Logistic Regression

CF Jeff Lin, MD., PhD.

May 2, 2006

References

1. D. Collett (2002),
   Modelling Binary Data, 2nd ed.,
   Chapman and Hall.

2. D. W. Hosmer and S. Lemeshow (2000),
   Applied Logistic Regression,
   John Wiley.

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

Introduction

1. Let Y denote a binary response variable.

2. For instance, Y might indicate vote in an election (Democrat, Republican), event occurrence present or not.

3. Denoting the two outcomes by 0 and 1 gives the Bernoulli random variable with mean

$$\mathcal{E}(Y) = 1 \times P(Y = 1) + 0 \times P(Y = 0) = \pi(\underline{\mathbf{x}}) \tag{1}$$

4. Denote $\pi(\underline{\mathbf{x}})$ as "success" probability

5. Depend on variables of explanatory variable $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$.

Binary Random Variable

1. Y denote a binary response variable as 0 and 1 gives the Bernoulli random variable

$$\begin{aligned} \mathcal{E}(Y) &= 1 \times P(Y=1) + 0 \times P(Y=0) = \pi(\underline{\mathbf{x}}) && (2) \\ \mathcal{E}(Y^2) &= 1^2\pi(\underline{\mathbf{x}}) + 0^2[1 - \pi(\underline{\mathbf{x}})] = \pi(\underline{\mathbf{x}}) && (3) \end{aligned}$$

2. The variance of Y is

$$\mathbf{Var}(Y) = \mathcal{E}(Y^2) - [\mathcal{E}(Y)]^2 = \pi(\underline{\mathbf{x}})[1 - \pi(\underline{\mathbf{x}})] \tag{4}$$

Linear Probability Model with OLS

1. For a binary response, the regression model

$$\mathcal{E}(Y) = \pi(\underline{\mathbf{x}}) = \alpha + \beta x \tag{5}$$

2. If we regress π on x using **ordinary least square (OLS)**, the linear probability model has two major structural defects, **non-linearity and heteroscedasticity**.

Linear Probability Model with OLS

3. Probabilities must fall between 0 and 1,

   OLS predicts $\pi < 0$ and $\pi > 1$

4. Heteroscedastic and not constant variance of Y is

   $\mathbf{Var}(Y) = \pi(1 - \pi)$

Linear Probability Model with OLS

1. Transformation $\sin^{-1}\sqrt{\pi}$

2. Makes the variance approximately constant

3. Solves the second problem but not the first

4. Hard to interpret model

Linear Probability Model with GLIM

1. For a binary response, the regression model

$$\mathcal{E}(Y) = \pi(\underline{\mathbf{x}}) = \alpha + \beta x \tag{6}$$

2. If Ys are independent, consider a **GLM** with **identity** link function.

Logistic Regression Model

- $y_i \sim Bin(n_i, \pi_i)$ $i = 1, \ldots, C$,

- π_i depends on covariates $\underline{\mathbf{x}}_i = (x_{1i}, x_{2i}, \ldots x_{pi})^T$

- The relationship is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} \tag{7}$$

where $\underline{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is to be estimated.

Logistic Regression Model and EL_{50}

1. If we only have single covariate, say x, as continuous variable.

2. Let

$$\mathrm{logit}[\pi_i(x_i)] \quad = \quad \alpha + \beta x_i. \tag{8}$$

3. The parameter β determines the rate of increase of decrease of the S-shaped curve.

4. The sign of β indicates whether the curve ascends or descends, and the rate of change increase as $|\beta|$ increases.

Logistic Regression Model and EL_{50}

5. A straight line drawn tangent to the curve at a particular x value describes the rate of change at that point.

6. For logistic regression parameter β, that line has **slope** equal to $\beta[\pi(x)(1-\pi(x))]$.

Logistic Regression Model and $\mathbf{EL}_{50}$

7. The steepest slope of the curve occurs at x for which $\pi(x) = 0.5$; that x value is $-\frac{\alpha}{\beta}$.

8. This value is sometimes called the **median effective level** and is denoted $\mathbf{EL}_{50}$.

9. It represents the level at which each outcome has a 50% change.

Logistic Regression Model and Odds Ratio

$$\pi_i = \frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}} \tag{9}$$

1. logit of π_i is **log odds**

2. coefficient are **log odds ratios**

3. β_j is the additive increase in log odds resulting from a one unit increase in x_{ij}.

Logistic Regression Model and Odds Ratio

1. Single covariate, x, two levels 0, 1.

$$
\begin{aligned}
\pi_i(x_i = 1) &= \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} && (10) \\
\pi_i(x_i = 0) &= \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{e^{\alpha}}{1+e^{\alpha}} && (11) \\
\text{logit}[\pi_i(x_i)] &= \alpha + \beta x_i && (12) \\
\frac{\pi_i(x_i)}{1-\pi_i(x_i)} &= \exp(\alpha + \beta x_i) && (13) \\
\log(\text{Odds Ratio}) &= \text{logit}(\pi(x_i = 1)) - \text{logit}(\pi(x_i = 0)) = \beta
\end{aligned}
$$

2. β is just the **log odds ratio** of $x_i = 1$ respect to $x_i = 0$.

Logistic Regression Model

1. Advantage for both prospective or retrospective studies

2. Empirical (sample) logit vs. grouped predictor(s) plot

$$\text{empirical logit} = \log\left(\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}\right) = \mathbf{X}\underline{\boldsymbol{\beta}} \tag{14}$$

3. "Quick and dirty" method for estimating $\underline{\boldsymbol{\beta}}$ is to regress the **empirical logit** on x_i by **OLS**

4. X is continuous, group the data before calculating empirical logit

Likelihood

For n independent observations, the log likelihood is

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{15}$$

$$\boldsymbol{\ell}(\underline{\boldsymbol{\beta}}) = \sum_i \log f(y_i;\theta_i,\phi) = \sum_i \ell_i \tag{16}$$

Likelihood

We consider the proportion as a Binomial distribution

$$n_i y_i \sim Bin(n_i, \pi_i) \tag{17}$$

in the generalized linear model, (Bernoulli distribution is a special case of Binomial Distribution).

Likelihood

$$
\begin{aligned}
& f(y_i;\theta_i,\phi) \\
= \;& n_i \binom{n_i}{n_i y_i} \pi^{n_i y_i} (1-\pi_i)^{n_i - n_i y_i} \\
= \;& \exp\left[n_i y_i \log \pi_i + (n_i - n_i y_i) \log(1-\pi_i) + \log\left(n_i \binom{n_i}{n_i y_i} \right) \right] \\
= \;& \exp\left[\frac{y_i \log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i)}{1/n_i} + \log\left(n_i \binom{n_i}{n_i y_i} \right) \right] \\
= \;& \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \right]
\end{aligned}
\qquad (18)
$$

Likelihood

$$\theta_i = \log(\frac{\pi_i}{1-\pi_i}) \qquad (19)$$

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \qquad (20)$$

$$b(\theta_i) = -\log(1-\pi_i) = \log[1+e^{\theta_i}] \qquad (21)$$

$$c(y_i,\phi) = \log\left(n_i\binom{n_i}{n_iy_i}\right) \qquad (22)$$

$$a(\phi) = \phi/n_i \qquad (23)$$

$$\phi = 1 \qquad (24)$$

Likelihood

$$\mathcal{E}(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \pi_i \tag{25}$$

$$\mathbf{Var}(Y_i) = b''(\theta_i)a(\phi) = \frac{e^{\theta_i}(1+e^{\theta_i}) - e^{\theta_i}(e^{\theta_i})}{(1+e^{\theta_i})^2}a(\phi) \tag{26}$$

$$= \frac{e^{\theta_i}}{(1+e^{\theta_i})^2}a(\phi) = \frac{\pi_i(1-\pi_i)}{n_i} \tag{27}$$

Estimation

Again we assume $a(\phi)$ the same for all observations and now let $\pi_i = \pi(\underline{\mathbf{x}}_i), i = 1, 2, \ldots, C$ such that

$$\pi_i = \pi(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}) = \text{expit}(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}) = \frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}} \tag{28}$$

$$\eta_i = g(\mu_i) = \theta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{(1 - \pi_i)} = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} \tag{29}$$

Estimation

The most common iterative methods are

Newton-Raphson and **Fisher's scoring**.

These are closely related.

Estimation: Newton-Raphson Method

1. Let $\underline{\boldsymbol{\beta}}^{(m)}$ denote the m^{th} approximation for the ML estimate $\hat{\underline{\boldsymbol{\beta}}}$.

2. For the **Newton-Raphson method**,

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \underline{\boldsymbol{\beta}}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathfrak{U}^{(m)} \tag{30}$$

3. where $\mathfrak{I}$ is the information matrix having elements $\partial^2\boldsymbol{\ell}(\underline{\boldsymbol{\beta}})/\partial\beta_h\partial\beta_j$,

4. $\mathfrak{U}$ is the score function, a vector having elements $\partial\boldsymbol{\ell}(\underline{\boldsymbol{\beta}})/\partial\beta_j$.

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

Estimation: Fisher's Scoring Method

1. The formula for Fisher's scoring is

$$\begin{aligned} \underline{\boldsymbol{\beta}}^{(m+1)} &= \underline{\boldsymbol{\beta}}^{(m)} + (\boldsymbol{I}^{(m)})^{-1}\boldsymbol{\mathfrak{U}}^{(m)} & (31) \\ \text{or} \quad \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m+1)} &= \boldsymbol{I}^{(m)}\underline{\boldsymbol{\beta}}^{(m)} + \boldsymbol{\mathfrak{U}}^{(m)} & (32) \end{aligned}$$

2. where $\boldsymbol{I}$ is expected information matrix $\boldsymbol{I} = \mathcal{E}(\boldsymbol{\mathfrak{I}})$, the Fisher's information matrix.

3. $\boldsymbol{I}^{(m)}$ is the m^{th} approximation for the estimated expected information matrix; that is, $\boldsymbol{I}^{(m)}$ has elements - $\mathcal{E}[\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})/\partial\beta_h\partial\beta_j]$, evaluated at $\underline{\boldsymbol{\beta}}^{(m)}$.

Estimation

Certain simplifications occur when a **GLIM** uses the canonical link.

For the **canonical link** (logit) in Binomial distribution,

$$\eta_i = \theta_i = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} = \sum_j x_{ij}\beta_j \tag{33}$$

Estimation

$$\mathbf{L}(\underline{\boldsymbol{\beta}}) \propto \prod_{1}^{C} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n-y_i} \tag{34}$$

$$\boldsymbol{\ell}_i(\underline{\boldsymbol{\beta}}) \propto y_i \log\left(\frac{\pi_i}{(1-\pi_i)}\right) - n_i \log(1-\pi_i) \tag{35}$$

$$\boldsymbol{\ell}(\underline{\boldsymbol{\beta}}) \propto \sum_{1}^{C} \boldsymbol{\ell}_i(\underline{\boldsymbol{\beta}}) = \sum_{1}^{C} y_i \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} - \sum_{1}^{C} n_i \log\left(1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}\right) \tag{36}$$

Estimation

The score function, $\mathcal{U}(\underline{\boldsymbol{\beta}})$, is

$$
\begin{aligned}
\mathcal{U}(\beta_j) &= \frac{\partial \boldsymbol{\ell}}{\partial \beta_j} && (37) \\
&= \sum_1^C y_i x_{ij} - \sum_1^C n_i \left(\frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}} \right) x_{ij} && (38) \\
&= \sum_1^C (y_i - \mu_i) x_{ij} && (39)
\end{aligned}
$$

where $\mu_i = \mathcal{E}(Y_i) = n_i \pi_i$.

Estimation

The second derivatives are

$$\mathfrak{I}(\beta_j, \beta_h) = -\frac{\partial^2 \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})}{\partial\beta_h \partial\beta_j} \tag{40}$$

$$= \sum_1^C n_i x_{ij} \frac{\partial}{\partial\beta_h}\left(\frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}\right) \tag{41}$$

$$= \sum_1^C n_i \pi_i (1 - \pi_i) x_{ij} x_{ih} \tag{42}$$

Estimation

1. Let $\underline{\mathbf{y}}_{C\times 1} = (y_1, \ldots, y_C)^T$, $\underline{\boldsymbol{\mu}}_{C\times 1} = (\mu_1, \ldots, \mu_N)^T$, $\mathsf{X}_{C\times p} = (\underline{\mathbf{x}}_1^T, \ldots, \underline{\mathbf{x}}_C^T)$, and the elements of $\underline{\boldsymbol{\mu}}$ are non-linear functions of an assumed value $\underline{\boldsymbol{\beta}}$.

2. Also define

$$\mathsf{W}_{C\times C} = \mathbf{Diag}[n_i \pi_i (1 - \pi_i)] \tag{43}$$

Then it is easy to show that

$$\mathfrak{U}(\underline{\boldsymbol{\beta}}) = \mathsf{X}^T(\underline{\mathbf{y}} - \underline{\boldsymbol{\mu}}) \tag{44}$$

$$\mathfrak{I}(\underline{\boldsymbol{\beta}}) = \mathsf{X}^T\mathsf{W}\mathsf{X} \tag{45}$$

Estimation

1. For one step of Newton-Raphson method, we use $\underline{\boldsymbol{\beta}}^{(m)}$ the current estimate of $\underline{\boldsymbol{\beta}}$ to calculate $\underline{\boldsymbol{\mu}}^{(m)}$ and $\mathsf{W}^{(m)}$.

2. The new estimate of $\underline{\boldsymbol{\beta}}^{(m+1)}$ is then

$$\underline{\boldsymbol{\beta}}^{(m+1)} = \underline{\boldsymbol{\beta}}^{(m)} + (\ \mathsf{X}^T\,\mathsf{W}^{(m)}\,\mathsf{X}\ )^{-1}\ \mathsf{X}(\underline{\mathbf{y}} - \underline{\boldsymbol{\mu}}^{(m)}) \tag{46}$$

Inference of Coefficients

With large samples, a reasonable approximation for inference is

$$\hat{\underline{\boldsymbol{\beta}}} \;\sim\; N\big(\,\underline{\boldsymbol{\beta}}, \widehat{\mathbf{Var}}(\hat{\underline{\boldsymbol{\beta}}})\,\big) \tag{47}$$

$$\text{where} \quad \widehat{\mathbf{Var}}(\hat{\underline{\boldsymbol{\beta}}}) \;=\; \big(\; \mathsf{X}^T\, \hat{\mathsf{W}}\, \mathsf{X}\; \big)^{-1} \tag{48}$$

$\big(\, \mathsf{X}^T\, \hat{\mathsf{W}}\, \mathsf{X}\, \big)^{-1}$ comes from the final step of NR.

Inference of Coefficients: Wald Test

1. To test $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$

$$\mathsf{z} = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim N(0,1) \tag{49}$$

   where $s.e.(\hat{\beta}_j)$ is the square root of the $(j,j)^{th}$ element of $\widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\beta}}})$.

2. An approximation $(1-\alpha) \times 100\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm Z_{1-\alpha/2}\, s.e.(\hat{\beta}_j) \tag{50}$$

Inference of Coefficients: LR Test

1. To test $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$

   (a) Remove the j^{th} column from the design matrix $\mathbf{X}$

   (b) Re-fit the model

   (c) Compare the drop in $2 \times \boldsymbol{\ell}(\underline{\boldsymbol{\beta}})$ to $\boldsymbol{\chi}_1^2$

2. This is a **likelihood ratio (LR) test** for single coefficient.

Inference of Coefficients

1. For a linear function $\theta = \underline{\mathbf{a}}^T \underline{\boldsymbol{\beta}}$, a reasonable approximation is

$$\underline{\mathbf{a}}^T \underline{\hat{\boldsymbol{\beta}}} \sim N\big(\underline{\mathbf{a}}^T \underline{\boldsymbol{\beta}}, \underline{\mathbf{a}}^T \widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\beta}}})\underline{\mathbf{a}}\big) \tag{51}$$

2. For a nonlinear function $\underline{\boldsymbol{\theta}} = g(\underline{\boldsymbol{\beta}})$, the delta method approximation is

$$g(\underline{\hat{\boldsymbol{\beta}}}) \sim N\big(g(\underline{\boldsymbol{\beta}}), g'(\underline{\hat{\boldsymbol{\beta}}})^T \widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\beta}}}) g'(\underline{\hat{\boldsymbol{\beta}}})\big) \tag{52}$$

Inference of Coefficients

3. One important nonlinear function is

$$\pi_i = \pi_i(\underline{\boldsymbol{\beta}}) = \frac{e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}}{1 + e^{\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}}} \quad (53$$

$$\frac{\partial \pi_i}{\partial \beta_j} = \pi_i(1 - \pi_i)x_{ij} \quad (54$$

$$\frac{\partial \pi_i}{\partial \underline{\boldsymbol{\beta}}} = \pi_i(1 - \pi_i)\underline{\mathbf{x}}_i \quad (55$$

4. There we estimate variance for the fitted probability $\hat{\pi}_i = \pi_i(\hat{\underline{\boldsymbol{\beta}}})$ is

$$\widehat{\mathbf{Var}}(\hat{\pi}_i) = \pi_i^2 \, (1 - \pi_i)^2 \, \underline{\mathbf{x}}_i^T \, \widehat{\mathbf{Var}}(\hat{\underline{\boldsymbol{\beta}}}) \, \underline{\mathbf{x}}_i \quad (56$$

Inference of Coefficients

5. In general, from model with $\hat{\pi}_i$,

$$\widehat{\mathbf{Var}}(\hat{\pi}_i) = \pi_i^2\,(1-\pi_i)^2\,\underline{\mathbf{x}}_i^T\,\widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\beta}}})\,\underline{\mathbf{x}}_i \qquad (57$$

   this should be **smaller** than the estimated variance of $\hat{p}_i$ (sample proportion as y_i/n_i,

$$\widehat{\mathbf{Var}}(\hat{p}_i) = \hat{p}_i(1-\hat{p}_i)/n_i). \qquad (58$$

6. Because it borrows information across the entire dataset $\underline{\mathbf{y}}$.

7. Note: if the model is not true, the $\underline{\hat{\boldsymbol{\pi}}}$ could be substantially biased. Therefore, it is important to check whether the model fits the data well.

Inference of Coefficients

8. Now consider the entire vector of probabilities

$$\underline{\boldsymbol{\pi}} = (\pi_1, \ldots, \pi_C)^T. \tag{59}$$

9. The $C \times p$ Jacobian (first -derivative) matrix for $\underline{\boldsymbol{\pi}} = \underline{\boldsymbol{\pi}}(\underline{\boldsymbol{\beta}})$ is

$$\frac{\partial \underline{\boldsymbol{\pi}}}{\partial \underline{\boldsymbol{\beta}}} = \mathbf{QX} \quad \text{where } \mathbf{Q} = \mathbf{Diag}(\pi_i(1-\pi_i)). \tag{60}$$

10. By multivariate delta method, an approximate covariance matrix is

$$\widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\pi}}}) = \mathbf{QX}\ \widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\beta}}})\mathbf{X}^T\ \mathbf{Q} = \mathbf{Q}\,\mathbf{X}\ \left(\ \mathbf{X}^T\,\hat{\mathbf{W}}\,\mathbf{X}\ \right)^{-1}\ \mathbf{X}^T\,\mathbf{Q} \tag{61}$$

11. $\widehat{\mathbf{Var}}(\underline{\hat{\boldsymbol{\pi}}})$ is $C \times C$, its rank is p, same as $\left(\ \mathbf{X}^T\,\hat{\mathbf{W}}\,\mathbf{X}\ \right)^{-1}$.

Goodness-of-Fit Test

If the n_i's are sufficiently large, compare

p-**parameter restricted model** $\mathrm{logit}(\pi_i) = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}$

to C-**parameter saturated model** (an intercept plus $C - 1$ dummy variables to distinguish among the C observation units)

Goodness-of-Fit Test

1. In restricted model, $\hat{\pi}_i$ and $\hat{\mu}_i = n_i \hat{\pi}_i$ be the ML fitted values.

2. In saturated model, $\hat{p}_i = y_i / n_i$ and y_i are the ML estimates for π_i and μ_i.

3. G^2, **deviance**, measures the distance from $\underline{\hat{\boldsymbol{\pi}}}$ to $\hat{p}$

4. G^2 is the likelihood ratio statistic for testing

$$\begin{aligned} H_0 &: \text{ restricted model is true} \\ H_A &: \text{ saturated model is true} \end{aligned}$$

Goodness-of-Fit Test

The statistic is

$$\begin{aligned} G^2 &= 2[\, \boldsymbol{\ell}(\hat{p}; \underline{\mathbf{Y}}) - \boldsymbol{\ell}(\underline{\hat{\boldsymbol{\pi}}}; \underline{\mathbf{Y}})\,] = 2\left[\, \sum_{i=1}^{C} Y_i \, \log \hat{p}_i - \sum_{i=1}^{C} Y_i \log \hat{\pi}_i \,\right] \\ &= 2 \sum_{i=1}^{C} Y_i \, \log \frac{Y_i}{n\hat{\pi}_i} \quad \sim \boldsymbol{\chi}^2_{C-p} \quad \text{(asymptotically)} \end{aligned} \tag{62}$$

Under the null hypothesis, the limiting distribution of G^2 is $\boldsymbol{\chi}^2$ with degree of freedom, $(C - p)$ given by the number of parameters under the alternative C minus p, the number of parameters under the null hypothesis. G^2 has the same limiting distribution as Pearson X^2_p.

Testing Nested Models

1. To test:

$$H_0 \quad : \quad \mathcal{M}_2 \quad \text{simpler model is true} \tag{63}$$

$$H_1 \quad : \quad \mathcal{M}_1 \quad \text{more complicated model is true} \tag{64}$$

2. Suppose model $\mathcal{M}_2$ is a special case of model $\mathcal{M}_1$.

Testing Nested Models

3. Let s_1 and s_2 denote the numbers of parameters in the two models.

4. Let $\underline{\hat{\boldsymbol{\pi}}}_1$ and $\underline{\hat{\boldsymbol{\pi}}}_2$ denote ML estimators of cell probabilities for the two models.

Testing Nested Models

5. Then

$$\Delta G^2_{\mathcal{M}_2-\mathcal{M}_1} = G^2(\mathcal{M}_2) - G^2(\mathcal{M}_1) = 2n\sum \hat{p}_i \log\left(\frac{\hat{\pi}_{1i}}{\hat{\pi}_{i2}}\right) \qquad (65$$

$$\Delta G^2_{\mathcal{M}_2-\mathcal{M}_1} \sim \chi^2_{s_1-s_2}, \qquad (66$$

has the form of -2(log likelihood ratio) for testing the hypothesis $H_0 : \mathcal{M}_2$ holds against $H_A : \mathcal{M}_1$ holds.

6. When the simpler model hold, $G^2(\mathcal{M}_2) - G^2(\mathcal{M}_1)$ is asymptotically chi-squared distributed with $s_1 - s_2$ degrees of freedom.

Testing Non-Nested Models

1. AIC: Akaike information criterion

$$\text{AIC } = -2\left(\text{maximized log likelihood - number of parameters in model}\right) \tag{67}$$

2. With models for categorical Y, this ordering is equivalent to one based on an adjustment of deviance $[G^2 - 2(\text{df})]$, by twice its residual df.

3. The smaller AIC, the better

Horseshoe Crab Data

1. Table 1 (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 1: 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 1 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.

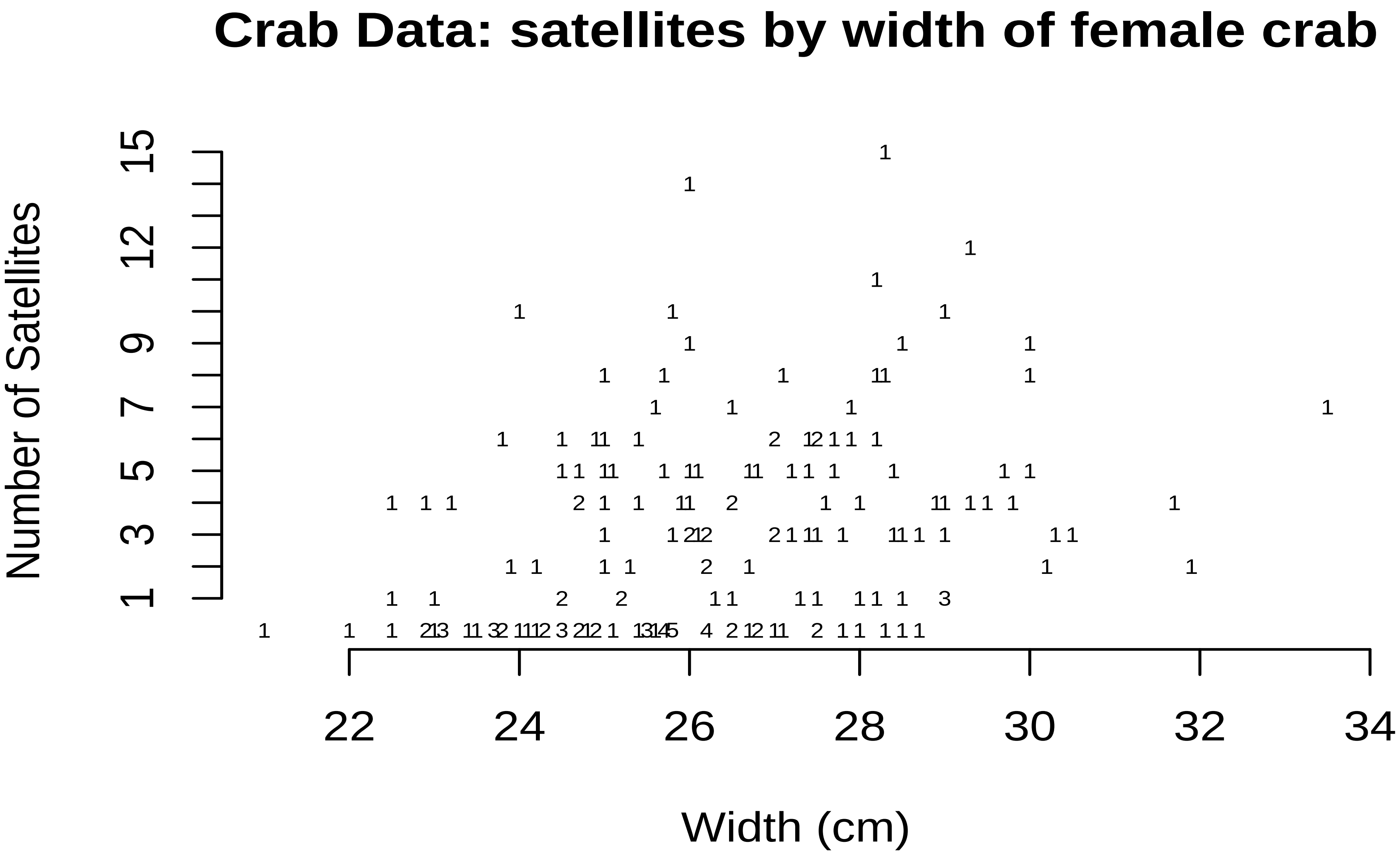

Figure 1: Number of satellites by width of female crab.

Crab Data

1. We use the binary response of whether a female crab has any satellites present

2. $Y = 1$ if a female crab has at least one satellite, and $Y = 0$ if she has no satellite.

3. We first use the crab's width as the sole predictor.

Crab Data: Linear Probability Model

1. For the ungrouped data, let $\pi(x)$ denote the probability that a female horseshoe crab of width x as a satellite.

2. The simplest model to interpret is the linear probability model,

$$\pi(x) = \alpha + \beta x. \tag{68}$$

Crab Data: Linear Probability Model

3. Ordinary least squares fitting yields

$$\hat{\pi}(x) = -1.766 + 0.092x. \tag{69}$$

.

4. At maximum width in this sample of 33.5 cm, its predicted probability equals $\hat{\pi}(x) = -1.766 + 0.092(33.5) = 1.3$ which falls outside the legitimate range for a binomial parameter (so ML fitting fails).

Crab Data: Logistic Model

1. The ML parameter estimates for the logistic regression model are

$$\operatorname{logit}(\pi(x)) = -12.315 + 0.497x. \tag{70}$$

2. The predicted probability of a satellite is the sample analog

$$\hat{\pi} = \frac{\exp(-12.315 + 0.497x)}{1 + \exp(-12.315 + 0.497x)}. \tag{71}$$

Crab Data: Interpretation for Logit Model

3. Since $\hat{\beta} > 0$, the predicted probability $\hat{\pi}$ is higher at larger width values.

4. At the minimum width in this sample of 21.0 cm, the predicted probability is 0.129.; at the maximum width of 33.5 cm the predicted probability equals 0.987.

5. The medial effective level is the width at which the predicted probability equals 0.5, which is $x = \text{EL}_{50} = 1\frac{\hat{\alpha}}{\hat{\beta}} = \frac{12.351}{0.497} = 24.8$.

Crab Data: Interpretation for Logit Model

6. So, the estimated odds of having a satellite increase by 1.64 for each 1 cm increase in width (a 64% increase).

7. Figure 2, (Agresti 5.2) (1996, Fig5.2), is created using similar steps as before, except now we add the predicted logistic regression curve.

Crab Data: Interpretation for Logit Model

8. At the sample mean width of 26.3 cm, the predicted probability of a satellite equals 0.674.

9. The incremental rate of change in the fitted probability at that point is $\hat{\beta}\hat{\pi}(1-\hat{\pi}) = 0.497(0.674)(1-0.674)$.

10. Unlike the linear probability model, the logistic regression model permits the rate of change to vary as x varies.

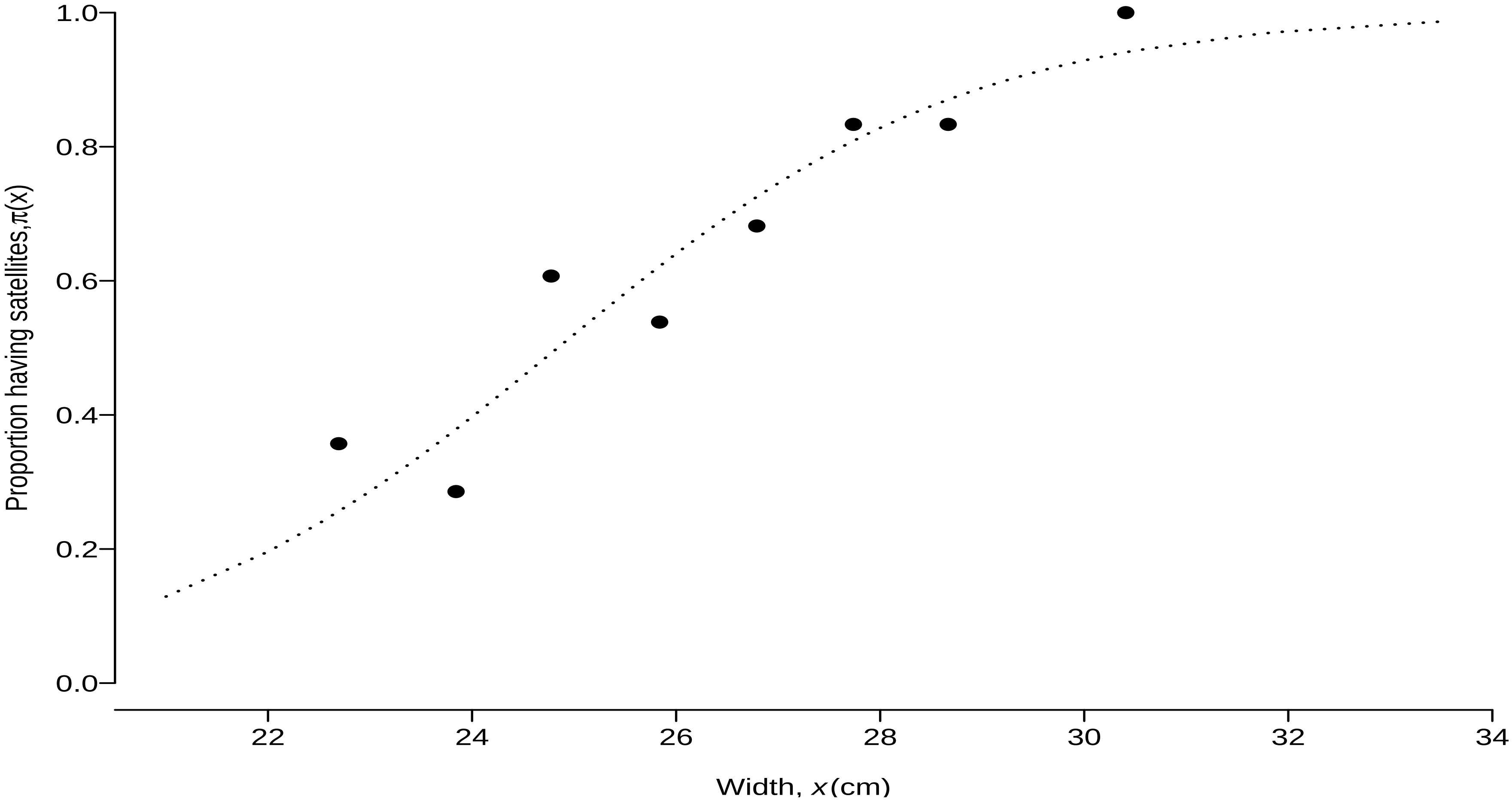

Figure 2: Observed and fitted proportions of satellites by width of female crab.

Crab Data: Interpretation for Logit Model

1. For the female horseshoe crabs, the estimated odds of a satellite multiply by $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ for each centimeter increase in width.

2. There is a 64% increase; the estimated odds of having a satellite increase by 1.64 for each 1 cm increase in width.

Crab Data: Interpretation for Logit Model

3. To illustrate, the mean width value of $x = 26.3$ has a predicted probability of a satellite equal to 0.674, and odds $0.674/(1 - 0.674) = 2.07$.

4. At $x = 27.3 = 26.3 + 1.0$, on can check that the predicted probability equal 0.773, and odds $0.773(1 - 0.773) = 3.40$.

5. But this is a 64% increase; that is, $3.40 = 2.07(1.64)$.

Crab Data: Inference for Logit Model

1. Inference for the logistic regression is asymptotic.

2. Parameter estimators in logistic regression models have large-sample normal distributions.

3. Thus, inference can use the Wald, likelihood-ratio, score triad of methods

4. Standard errors via the inverse of observed Fisher's information can be obtained.

Crab Data: Inference for Logit Model

5. We illustrate logistic regression inferences with the model for the probability a horseshoe crab has a satellite, with width as the predictor.

6. Table 2 showed the fit and standard errors.

Table 2: Computer Output for Logistic Regression Model with Horseshoe Crab Data

```
(From SAS PROC LOGISTIC)
   Criteria For Assessing Goodness Of Fit
Criterion             DF  Value   Value/DF
Deviance             171  194.45     1.13
Pearson Chi-Square 171  165.14     0.96
Log Likelihood            -97.22
```

Table 3: Computer Output for Logistic Regression Model with Horseshoe Crab Data

```
                   Analysis Of Parameter Estimates
                      Standard Wald 95% Confidence Likelihood Ratio  Chi-
Parameter DF  Estimate  Error    Limits            95% Confidence   Square
Intercept  1  12.350   2.628    7.198    17.50      7.45    17.80    22.07
width      1  -0.497   0.101   -0.696    -0.29     -0.70    -0.30    23.89
Scale      0   1.000   0.000    1.000     1.00      1.00     1.00
```

Crab Data: Inference for Logit Model

1. The estimated effect of width in the fitted equation for the probability of a satellite is $\hat{\beta} = 0.497$, with ASE $= 0.102$,

2. The Wald 95% confidence interval for $\hat{\beta}$ is $0.497 \pm 1.96(0.102)$, or $(0.298, 0.697)$.

3. Table 2 reports a likelihood-ratio confidence interval of $(0.308, 0.709)$, based on the profile likelihood function.

4. The confidence interval for the effect on the odds per 1-cm increase in width equals $e^(0.298), e^{0.697} = (1.36, 2.01)$.

Crab Data: Inference for Logit Model

5. We infer that a 1-cm increase in width has at least a 36% increase and at most a doubling in the odds of a satellite.

6. The statistic z $= \dfrac{\hat{\beta}}{\mathsf{ASE}(\hat{\beta})} = 0.497/0.102 = 4.9$ provides strong evidence of a positive width effect $(p < 0.0001)$.

7. The equivalent Wald chi-squared statistic, $\mathsf{z}^2 = 23.9$, has df $= 1$.

Crab Data: Inference for Logit Model

8. The maximized log likelihoods equal $\boldsymbol{\ell}_0 = 112.88$ under $H_0 : \beta = 0$ and $\boldsymbol{\ell}_1 = 97.23$ for the full model.

9. The likelihood-ratio statistic equals $-2(\boldsymbol{\ell}_0 - \boldsymbol{\ell}_1) = 112.88 - 97.23 = 31.3$, with df $= 1$.

10. This provides even stronger evidence than the Wald test.

Crab Data: Predicting $\hat{\pi}(x)$ for Logit Model

1. Most software for logistic regression also reports estimates and confidence intervals for $\hat{\pi}(x)$ (e.g., PROC GENMOD in SAS with the OBSTATS option).

2. Software reports

$$\widehat{\mathbf{Var}}(\hat{\alpha}) = 6.910, \widehat{\mathbf{Var}}(\hat{\beta}) = 0.01035, \widehat{\mathbf{Cov}}(\hat{\alpha}, \hat{\beta}) = -0.2688. \tag{72}$$

Crab Data: Predicting $\hat{\pi}(x)$ for Logit Model

3. The estimated logit $\hat{\pi}(x)$ has large-sample ASE given by the estimated square root of

$$\mathbf{Var}(\hat{\alpha}+\hat{\beta}) = \widehat{\mathbf{Var}}(\hat{\alpha}) + x^2\widehat{\mathbf{Var}}(\hat{\beta}) + 2x\widehat{\mathbf{Cov}}(\hat{\alpha},\hat{\beta}). \tag{73}$$

Crab Data: Predicting $\hat{\pi}(x)$ for Logit Model

4. Consider this for crabs of width $x = 26.5$, near the mean width.

5. The estimated logit is $-12.351 + 0.497(26.5) = 0.825$, and $\hat{\pi}(x) = 0.695$.

6. The estimated predicted logit equals 0.038.

7. The 95% confidence interval for true logit equals $0.825 \pm 1.96(\sqrt{0.038})$ or $(0.44, 1.21)$. This translates to the interval

$$\Big(\frac{\exp(0.44)}{1+\exp(0.44)}, \frac{\exp(1.21)}{1+\exp(1.21)}\Big) = (0.61, 0.77) \qquad (74$$

for the probability of satellites at width 26.5 cm.

$\hat{\pi}(x)$ in Logit Model and Sample Proportion

1. When the logistic regression model truly holds, the model-based estimator of a probability is considerably better than the sample proportion.

2. The model has only two parameters to estimate, whereas the saturated model has a separate parameter for every distinct value of x.

$\hat{\pi}(x)$ in Logit Model and Sample Proportion

3. For instance, at $x = 26.5$, software reports ASE $= 0.04$ for the model-based estimate 0.695, whereas the non-model-based ASE is $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{(0.67)(1-0.67)/6} = 0.19$ for the sample proportion of 0.67 with only 6 observations with $x = 26.5$ and 4 female crabs of 6 had satellites.

4. The 95% confidence intervals are $0.61, 0.77$ using the model versus $(0.30, 0.90)$ using the sample proportion.

$\hat{\pi}(x)$ in Logit Model and Sample Proportion

5. Instead of using only 6 observations, the model uses the information that all 173 observations provide in estimating the two model parameters.

6. The result is a much more precise estimate.

$\hat{\pi}(x)$ in Logit Model and Sample Proportion

1. Reality is a bit more complicated. In practice, the model is not exactly true relationship between $\pi(x)$ and x.

2. However, if it approximates the true probabilities decently, its estimator still tends to be closer than the sample proportion to the true value.

3. The model smooths the sample data, somewhat dampening the observed variability.

4. The resulting estimators tend to be better unless each sample proportion is based on an extremely large sample.

Crab Data: GOF for Logit Model

1. We next consider overall goodness-of-fit analysis for the model using $x =$ width to predict the probability that a female crab has a satellite.

$$\operatorname{logit} \pi(x) = \alpha + \beta x. \tag{75}$$

2. Width takes 66 distinct values for the 173 crabs, with few observations at most widths.

3. One can view the data as a 66×2 contingency table.

4. The two cells in each row count the number of crabs with satellites and the number of crabs without satellites, at that width.

Crab Data: GOF for Logit Model

5. The chi-squared theory for X^2 and G^2 applies when the number of levels of x is fixed, and the number of observations at each level grows.

6. Although we grouped the data using the distinct width values rather than using 173 separate binary responses, this theory is violated here in two ways.

7. First, most fitted counts are very small.

8. Second, when more data are collected, additional width values would occur, so the contingency table would contain more cells rather than a fixed number.

Crab Data: GOF for Logit Model

9. Because of this, X^2 and G^2 for logistic regression models with continuous or nearly continuous predictors do not have approximate chi-squared distributions.

10. Normal approximations can be more appropriate, but no single method has received much attention.

11. One could use X^2 and G^2 to compare the observed and fitted values in grouped form.

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 3 plots these sample mean against the sample mean width in each category.

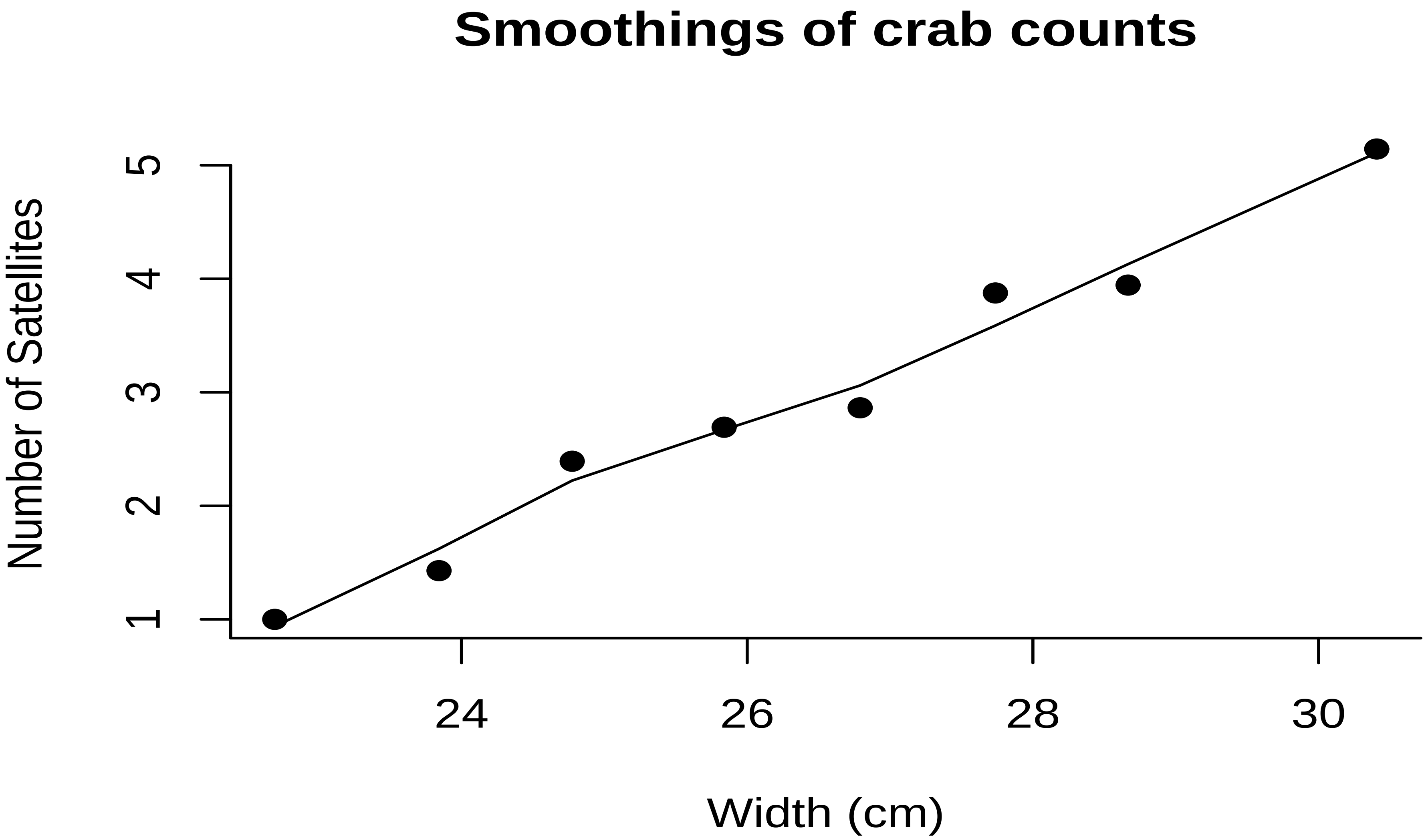

Figure 3: Number of satellites by width of female crab.

Crab Data: Grouped and Ungrouped

1. Table 7 uses the groupings of Crab Data, giving an 8×2 table.

2. In each width category, the fitted value for a yes response is the sum of the estimated probabilities $\hat{\pi}(x)$ for all crabs having width in that category; the fitted value for a no response is the sum of $1 - \hat{\pi}(x)$ for those crabs.

3. The fitted values are then much larger.

4. Then, X^2 and G^2 have better validity, although the chi-squared theory still is not perfect since $\pi(x)$ is not constant in each category.

Table 4: Computer Output for Logistic Regression Model with Grouped Horseshoe Crab Data

```
(From SAS PROC LOGISTIC)
         Model Fit Statistics
                                     Intercept
                  Intercept                and
Criterion              Only         Covariates
AIC                 227.759            201.694
SC                  230.912            208.001
-2 Log L            225.759            197.694
```

Table 5: Computer Output for Logistic Regression Model with Grouped Horseshoe Crab Data

```
        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square   DF Pr > ChiSq
Likelihood Ratio        28.0644    1    <.0001
Score                   25.6828    1    <.0001
Wald                    22.2312    1    <.0001
```

Table 6: Computer Output for Logistic Regression Model with Grouped Horseshoe Crab Data

```
           Analysis of Maximum Likelihood Estimates
                     Standard        Wald
Parameter DF Estimate   Error  Chi-Square Pr > ChiSq
Intercept  1 -11.5128  2.5488     20.4031     <.0001
width      1   0.4646  0.0985     22.2312     <.0001

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
width        1.591      1.312        1.930
```

Table 7: Sample mean and variance of numbers of Satellites

Width (cm)	Numbers of Cases	Numbers of Satellites	Fitted Yes	Fitted No	Sample Mean	Sample Variance
≤ 23.3	14	14	3.64	10.36	1.00	2.77
(23.3, 24.3]	14	20	5.31	8.69	1.43	8.88
(24.3, 25.3]	28	67	13.78	14.22	2.39	6.54
(25.3, 26.3]	39	105	24.23	14.77	2.69	11.38
(26.3, 27.3]	22	63	15.94	6.06	2.86	6.88
(27.3, 28.3]	24	93	19.38	4.62	3.87	8.81
(28.3, 29.3]	18	71	15.65	2.35	3.94	16.88
≥ 29.3	14	72	13.08	0.92	5.14	8.28

Grouped and Ungrouped

1. Let

$$X^2 = \sum \frac{(\text{observed} \ - \ \text{fitted})^2}{\text{fitted}} \tag{76}$$

$$G^2 = 2\sum(\text{observed}) - \log\left(\frac{\text{observed}}{\text{fitted}}\right). \tag{77}$$

2. Their values are $X^2 = 5.3$ and $G^2 = 6.2$.

Grouped and Ungrouped

3. Table 7 has eight binomial samples, one for each width setting; the model has two parameters, so df $= 8 - 2 = 6$.

4. Neither X^2 nor G^2 shows evidence of lack of fit. $p -$ value > 0.4.

5. Thus, we can feel more comfortable about using the model for the original ungrouped data.

Homser and Lemeshow: GOF Test

1. As just noted, with ungrouped data or with continuous or nearly continuous predictors, X^2 and G^2 do not have limiting chi-squared distributions.

2. They are still useful for comparing models, as done above for checking a quadratic term and as we will comparing the nested models.

3. Also, as just noted, one can apply them in an approximate manner to grouped observed and fitted values for a partition of the space of x values.

Homser and Lemeshow: GOF Test

4. As the number of explanatory variables increases, however, simultaneous grouping of values for each variable can produce a contingency table with a large number of cells, most of which have small counts.

5. Regardless of the number of predictors, one can partition observed and fitted values according to the estimated probabilities of success using the original ungrouped data.

Homser and Lemeshow: GOF Test

6. One common approach forms the groups in the partition so they have approximately equal size.

7. With 10 groups, the first pair of observed counts and corresponding fitted counts refers to the $n/10$ observations having the highest estimated probabilities, the next pair refers to the $10/n$ observations having the second decile of estimated probabilities, and so on.

Homser and Lemeshow: GOF Test

8. Each group has an observed count of subjects with each outcome and a fitted value for each outcome.

9. The fitted value for an outcome is the sum of the estimated probabilities for that outcome for all observations in that group.

Homser and Lemeshow: GOF Test

10. This construction is the basis of a test due to Hosmer and Lemeshow (1980, 1989, p.140).

11. They proposed a Pearson statistic comparing the observed and fitted counts for this partition. Let y_{ij} denote the binary outcome for observation j in group i of the partition, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$.

12. Let $\hat{\pi}_{ij}$ denote the corresponding fitted probability for the model fitted to the ungrouped data.

Homser and Lemeshow: GOF Test

13. Their statistic equals

$$\sum_{i=1}^{g} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}. \qquad (78$$

14. When many observations have the same estimated probability, there is some arbitrariness in forming the groups, and different software may report somewhat different values.

Homser and Lemeshow: GOF Test

15. Their Pearson-like statistic does not actually have a chi-square distribution.

16. This statistic does not have a limiting chi-squared distribution, because the observations in a group are not identical trials, since they do not share a common success probability.

17. However, Hosmer and Lemeshow noted that when the number of distinct patterns of covariate values equals the sample size, the null distribution is approximated by chi-squared with df $= g - 2$.

Homser and Lemeshow: GOF Test

18. SAS PROC LOGISTIC calculates the Hosmer-Lemeshow goodness-of-fit test when the LACKFIT option is specified on the MODEL statement.

LRT for Nested Model

1. One can also detect lack of fit by using a likelihood-ratio test to compare working model to more complex ones.

2. If we do not find a more complex model that provides a better fit, this provides same assurance that our fitted model is reasonable.

3. This is more useful in scientific perspective.

4. A large goodness-of-fit statistic is simply indicates there is some lack of fit, but provides no insight about its nature. Comparing a model to a more complex model, on the other hand, indicates whether lack of fit exists of a particular type.

LRT for Nested Model

1. We illustrated this comparison for two models fitted to the grouped crab data.

2. Denote the logistic regression model with width as the sole predictor by M_1 and the simpler model having only an intercept parameter as M_0.

3. That simpler model posits independence of width and having a satellite, and the G^2 goodness-of-fit statistic for testing it is simply the G^2 statistic for testing independence in a two-way contingency table.

LRT for Nested Model

4. For the observed count in the 8×2 , it equals $G^2(M_0) = 34.0$, based on df $= 7$.

5. Since the fit of the model with width as a predictor has $G^2(M_1) = 6.0$ based on df $= 6$,

LRT for Nested Model

6. The comparison statistic for the two models is $G^2(M_0 \mid M_1) = G^2(M_0) - G^2(M_1) = 34.0 - 6.0 = 28.0$, with $\text{df} = 7 - 6 = 1$.

7. In fact, this equals the likelihood-ratio statistic $-2(\boldsymbol{\ell}_0 - \boldsymbol{\ell}_1) =$ for testing that $\beta = 0$ in the logistic regression model fitted to the grouped data of Table 7.

Categorical Predictors

1. Like ordinary regression, logistic regression extends to include qualitative explanatory variables, often called **factors**.

2. In this section we use dummy variables to do this.

ANOVA-Type Representation of Factors

1. For simplicity, we first consider a single factor X, with I categories. In row i of the $I \times 2$ table, y is the number of outcomes in the first column successes out of n trials.

2. We treat y as binomial with parameter π_i.

3. The logit model with a factor is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i. \tag{79}$$

4. The higher β_i is, the higher the value of π_i.

ANOVA-Type Representation of Factors

5. The right-hand side of (79) resembles the model formula for cell means in one-way ANOVA.

6. As in ANOVA, the factor has as many parameters $\{\beta_i\}$ as categories, but one is redundant.

7. With I categories, X has $I - 1$ nonredundant parameters.

ANOVA-Type Representation of Factors

8. One parameter can be set to 0, say $\beta_i = 0$ (last level $= 0$) or $\beta_1 = 0$ (first level $= 0$.

9. If the values do not satisfy this, we can recode so that it is true.

ANOVA-Type Representation of Factors

10. For any $\{\pi_i > 0\}$, $\{\beta_i\}$ exist such that model (79) holds.

11. The model has as many parameters I as binomial observations and is saturated.

12. When a factor has no effect, $\beta_1 = \beta_2 = \cdots = \beta_I = 0$.

13. Since this is equivalent to $\pi_1 = \pi_2 = \cdots = \pi_I$, this model with only an intercept term specifies statistical independence of X and Y.

Categorical Predictors: Dummy Variables

1. An equivalent expression of model (79) uses **“dummy variables”** (**dummy coding**).

2. Let $x_i = 1$ for i observations in row i and $x_i = 0$ otherwise, $i = 1, 2, \ldots, I - 1$.

3. The model is

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \beta_1 x_1 + + \beta_2 x_2 + \cdots + \beta_{I-1} x_{I-1}. \tag{80}$$

Categorical Predictors: Dummy Variables

4. This accounts for parameter redundancy by not forming a dummy variable for category I.

5. The constraint $\beta_I = 0$ in (79) corresponds to this form of dummy variable.

6. The choice of category to exclude for the dummy variable is arbitrary.

7. Some software sets $\beta_1 = 0$; this corresponds to a model with dummy variables for categories 2 through I, but not category 1.

Categorical Predictors: Effect Coding and Zero-Sum Coding

1. Another way to impose constraints sets $\sum_i \beta_i = 0$.

2. Suppose that X has $I = 2$categories, so $\beta_1 = -\beta_2$.

3. This results from **"effect coding"** (**"zero-sum"**) for a dummy variable, $x = 1$ in category 1 and $x = 1$- in category 2.

4. For model (79), regardless of the constraint for $\{\beta_i\}$ and hence $\{\pi_i\}$ are the same.

Categorical Predictors: Effect Coding and Zero-Sum Coding

5. The differences $\hat{\beta}_a - \hat{\beta}_b$ for pairs (a, b) of categories of X are identical and represent estimated log odds ratios.

6. Thus, $\exp(\hat{\beta}_a - \hat{\beta}_b)$ is the estimated odds of success in category a of X divided by the estimated odds of success in category b of X.

7. Reparameterizing a model may change parameter estimates but does not change the model fit or the effects of interest.

Categorical Predictors: Coding

1. The value β_i or $\hat{\beta}_i$ for a single category is irrelevant.

2. Different constraint systems result in different values.

3. For a binary predictor, for instance, using dummy variables with reference value $\beta_2 = 0$, the log odds ratio equals $\beta_1 - \beta_2 = \beta_1$; by contrast, for effect coding with ± 1 dummy variable and hence $\beta - 1 + \beta_2 = 0$, the log odds ratio equals $\beta - 1 + \beta_2 = 2\beta_1 = -2\beta - 2$.

4. A parameter or its estimate makes sense only by comparison with one for another category.

AZT and AIDS Example

1. Table 8 is based on a study escribed in the New York Times (Feb 15, 1991) on the effects of AZT in showing the development of AIDS symptoms.

2. In the study, 338 veterans whose immune system were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until T cells showed severe immune weakness.

3. Table 8 cross-classifies the veterans' race, whether they received AZT immediately, and whether they developed AIDS symptoms during the 3-year study.

Table 8: Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

AZT and AIDS Example: Dummy Variables

1. We identify X with AZT treatment ($x_1 = 1$ for immediate AZT use, $x_2 = 0$ otherwise) and Z with race ($z_1 = 1$ for whites, $z_1 = 0$ for blacks), for predicting the probability that AIDS symptoms developed.

2. Thus, α is he log odds of developing AIDS symptoms for black subject without immediate AZT use,

3. β_1 is the increment to the log odds for those with immediate AZT use, and

4. β_2 is the increment to the log odds for white subjects.

AZT and AIDS Example: Model

5. The model for $\pi(\underline{\mathbf{x}}) = P(Y = 1)$, and Y denote as the symptoms occurs.

6. We will represent the predictors as factors with two levels each.

7. This ensures that we have the correct level specifications.

8. The model that is fit is the "main effects" model

$$\begin{aligned} \text{logit}[P(Y=1)] &= \alpha + \beta_i^X + \beta_k^Z && (81) \\ &= \alpha + \beta_{\text{yes}}^{\text{AZT}} + \beta_{\text{white}}^{\text{race}} && (82) \\ &= \alpha + \beta_1 + \beta_2 && (83) \end{aligned}$$

Table 9: Results for Logit Model Fitted AZT and AIDS Example

```
           Criteria For Assessing Goodness Of Fit
Criterion                  DF          Value        Value/DF
Deviance                    5       335.1512         67.0302
Scaled Deviance             5       335.1512         67.0302
Pearson Chi-Square          5       338.3142         67.6628
Scaled Pearson X2           5       338.3142         67.6628
Log Likelihood                     -167.5756
Algorithm converged.
```

Table 10: Results for Logit Model Fitted AZT and AIDS Example

```
                              Analysis Of Parameter Estimates

                                 Standard    Wald 95% Confidence      Chi-
Parameter    DF    Estimate         Error          Limits            Square    Pr > ChiSq

Intercept     1     -1.0736        0.2629    -1.5889     -0.5582      16.67        <.0001
race1         1      0.0555        0.2886    -0.5102      0.6212       0.04        0.8476
azt1          1     -0.7195        0.2790    -1.2662     -0.1727       6.65        0.0099
Scale         0      1.0000        0.0000     1.0000      1.0000
```

AZT and AIDS Example: Estimation

1. The ML estimate of the effect of AZT is $\hat{\beta}_1 = -0.7195$, (ASE $= 0.279$).

2. Thus, the estimated odds ratio between immediate AZT use and development of AIDS is around $\exp(-0.7195) = 0.487$.

3. The model has four sample logits, one for each binomial response distributions of AZT use and race.

4. Further analysis suggests that an even simpler model may be adequate, since the effect of race is not significant.

Table 11: Parameter Estimates for Logit Model Fitted AZT and AIDS Example

	Definition of Parameters		
Parameter	Last = Zero	First = Zero	Sum = Zero
Intercept	-1.0736	-1.7375	-1.4056
Race–White	0.0555	0.0000	-0.3597
Race–Black	0.0000	-0.0555	0.3597
AZT–Yes	-0.7195	0.0000	0.0277
AZT–No	0.0000	0.7195	-0.0277

AZT and AIDS Example: Estimation

1. The ML estimate of the effect of AZT is $\hat{\beta}_1 = -0.7195$, (ASE $= 0.279$).

2. Thus, the estimated odds ratio between immediate AZT use and development of AIDS is around $\exp(-0.7195) = 0.487$.

3. The model has four sample logits, one for each binomial response distributions of AZT use and race.

4. Further analysis suggests that an even simpler model may be adequate, since the effect of race is not significant.

Alcohol and Infant Malformation

1. Table 12 (Graubard and Korn 1987) illustrates the potential dependence.
2. It refers to a prospective study of maternal drinking and congenital malformations.
3. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption.
4. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations.

Alcohol and Infant Malformation

Table 12: Maternal Drinking and Childbirth Sex Organ Malformation

	Alcohol Consumption (average number of drinks per day)				
Malformation	0	< 1	$1 - 2$	$3 - 5$	≥ 6
Present	48	38	5	1	1
Absent	17,066	14,464	788	126	37

Alcohol and Infant Malformation

5. When a variable is nominal but has only two categories, statistics that treat it as ordinal are still valid.

6. Alcohol consumption, measured as the average number of drinks per day, is an ordinal explanatory variable.

7. This groups a naturally continuous variable.

Table 13: Results for Logit Model Fitted for Alcohol and Infant Malformation Data with Categorical Predictors

```
                              Analysis Of Parameter Estimates

                                    Standard   Wald 95% Confidence      Chi-
Parameter         DF   Estimate        Error          Limits           Square   Pr > ChiSq
Intercept          1    -3.6109       1.0134    -5.5972    -1.6246     12.70       0.0004
Alc         0      1    -2.2627       1.0237    -4.2691    -0.2564      4.89       0.0271
Alc         0.5    1    -2.3309       1.0264    -4.3425    -0.3193      5.16       0.0231
Alc         1.5    1    -1.4491       1.1083    -3.6213     0.7231      1.71       0.1910
Alc         4      1    -1.2254       1.4265    -4.0213     1.5706      0.74       0.3903
Alc         7      0     0.0000       0.0000     0.0000     0.0000       .           .
Scale              0     1.0000       0.0000     1.0000     1.0000
```

Alcohol and Infant Malformation: Categorical Predictor

1. For model (79), we treat malformations as the response and alcohol consumption as an explanatory factor.

2. Regardless of the constraint for β_i, $\{\hat{\alpha}+\hat{\beta}\}$ are the sample logits, reported in Table **??**.

3. For instance,

$$\operatorname{logit}(\hat{\pi}_1) = \hat{\alpha}+\hat{\beta} = \log(48/17,066) = -5.87. \tag{84}$$

Alcohol and Infant Malformation: Categorical Predictor

4. For the coding that constrains $\beta_5 = 0$, $\hat{\alpha} = -3.61$ and $\hat{\beta}_1 = -2.66$.

5. For the coding $\beta_1 = 0$, $\hat{\alpha} = -5.87$.

6. Table 12 and **??** show that except for the slight reversal between the first and second categories of alcohol consumption, the logits and hence the sample proportions of malformation cases increase as alcohol consumption increases.

Alcohol and Infant Malformation: Categorical Predictor

7. The simpler model with all $\beta_i = 0$ specifies independence.

8. For it, $\hat{\alpha}$ equals the logit for the overall sample proportion of malformations, or $\log(93/32481) = -5.86$.

9. To test H_0 : independence(df $= 4$, the Pearson statistic is $X^2 = 12.1$ ($p -$ value $= 0.02$),

10. Likelihood-ratio statistic is $G^2 = 6.2$ ($p -$ value $= 0.19$).

11. These provide mixed signals.

Alcohol and Infant Malformation: Categorical Predictor

12. Table 12 has a mixture of very small, moderate, and extremely large counts.

13. Even though $n = 32,574$, the null sampling distributions of X^2 or G^2 may not be close to chi-squared.

14. The P-values using the exact conditional distributions of X^2 and G^2 are 0.03 and 0.13.

15. These are closer, but still give differing evidence.

Alcohol and Infant Malformation: Categorical Predictor

16. In any case, these statistics ignore the ordinality of alcohol consumption.

17. The sample suggests that malformations may tend to be more likely with higher alcohol consumption.

18. The first two percentages are similar and the next two are also similar, however, and any of the last three percentages changes substantially with the addition or deletion of one malformation case.

Alcohol and Infant Malformation: Ordinal (Linear) Predictor

1. Model (79) treats the explanatory factor as nominal, since it is invariant to the ordering of categories.

2. For ordered factor categories, other models are more parsimonious than this, yet more complex than the independence model.

Alcohol and Infant Malformation: Ordinal (Linear) Predictor

3. For instance, let scores $\{x_1, x_2, \ldots, x_I\}$,describe distances between categories of X. When one expects a monotone effect of X on Y, it is natural to fit the linear logit model

$$\text{logit}(\pi_i) = \alpha + \beta x_i. \tag{85}$$

4. The independence model is the special case $\beta = 0$.

Table 14: Results for Logit Model Fitted for Alcohol and Infant Malformation Data with Categorical Predictors

```
                           Analysis Of Parameter Estimates

                                 Standard   Wald 95% Confidence      Chi-
Parameter         DF   Estimate     Error          Limits          Square   Pr > ChiSq
Intercept          1    -3.6109    1.0134    -5.5972    -1.6246     12.70       0.0004
Alc         0      1    -2.2627    1.0237    -4.2691    -0.2564      4.89       0.0271
Alc         0.5    1    -2.3309    1.0264    -4.3425    -0.3193      5.16       0.0231
Alc         1.5    1    -1.4491    1.1083    -3.6213     0.7231      1.71       0.1910
Alc         4      1    -1.2254    1.4265    -4.0213     1.5706      0.74       0.3903
Alc         7      0     0.0000    0.0000     0.0000     0.0000       .           .
Scale              0     1.0000    0.0000     1.0000     1.0000
```

Alcohol and Infant Malformation: Ordinal (Linear) Predictor

1. The near-monotone increase in sample logits in Table **??** indicates that the linear logit model (85) may fit better than the independence model.

2. As measured, alcohol consumption groups a naturally continuous variable.

3. With scores $\{x_1 = 0, x_2 = 0.5, x_3 = 1.5, x_4 = 4.0, x_5 = 7.0\}$, the last score being somewhat arbitrary,

Alcohol and Infant Malformation: Ordinal (Linear) Predictor

4. Table 15 shows results.

5. The estimated multiplicative effect of a unit increase in daily alcohol consumption on the odds of malformation is $\exp(0.317) = 1.37$.

6. Table 16 shows the observed and fitted proportions of malformation.

7. The model seems to fit well, as statistics comparing observed and fitted counts are $G^2 = 1.95$ and $X^2 = 2.05$, with $=$ df $= 3$.

Table 15: Results for Linear Logit Model Fitted for Alcohol and Infant Malformation Data

```
                               Analysis Of Parameter Estimates
                                  Standard     Wald 95% Confidence        Chi-
Parameter    DF     Estimate         Error            Limits            Square    Pr > ChiSq
Intercept     1      -5.9605        0.1154     -6.1867      -5.7342    2666.41        <.0001
Alc           1       0.3166        0.1254      0.0707       0.5624       6.37        0.0116
Scale         0       1.0000        0.0000      1.0000       1.0000
```

Table 16: Maternal Drinking and Childbirth Sex Organ Malformation

	Alcohol Consumption (average number of drinks per day)				
Malformation	0	< 1	$1-2$	$3-5$	≥ 6
Present	48	38	5	1	1
Absent	17,066	14,464	788	126	37
Logit	-5.87	-5.94	-506	4.84	-3.61
Proportional Malformed					
Observed	0.0028	0.0026	0.0063	0.0079	0.0263
Fitted	0.0026	0.0030	0.0041	0.0091	0.0231

Alcohol and Infant Malformation: Ordinal (Linear) Predictor

1. The Cochran-Armitage trend test i.e., the score test usually gives results similar to the Wald or likelihood-ratio test of $H_0 : \beta = 0$ in the linear logit (85) model.

2. The highly unbalanced counts suggest that it is safest to use the likelihood function through the likelihood-ratio approach.

3. This is also true for estimation.

4. The profile likelihood 95% confidence interval of $(0.02, 0.52)$ for β reported is preferable to the Wald interval of $(0.317 \pm 1.96(0.125) = (0.07, 0.56)$.

Multiple Logistic Regression

1. Like ordinary regression, logistic regression extends to models with multiple explanatory variables.

2. For instance, the model for $\pi(\underline{\mathbf{x}}) = P(Y = 1)$ at values $\underline{\mathbf{x}} = (x_1, \ldots, x_p)^T$ of p predictors is

$$\text{logit}[\pi(\underline{\mathbf{x}})] = \alpha + \beta - 1x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{86}$$

3. The alternative formula, directly specifying $\pi(\underline{\mathbf{x}})$, is

$$\pi(\underline{\mathbf{x}}) = \frac{\exp(\alpha + \beta - 1x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta - 1x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}. \tag{87}$$

Multiple Logistic Regression

4. The parameter β_i refers to the effect of x_i on the log odds that $Y = 1$, controlling the other $_j$.

5. For instance, $\exp(\beta_i)$ is the multiplicative effect on the odds of a 1-unit increase in x, at fixed levels of other xs.

6. An explanatory variable can be qualitative, using dummy variables for categories.

Multiple Logistic Regression: Crab Data

1. Like ordinary regression, logistic regression can have a mixture of quantitative and qualitative predictors.

2. We illustrate with the horseshoe crab data, using the female crab width and color as predictors.

3. Color has five categories: light, medium light, medium, medium dark, dark.

4. It is a surrogate for age, older crabs tending to be darker.

5. The sample contained no light crabs, so our models use only the other four categories.

Crab Data: Color as a Categorical Variable

6. We first treat color as qualitative.

7. The four categories use three dummy variables.

Crab Data: Color as a Categorical Variable

8. The model is

$$\text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x. \qquad (88)$$

where $\pi = P(Y = 1)$, $x =$ width, in centimeters, and

$$\begin{aligned} c_1 &= \text{1 for medium-light color, and 0 otherwise,} \\ c_2 &= \text{1 for medium color, and 0 otherwise,} \\ c_3 &= \text{1 for medium-dark color, and 0 otherwise.} \end{aligned}$$

9. The crab color is dark (category 4) when $c_1 = c_2 = c_3 = 0$.

Table 17: Results for Multiple Logit Model Fitted Crab Data

```
             Analysis of Maximum Likelihood Estimates
                                  Standard          Wald
Parameter    DF    Estimate          Error    Chi-Square    Pr > ChiSq
Intercept     1    -12.7151         2.7618       21.1965        <.0001
c1            1      1.3299         0.8525        2.4335        0.1188
c2            1      1.4023         0.5484        6.5380        0.0106
c3            1      1.1061         0.5921        3.4901        0.0617
width         1      0.4680         0.1055       19.6573        <.0001
```

Table 18: Results for Multiple Logit Model Fitted Crab Data

```
           Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits
c1           3.781       0.711      20.102
c2           4.065       1.387      11.909
c3           3.023       0.947       9.646
width        1.597       1.298       1.964
```

Crab Data: Color as a Categorical Variable

1. Table 17 shows the ML parameter estimate.

2. For instance, for dark crabs, $\text{logit}(\hat{\pi}) = -12.715 + 0.468x$; by contrast, for medium-light crabs, $c_1 = 1$, and $\text{logit}(\hat{\pi}) = (-12.715 + 1.330) + 0.468x$.

3. At the average width of 26.3 cm, $\hat{\pi} = 0.399$ for dark crabs and 0.715 for medium-light crabs.

Crab Data: Color as a Categorical Variable

4. The model assumes a lack of interaction between color and width in their effects.

5. Width has the same coefficient 0.468 for all colors, so the shapes of the curves relating width to π are identical.

6. For each color, a 1-cm increase in width has a multiplicative effect of $\exp(0.468) = 1.60$ on the odds that $Y = 1$.

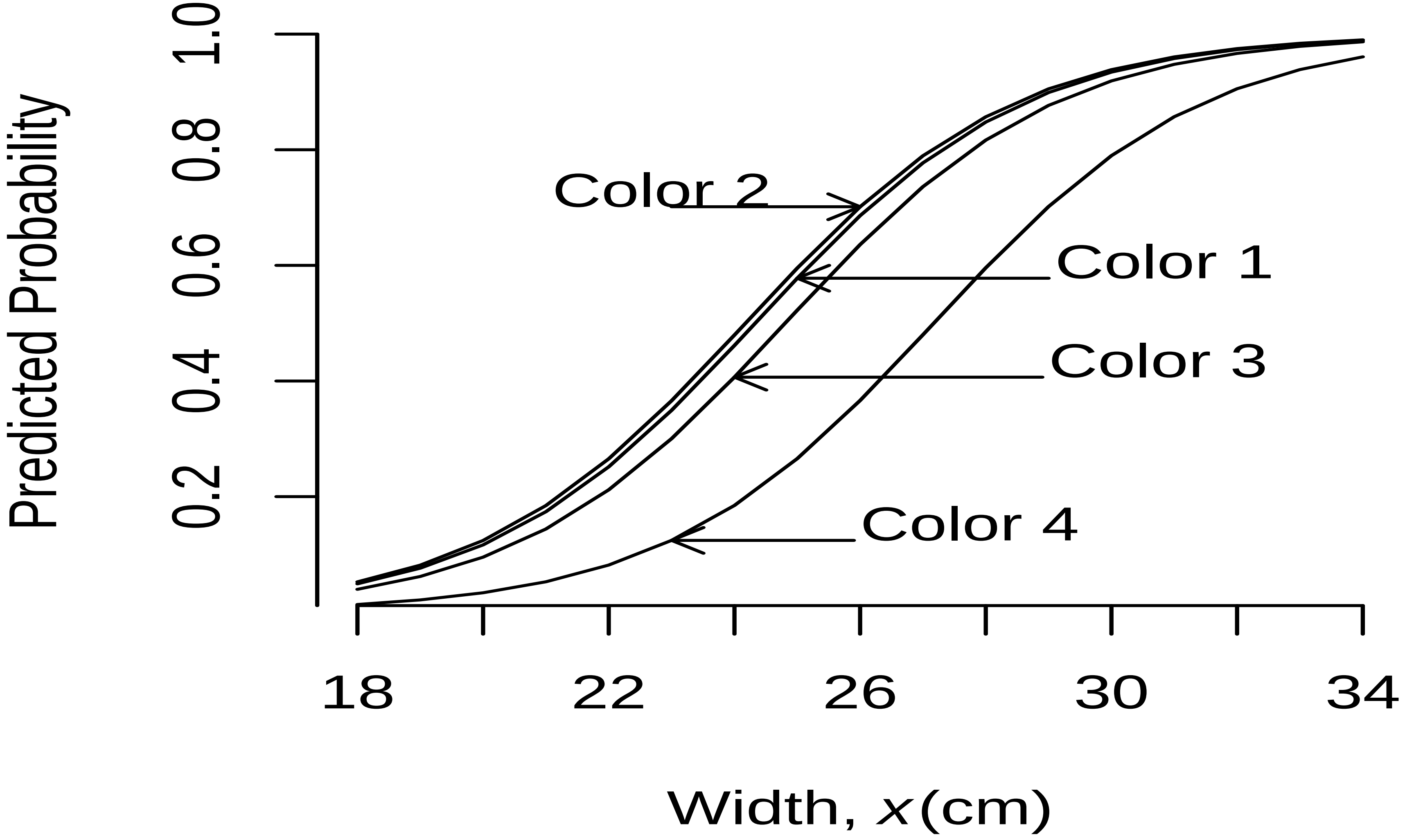

Figure 4: Logistic regression model using width and color predictors of satellite presence

Crab Data: Color as a Categorical Variable

1. Figure 4 displays the fitted model.

2. Any one curve equals any other curve shifted to the right or left.

3. The parallelism of curves in the horizontal dimension implies that any two curves never cross.

4. At all width values, color 4 dark has a lower estimated probability of a satellite than the other colors.

Crab Data: Color as a Categorical Variable

5. There is a noticeable positive effect of width.

6. The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors.

7. For instance, the difference for medium-light crabs and dark crabs equals 1.330.

Crab Data: Color as a Categorical Variable

8. At any given width, the estimated odds that a medium-light crab has a satellite are $\exp(1.330) = 3.8$ times the estimated odds for a dark crab.

9. At width $x = 26.3$, the odds equal $0.715/0.285 = 2.51$ for a medium-light crab and $0.399/0.601 = 0.66$ for a dark crab, for which $2.51/0.66 = 3.8$.

Crab Data: Model Comparison

1. To test whether color contributes significantly to model (88), we test $H_0 : \beta + 1 = \beta_2 = \beta_3 = 0$.

2. This states that controlling for width, the probability of a satellite is independent of color.

3. We compare the maximized log-likelihood $\boldsymbol{\ell}_1$ for the full model (88) to $\boldsymbol{\ell}_0$ for the simpler model.

4. The test statistic $-2(\boldsymbol{\ell}_0 - \boldsymbol{\ell}_1) = 7.0$ has df $= 3$, the difference between the numbers of parameters in the two models.

5. The chi-squared p-value of 0.07 provides slight evidence of a color effect.

Crab Data: Model Comparison

6. The more complex model allowing color × width interaction has three additional terms, the cross-products of width with the color dummy variables.

7. Fitting this model is equivalent to fitting logistic regression with width predictor separately for crabs of each color.

8. Each color then has a different shaped curve relating width to $P(Y=1)$, so a comparison of two colors varies according to the width value.

Crab Data: Model Comparison

9. The likelihood-ratio statistic comparing the models with and without the interaction terms equals 4.4, with df $= 3$.

10. The evidence of interaction is weak $p -$ value $= 0.22$.

Crab Data: Color as a Ordinal Variable

1. Color has ordered categories, from lightest to darkest. A simpler model yet treats this predictor as quantitative.

2. Color may have a linear effect, for a set of monotone scores.

3. To illustrate, for scores $c = \{1, 2, 3, 4\}$ for the color categories

Table 19: Crab Data: Color as Ordinal Variable

```
                   Analysis of Maximum Likelihood Estimates

                                         Standard              Wald
Parameter     DF     Estimate               Error        Chi-Square     Pr > ChiSq

Intercept      1     -10.0708              2.8069           12.8733         0.0003

color          1      -0.5090              0.2237            5.1791         0.0229

width          1       0.4583              0.1040           19.4129         <.0001
```

Table 20: Crab Data: Color as Ordinal Variable

```
          Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
color        0.601       0.388       0.932
width        1.581       1.290       1.939
```

Crab Data: Color as a Ordinal Variable

1. The model

$$\text{logit}(\pi) = \alpha + \beta_1 c + \beta_4 x. \tag{89}$$

has $\hat{\beta}_1 = -0.509(\text{SE} = 0.224)$ and $\hat{\beta}_2 = 0.458(\text{SE} = 0.104)$.

2. This shows strong evidence of an effect for each.

3. At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by $\exp(-0.509) = 0.60$.

Crab Data: Color as a Ordinal Variable

1. The likelihood-ratio statistic comparing this fit to the more complex model (88) having a separate parameter for each color equals 1.7 df $= 2$.

2. This statistic tests that the simpler model (89) is adequate, given that model (88) holds.

3. It tests that when plotted against the color scores, the color parameters in (88) follow a linear trend.

4. The simplification seems permissible $p -$ value $= 0.44$.

Crab Data: Color with Two Levels

1. The color parameter estimates in the qualitative-color model (88) are $(1.33, 1.40, 1.11, 0)$, the 0 value for the dark category reflecting its lack of a dummy variable.

2. Although these values do not depart significantly from a linear trend, the first three are quite similar compared to the last one.

3. Thus, another potential color scoring for model (89) is $(1, 1, 1, 0)$; that is, scores $= 0$ for dark-colored crabs, and scores1 otherwise.

Table 21: Crab Data: Color as Two Levels

```
             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF     Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1     -12.9795      2.7272       22.6502        <.0001
cscore        1       1.3005      0.5259        6.1162        0.0134
width         1       0.4782      0.1041       21.0841        <.0001
```

Table 22: Crab Data: Color as Two Levels

```
           Odds Ratio Estimates

             Point          95% Wald

Effect     Estimate      Confidence Limits

cscore        3.671        1.310      10.290

width         1.613        1.315       1.979
```

Crab Data: Color as Two Levels

1. The likelihood-ratio statistic comparing model (89) with these binary scores to model (88) equals 0.5 $(\text{df} = 2)$, showing that this simpler model is also adequate.

2. Its fit is

$$\text{logit}(\hat{\pi}) = -12.980 + 1.300 + 0.478x, \tag{90}$$

   with standard errors 0.526 and 1.104.

3. At a given width, the estimated odds that a lighter-colored crab has a satellite are $\exp(1.300) = 3.7$ times the estimated odds for a dark crab.

Crab Data: Summary

1. In summary, the qualitative-color model, the quantitative-color model with scores $\{1, 2, 3, 4\}$, and the model with binary color scores $\{1, 1, 1, 0\}$ all suggest that dark crabs are least likely to have satellites.

2. A much larger sample is needed to determine which color scoring is most appropriate.

3. It is advantageous to treat ordinal predictors in a quantitative manner when such models fit well.

4. The model is simpler and easier to interpret, and tests of the predictor effect are more powerful when it has a single parameter rather than several parameters.

Model Building

1. Having studied the basics of fitting and interpreting logistic regression models, we now turn our attention to building and applying them.

2. With several explanatory variables, there are many potential models.

3. Model selection for logistic regression faces the same issues as for ordinary regression.

4. The selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions.

Model Building

There are two competing goals:

1. The model should be complex enough to fit the data well.

2. On the other hand, it should be simple to interpret, smoothing rather than overfitting the data.

Model Building

1. Most studies are designed to answer certain questions.

2. Those questions guide the choice of model terms.

3. Confirmatory analyses then use a restricted set of models.

4. For instance, a study hypothesis about an effect may be tested by comparing models with and without that effect.

5. For studies that are exploratory rather than confirmatory, a search among possible models may provide clues about the dependence structure and raise questions for future research.

Model Building

6. In either case, it is helpful first to study the effect on Y of each predictor by itself using graphics incorporating smoothing for a continuous predictor or a contingency table for a discrete predictor.

7. This gives a “feel” for the marginal effects.

8. Unbalanced data, with relatively few responses of one type, limit the number of predictors for the model.

Model Building: Sample Sizes and Variables

9. One guideline suggests at least 10 outcomes of each type should occur for every predictor (Peduzzi et al., 1996).

10. If ys 1 only 30 times out of $n = 1000$, for instance, the model should contain no more than about three x terms.

11. Such guidelines are approximate, and this does not mean that if you have 500 outcomes of each type you are well served by a model with 50 predictors.

Model Building: Multicollinearity

1. Many model selection procedures exist, no one of which is always best. Cautions that apply to ordinary regression hold for any generalized linear model.

2. For instance, a model with several predictors may suffer from "multicollinearity"-correlations among predictors making it seem that no one variable is important when all the others are in the model.

Model Building: Multicollinearity

3. A variable may seem to have little effect because it overlaps considerably with other predictors in the model, itself being predicted well by the other predictors.

4. Deleting such a redundant predictor can be helpful, for instance to reduce standard errors of other estimated effects.

Crab Data: All Main Effect

1. The horseshoe crab data set in Table 1 has four predictors: color (four categories), spine condition (three categories), weight, and width of the carapace shell.

2. We now fit a logistic regression model using all these to predict whether the female crab has satellites ($y = 1$).

Crab Data: All Main Effect

3. We start by fitting a model containing main effects,

$$\begin{aligned} & \text{logit}[P(Y=1)] \\ = \; & \alpha + \beta_1 \text{weight} + \beta - 2\text{width} + \beta_3 c - 1 + \beta_4 c_2 + \beta_5 c_3 \\ & \quad + \beta_6 s_1 + \beta_7 s_2. \end{aligned} \tag{91}$$

treating color c_i and spine condition s_j as qualitative factors, with dummy variables for the first three colors and the first two spine conditions.

Table 23: Crab Data: All Main Effect

```
        Testing Global Null Hypothesis: BETA=0
Test                     Chi-Square       DF     Pr > ChiSq
Likelihood Ratio            40.5565        7         <.0001


        Type 3 Analysis of Effects
                        Wald
Effect       DF    Chi-Square    Pr > ChiSq
weight        1        1.3765        0.2407
width         1        1.8152        0.1779
color         3        7.1610        0.0669
spine         2        1.0105        0.6034
```

Table 24: Crab Data: All Main Effect

```
              Analysis of Maximum Likelihood Estimates
                                  Standard          Wald
Parameter      DF     Estimate       Error    Chi-Square    Pr > ChiSq
Intercept       1      -9.2734      3.8378        5.8386        0.0157
weight          1       0.8258      0.7038        1.3765        0.2407
width           1       0.2631      0.1953        1.8152        0.1779
color     1     1       1.6087      0.9355        2.9567        0.0855
color     2     1       1.5058      0.5667        7.0607        0.0079
color     3     1       1.1198      0.5933        3.5624        0.0591
spine     1     1      -0.4003      0.5027        0.6340        0.4259
spine     2     1      -0.4963      0.6292        0.6222        0.4302
```

Crab Data: All Main Effect

1. A likelihood-ratio test that Y is jointly independent of these predictors simultaneously tests $H_0 : \beta_1 = \cdots = \beta_7 = 0$.

2. The test statistic equals 40.6 with df $= 7$, $p -$ value < 0.0001.

3. This shows extremely strong evidence that at least one predictor has an effect.

4. Although the overall test is highly significant, the Table **??** results are discouraging.

Crab Data: All Main Effect

5. The estimates for weight and width are only slightly larger than their SE values.

6. The estimates for the factors compare each category to the final one as a baseline.

7. For color, the largest difference is less than two standard errors; for spine condition, the largest difference is less than a standard error.

Crab Data: All Main Effect

8. The small p-value for the overall test, yet the lack of significance for individual effects, is a warning sign of multicollinearity.

9. We have showed strong evidence of a width effect.

10. Controlling for weight, color, and spine condition, little evidence remains of a partial width effect.

Crab Data: All Main Effect

11. However, weight and width have a strong correlation (0.887).

12. For practical purposes they are equally good predictors, but it is nearly redundant to use them both.

13. It is also not usually sensible to consider a model with interaction but not the main effects that make up that interaction.

Selection Algorithms

1. In exploratory studies, an algorithmic method for searching among models can be informative if we use results cautiously.

2. Goodman (1971) proposed methods analogous to forward selection and backward elimination in ordinary regression.

Forward Selection

1. **Forward selection** adds terms sequentially until further additions do not improve the fit.

2. At each stage it selects the term giving the greatest improvement in fit.

3. The minimum p-value for testing the term in the model is a sensible criterion, since reductions in deviance for different terms may have different df values.

4. A stepwise variation of this procedure retests, at each stage, terms added at previous stages to see if they are still significant.

Backward Elimination

1. **Backward elimination** begins with a complex model and sequentially removes terms.

2. At each stage, it selects the term for which its removal has the least damaging effect on the model e.g., largest p-value.

3. The process stops when any further deletion leads to a significantly poorer fit.

Stepwise Selection

1. **Stepwise selection procedure** combines the forward and backward procedures.

2. At each stage, a variable is either added, dropped, or interchanged with another variable, according to a set of rules until a stopping criterion is met.

Selection Algorithm: Summary

1. With either approach, for qualitative predictors with more than two categories, the process should consider the entire variable at any stage rather than just individual dummy variables.

2. Add or drop the entire variable rather than just one of its dummies.

3. Otherwise, the result depends on the coding.

4. The same remark applies to interactions containing that variable.

Selection Algorithm: Summary

5. **Many statisticians prefer backward elimination** over forward selection, feeling it safer to delete terms from an overly complex model than to add terms to an overly simple one.

6. Forward selection can stop prematurely because a particular test in the sequence has low power.

7. Neither strategy necessarily yields a meaningful model.

Selection Algorithm: Summary

8. Use variable selection procedures with caution!

9. When you evaluate many terms, one or two that are not important may look impressive simply due to chance.

10. For instance, when all the true effects are weak, the largest sample effect may substantially overestimate its true effect. See Westfall and Wolfinger (1997) and Westfall and Young (1993) for ways to adjust P-values to take multiple tests into account.

Selection Algorithm: Softwares

1. Some software has additional options for selecting a model.
2. One approach attempts to determine the best model with some fixed number of terms, according to some criterion.
3. If such a method and backward and forward selection procedures yield quite different models, this is an indication that such results are of dubious use.
4. Another such indication would be when a quite different model results from applying a given procedure to a bootstrap sample of the same size from the sample distribution.

Selection Algorithm: Softwares

5. Finally, statistical significance should not be the sole criterion for inclusion of a term in a model.

6. It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant.

Selection Algorithm: Softwares

7. Keeping it in the model may help reduce bias in estimated effects of other predictors and may make it possible to compare results with other studies where the effect is significant perhaps because of a larger sample size.

8. Algorithmic selection procedures are no substitute for careful thought in guiding the formulation of models.

Model Building

1. In selecting a model, we are mistaken if we think that we have found the true one.

2. Any model is a simplification of reality.

3. For instance, width does not exactly have a linear effect on the probability of satellites, whether we use the logit link or the identity link.

Model Building

4. What is the logic of testing the fit of a model when we know that it does not truly hold?

5. A simple model that fits adequately has the advantages of model parsimony.

6. If a model has relatively little bias, describing reality well, it tends to provide more accurate estimates of the quantities of interest.

Akaike information criterion (AIC)

1. Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest.

2. The best known is the **Akaike information criterion (AIC)**.

3. It judges a model by how close its fitted values tend to be to the true values, in terms of a certain expected value.

Akaike information criterion (AIC)

4. Even though a simple model is farther from the true model than is a more complex model, it may be preferred because it tends to provide better estimates of certain characteristics of the true model, such as cell probabilities.

5. Thus, the optimal model is the one that tends to have fit closest to reality.

6. Given a sample, Akaike showed that this criterion selects the model that minimizes

$$\text{AIC} = -2(\text{maximized log likelihood} - \text{number of parameters in model}).($$

7. This penalizes a model for having many parameters.

8. With models for categorical Y, this ordering is equivalent to one based on an adjustment of the deviance, $G^2 - 2(\text{df})]$, by twice its residual df.

9. For cogent arguments supporting this criterion, see Burnham and Anderson (1998).

Crab Data

1. We illustrate AIC for model selection. Of models using the three basic variables, AIC is smallest AIC$= 197.5$ for $(C + W)$, having main effects of color and width.

2. The simpler model having a dummy variable for whether a crab is dark fares better yet AIC$= 194.0$.

3. Either model seems reasonable.

4. We should balance the lower AIC for the simpler model against its having been suggested by the fit of $C + W$.

New Model Building Strategies for Data Mining

1. As computing power continues to explode, enormous data sets are more common.

2. A financial institution that markets credit cards may have observations for millions of subjects to whom they sent advertising, on whether they applied for a card.

3. For their customers, they have monthly data on whether they paid their bill on time plus information on many variables measured on the credit card application.

New Model Building Strategies for Data Mining

4. The analysis of huge data sets is called **data mining**.

5. Model building for huge data sets is challenging.

6. There is currently considerable study of alternatives to traditional statistical methods, including automated algorithms that ignore concepts such as sampling error or modelling.

7. Significance tests are usually irrelevant, as nearly any variable has a significant effect if n is sufficiently large.

8. Model-building strategies view some models as useful for prediction even if they have complex structure.

New Model Building Strategies for Data Mining

9. Nonetheless, a point of diminishing returns still occurs in adding predictors to models.

10. After a point, new predictors tend to be so correlated with a linear combination of ones already in the model that they do not improve predictive power.

11. For large n, inference is less relevant than summary measures of predictive power.