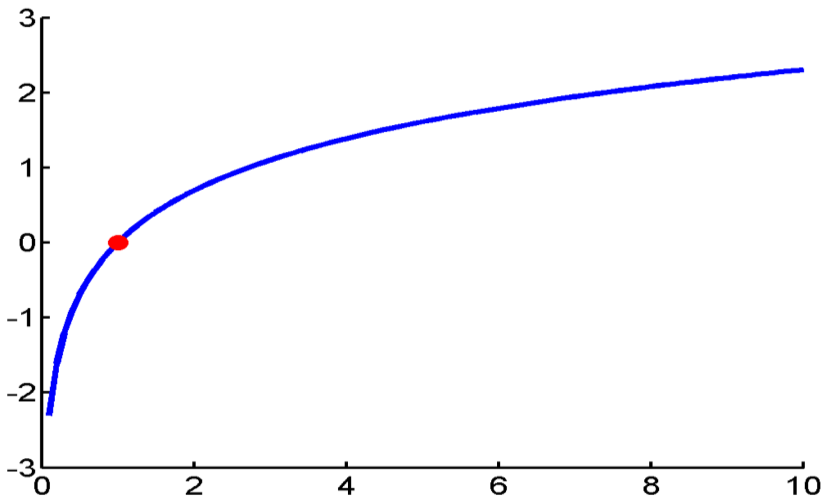
FIGURE 15.8 *More appropriate regression line for predicting outcome*

<u>Regression Curve</u>:
Sigmoid function!

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

Logit Transform

- The logit is the natural log of the odd



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

Logistic Regression

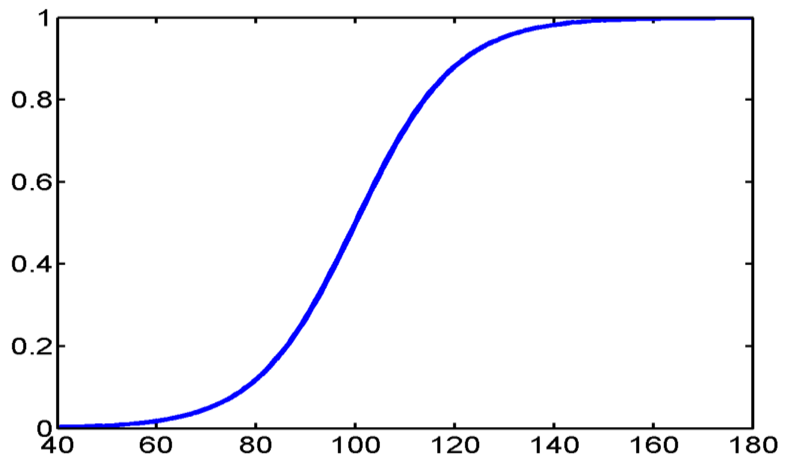
- In logistic regression, we seek a model:

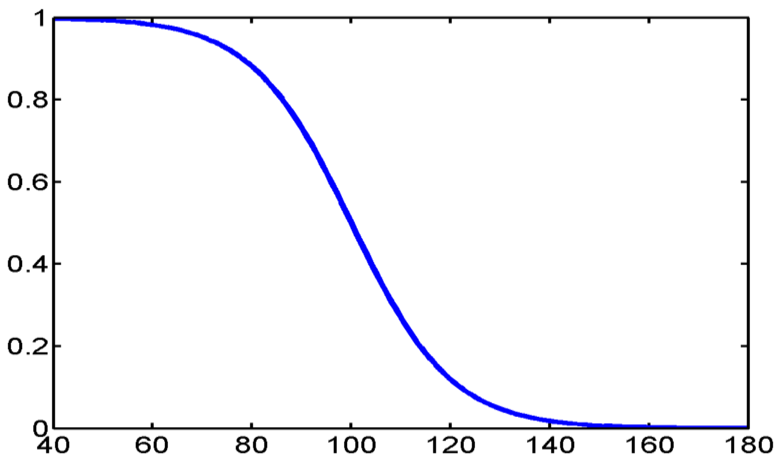
$$\mathrm{logit}(p) = b_0 + b_1 X$$

- 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!

Logistic Response Function

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





Interpretation of β1

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

$$\frac{odds2}{odds1} = \frac{e^{b_0 + b_1(X+1)}}{e^{b_0 + b_1 X}}$$

$$= \frac{e^{(b_0 + b_1 X) + b_1}}{e^{b_0 + b_1 X}} = \frac{e^{(b_0 + b_1 X)} e^{b_1}}{e^{b_0 + b_1 X}} = e^{b_1}$$

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

Sample Calculations

- 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 patient
- Using probabilities
  - p40 = 0.6374 and p41 = 0.6560
  - Improvements appear different with odds and with *p*

Dichotomous Predictor (+1/-1 coding)

Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

$$\frac{P}{1-P} = e^{\beta_o + \beta_1 X} \begin{cases} \text{Odds for Disease with Risk Present} = \dfrac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)} = e^{\beta_o + \beta_1} \\ \text{Odds for Disease with Risk Absent} = \dfrac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=-1)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=-1)} = e^{\beta_o - \beta_1} \end{cases}$$

Therefore the odds ratio (OR)

$$= \frac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \frac{e^{\beta_o + \beta_1}}{e^{\beta_o - \beta_1}} = e^{2\beta_1}$$

Dichotomous Predictor (+1/-1 coding)

- Therefore, for the odds ratio associated with risk presence we have
- Taking the natural logarithm we have

$$OR = e^{2\beta_1}$$

thus twice the estimated regression coefficient associated with a +1 / -1 coded dichotomous predictor is the natural log of the OR associated with risk presence!!!

$$\ln(OR) = 2\beta_1$$

Example: Smoking and Low Birth Weight

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.0608189	0.0127482	26133	0.0000*
Smoking Status[Cig]	0.33493469	0.0127482	690.28	<.0001*

For log odds of Low/Norm

$$\hat{\beta}_1 = .335$$

$$OR = e^{2\hat{\beta}_1} = e^{.670} = 1.954$$

Find a 95% CI for OR

1st Find a 95% CI for β_1

$$\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1) = .335 \pm 1.96 \cdot (.013) = .335 \pm .025 = \underbrace{(.310,.360)}_{\text{(LCL,UCL)}}$$

2nd Compute CI for OR = (e^{2LCL}, e^{2UCL})

$$(e^{2\times310}, e^{2\times.360}) = (1.86\,,\, 2.05)$$

We estimate that the odds for having a low birth weight infant are between 1.86 and 2.05 times higher for smokers than non-smokers, with 95% confidence.

Logistic Regression with 1 Predictor

- α, β are unknown parameters and must be estimated using statistical software
- Primary interest in estimating and testing hypotheses regarding β
  - Large-Sample test (Wald Test):
  - H_0: $\beta = 0$ H_A: $\beta \neq 0$

$$T.S.: X^2_{obs} = \left(\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \right)^2$$

$$R.R.: X^2_{obs} \geq \chi^2_{\alpha,1}$$

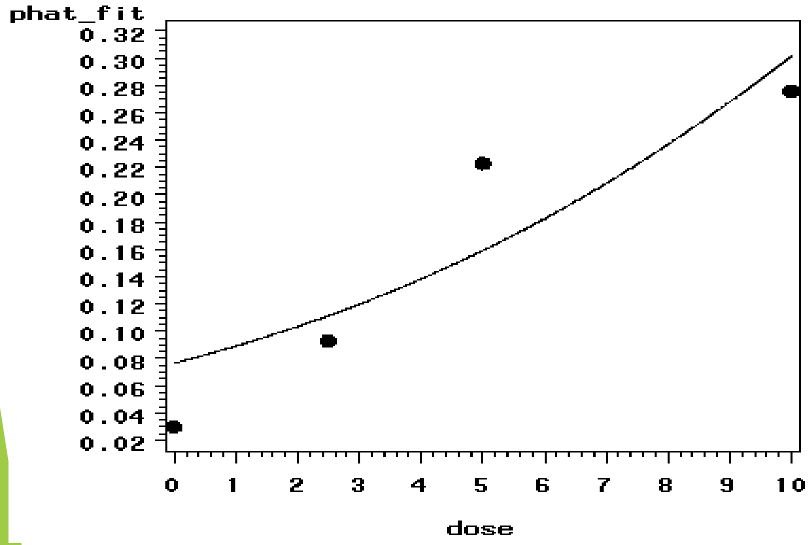
$$P-val: P(\chi^2 \geq X^2_{obs})$$

Example - Rizatriptan for Migraine

- Response - Complete Pain Relief at 2 hours (Yes/No)
- Predictor - Dose (*mg*): Placebo (0),2.5,5,10

Dose	# Patients	# Relieved	% Relieved
0	67	2	3.0
2.5	75	7	9.3
5	130	29	22.3
10	145	40	27.6

Example - Rizatriptan for Migraine (SPSS)



$$\hat{\pi}(x) = \frac{e^{-2.490+0.165x}}{1+e^{-2.490+0.165x}}$$

$$H_0 : \beta = 0 \quad H_A : \beta \neq 0$$

$$T.S.: X^2_{obs} = \left(\frac{0.165}{0.037}\right)^2 = 19.819$$

$$RR: X^2_{obs} \geq \chi^2_{.05,1} = 3.84$$

$$P - val : .000$$

95% Confidence Interval for Odds Ratio

- Step 1: Construct a 95% CI for β:

$$\hat{\beta} \pm 1.96\hat{\sigma}_{\hat{\beta}} \equiv \left(\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}},\ \hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}}\right)$$

- Step 2: Raise *e* = 2.718 to the lower and upper bounds of the CI:

$$\left(e^{\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}}},\ e^{\hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}}}\right)$$

- If entire interval is above 1, conclude positive association
- If entire interval is below 1, conclude negative association
- If interval contains 1, cannot conclude there is an association

Example - Rizatriptan for Migraine

- 95% CI for β :

$$\hat{\beta} = 0.165 \qquad \hat{\sigma}_{\hat{\beta}} = 0.037$$

$$95\%\ CI: \quad 0.165 \pm 1.96(0.037) \quad \equiv \quad (0.0925\,,\, 0.2375)$$

- 95% CI for population odds ratio:

$$\left(e^{0.0925}\ ,\ e^{0.2375}\right) \equiv (1.10\,,\, 1.27)$$

- Conclude positive association between dose and probability of complete relief

Multiple Logistic Regression

- Extension to more than one predictor variable (either numeric or dummy variables).
- With *k* predictors, the model is written:

$$\pi = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

- Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

- Many models have nominal/ordinal predictors, and widely make use of dummy variables

Example - ED in Older Dutch Men

- Response: Presence/Absence of ED (n=1688)
- Predictors: (p=12)
  - Age stratum (50-54*, 55-59, 60-64, 65-69, 70-78)
  - Smoking status (Nonsmoker*, Smoker)
  - BMI stratum (<25*, 25-30, >30)
  - Lower urinary tract symptoms (None*, Mild, Moderate, Severe)
  - Under treatment for cardiac symptoms (No*, Yes)
  - Under treatment for COPD (No*, Yes)

* Baseline group for dummy variables

Example - ED in Older Dutch Men

Predictor	b	s_b	Adjusted OR (95% CI)
Age 55-59 (vs 50-54)	0.83	0.42	2.3 (1.0 – 5.2)
Age 60-64 (vs 50-54)	1.53	0.40	4.6 (2.1 – 10.1)
Age 65-69 (vs 50-54)	2.19	0.40	8.9 (4.1 – 19.5)
Age 70-78 (vs 50-54)	2.66	0.41	14.3 (6.4 – 32.1)
Smoker (vs nonsmoker)	0.47	0.19	1.6 (1.1 – 2.3)
BMI 25-30 (vs <25)	0.41	0.21	1.5 (1.0 – 2.3)
BMI >30 (vs <25)	1.10	0.29	3.0 (1.7 – 5.4)
LUTS Mild (vs None)	0.59	0.41	1.8 (0.8 – 4.3)
LUTS Moderate (vs None)	1.22	0.45	3.4 (1.4 – 8.4)
LUTS Severe (vs None)	2.01	0.56	7.5 (2.5 – 22.5)
Cardiac symptoms (Yes vs No)	0.92	0.26	2.5 (1.5 – 4.3)
COPD (Yes vs No)	0.64	0.28	1.9 (1.1 – 3.6)

Interpretations: Risk of ED appears to be:

- Increasing with age, BMI, and LUTS strata
- Higher among smokers
- Higher among men being treated for cardiac or COPD

Example : Race and Low Birth Weight

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1979794	0.0165809	17572	0.0000*
Race[Black]	0.41029325	0.0190908	461.89	<.0001*
Race[Other]	-0.0890288	0.030963	8.27	0.0040*

For log odds of Low/Norm

$$Race[Black] = \begin{cases} +1 & \text{for race} = \text{black} \\ -1 & \text{for race} = \text{white} \end{cases}$$

$$Race[Other] = \begin{cases} +1 & \text{for race} = \text{other} \\ -1 & \text{for race} = \text{white} \end{cases}$$

Calculate the odds for low birth weight for each race (Low, Norm)

White Infants (reference group, missing in parameters)

$$e^{-2.198+.410(-1)-.089(-1)} = e^{-2.198-.410+.089} = .0805$$

Black Infants

$$e^{-2.198+.410(+1)-.089(0)} = .167$$

Other Infants

$$e^{-2.198+.410(0)-.089(+1)} = .102$$

OR for Blacks vs. Whites
= .167/.0805 = 2.075

OR for Others vs. Whites
= .102/.0805 = 1.267

OR for Black vs. Others
= .167/.102 = 1.637

Summery

- Basic Idea:
- Logistic regression is the type of regression we use for a response variable (Y) that follows a binomial distribution
- Linear regression is the type of regression we use for a continuous, normally distributed response (Y) variable
- Remember the Binomial Distribution?

Review of the Binomial Model

- Y ~ Binomial(n,p) n independent trials (e.g., coin tosses)
- p = probability of success on each trial (e.g., p = ½ = Pr of heads)
- Y = number of successes out of n trials (e.g., Y= number of heads)

Why can’t we use Linear Regression to model binary responses?

- The response (Y) is NOT normally distributed
- The variability of Y is NOT constant
- Variance of Y depends on the expected value of Y
- For a Y~Binomial(n,p) we have Var(Y)=pq which depends on the expected response, E(Y)=p
- The model must produce predicted/fitted probabilities that are between 0 and 1
- Linear models produce fitted responses that vary from -∞ to ∞