Exponential family of distributions

- Probability distribution on Y that has the form,

$$f(y|\theta) = exp\left[a(y)b(\theta) + c(\theta)d(y)\right]$$

- $a(\cdot)$, $d(\cdot)$ are functions of y.
- $b(\cdot)$, $c(\cdot)$ are functions of θ.
- $b(\theta)$ is called the *natural parameter*.
- In fact,

$$E(a(Y)) = \frac{-c'(\theta)}{b'(\theta)}; Var(a(Y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

- Normal, Poisson, Binomial, are members of this family.

Generalized linear models (GLM)

- Independent random variables $Y_1, Y_2, \ldots, Y_n$ where each Y_i

$$f(y_i|\theta_i) = exp\left[a(y_i)b(\theta_i) + c(\theta_i)d(y_i); i = 1, 2, \ldots, n\right]$$

- $E(Y_i) = \mu_i$ is a function of θ_i.
- For a GLM,

$$g(\mu_i) = x_i^t \beta$$

- where $x_i^t = (x_{i1}, x_{i2}, \ldots, x_{ip})$ set of covariates (predictors) for Y_i.
- $\beta^t = (\beta_1, \beta_2, \ldots, \beta_p)$ set of regression coefficients.
- $g(\cdot)$ is the *link function*. Monotone and differentiable.
- $g(\cdot)$ is a *modeling choice*. Linear regression: $g(\mu_i) = \mu_i$.

Example 3.4

- $Y_1, Y_2, \ldots, Y_n$ are n-independent success-failure trials.
- $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$., π is the probability of success.
- Probability function of Y_i is

$$f(y_i|\pi) = \pi^{y_i}(1-\pi)^{1-y_i}; y_i = 0, 1$$

- $E(Y_i) = \pi$ but π is between 0 and 1.
- Define a link function as

$$g(\pi) = log(\pi/(1-\pi)) = x^t\beta$$

- Maps $(0, 1)$ into $(-\infty, \infty)$.

- Solve equation for π, we get the *logistic* function.

$$\pi = \frac{exp(x^t\beta)}{1 + exp(x^t\beta)}$$

- For one-predictor X,

$$\pi = \frac{exp(\beta_1 + \beta_2 X)}{1 + exp(\beta_1 + \beta_2 X)}$$

- If $F(\cdot)$ is another cumulative distribution function (CDF), $\pi = F(\beta_1 + \beta_2 X)$.
- *Probit* function:

$$\pi = \Phi(\beta_1 + \beta_2 X)$$

where $\Phi(\cdot)$ is the CDF of $N(0, 1)$ (pnorm in R).

- Then, Φ^{-1} is the link function (qnorm in R) so

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 X$$

Comparision logistic/probit. $\beta_1 = 0.5$; $\beta_2 = 0.5, -0.5$

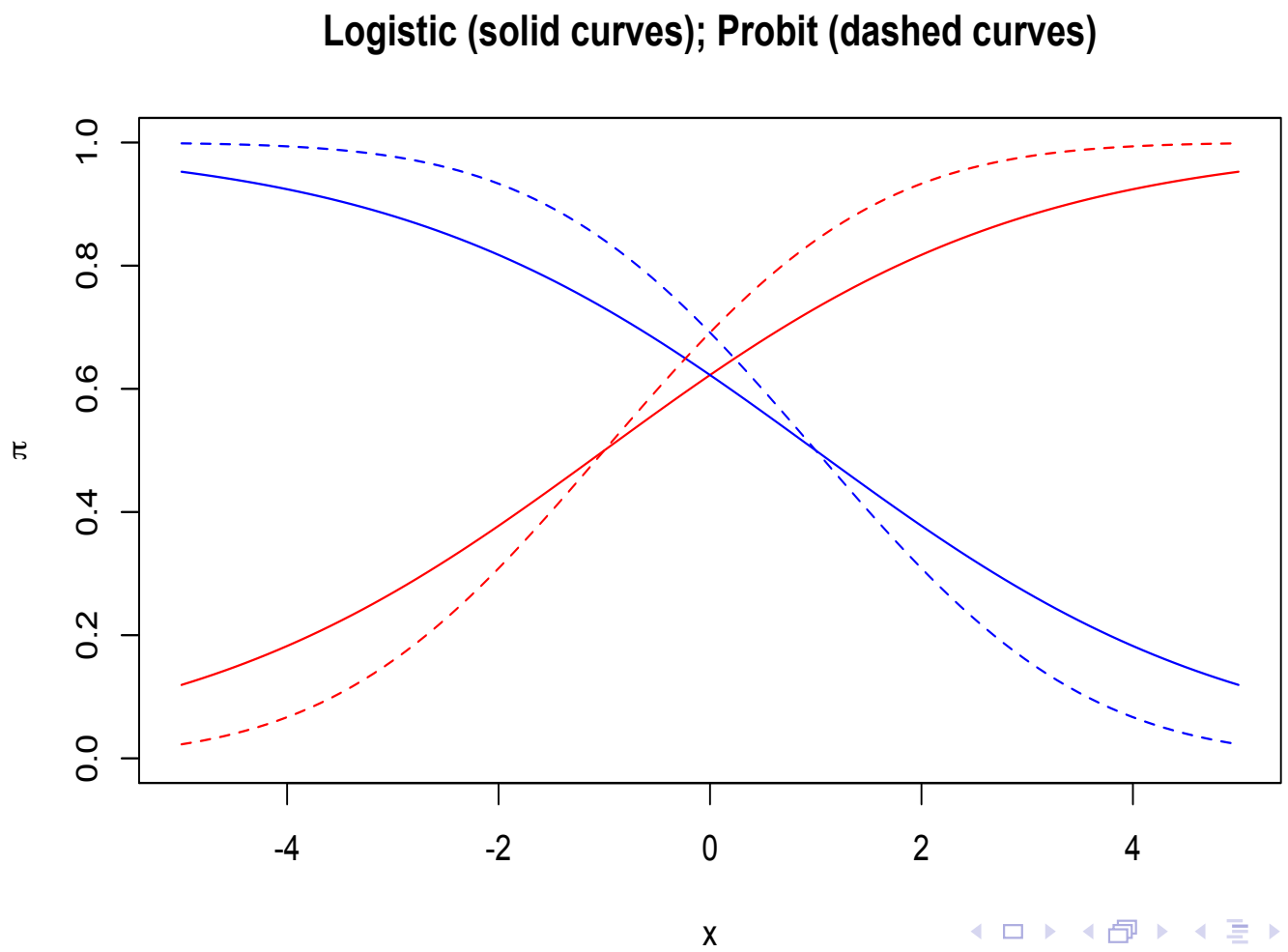


Table 3.2 example

- Number of deaths from coronary heart disease of men.
- Population sizes and several age groups
- $Y_1, Y_2, \ldots, Y_n$ represent no of deaths occurring in successive age groups.
- Possible model,

$$E(Y_i) = \mu_i = n_i exp(\theta * i)$$

- n_i is the population size for group i and θ is some rate increase.
- $i = 1, 2, \ldots, 8$ represents group age. As a GLM,

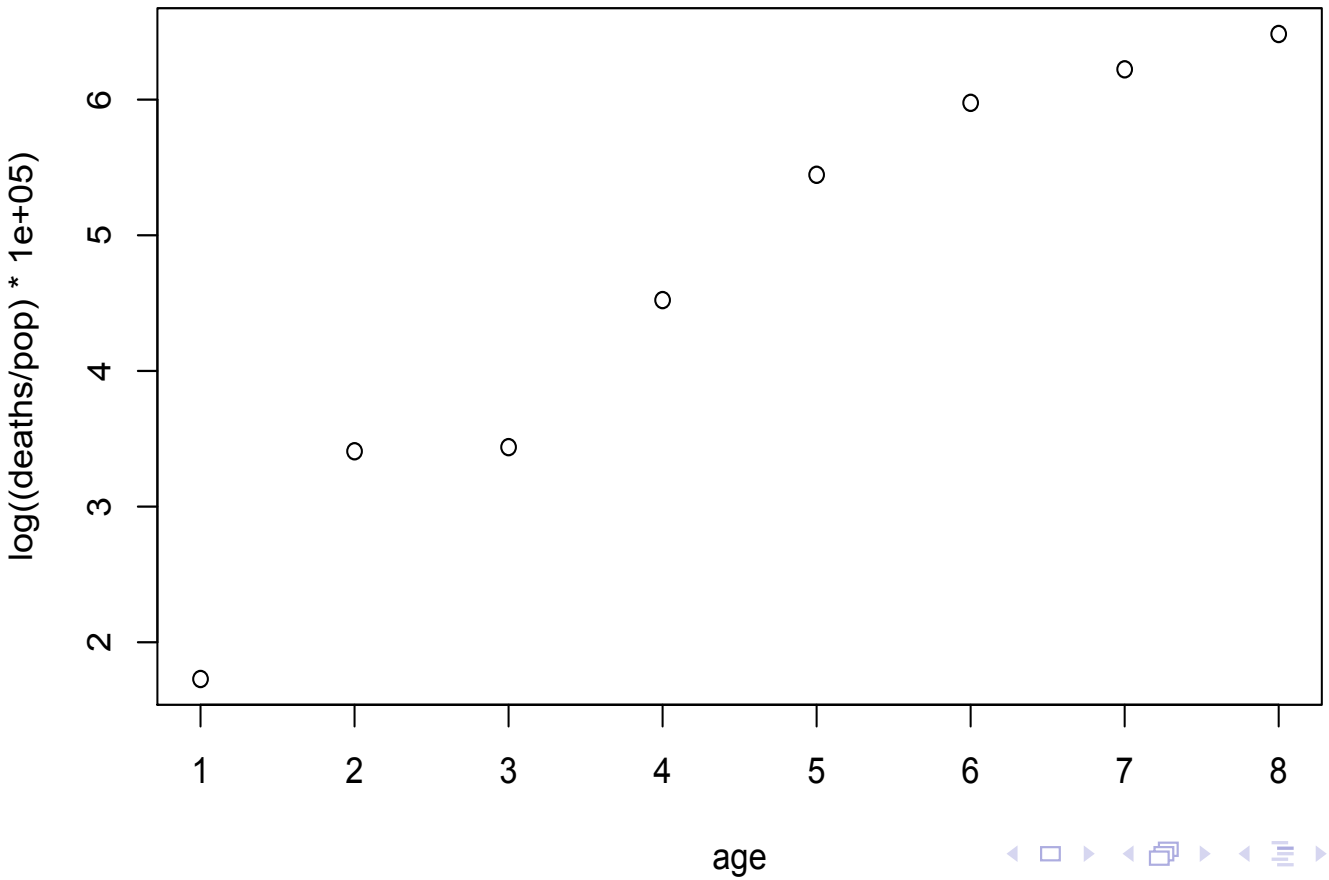
$$log(\mu_i) = \log(n_i) + \theta * i$$

Table 3.2 Mortality rates data

Table 3.2 Mortality rates

age	deaths	population
30–34	1	17742
35–39	5	16554
40–44	5	16059
45–49	12	13083
50–54	25	10784
55–59	38	9645
60–64	54	10706
65–69	65	9933

Scatter plot of data



Fit of model with glm R-function

```
glm(deaths~age,offset=log(pop/100000),
family=poisson(link="log"))
Coefficients:
(Intercept)           age
      2.497         0.522
Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
 AIC: 54.37

glm(deaths~age-1,offset=log(pop/100000),
family=poisson(link="log"))
Coefficients:
   age
0.8682
Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
AIC: 115.9
```