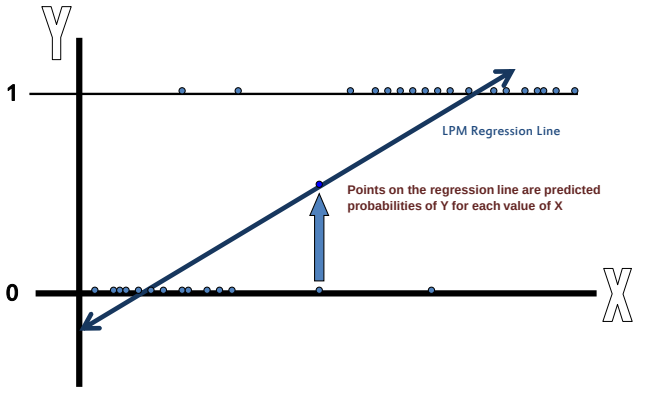
Analysis of case/control studies

- Case/control studies are designed to consider observed genotype as the random variable, and compare its distribution between cases and controls
- The analysis and interpretation is easier if we consider disease status (case vs. control) as a random outcome, predicted by genotype
- These models lend themselves to analysis via logistic regression

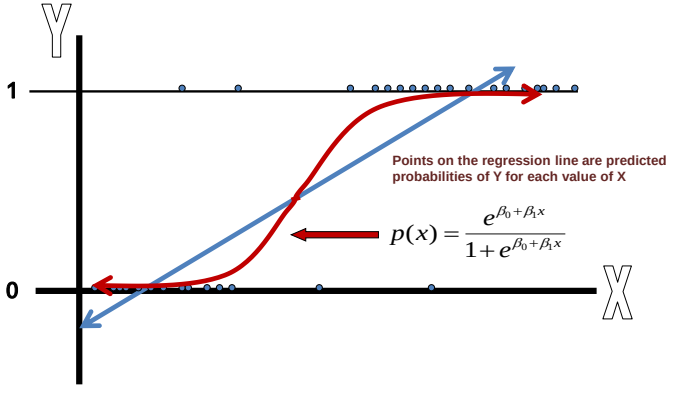
Linear Probability Model (LPM)



LPM problems

- Predicted probabilities can be >1 or <0
- The error terms vary based on the size of X
- The errors are not normally distributed since Y takes on only two values

Logistic Function



Logistic Regression

- Response: Presence/Absence of a characteristic or disease
- Predictor: Numeric variable observed for each case
- Model: $\pi(x)$ = Prob. of presence at predictor level x *[sometimes p(x)]*

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

- $\beta = 0 \Rightarrow$ P(Presence) is the same at each level of x
- $\beta > 0 \Rightarrow$ P(Presence) increases as x increases
- $\beta < 0 \Rightarrow$ P(Presence) decreases as x increases

Logistic Regression – Statistical Details

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad \leftrightarrow \quad \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_i$$

Data: $Y_i \in (0,1) \;\; i = 1,\ldots,n \qquad p = P(Y_i = 1)$

Model for y_i: $f(y_i) = p^{y_i}(1-p)^{1-y_i} \;\; i = 1,\ldots,n$

Allelic Odds Ratio - revisited

Allele	Disease: Yes	Disease: No
A	π_A	$1-\pi_A$
a	π_a	$1-\pi_a$

$$OR_A = \frac{\pi_A/(1-\pi_A)}{\pi_a/(1-\pi_a)}$$

π_A represents the probability that an allele/chromosome drawn at random from the A alleles/chroms is from an individual with disease (a case subject)

Logistic Regression & Allelic Odds Ratios

		Case allele or control allele?			
Allele	x	Probability	Odds	Log Odds	$= \alpha + \beta x$
a	0	π_a	$\pi_a/(1-\pi_a)$	$\log\{\pi_a/(1-\pi_a)\}$	α
A	1	π_A	$\pi_A/(1-\pi_A)$	$\log\{\pi_A/(1-\pi_A)\}$	$\alpha+\beta$

- We can fit a logistic regression model predicting the origin (case or control) of an allele using ***x*** - an indicator variable for allele: (0,1) = (A, a)
- Testing H_o: $\beta=0$ is equivalent to testing if the $OR=1$

Genotype Odds Ratios - revisited

	Disease: Yes	Disease: No
A/A	$\pi_{A/A}$	1-$\pi_{A/A}$
A/a	$\pi_{A/a}$	1-$\pi_{A/a}$
a/a	$\Pi_{a/a}$	1-$\pi_{a/a}$

$$OR_{AA} = \frac{\pi_{AA}/(1-\pi_{AA})}{\pi_{aa}/(1-\pi_{aa})}$$

$$OR_{Aa} = \frac{\pi_{Aa}/(1-\pi_{Aa})}{\pi_{aa}/(1-\pi_{aa})}$$

Logistic Regression & Genotype Odds Ratios

			Subject is a case?		
Genotype	x_1	x_2	Probability	Odds	Log odds $= \alpha + \beta_1 x_1 + \beta_2 x_2$
a/a	0	0	$\pi_{a/a}$	$\pi_{a/a}/(1-\pi_{a/a})$	α
A/a	1	0	$\pi_{A/a}$	$\pi_{A/a}/(1-\pi_{A/a})$	$\alpha+\beta_1$
A/A	0	1	$\pi_{A/A}$	$\pi_{A/A}/(1-\pi_{A/A})$	$\alpha+\beta_2$

$\log(\text{Odds ratio, } A/A \text{ vs } a/a) = \beta_2, \qquad \log(\text{Odds ratio, } A/a \text{ vs } a/a) = \beta_1$

- For the Genotype OR, we need 2 indicator variables, x_1 & x_2, to represent the genotype categories (i.e., 2 d.f.).
- The hypothesis of no association is tested with H_o: $\beta_1=\beta_2=0$

Multiple Logistic Regression

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

$$\pi(x_1,\cdots,x_p) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

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

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

- Inferences on β_i and OR_i are conducted as in the case of a single predictor

95% Confidence Interval for Odds Ratio

- Construct a 95% CI for β:

$$\hat{\beta} \pm 1.96 * SE_{\hat{\beta}} \equiv \left(\hat{\beta} - 1.96 * SE_{\hat{\beta}} \, , \; \hat{\beta} + 1.96 * SE_{\hat{\beta}} \right)$$

- Exponentiate both endpoints of the CI for β:

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

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

$$= \exp\left[\left(\ln p\right)^{\sum y_i} + \left(\ln(1-p)\right)^{\sum (1-y_i)} \right]$$

$$= \exp\left[\sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{p}{1-p} \right) + \ln(1-p) \right\} \right]$$

1-parameter (no-intercept) Model

Reparameterize the Likelihood function using:

$$\ln\left(\frac{p}{1-p} \right) = \beta x_i \qquad (1-p) = \frac{1}{1+\exp(\beta x_i)}$$

$$L(\beta) = \exp\left[\sum_{i=1}^{n} \left\{ y_i \left(\beta x_i\right) + \ln\left(\frac{1}{1+\exp(\beta x_i)} \right) \right\} \right]$$

1-parameter (no-intercept) Model

Likelihood: $$L(\beta) = \exp\left[\sum_{i=1}^{n} \left\{ y_i \left(\beta x_i\right) + \ln\left(\frac{1}{1+\exp(\beta x_i)} \right) \right\} \right]$$

Log Likelihood: $$l(\beta) = \sum_{i=1}^{n} \left[y_i \beta x_i - \ln(1+\exp(\beta x_i)) \right]$$

$\frac{dl}{d\beta} = 0$ → Solve for the MLE of β

2-parameter Model (with Intercept)

Log Likelihood:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - \ln(1+\exp(\beta_0 + \beta_1 x_i))]$$

Solve for: $\frac{\partial l}{\partial \beta_0} = 0 \qquad \frac{\partial l}{\partial \beta_1} = 0$

$$\underset{\sim}{f}(\beta_0, \beta_1) = \begin{bmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \end{bmatrix} = \begin{bmatrix} f_1(\beta_0, \beta_1) \\ f_2(\beta_0, \beta_1) \end{bmatrix}$$

Hessian Matrix = $\underset{\sim}{f}'(\beta_0, \beta_1) = \begin{bmatrix} \frac{\partial f_1}{\partial \beta_0} & \frac{\partial f_1}{\partial \beta_1} \\ \frac{\partial f_2}{\partial \beta_0} & \frac{\partial f_2}{\partial \beta_1} \end{bmatrix}$

$$\begin{bmatrix} \beta_0^{n+1} \\ \beta_1^{n+1} \end{bmatrix} = \begin{bmatrix} \beta_0^{n} \\ \beta_1^{n} \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} \frac{-\exp(\beta_0+\beta_1 x_i)}{(1+\exp(\beta_0+\beta_1 x_i))^2} & \sum_{i=1}^{n} \frac{-x_i \exp(\beta_0+\beta_1 x_i)}{(1+\exp(\beta_0+\beta_1 x_i))^2} \\ \sum_{i=1}^{n} \frac{-x_i \exp(\beta_0+\beta_1 x_i)}{(1+\exp(\beta_0+\beta_1 x_i))^2} & \sum_{i=1}^{n} \frac{-x_i^2 \exp(\beta_0+\beta_1 x_i)}{(1+\exp(\beta_0+\beta_1 x_i))^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i - \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \\ \sum_{i=1}^{n} y_i x_i - \frac{x_i \exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \end{bmatrix}$$

$$\beta^{n+1} = \beta^{n} - f'(\beta^{n})^{-1} f(\beta^{n})$$

In general, Newton-Raphson's method can be expanded to *k* possible parameters. As a result, the Hessian matrix is always of dimension *k*x*k* and all other vectors are of dimension *k*x1.