Generalized Linear Model: Introduction

CF Jeff Lin, MD., PhD.

April 9, 2006, 2004

References

1. Nelder and Wedderburn (1972).
   Generalized linear models.
   J Roy. Statist. Soc. A135:370-384.

2. McCullagh and Nelder (1989).
   Generalized Linear Models, 2nd ed.
   Chapman and Hall.

Outlines

Statistical Models: systematic and random components

1. Suppose n observations with $\underline{\boldsymbol{X}} = (X_1, \ldots, X_n)^T$ and $\underline{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)^T$ of two variables X and Y.

2. $y_i = \alpha + \beta x, \;\; i = 1, \ldots, n.$

3. Approximately linear relationship between y & x

4. makes the $\underline{\hat{\mathbf{y}}} = (\hat{y}_1, \ldots, \hat{y}_n)$, (or fitted values $\underline{\hat{\boldsymbol{\mu}}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$), close to the observed data.

Modelling: Science and Art

1. All models are wrong; some, though, are more useful than others

2. Not to fall in love with one model to the excluding of alternatives

3. Checks on the fit of a model to the data

4. Any statistical model has its own statistical assumptions

General Linear Model (GLM)

ANOVA, ANCOVA, Regression

$$\underline{\mathbf{y}} = \mathbf{X}\underline{\boldsymbol{\beta}} + \underline{\mathbf{e}} \tag{1}$$

$$\underline{\mathbf{e}} = \underline{\mathbf{y}} - \mathbf{X}\underline{\boldsymbol{\beta}} \tag{2}$$

$$y_i = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} + e_i \tag{3}$$

$$y_i \sim N(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}, \sigma^2), \quad i = 1, 2, \ldots, n \tag{4}$$

Minimizing $\underline{\mathbf{e}}^T \underline{\mathbf{e}} = \sum_i e_i^2$

Generalized Linear Model (GLIM)

1. Response variable $y_i \sim$ an exponential family distribution

2. $\mathcal{E}[Y_i] = \mu_i$

3. $\underline{\boldsymbol{\eta}} = g(\underline{\boldsymbol{\mu}}) = \mathbf{X}\underline{\boldsymbol{\beta}}$

GLIM Terminology

1. Random component:

   $\underline{\boldsymbol{Y}}$ are i.i.d. with

   $\mathcal{E}[\underline{\boldsymbol{Y}}] = \underline{\boldsymbol{\mu}}$ and constant variance $\mathbf{Var}[\underline{\boldsymbol{Y}}] = \sigma^2(\theta, \phi)$

2. Systematic component:

   Covariates: $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_p$.

   Linear predictor: $\underline{\boldsymbol{\eta}}$

   Relationship: $\underline{\boldsymbol{\eta}} = \sum_1^p \underline{\mathbf{x}}_j \beta_j$

3. Link function: $g(\underline{\boldsymbol{\mu}}) = \underline{\boldsymbol{\eta}}$

Types of Link Functions

1. **logit**: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$

2. **probit**: $\eta = \Phi^{-1}(\mu)$

3. **complementary log-log**: $\eta = \log\Big[-\log\Big(1-\mu\Big)\Big]$

4. **power family**:

$$\eta = \begin{cases} \frac{\mu^\lambda - 1}{\lambda}; \\ \log\mu; \quad \text{as } \lambda \to 0 \end{cases} \quad \text{or} \quad \eta = \begin{cases} \mu^\lambda; & \lambda \neq 0, \\ \log\mu; & \lambda = 0. \end{cases}$$

Mean and Variance Functions of Exponential Family

1. $\underline{Y}$ has a distribution in the exponential family,

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{5}$$

2. θ: canonical parameter

3. $b(\theta)$: cumulant function

Mean and Variance Functions of Exponential Family

Normal distribution

$$
\begin{aligned}
f_Y(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{(y\mu-\mu^2/2)}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right)\right]
\end{aligned}
\tag{6}
$$

so $\theta = \mu$, $\phi = \sigma^2$, and

$$
a(\phi)=\phi, \quad b(\theta)=\frac{\theta^2}{2}, c(y,\phi)=-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right)
$$

Mean and Variance Functions of Exponential Family

MLE and UMVUE

$$\begin{aligned}
\boldsymbol{\ell}(\theta, \phi; y) = \log f_Y(y; \theta, \phi) &= \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \\
\mathcal{E}\left(\frac{\partial \boldsymbol{\ell}}{\partial \theta}\right) &= 0 \qquad (7) \\
\mathcal{E}\left(\frac{\partial^2 \boldsymbol{\ell}}{\partial \theta^2}\right) + \mathcal{E}\left(\frac{\partial \boldsymbol{\ell}}{\partial \theta}\right)^2 &= 0 \qquad (8) \\
-\mathcal{E}\left(\frac{\partial^2 \boldsymbol{\ell}}{\partial \theta^2}\right) &= \mathcal{E}\left(\frac{\partial \boldsymbol{\ell}}{\partial \theta}\right)^2 \qquad (9)
\end{aligned}$$

Mean and Variance Functions of Exponential Family

$$\boldsymbol{\ell}(\theta, \phi; y) = \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \tag{10}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \equiv 0 \tag{11}$$

$$\mathcal{E}\left(\frac{\partial \boldsymbol{\ell}}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\phi)} \equiv 0 \tag{12}$$

$$\frac{\partial^2 \boldsymbol{\ell}}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)} \tag{13}$$

$$\mathcal{E}\left(\frac{\partial \boldsymbol{\ell}}{\partial \theta}\right)^2 = \mathcal{E}\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 = \frac{\mathbf{Var}[Y]}{a^2(\phi)} \tag{14}$$

Mean and Variance Functions of Exponential Family

$$\mathcal{E}[Y] = \mu = b'(\theta) \tag{15}$$

$$-\frac{b''(\theta)}{a(\phi)} + \frac{\mathbf{Var}[Y]}{a^2(\phi)} \equiv 0 \tag{16}$$

$$\mathbf{Var}[Y] = b''(\theta)a(\phi) = \mathbf{V}(\mu) \tag{17}$$

$a(\phi)$: common form $a(\phi) = \frac{\phi}{w} = \frac{\sigma^2}{w}$

$\phi = \sigma^2$: dispersion parameter

w: prior weight

Canonical Links

Sufficient statistic equal in dimension of $\underline{\boldsymbol{\beta}}$ such that $\theta = \eta$

Normal	$\eta = \mu$
Poisson	$\eta = \log \mu$
Binomial	$\eta = \log[\pi/(1-\pi)]$
Gamma	$\eta = \mu^{-1}$

For the canonical links, the sufficient statistic is $\mathbf{X}^T\underline{\boldsymbol{Y}}$

Common Univariate Distribution

Table 1: Common Exponential Family Distribution

	Normal	Poisson	Binomial	Gamma
Notation	$N(\mu, \sigma^2)$	$\mathbf{P}(\mu)$	$\mathbf{Bin}(n, \pi)/n$	$\mathbf{G}(\mu, v)$
Dispersion: ϕ	$\phi = \sigma^2$	1	$1/n$	$\phi = v^{-1}$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+e^\theta)$	$-\log(-\theta)$
$c(y;\phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y$	$\log \binom{n}{ny}$	$v\log(vy) - \log y - \log\Gamma(v)$
$\mu(\theta) = \mathcal{E}(Y)$	θ	$\exp(\theta)$	$e^\theta/(1+e^\theta)$	$-1/\theta$
Canonical link: $\theta(\mu)$	identity	log	logit	reciprocal
Variance function: $\mathbf{V}(\mu)$	1	μ	$\mu(1-\mu)$	μ^2

Likelihood Equations

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{18}$$

$$\boldsymbol{\ell}(\underline{\boldsymbol{\beta}}) = \sum_i \log f(y_i;\theta_i,\phi) = \sum_i \ell_i(\underline{\boldsymbol{\beta}}) \tag{19}$$

To obtain the likelihood equations, we calculate

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\,\frac{\partial \theta_i}{\partial \mu_i}\,\frac{\partial \mu_i}{\partial \eta_i}\,\frac{\partial \eta_i}{\partial \beta_j} \tag{20}$$

Likelihood Equations

Since $\mu_i = b'(\theta)$ and $\mathbf{Var}(Y_i) = b''(\theta_i)a(\phi)$, then

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \tag{21}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathbf{Var}(Y_i)}{a(\phi)} \tag{22}$$

Also , since $\eta_i = \sum_j \beta_j x_{ij}$,

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \tag{23}$$

Likelihood Equations

Finally, $\eta_i = g(\mu_i)$,

and $\partial\mu_i/\partial\eta_i$ depends on link function

In summary,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\mathbf{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \tag{24}$$

Likelihood Equations

$$\sum_{i=1}^{C} \frac{(y_i - \mu_i)\, x_{ij}}{\mathbf{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \ldots p. \tag{25}$$

Likelihood Equations

Fisher's information matrix:

$$\boldsymbol{I} = -\mathcal{E}\left(\frac{\partial^2 \ell_i}{\partial\beta_h \partial\beta_j}\right) = -\mathcal{E}\left(\frac{\partial \ell_i}{\partial\beta_h}\frac{\partial \ell_i}{\partial\beta_j}\right) \tag{26}$$

$$= -\mathcal{E}\left[\frac{(Y_i-\mu_i)x_{ih}}{\mathbf{Var}(Y_i)}\ \frac{\partial\mu_i}{\partial\eta_i}\ \frac{(Y_i-\mu_i)x_{ij}}{\mathbf{Var}(Y_i)}\ \frac{\partial\mu_i}{\partial\eta_i}\right] \tag{27}$$

$$= \frac{-x_{ih}x_{ij}}{\mathbf{Var}(Y_i)}\ \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \text{ so that} \tag{28}$$

$$\boldsymbol{I} = \mathcal{E}\left(\frac{\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial\beta_h \partial\beta_j}\right) = -\sum_{i=1}^{C} \frac{x_{ih}x_{ij}}{\mathbf{Var}(Y_i)}\ \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{29}$$

Likelihood Equations

The Fisher's information matrix $\boldsymbol{I}$, which has elements

$$\boldsymbol{I} = \mathcal{E}\left[\frac{-\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial \beta_h \partial \beta_j}\right] = \mathbf{X}^T \mathbf{W} \mathbf{X} \tag{30}$$

where $\mathbf{W}$ is the diagonal matrix with elements

$$w_i = \frac{\left(\partial \mu_i / \partial \eta_i\right)^2}{\mathbf{Var}(Y_i)} \tag{31}$$

on the main diagonal.

Estimation

Newton-Raphson method,

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \underline{\boldsymbol{\beta}}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathfrak{U}^{(m)} \tag{32}$$

$\mathfrak{I}$ is the sample information matrix having elements

$$\partial^2 \ell(\underline{\boldsymbol{\beta}})/\partial\beta_h \partial\beta_j, \tag{33}$$

$\mathfrak{U}$ is the score vector having elements

$$\partial \ell(\underline{\boldsymbol{\beta}})/\partial\beta_j, \tag{34}$$

$\mathfrak{I}^{(m)}$ and $\mathfrak{U}^{(m)}$ are $\mathfrak{I}$ and $\mathfrak{U}$ evaluated at $\underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\beta}}^{(m)}$.

Estimation

Fisher scoring method

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \underline{\boldsymbol{\beta}}^{(m)} + (\boldsymbol{I}^{(m)})^{-1}\boldsymbol{\mathcal{U}}^{(m)} \quad (35)$$

$$\text{or} \quad \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m+1)} = \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m)} + \boldsymbol{\mathcal{U}}^{(m)} \quad (36)$$

where $\boldsymbol{I}^{(m)}$ is the m^{th} approximation for the estimated Fisher's information matrix.

$\boldsymbol{I}^{(m)}$ has elements - $\mathcal{E}[\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})/\partial\beta_h \partial\beta_j]$,

evaluated at $\underline{\boldsymbol{\beta}}^{(m)}$.

Estimation

$$\text{Recall:} \quad \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m+1)} = \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m)} + \mathcal{U}^{(m)} \tag{37}$$

The right-hand side of (37) is the vector elements

$$\sum_j \left[\sum_i \mu_i \frac{x_{ih}x_{ij}}{\mathbf{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \beta_j^{(m)} \right] + \sum_i \frac{(y_i - \mu_i^{(m)})x_{ih}}{\mathbf{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

where μ_i and $(\partial \mu_i / \partial \eta_i)$ are evaluated at $\underline{\boldsymbol{\beta}}^{(m)}$.

Estimation

Thus; $\boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m)} + \boldsymbol{\mathcal{U}}^{(m)} = \mathbf{X}^T\mathbf{W}^{(m)}\underline{\mathbf{z}}^{(m)}$, (38)

where $\mathbf{W}^{(m)}$ is $\mathbf{W}$ evaluated at $\underline{\boldsymbol{\beta}}^{(m)}$ and $\underline{\mathbf{z}}^{(m)}$ has elements

$$z_i^{(m)} = \sum_j x_{ij}\beta_j^{(m)} + (y_i - \mu_i^{(m)})\left(\frac{\partial \eta_i^{(m)}}{\partial \mu_i^{(m)}}\right) \quad (39)$$

$$= \eta_i^{(m)} + (y_i - \mu_i^{(m)})\left(\frac{\partial \eta_i^{(m)}}{\partial \mu_i^{(m)}}\right) \quad (40)$$

Estimation

Fisher scoring have form

$$\left(\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \right) \underline{\boldsymbol{\beta}}^{(m+1)} = \mathbf{X}^T \mathbf{W}^{(m)} \underline{\mathbf{z}}^{(m)}. \tag{41}$$

The equations have solution

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \left(\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \underline{\mathbf{z}}^{(m)}. \tag{42}$$

The vector $\underline{\mathbf{z}}$ in this formulation is a linearized form of the link function a $\underline{\boldsymbol{\mu}}$, evaluated at $\underline{\mathbf{y}}$,

$$\begin{aligned} g(y_i) &\approx g(\mu_i) + (y_i - \mu_i) g'(\mu_i) + \cdots \\ &\approx \eta_i + (y_i - \mu_i)(\partial \eta_i / \partial \mu_i) = z_i \end{aligned} \tag{43}$$

ML Estimation:

Iterative Re-Weighted Least Squares (IRWLS)

This z_i "**adjusted**" or "**working**" **dependent variable** $\underline{\mathbf{z}}$ has i^{th} element approximated by $z_i^{(m)}$ for the m^{th} cycle of the iterative scheme.

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \left(\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \underline{\mathbf{z}}^{(m)}. \tag{44}$$

We regress $\underline{\mathbf{z}}^{(m)}$ on the $\mathbf{X}$ with weight $\mathbf{W}^{(m)}$ to obtained a new estimate $\underline{\boldsymbol{\beta}}^{(m+1)}$. This estimate yields a new linear predictor value $\underline{\boldsymbol{\eta}}^{(m+1)} = \mathbf{X}\underline{\boldsymbol{\beta}}^{(m+1)}$ and a new-adjusted-dependent-variable value $\underline{\mathbf{z}}^{(m+1)}$ for the next cycle.

ML Estimation

The asymptotic covariance matrix of $\underline{\hat{\beta}}$ is the inverse of the information matrix, estimated by

$$\widehat{\mathbf{Cov}}(\underline{\hat{\beta}}) = \left(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X} \right)^{-1} \tag{45}$$

where $\widehat{\mathbf{W}}$ is evaluated at $\underline{\hat{\beta}}$. The form of $\mathbf{W}$ depends on the link chosen for the model.

Simplification for Canonical Links

$$\eta_i = \theta_i = \sum_j x_{ij}\beta_j \tag{46}$$

when $a(\phi)$ in the likelihood function is identical for all observations, the **kernel** of the log likelihood is $\sum y_i\theta_i$, which simplifies to

$$\sum_i y_i\left(\sum_j x_{ij}\beta_j\right) = \sum_j \beta_j\left(\sum_i y_i x_{ij}\right) \tag{47}$$

Sufficient statistics for estimating $\underline{\boldsymbol{\beta}}$ in the **GLIM** are then

$$\sum_i y_i x_{ij}, \quad j = 1, \ldots, p \tag{48}$$

Simplification for Canonical Links

For the canonical link,

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) \tag{49}$$

so we simplify the likelihood equation as

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\mathbf{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \tag{50}$$

$$= \frac{(y_i - \mu_i)}{\mathbf{Var}(Y_i)} b''(\theta_i) x_{ij} \tag{51}$$

$$= \frac{(y_i - \mu_i) x_{ij}}{a(\phi)} \tag{52}$$

Simplification for Canonical Links

The second derivatives of the log likelihood

$$\frac{\partial^2 \ell_i}{\partial \beta_h \partial \beta_j} = -\frac{x_{ij}}{a(\phi)}\left(\frac{\partial \mu_i}{\partial \beta_h}\right) \tag{53}$$

These do not depend on the observations y_i, so

$$\frac{\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial \beta_h \partial \beta_j} = \mathcal{E}\left[\frac{\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial \beta_h \partial \beta_j}\right] \implies \boldsymbol{\mathcal{I}} = \boldsymbol{I} \tag{54}$$

The Newton-Raphson and Fisher scoring algorithm are identical.

Goodness-of-Fit Tests

A **saturated GLIM** has many parameters as observations, giving a perfect fit. (vs. **nested unsaturated (reduced) model**)

- Let $\tilde{\theta}$ denote the estimate of θ for all the **saturated model**
- Let $\hat{\theta}$ denote the estimate of θ for the **reduced model**
- Then the ratio

$$-2\log\left(\frac{\text{maximum likelihood under reduced model}}{\text{maximum likelihood under saturated model}}\right)$$

describes lack of fit.

Goodness-of-Fit

When the random component has $a(\phi) = \phi/w$, this measure equals

$$\begin{aligned} & -2\log(\text{likelihood ratio}) \\ = \quad & 2\sum w_i \left[\frac{y_i(\tilde{\theta}_i - \hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta}))}{\phi}\right] \qquad (55) \\ = \quad & \frac{D(\underline{\mathbf{y}}; \underline{\hat{\boldsymbol{\mu}}})}{\phi} \qquad (56) \end{aligned}$$

$\left(\frac{D(\underline{\mathbf{y}};\underline{\hat{\boldsymbol{\mu}}})}{\phi}\right)$ is called the **scaled deviance**.

$D(\underline{\mathbf{y}}; \underline{\hat{\boldsymbol{\mu}}})$ is called the **deviance**.

The greater the scaled deviance, the poorer the fit.

Goodness-of-Fit Tests

For two models, when the second is a special case of the first, the difference

$$\begin{aligned} & D(\underline{\mathbf{y}};\underline{\hat{\boldsymbol{\mu}}}_2) - D(\underline{\mathbf{y}};\underline{\hat{\boldsymbol{\mu}}}_1) \\ = \; & 2\sum w_i\left[y_i(\hat{\theta}_{1i} - \hat{\theta}_{2i}) - \left(b(\hat{\theta}_{1i} - b(\hat{\theta}_{2i})\right)\right] \sim \boldsymbol{\chi}^2 \end{aligned} \tag{57}$$

also has the form of the deviance.

Under regular conditions, **the difference in scaled deviances has approximately a chi-squared distribution**, with degrees of freedom equal to the difference between the number of parameters in the two models.

Testing Hypothesis

$\underline{\boldsymbol{\beta}} = (\underline{\boldsymbol{\beta}}_1^T, \underline{\boldsymbol{\beta}}_2^T)^T$ where $\underline{\boldsymbol{\beta}}_1$ is the $q \times 1$ vector of coefficients of interest and $\underline{\boldsymbol{\beta}}_2$ is the $(p-q) \times 1$ vector of other covariate coefficients.

Also we partition the information matrix into q and $(p-q)$ rows and columns as follows:

$$\mathfrak{I}(\underline{\boldsymbol{\beta}})_{p\times p} = \begin{pmatrix} \mathfrak{I}_{11}(\underline{\boldsymbol{\beta}}) & \mathfrak{I}_{12}(\underline{\boldsymbol{\beta}}) \\ \mathfrak{I}_{21}(\underline{\boldsymbol{\beta}}) & \mathfrak{I}_{22}(\underline{\boldsymbol{\beta}}) \end{pmatrix} \tag{58}$$

$$\text{where} \quad [\mathfrak{I}_{rs}(\underline{\boldsymbol{\beta}})] = \left(-\frac{\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial \beta_r \partial \beta_s} \right), \quad r, s = 1, \ldots p.$$

Testing Hypothesis

Let

$$\mathfrak{I}^{-1}(\underline{\boldsymbol{\beta}}) = \begin{pmatrix} \mathfrak{I}^{11}(\underline{\boldsymbol{\beta}}) & \mathfrak{I}^{12}(\underline{\boldsymbol{\beta}}) \\ \mathfrak{I}^{21}(\underline{\boldsymbol{\beta}}) & \mathfrak{I}^{22}(\underline{\boldsymbol{\beta}}) \end{pmatrix} \tag{59}$$

be the partition of its inverse.

Let $\underline{\hat{\boldsymbol{\beta}}}_{(p\times 1)} = (\underline{\hat{\boldsymbol{\beta}}}_1^T, \underline{\hat{\boldsymbol{\beta}}}_2^T)^T$ be the partitioned maximum likelihood estimate for $\underline{\boldsymbol{\beta}}$.

Testing Hypothesis

$H_0 : \underline{\boldsymbol{\beta}}_1 = \underline{\boldsymbol{\beta}}_{01}$ vs. $H_A : \underline{\boldsymbol{\beta}}_1 \neq \underline{\boldsymbol{\beta}}_{01}$

1. The Wald test statistic has the form

$$X^2_W = (\hat{\underline{\boldsymbol{\beta}}}_1 - \underline{\boldsymbol{\beta}}_{01})^T \, [\mathfrak{I}^{11}(\hat{\underline{\boldsymbol{\beta}}})]^{-1} \, (\hat{\underline{\boldsymbol{\beta}}}_1 - \underline{\boldsymbol{\beta}}_{01}) \sim \boldsymbol{\chi}^2_q \tag{60}$$

Note that this statistic depends upon the entire vector $\hat{\underline{\boldsymbol{\beta}}}$ in the inverse information calculation.

Testing Hypothesis

$H_0 : \underline{\beta}_1 = \underline{\beta}_{01} \quad \text{vs.} \quad H_A : \underline{\beta}_1 \neq \underline{\beta}_{01}$

2. The Likelihood ratio test statistic is

$$X^2_{LR} = 2 \{ \ell(\hat{\underline{\beta}}) - \ell(\underline{\beta}_{01}, \hat{\underline{\beta}}_2(\underline{\beta}_{01})) \} \sim \chi^2_q \tag{61}$$

where $\hat{\underline{\beta}}_2(\underline{\beta}_{01})$ be the maximum partial likelihood estimate of $\underline{\beta}_2$ with $\underline{\beta}_1$ fixed at $\underline{\beta}_{01}$.

Testing Hypothesis

$$H_0 : \underline{\boldsymbol{\beta}}_1 = \underline{\boldsymbol{\beta}}_{01} \quad \text{vs.} \quad H_A : \underline{\boldsymbol{\beta}}_1 \neq \underline{\boldsymbol{\beta}}_{01}$$

3. The score test statistic is

$$\begin{aligned} X^2_{SC} &= \mathcal{U}_1[\underline{\boldsymbol{\beta}}_{01}, \hat{\underline{\boldsymbol{\beta}}}_2(\underline{\boldsymbol{\beta}}_{01})]^T \, [\mathcal{I}^{11}[\underline{\boldsymbol{\beta}}, \hat{\underline{\boldsymbol{\beta}}}_2(\underline{\boldsymbol{\beta}}_{01})] \, \mathcal{U}_1[\underline{\boldsymbol{\beta}}_{01}, \hat{\underline{\boldsymbol{\beta}}}_2(\underline{\boldsymbol{\beta}}_{01})] \\ &\sim \boldsymbol{\chi}^2_q \end{aligned} \tag{62}$$

where $\mathcal{U}_1[\underline{\boldsymbol{\beta}}_{01}, \hat{\underline{\boldsymbol{\beta}}}_2(\underline{\boldsymbol{\beta}}_{01})]$ is the $(q \times 1)$ of scores for $\underline{\boldsymbol{\beta}}_1$, evaluated at the hypothesized value of $\underline{\boldsymbol{\beta}}_{01}$ and at the restricted partial maximum likelihood estimator for $\underline{\boldsymbol{\beta}}_2$.

Testing Hypothesis

Asymptotic, as $n \longrightarrow \infty$ and $m \longrightarrow \infty$, all three of these statistics have an approximate chi-squared distribution with q degree of freedom, when the null hypothesis is true.

Processes in Model Fitting of GLIM

1. Model selection

2. Parameter estimation

3. Prediction of future values

We must anticipate that, cluster around the "best" model will be a set of alternatives almost as good and not statistically distinguishable.

Model Selection

1. Assume independent (or at least uncorrelated) observations
2. Error structure: a single error term in the model
3. Scale depends on the purpose
4. Additivity effect
5. Choice of independent ($\mathbf{X}$) variables (or covariates)

Estimation in Model Fitting

1. Maximizing the likelihood or log likelihood

2. Minimize the goodness-of-fit criterion

$$D^{\star}(\underline{\mathbf{y}};\underline{\boldsymbol{\mu}}) = 2[\boldsymbol{\ell}(\underline{\mathbf{y}};\underline{\mathbf{y}}) - \boldsymbol{\ell}(\underline{\boldsymbol{\mu}},\underline{\mathbf{y}})] \tag{63}$$

3. $D^{\star}(\underline{\mathbf{y}};\underline{\hat{\boldsymbol{\mu}}})$ the scaled deviance
   For normal-theory linear regression models with known variance σ^2, th deviance is identical to the residual of sum of squares and minimum deviance is synonymous with least squares.

Prediction in Model Fitting

1. Prediction

2. Calibration

Snoring and Heart Disease Example

1. Table 2 is from an epidemiological survey of 2484 subjects to investigate snoring as a risk factor for heart disease.

2. Those surveyed were classified according to their spouses' report how much they snored.

3. The model states that the probability of heart disease $\pi(x)$ is how much they snored related to the level of snoring x.

Snoring and Heart Disease Example

1. We treat the rows of the table as independent binomial samples with that probability as the parameter.

2. We use score $(0, 2, 4, 5)$ for the snoring categories, treating the last two levels as closer than the other adjacent.

Snoring and Heart Disease Example

Table 2: Relationship between Snoring and Heart Disease

Snoring	Heart Disease	
	yes	No
Never	24	1355
Occasionally	35	603
nearly every night	21	192
Every night	30	224

Linear Probability Model

1. One approach to modelling the effect of X uses the form of ordinary regression, by which the expected value of Y is a linear function of X. The model

$$\pi(x) = \alpha + \beta x \tag{64}$$

   is called a **linear probability model**, because the probability of success changes linearly in x.

2. The parameter β represents the change in the probability per unit change in x.

3. This is a GLIM with binomial random component and identity link function.

Linear Probability Model: Disadvantages

1. Unfortunately, this model has a major structural defect.

2. Probabilities fall between 0 and 1, whereas linear functions take values over the entire real line.

3. This model can be valid over a finite range of x values.

Linear Probability Model: Disadvantages

1. It looks like an ordinary regression, least squares estimators of the model parameters are not optimal.

2. The variance of the binary outcome for each subject, $\mathbf{Var}(Y) = \pi(x)[1 - \pi(x)]$, is not constant for all x, but rather depends on x through its influence on $\pi(x)$.

3. Because of the non-constant variance, maximum likelihood (ML) estimators for this model, like most GLIM, can have smaller standard errors than least squares estimators.

Linear Probability Model: Snoring Data

1. For the snoring data the linear probability model is

$$\hat{\pi}(x) = 0.0172 + 0.0198x. \tag{65}$$

2. The estimated probability of heart disease is about 0.02 for nonsnorers, it increases $2(0.0198) = 0.04$ for occasional snores.

3. The standard error of the slope estimate of 0.0198equals 0.0028.

Logistic Regression Model

1. In practice, nonlinear relationships between $\pi(x)$ and x are often monotonic, with $\pi(x)$ increasing continuously as x increases, or $\pi(x)$ decreasing continuously as x increases.

2. The S-shaped curves displayed are often realistic shapes for the relationship.

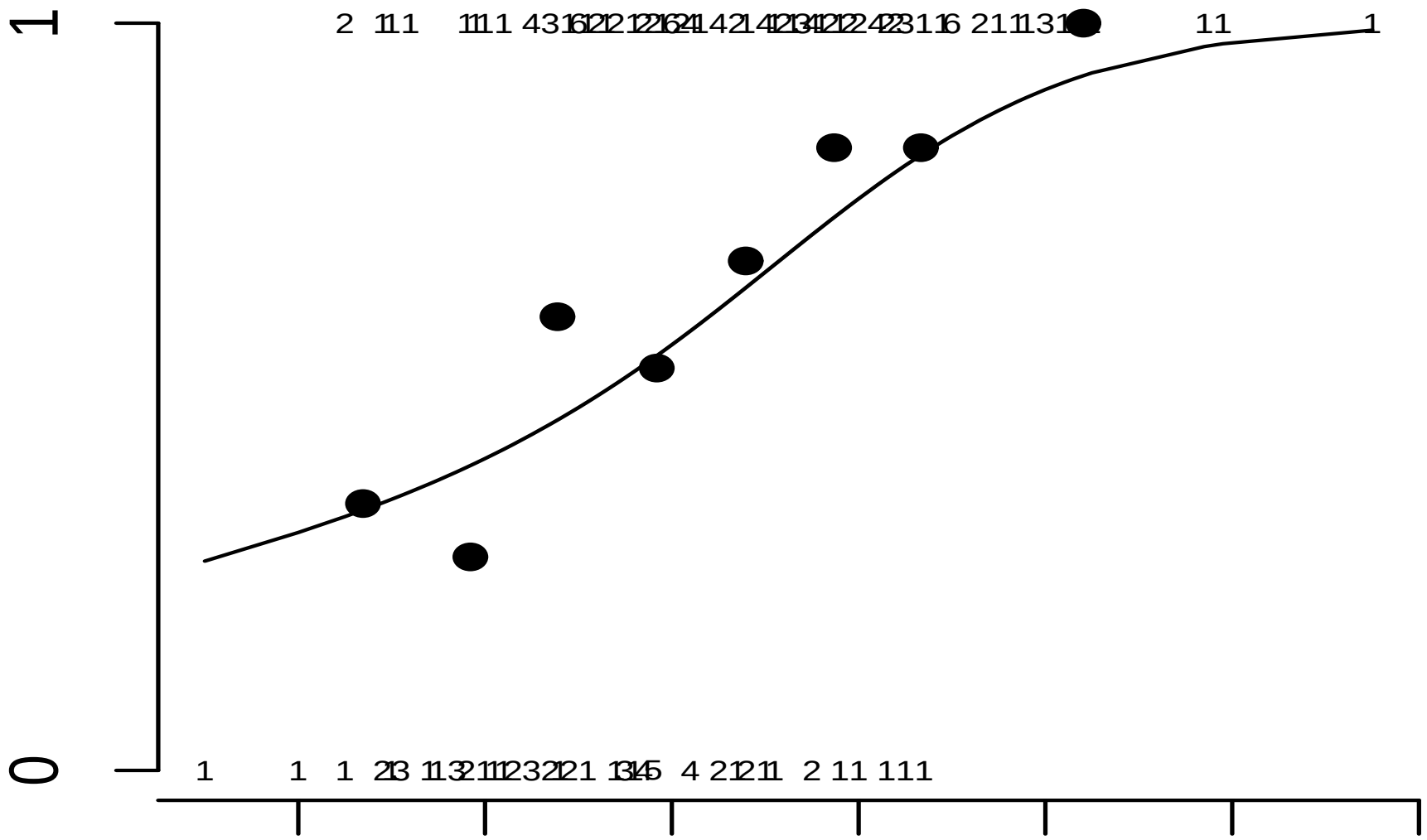

Figure 1: Logistic Regression Function

Logistic Regression Model

1. The most important function having this shape has the model form

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \tag{66}$$

Logistic Regression Model

1. This is called the **logistic regression function**, and is often called **logit model**.

2. In GLIM ,the random component for the (success, failure) determinations is binomial.

3. The link function is the logit transformation $\log[\pi/(1-\pi)]$, symbolized by logit (π).

4. The logit is the natural parameter of the binomial distribution, so the logit link is its canonical link.

Logistic Regression Model: Snoring Data

1. For the snoring data the logistic regression model is

$$\text{logit}\left[\hat{\pi}(x)\right] = -3.87 + 0.397x. \tag{67}$$

2. This gives a model of

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x, \tag{68}$$

where $x =$ snore with ML estimates $\alpha = .3.87$ and $\beta = 0.397$.

Logistic Regression Model: Snoring Data

1. The positive value of $\hat{\beta} = 0.40$ reflects the increased chance of heart disease at higher levels of snoring.

2. Since $\pi(x) = \frac{\exp(\alpha+\beta x)}{1\exp(\alpha+\beta x)}$, the estimated probability of heart disease is about 0.02 for nonsnorers, it increases $2(0.0198) = 0.04$ for occasional snores.

Tolerance Regression Model

1. Let X denote random variable, and let x denote a potential value for X.

2. The cumulative distribution function (cdf) $F(x)$ for X is defined as

$$F(x) = P(X \leq x), \quad -\infty < x < \infty. \tag{69}$$

3. Such a function, plotted as a function of x, has appearance, like S-shaped.

Tolerance Regression Model

1. As x increases, $F(x)$ increases gradually from 0 to 1, since $P(X \leq x)$ increases as x increases.

2. This subsets a class of models for binary responses whereby the dependence of $\pi(x)$ on x has form

$$\pi(x) = F(x), \tag{70}$$

where F is a cdf for some distribution.

Tolerance Regression Model

1. The logistic regression curve has this form.

2. When $\beta > 0$, $F(x)$ is the cdf of a two-parameter logistic distribution.

3. When $\beta < 0$, the formula for $1 - \pi(x)$ has the logistic cdf appearance.

4. Each choice of α and of $\beta > 0$ corresponds distribution with a symmetric,bell shape.

5. In fact, it looks similar to a normal distribution but with slightly thicker tail. Model from (70) occurs naturally when a **tolerance distribution** applies to subjects' response.

Tolerance Regression Model

1. For instance, in a toxicology study, suppose that researchers spray an insectide at various dosage levels on batches of mosquitoes.

2. For each mosquito, the response is whether it dies.

3. Each mosquito may have a certain tolerance to the insecticide, such that it dies if the dosage level exceeds its tolerance and survives if the dosage level is less than its tolerance

Tolerance Regression Model

1. Tolerance would vary among mosquitoes.

2. If a $cdf\ F$ describes the distribution of tolerances, then the model for the probability $\pi(x)$ of death at dosage level x necessarily has form (70).

Probit Regression Model

1. When F is the cdf of a normal distribution, model type (70) is called the **probit model**.

2. The link function for the model is then called the **probit link**.

3. The probit model has alternative expression

$$\text{probit}\left[\pi(x)\right] = \alpha + \beta x. \tag{71}$$

4. The probit link applied to a probability $\pi(x)$ transforms it to the standard normal z-score at which the left-tail probability equal to $\pi(x)$.

Probit Regression Model

1. For instance, probit $(0.05) = -1.645$, probit $(0.5) = 0$, probit $(0.95) = 1.645$, and probit $(0.975) = 1.96$.

2. The probit model is a GLM with binomial random component and probit link.

Probit Regression Model

1. For the snoring data the probit regression model is

$$\text{probit}\left[\hat{\pi}(x)\right] = -2.061 + 0.188x. \tag{72}$$

2. At snoring level $x = 0$, the fitted probit equals $-2.061 + 0188(0) = -2.061$.

3. The fitted probability $\hat{\pi}(x)$ is the left-tail probability for the standard normal distribution at -2.061, which equals 0.02.

4. At snoring level $x = 5$, the fitted probability equals $-2.061 + 0188(5) = -1.12$, which corresponding to a fitted probability 0.131.

Probit Regression Model

1. For practical purposes, probit and logistic regression curves Look the same.

2. It is rare, and requires enormous sample sizes, to find data for which a logistic regression model fits well but the probit model fits poorly, or conversely.

3. When both model fit well, slope estimates in logistic regression models are roughly about $1.6 - 2.0$ times those in probit model.

Probit Regression Model

1. The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1-\pi(x)$), when $\beta < 0$ has appearance of the normal cdf with mean $-\frac{\alpha}{\beta}$ and standard deviation $\sigma = \frac{1}{|\beta|}$.

2. For the snoring data, the probit fit corresponds to a normal cdf having mean of $-\frac{\hat{\alpha}}{\hat{\beta}} = \frac{2.061}{0.188} = 11.0$ and standard deviation of $\frac{1}{|\beta|} = \frac{1}{0.188} = 5.3$.

Probit Regression Model

1. The predicted probability of heart disease equals $\frac{1}{2}$ at snoring level $x = 11.0$; that is $x = 11.0$ has a fitted probit $-2.061 + 0.188(11) = 0$, which is the z-score corresponding to a left-tail probability of $\frac{1}{2}$.

2. The fitted probit value of -2.061 at $x = 0$ means that 0 is 2.06 standard deviations below the mean of a normal distribution with mean 11.0 and standard deviation 5.3.

3. Since snoring level is restricted to the range $0 - 5$, for these data, well below 11.0, then fitted probability over this range are quite small.

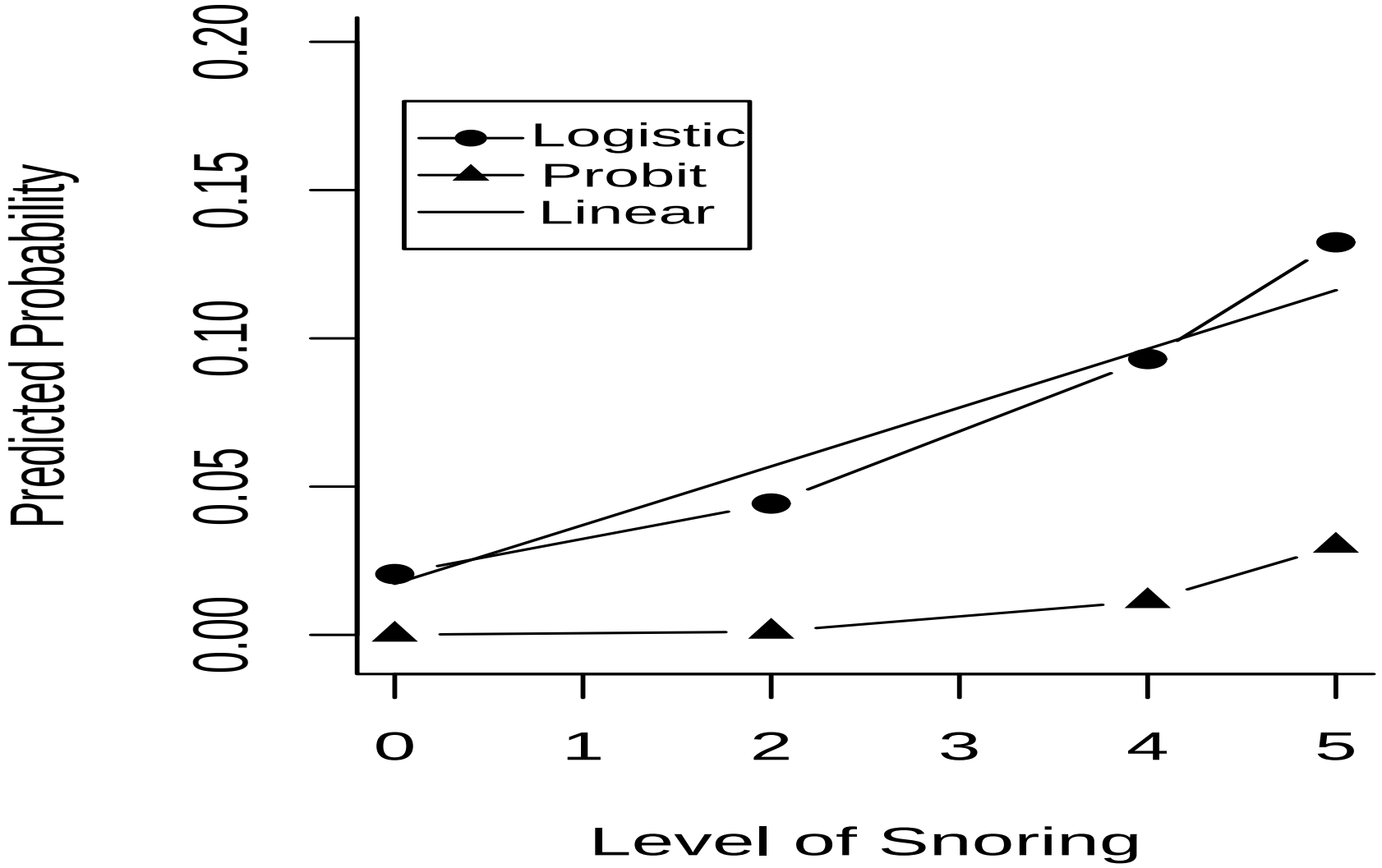

Figure 2: Predicted probabilities for logistic, probit and linear regression models.

Poisson Regression Model

1. The Poission distribution has a ppositive mean, it is more common to model the log of the mean.

2. Like the linear predictor $\alpha + \beta x$, the log of the mean can take any real value.

3. The log mean is the natural parameter for the poisson distribution, and the log link is the canonical link with a Poisson random component.

4. A **Poisson loglinear model** is a GLIM that assume a Poisson distribution for Y and uses the log link.

Poisson Regression Model

1. Let μ denote the expected value for a Poisson variate, Y, and let X denote an explanatory variable.

2. The Poisson loglinear model has form

$$\log(\mu) = \alpha + \beta x. \tag{73}$$

3. For this model, the man satisfies the exponential relationship

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}. \tag{74}$$

Poisson Regression Model

1. A one-unit increase in X has a multiplicative impact of e^{β} on μ.

2. The mean of Y at $x = 1$ equals the mean of Y at x multiplied by e^{β}.

3. If $\beta = 0$, then $e^{\beta} = e^{0} = 1$ and the multiplicative factor is 1; that is, the mean of Y does not changes as X changes.

4. If $\beta > o$, then $e^{\beta} > 1$, and the mean of Y increases as X increases.

5. If $\beta < o$, then $e^{\beta} < 1$, and the mean of Y decreases as X increases.

Horseshoe Crab Data

1. Table 3 (in file crab.txt) is a study of nesting horseshoe crabs is that each female horseshoe crab had a male crab resident in her nest.

2. The study investigated factors affecting whether the female crab had any other males, called **satellites**, residing nearby.

3. The response outcome for each female crab is her number of satellites.

4. Explanatory variables are the female crab's color, spine condition, carapace width, and weight.

5. This data set comes from a study on 173 female horseshoe crabs.

Horseshoe Crab Data

Table 3: Variables descriptions of Crab Data

Variable	Description
C	= color (light-medium, medium, dark-medium)
S	= spine condition (both good, one worn or broken, both broken)
W	= width of carapace in cm
Wt	= weight in kg
Sa	= number of satellites (male residing nearby)

Horseshoe Crab Data

1. Figure 3 plots the response counts of satellites against width, with numbered symbols indicating the number of observations at each point.

2. The substantial variability makes it difficult to discern a clear trend.

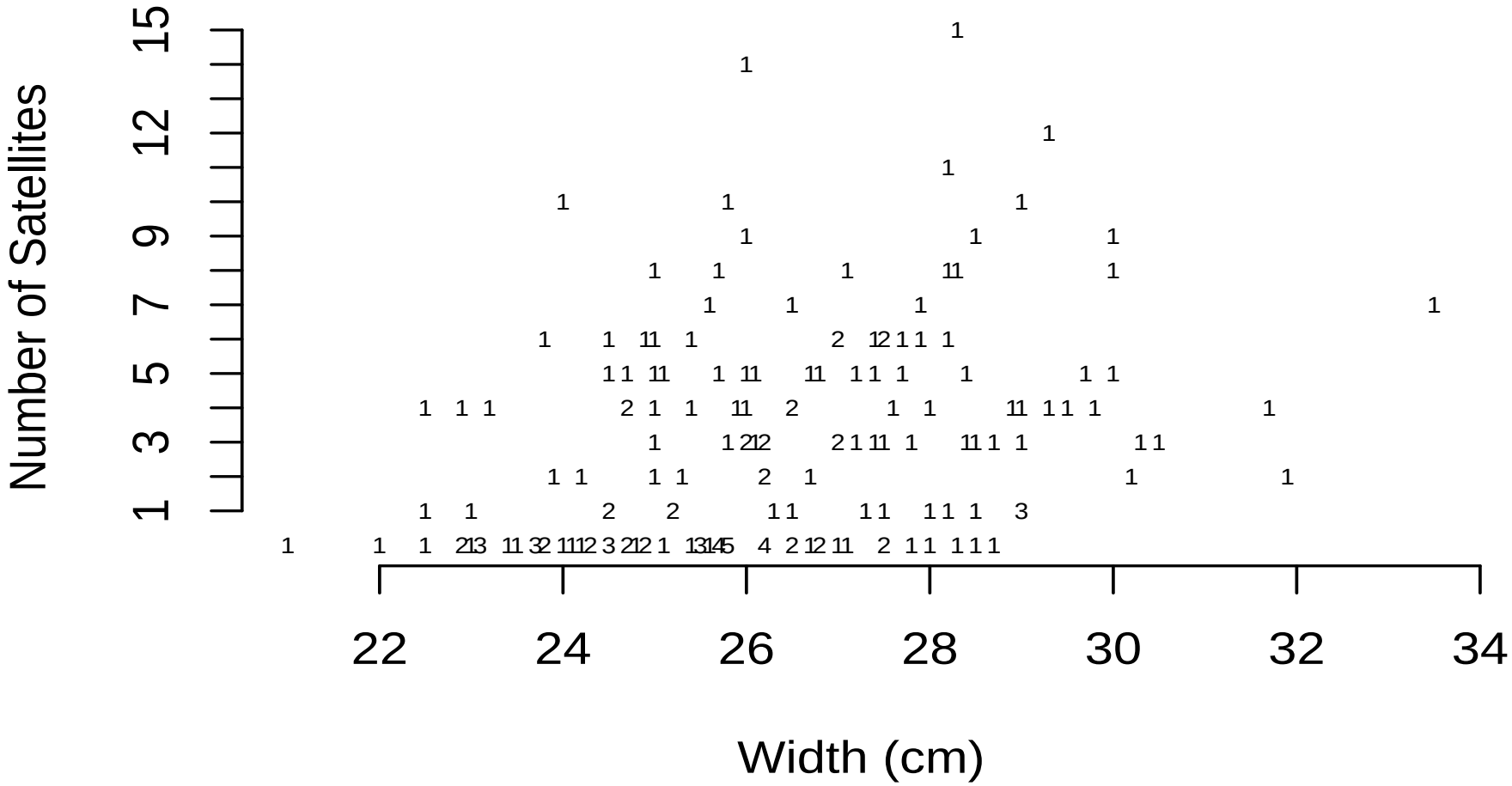

Figure 3: Number of satellites by width of female crab.

Horseshoe Crab Data

1. To get a clear picture, we grouped the female crabs into width categories $(\leq 23.25, 23.25-24.25, 24.25-25.25, 25.25-26.25, 26.25-27.25, 27.25-28.25, 28.25-29.25, > 29.25)$ and calculated the sample mean number of satellites for female crabs in each category.

2. Figure 4 plots these sample mean against the sample mean width in each category.

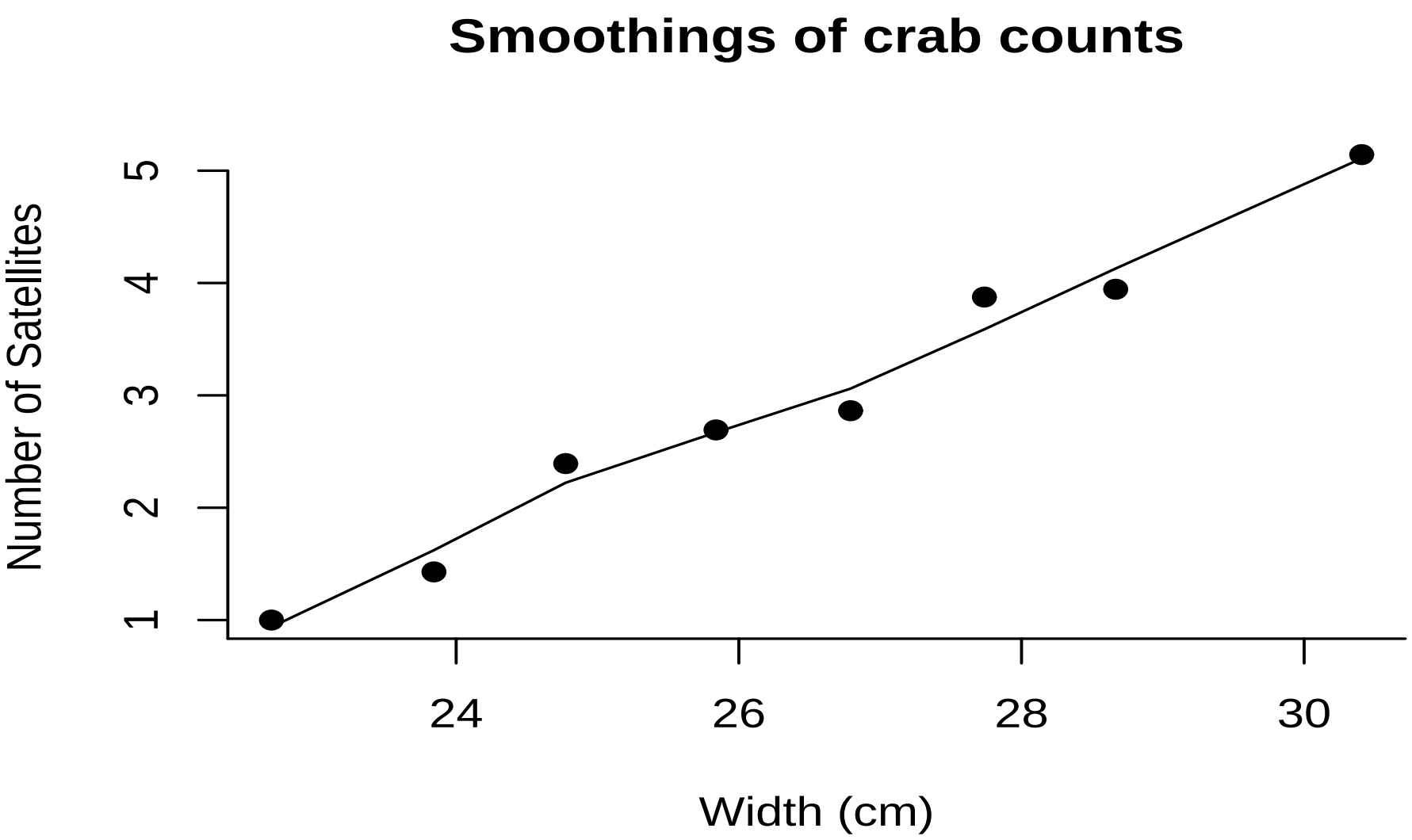

Figure 4: Number of satellites by width of female crab.

Poisson Model: Horseshoe Crab Data

1. The Goal of this study was to find a model for the number of satellites.

2. Let μ denote the expected number of satellites for a female crab, and let x denote her width.

3. Assuming that the number of satellites has a Poisson distribution, we get a model (73) (using the log link) of

$$\log(\mu) = \alpha + \beta x, \tag{75}$$

where $\alpha = -3.305$ and $\beta = 0.164$.

Poisson Model: Horseshoe Crab Data

1. The effect $\hat{\beta} = 0.164$ of width has an asymptotic (large-sample) standard error [s.e. $= 0.020$.

2. Since $\hat{\beta} > 0$, width has a positive estimated effect on the number of satellites.

Poisson Model: Horseshoe Crab Data

1. The model fitted value at any width level is an estimated mean number of satellites, $\hat{\mu}$.

2. The **fitted value**, $\hat{\mu}$ from $\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}x)$, at the mean width of $x = 26.3$ is

$$\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74. \tag{76}$$

Poisson Model: Horseshoe Crab Data

1. For this model, $\exp[\hat{\beta}] = \exp[0.164] = 1.18$ represents the multiplicative effect on the fitted value for each 1-unit increases in x.

2. For instance, the fitted value at $x = 27.3 = 26.3 + 1$ is $\exp[-3.305 + 0.164(27.3)] = 3.23$, which equals $1.18 \times 2.74 = 3.23$.

3. A 1-cm increase in width yields an 18% increase in the estimated mean number of satellites.

Poisson Model: Horseshoe Crab Data

1. Figure 4 shows that $\mathcal{E}(Y)$ may grow approximately linearly with width.

2. This suggests the Poisson GLIM identity link.

3. It has ML fit

$$\mu = \alpha + \beta x, \tag{77}$$

where $\alpha = -11.525$ and $\beta = 0.55(\text{s.e.} = 0.058)$.

Poisson Model: Horseshoe Crab Data

1. The effect of X on μ in this model is additive, rather than multiplicative.

2. A 1-cm increase in width has a predicted increase of $\hat{\beta} = 0.55$ in the expected number of satellites.

3. For instance, the fitted value at the mean width of $x = 26.3$ is $\hat{\mu} = -11.53 + 0.55(26.3) = 2.93$; at $x = 27.3$, it is $2.93 + 0.55 = 3.48$.

Poisson Model: Horseshoe Crab Data

1. The fitted value are positive at all widths observed in the sample, and the model provides a simple description of the effect of width on the number of satellites:

2. Approximately 2 cm increase in width give an additional satellite male.

Poisson Model: Horseshoe Crab Data

1. A comparison of the two models' predictions, Figure 5 plots $\hat{\mu}$ against width for the models with log link and identity link.

2. Although they diverge somewhat for relatively small and large widths, they provide similar predictions over the width range in which most observations occur.

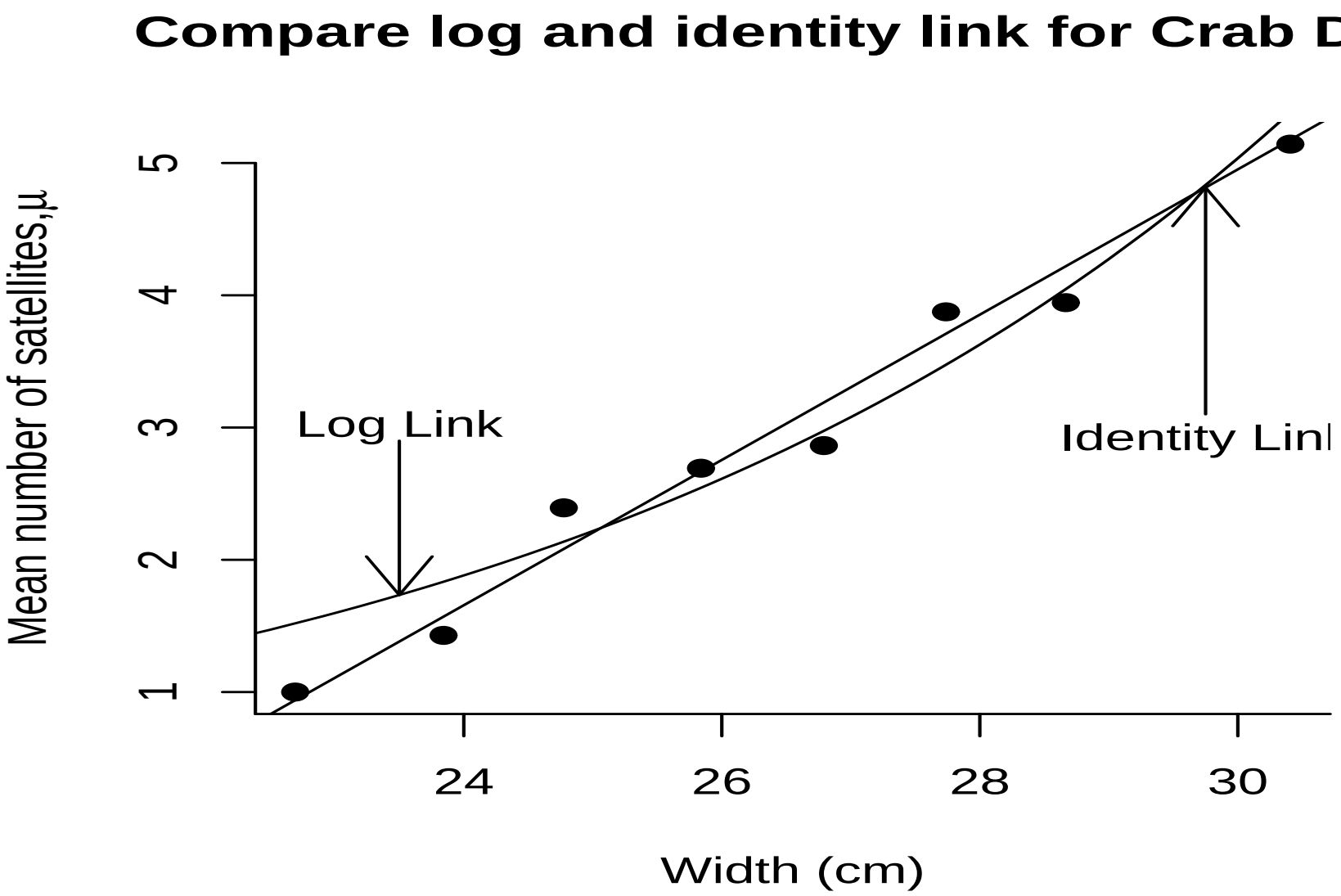

Figure 5: Estimated mean number of satellites for log and identity links for Crab Data.