Logistic regression

Logistic regression analysis is a popular and widely used analysis that is similar to linear regression analysis except that the outcome is dichotomous (e.g., success/failure or yes/no or died/lived).

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Simple logistic regression analysis refers to the regression application with one dichotomous outcome and one independent variable.

Multiple logistic regression analysis applies when there is a single dichotomous outcome and more than one independent variable.

Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?

The Logistic Regression Model

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.

- The estimated probability is:

$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then p = .50
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

Since $0 \leq P \leq 1$

Odds = P/(1-P)

Odds has no “ceiling” but has “floor” of zero.

So we use the logit transformation

ln(P/(1-P)) = ln(odds) = logit(P)

Logit does not have a floor or ceiling.

Model:

$$\ln(P/(1-P))=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

or

$$\text{Odds}=e^{(\beta0+\beta1X1+\beta2X2+\ldots+\beta kXk)}=e^{\text{logit}}$$

Since $P = \text{odds}/(1 + \text{odds})$ & $\text{odds} = e^{\text{logit}}$

$$P = e^{\text{logit}}/(1 + e^{\text{logit}}) = 1/(1 + e^{-\text{logit}})$$

If $\ln(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

then

$\text{odds} = (e^{\beta 0})\,(e^{\beta 1 X1})\,(e^{\beta 2 X2}) \ldots (e^{\beta k Xk})$

or

$\text{odds} = (\text{base odds})\ OR_1\ \ OR_2\ \ldots\ OR_k$

Model is multiplicative on the odds scale

(Base odds are odds when all Xs=0)

OR_i = odds ratio for the i^{th} X

Interpreting β coefficients

Example: Dichotomous X

X = 0 for males, X=1 for females

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

M: X=0, $\text{logit}(P_m) = \beta_0$

F: X=1, $\text{logit}(P_f) = \beta_0 + \beta_1$

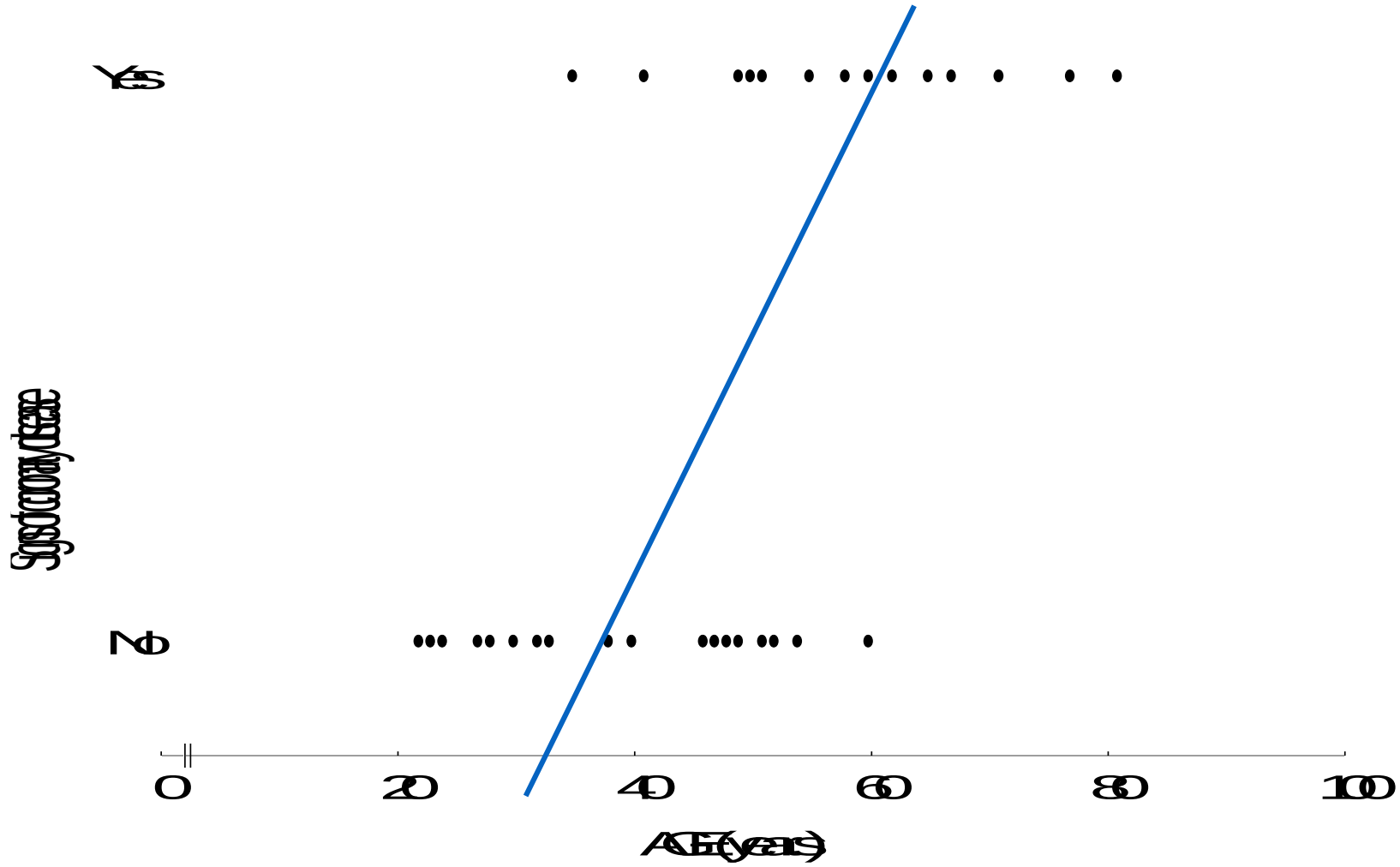
$\text{logit}(P_f) - \text{logit}(P_m) = \beta_1$

$\log(OR) = \beta_1$, $e^{\beta 1} = OR$

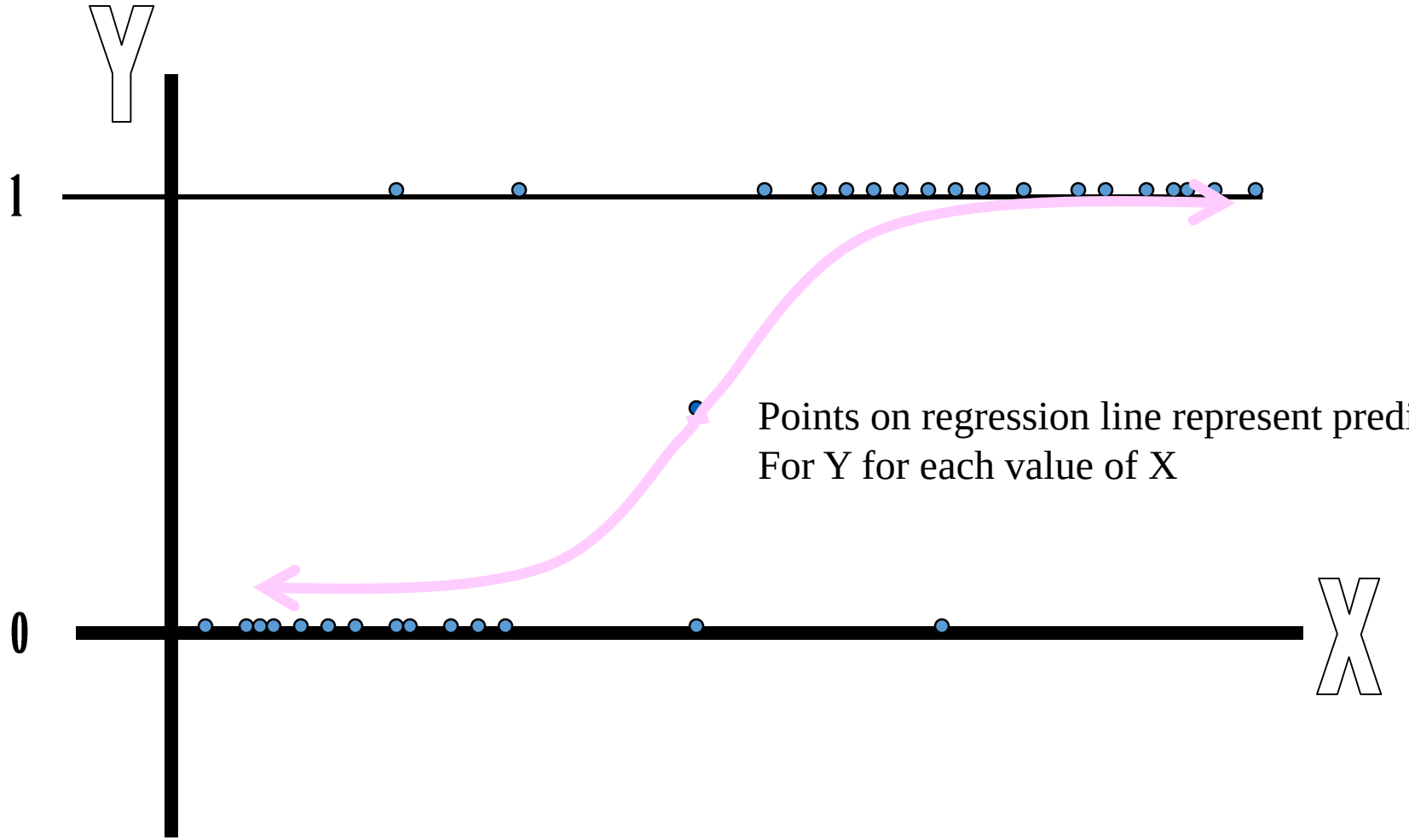
Table Age and signs of coronary heart disease (CD)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Dot-plot:



Picture of Logistic Regression



Points on regression line represent predicted probabilities
For Y for each value of X

Multiple logistic regression

- More than one independent variable
  - Dichotomous, ordinal, nominal, continuous …

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

- Interpretation of b_i :
  - Increase in log-odds for a one unit increase in x_i with all the other x_is constant
  - Measures association between x_i and log-odds adjusted for all other x_i

Example: P is proportion with disease

$\text{logit}(P) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ sex}$

"sex" is coded 0 for M, 1 for F

OR for F vs M for disease is $e^{\beta 2}$ if both are the same age.

$e^{\beta 1}$ is the increase in the odds of disease for a one year increase in age.

$(e^{\beta 1})^k = e^{k\beta 1}$ is the OR for a 'k' year change in age in two groups with the same gender.

Estimation of parameter

- Coefficients in the regression model are estimated by minimizing the sum of squared errors

- Since, p is non-linear in the parameter estimates we need a non-linear estimation technique
  - **Maximum-Likelihood Approach**
  - Non-Linear Least Squares