MSAN 601
Linear Regression Analysis

Paul Intrevado

Introduction to Nonlinear Regression

October 6, 2016

UNIVERSITY OF
SAN FRANCISCO

Master of Science
in Analytics

- There are occasions when an empirically indicated or theoretically justified nonlinear regression model is more appropriate than its linear counterpart
  - Growth from birth to maturity
  - Pharmacological dose-response relationships
- We will explore some nonlinear regression models, how to obtain estimates of the regression parameters in such models, and how to make inferences about said regression parameters

Section 1

Linear and Nonlinear Regression Models

- We previously considered linear regression models, i.e., models that are linear in the parameters, which can be generally represented as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Linear regression models also include more complex models that above, including polynomial variables, interaction terms and non-linear transforms of variables
- In general, we can state a linear regression model in the form

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i$$

where $\mathbf{X}_i$ is the vector of observations on the predictor variables for the i^{th} case

$$\mathbf{X}_i' = \begin{bmatrix} 1 & X_{i1} & \ldots & X_{i,p-1} \end{bmatrix}$$

- $\boldsymbol{\beta}$ is the vector of regression coefficients, and $f(\mathbf{X}_i, \boldsymbol{\beta})$ represents the expected value $E\{Y_i\}$, which for linear regression models equals $f(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i'\boldsymbol{\beta}$

- Nonlinear regression models are of the same basic form

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

- An observation Y_i is still the sum of a mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$ given by a nonlinear response function $f(\mathbf{X}, \boldsymbol{\gamma})$ and the error term ε_i
- Just as in linear regression, the error terms are assumed to have an expectation of zero, constant variance and to be uncorrelated
- Often, a normal error model is utilized, i.e., error terms are assumed independent normal random variables with constant variance

- Nonlinear regression models are of the same basic form

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

- An observation Y_i is still the sum of a mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$ given by a nonlinear response function $f(\mathbf{X}, \boldsymbol{\gamma})$ and the error term ε_i
- Just as in linear regression, the error terms are assumed to have an expectation of zero, constant variance and to be uncorrelated
- Often, a normal error model is utilized, i.e., error terms are assumed independent normal random variables with constant variance

n.b. The parameter vector here in the response function $f(\mathbf{X}_i, \boldsymbol{\gamma})$ is denoted by $\boldsymbol{\gamma}$ rather than $\boldsymbol{\beta}$ as a reminder that the response function is nonlinear in the parameters

- Exponential regression models are widely used class of non-linear regression model
- Assuming a singular predictor variable, one form of the regression model with normal error terms is

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$

where

- γ_0 and γ_1 are parameters
- X_i are known constants
- the ε_i are independent $\mathcal{N}(0, \sigma^2)$

- The response function for this model is

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 \exp(\gamma_1 X_i)$$

- A more general nonlinear exponential regression model in one predictor variable with normal, independent error terms with constant variance σ^2 is

$$Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \varepsilon_i$$

- The response function for this regression model is

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i)$$

- Exponential regression models are commonly used in growth studies where the rate of growth at a given time X is proportional to the amount of growth remaining as time increases, with γ_0 representing the maximum growth value

- Another important nonlinear regression model is the logistic regression model, which, with one predictor variable and normal error terms, is

$$Y_i = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \varepsilon_i$$

where the error terms ε_i are independent normal with constant variance σ^2

- The response function is

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)}$$

- Logistic regression models have been used in population studies to related the number of species (Y) to time (X)
- Logistic regression models are also widely used when the response variable is qualitative, e.g., will a household purchase a new car this year (yes/no)

General Form of Nonlinear Regression Models

- Each Y_i observation is postulated to be the sum of a mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$ based on a given nonlinear response function and a random error term ε_i
- The error terms ε_i are often assumed to be independent random normal variables with constant variance

General Form of Nonlinear Regression Models

- Each Y_i observation is postulated to be the sum of a mean response $f(\mathbf{X}_i, \gamma)$ based on a given nonlinear response function and a random error term ε_i
- The error terms ε_i are often assumed to be independent random normal variables with constant variance
- A key difference in non-linear regression is that the number of regression parameters is is not necessarily directly related to the number of X variables in the model
  - In linear regression, if there are $p-1$ regressors in the model, there there are p regression coefficients in the model
- Therefore, in nonlinear regression models, we will denote the number of regressors by q and continue to denote the number of regression parameters by p

E.g. In the exponential regression model

$$Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \varepsilon_i$$

$q = 1$ and $p = 3$

General Form of Nonlinear Regression Models cont'd

- We shall define the vector $\mathbf{X}_i$ of the observations on the X variables without the initial element 1
- The general form of the a nonlinear regression models therefore expressed as follows

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

where

$$\underset{q\times 1}{\mathbf{X}_i'} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iq} \end{bmatrix} \qquad \underset{p\times 1}{\boldsymbol{\gamma}} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix}$$

Subsection 1

Least Squares Estimation in Nonlinear Regression

Least Squares Estimation in Nonlinear Regression

- The concept of least-squares regression extends directly to nonlinear regression models
- The least-squares criterion is

$$\mathcal{Q} = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})]^2$$

  where $f(\mathbf{X}_i, \boldsymbol{\gamma})$ is the mean response for the i^{th} case according to nonlinear response function $f(\mathbf{X}, \boldsymbol{\gamma})$
- The least squares response criterion $\mathcal{Q}$ must be minimized with respect to the nonlinear regression parameters $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ to obtain the least squares estimates
- These estimates can be obtained via a direct numerical search or via the normal equations, although the normal equations themselves require a numerical search, as analytical solutions generally cannot be found

- To obtain the normal equations for a nonlinear regression model

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i$$

we need to minimize the least squares criterion $\mathcal{Q}$

$$\mathcal{Q} = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})]^2$$

with respect to $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$

- The partial derivative of $\mathcal{Q}$ with respect to γ_k is

$$\frac{\partial \mathcal{Q}}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})] \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k} \right]$$

- When the p partial derivates are each set equal to 0 and the parameters γ_k are replaced by the least squares estimates g_k, we obtain (after some simplification) the p normal equations

$$\sum_{i=1}^{n} Y_i \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\gamma=g} - \sum_{i=1}^{n} f(\mathbf{X}_i, \mathbf{g}) \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\gamma=g} = 0$$

for $k = 0, 1, \ldots, p-1$, and where $\mathbf{g}$ is the vector of the least squares estimates g_k

$$\underset{p\times 1}{\mathbf{g}} = \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_{p-1} \end{bmatrix}$$

- The Gauss-Newton, or linearization method uses a Taylor series expansion to approximate the nonlinear regression models with linear terms and then employs ordinary least squares to estimate the parameters
- The Gauss-Newton method being with initial or starting values for the regression parameters $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ denoted $g_0^{(0)}, g_1^{(0)}, \ldots, g_{p-1}^{(0)}$, where the bracketed superscript denotes the iteration number
- Starting values need not be any specific value, but should, when possible, be chosen intelligently, i.e., based on some theoretical or empirical information

- With starting values selected, mean responses of $f(\mathbf{X}_i, \boldsymbol{\gamma})$ for the n cases are approximated by the linear terms in the Taylor series expansion around the starting values $g_k^{(0)}$
- We obtain for the i^{th} case

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) \approx f(\mathbf{X}_i, \mathbf{g}^{(0)}) + \sum_{k=0}^{p-1} \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k} \right]_{\gamma = g^{(0)}} (\gamma_k - g_k^{(0)})$$

where

$$\underset{p \times 1}{\mathbf{g}^{(0)'}} = \begin{bmatrix} g_0^{(0)} & g_1^{(0)} & \cdots & g_{p-1}^{(0)} \end{bmatrix}$$

- Allowing for the following simplifications

$$f_i^{(0)} = f(\mathbf{X}_i, \mathbf{g}^{(0)}) \qquad \beta_k^{(0)} = \gamma_k - g_k^{(0)} \qquad D_{ik}^{(0)} = \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\gamma = g^{(0)}}$$

the Taylor approximation for the means response for the i^{th} case then becomes

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)}$$

and an approximation to the nonlinear regression model model is

$$Y_i \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} \varepsilon_i$$

- When we shift the $f_i^{(0)}$ term to the left and denote the difference $Y_i - f_i^{(0)}$ by $Y_i^{(0)}$, we obtain the following linear regression model approximation

$$Y_i^{(0)} \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i$$

where

$$Y_i^{(0)} = Y_i - f_i^{(0)}$$

n.b. The above regression model approximation is of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$Y_i^{(0)} \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i$$

- The responses $Y_i^{(0)}$ are the residuals, the X variables are the observations $D_{ik}^{(0)}$ and each regression coefficient $\beta_k^{(0)}$ represent the difference between the true regression parameter and the initial estimate of the parameter
- The regression coefficients represent the adjustment amounts by which the initial regression coefficients must be corrected
- The purpose of fitting the linear regression model approximation is therefore to estimate the regression coefficients $\beta_k^{(0)}$ and use these estimates to adjust the initial starting estimates of the regression parameters

n.b. In fitting the linear regression approximation, note that there is no intercept in the model, therefore, if using a computer package, be sure to specify this requirement

- At this point, we estimate the coefficients $\boldsymbol{\beta}^{(0)}$ by $\mathbf{b}^{(0)}$, and use those estimates to compute a revised estimate of the regression coefficients $g_k^{(1)}$

$$\mathbf{g}^{(1)} = \mathbf{g}^{(0)} + \mathbf{b}^{(0)}$$

- Confirm that $SSE^{(0)} > SSE^{(1)}$, and repeat the procedure, obtaining new estimates $\mathbf{g}_k^{(\cdot)}$, until the difference between subsequent SSE's becomes negligible
- The Gauss-Newton method can work effectively, but poor selection of initial starting values may result in slow convergence, convergence to a local (as opposed to global convergence) or even divergence
- Execute the Gauss-Newton method from various starting values to ensure it converges to the same solution
- For nonlinear regression, $SSTO$ is not necessarily equal to $SSR + SSE$, therefore, R^2 is not a meaningful descriptive statistic

- Other frequently used direct search procedures include steepest descent and the Marquardt algorithm
- Steepest descent is particularly effective when starting values $\mathbf{g}^{(0)}$ are poor
- The Marquardt algorithm seeks to utilize the best features of the Gauss-Newton method and the method of steepest descent, occupying a middle ground between the two

- Some computer packages may have a library of typical nonlinear functions, but prior knowledge of the expected functional form can be very useful
- The functional form of nonlinear regression models does not necessarily lend itself to adding and/or deleting predictor variables
- Validation of the the selected nonlinear regression model can be performed in the same fashion as for linear regression models

Subsection 2

Inferences about Nonlinear Regression Parameters

Inferences about Nonlinear Regression Parameters

- Exact inference procedures about nonlinear regression parameters are unfortunately not available
- Consequently, inferences about nonlinear regression parameters are usually based on large-sample theory
- This theory tells us that the least squares and maximum likelihood estimators for nonlinear regression models with normal error terms, when the sample size is large, are approximately normally distributed, are almost unbiased and have almost minimum variance

Estimate of the Error Term Variance

- This estimate is of the same form as for linear regression , the error sum of squares again being the sum of the squared residuals

$$MSE = \frac{SSE}{n-p} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p} = \frac{[Y_i - f(\mathbf{X}_i, \mathbf{g})]^2}{n-p}$$

  where $\mathbf{g}$ is the vector of final parameter estimates
- For nonlinear regression, MSE is not an unbiased estimator of σ^2, but the bias is small when the sample size is large

When is Large Sample Theory Applicable?

- There are no fixed rules that can tell us with certainty whether or not a sample size is sufficiently large for a given nonlinear regression application, but there are a few guidelines/indicators that can help
- Quick convergence of the iterative procedure in finding the estimates of the nonlinear regression parameter is often an indication that the linear approximation to the nonlinear regression model is a good one, and that the asymptotic properties of the regression estimates apply

Interval Estimation of a Single γ_k

- Based on large sample theory, the following approximate result holds when the sample size is large and the error terms are normally distributed

$$\frac{g_k - \gamma_k}{s\{g_k\}} \sim t_{(n-p)} \qquad k = 0, 1, \ldots, p-1$$

- The confidence limits for any single γ_k are

$$g_k \pm t_{(1-\alpha/2;n-p)} s\{g_k\}$$

Section 2

Logistic Regression

- In a variety of regression applications, the response variable of interest has only two possible qualitative outcomes, and therefore can be represented by a binary indicator variable taking on values of 0 and 1
- A binary response variable, exclusively taking on the values of 0 and 1, is said to involve binary responses or dichotomous responses

Response Function with Binary Outcomes

- Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = 0, 1$$

- The expected response $E\{Y_i\}$ has a special meaning in the is case (given $E\{\varepsilon_i\} = 0$)

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

- Consider Y_i to be a Bernoulli random variable for which we can state the probability distribution as follows

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

- Therefore π_i is the probability that $Y_i = 1$ and $1 - \pi_i$ is the probability that $Y_i = 0$

Response Function with Binary Outcomes cont'd

- The expected value of $E\{Y_i\}$ can therefore be computed to be

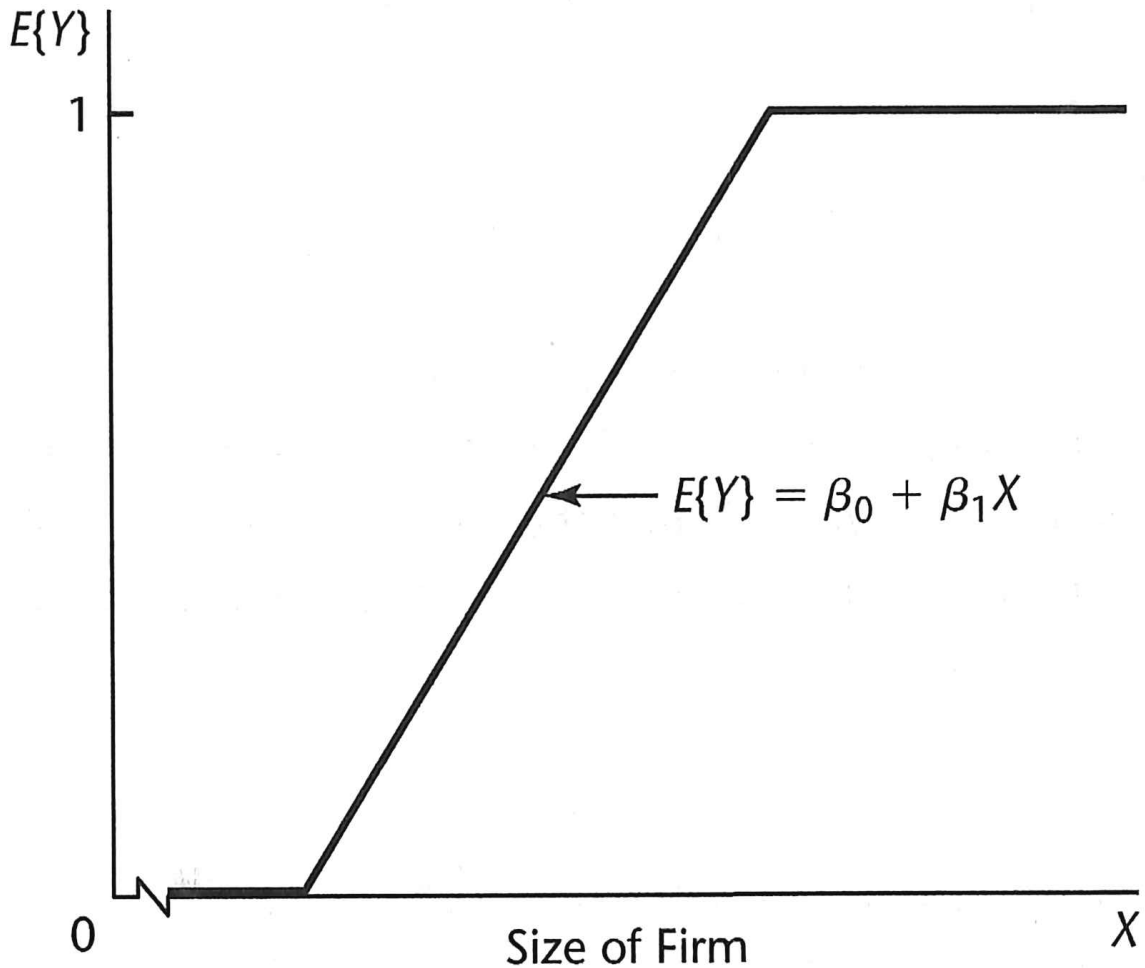
$$E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = P(Y_i = 1)$$

  which, when substituted into the simple linear regression equation becomes

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = P(Y_i = 1) = \pi_i$$

- Therefore, the interpretation of the mean response is simply the probability of Y_i when the level of the predictor variable is X_i

Response Function with Binary Outcomes cont'd



Special Problems when Response Variable is Binary

Non-normal error terms

- For a binary response variables, each error term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ can only take on two values
  - When $Y_i = 1$: $\varepsilon_i = 1 - \beta_0 - \beta_1 X_i$
  - When $Y_i = 0$: $\varepsilon_i = -\beta_0 - \beta_1 X_i$
- The normal error regression model, which assumes that the ε_i are normally distributed, is not appropriate

Special Problems when Response Variable is Binary

Nonconstant Error Variance

- For a binary response variables, the error terms do not have equal variances
- To illustrate this, we will derive $\sigma^2\{Y_i\}$ for the SLR model

$$\begin{aligned}\sigma^2\{Y_i\} &= E\{(Y_i - e\{Y_i\})^2\} = (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i) \\ &= \pi_i(1-\pi_i) \\ &= (E\{Y_i\})(1-E\{Y_i\})\end{aligned}$$

- The variance of the error terms is that same as that of Y_i because $\varepsilon_i = Y_i - \pi_i$, where π_i is a constant, therefore

$$\begin{aligned}\sigma^2\{\varepsilon_i\} &= \pi_i(1-\pi_i) \\ &= (E\{Y_i\})(1-E\{Y_i\}) \\ &= (\beta_0 + \beta_1 X_i)(1-\beta_0-\beta_1 X_i)\end{aligned}$$

n.b. $\sigma^2\{\varepsilon_i\}$ depends on X_i, and is therefore noncosntant, and OLS will no longer be optimal

Special Problems when Response Variable is Binary

Constraints on the Response Function

- As the response function represents probabilities when the outcome variables is binary, the mean responses should also be constrained as follows

$$0 \leq E\{Y\} = \pi \leq 1$$

- Many response functions do not automatically possess this constraint, e.g., a linear response function
- This restriction causes the most serious issues, e.g., non-normal error terms, even with large sample sizes and the method of least squares providing estimators that are asymptotically normal under quite general conditions

Subsection 1

Sigmoidal Response Functions for Binary Responses

Probit Mean Response Function

- Consider a health researcher studying the effect of a mother's alcohol use X (an index of use) on the duration of pregnancy, Y^c, where the superscript c is used to denote the fact that Y^c is a continuous response variable
- The SLR would be

$$Y_i^c = \beta_0^c + \beta_1^c X_i + \varepsilon_i^c$$

  assuming the ε_i^c are normally distributed with constant variance σ_c^2
- If researchers decided to parse the pregnancy duration, Y^c, into preterm ($\leq$ 38 weeks) and full-term ($>$ 38 weeks), i.e.,

$$Y_i = \begin{cases} 1 & \text{if } Y_i^c \leq 38 \text{ weeks (preterm)} \\ 0 & \text{if } Y_i^c > 38 \text{ weeks (full term)} \end{cases}$$

Given

$$Y_i = \begin{cases} 1 & \text{if } Y_i^c \leq 38 \text{ weeks (preterm)} \\ 0 & \text{if } Y_i^c > 38 \text{ weeks (full term)} \end{cases}$$

it follows

$$\begin{aligned} P(Y_i = 1) = \pi_i &= P(Y_i^c \leq 38) \\ &= P(\beta_0^c + \beta_1^c X_i + \varepsilon_i^c \leq 38) \\ &= P(\varepsilon_i^c \leq 38 - \beta_0^c - \beta_1^c X_i) \\ &= P\left(\frac{\varepsilon_i^c}{\sigma_c} \leq \frac{38 - \beta_0^c}{\sigma_c} - \frac{\beta_1^c}{\sigma_c} X_i\right) \\ &= P\left(Z \leq \beta_0^* + \beta_1^* X_i\right) \end{aligned}$$

where $\beta_0^* = (38 - \beta_0^c)/\sigma_c$, $\beta_1^* = -\beta_1^c/\sigma_c$ and $Z = \varepsilon_i^c/\sigma_c$ follows a standard normal distribution

- If we let $P(Z < z) = \Phi(z)$, we have

$$P(Y_i = 1) = \Phi(\beta_0^* + \beta_1^* X_i)$$

- The nonlinear regression function known as the probit mean response function

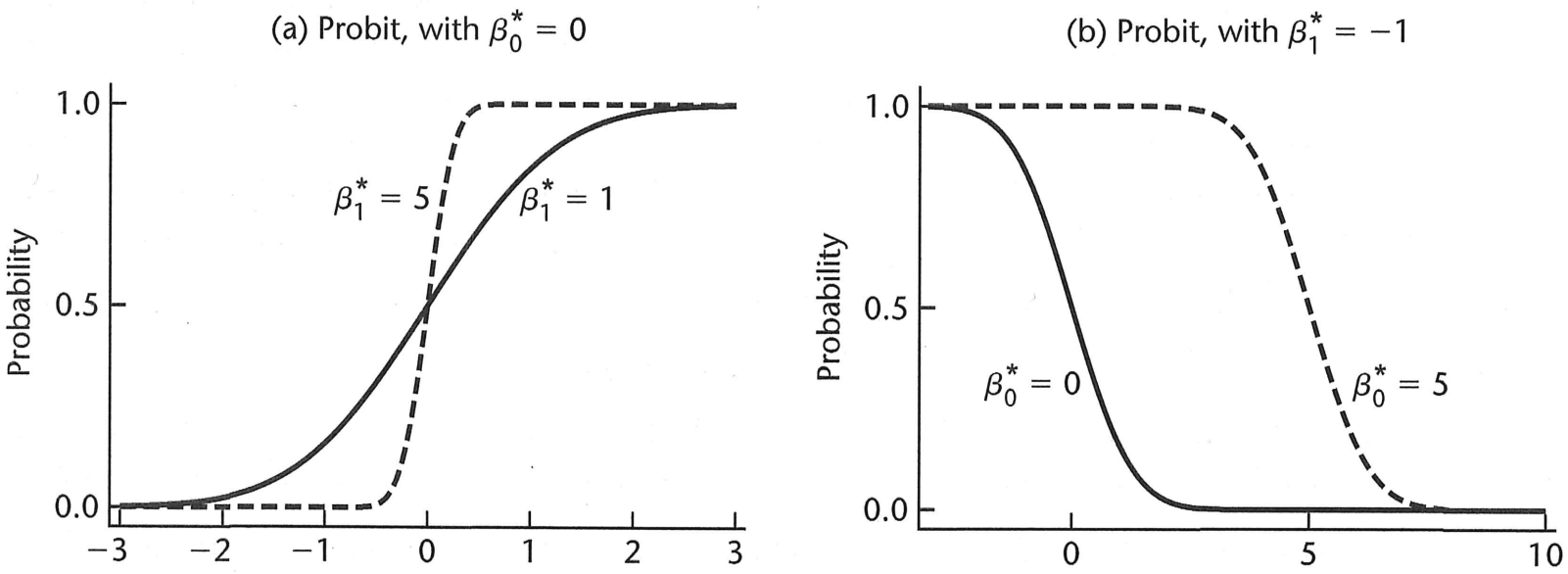
$$E\{Y_i\} = \pi_i = \Phi(\beta_0^* + \beta_1^* X_i)$$

- The inverse function, Φ^{-1}, of the standard normal cumulative distribution function Φ is sometimes called the probit transformation
- We solve for the linear predictor, $\beta_0^* + \beta_1^* X_i$ by applying the probit transformation to both sides of the expression

$$\Phi^{-1}(\pi_i) = \pi_i' = \beta_0^* + \beta_1^* X_i$$

- The expression $\pi_i' = \beta_0^* + \beta_1^* X_i$ is called the probit response function or the linear predictor

Various Probit Mean Response Functions



Notes on the Probit Mean Response Function

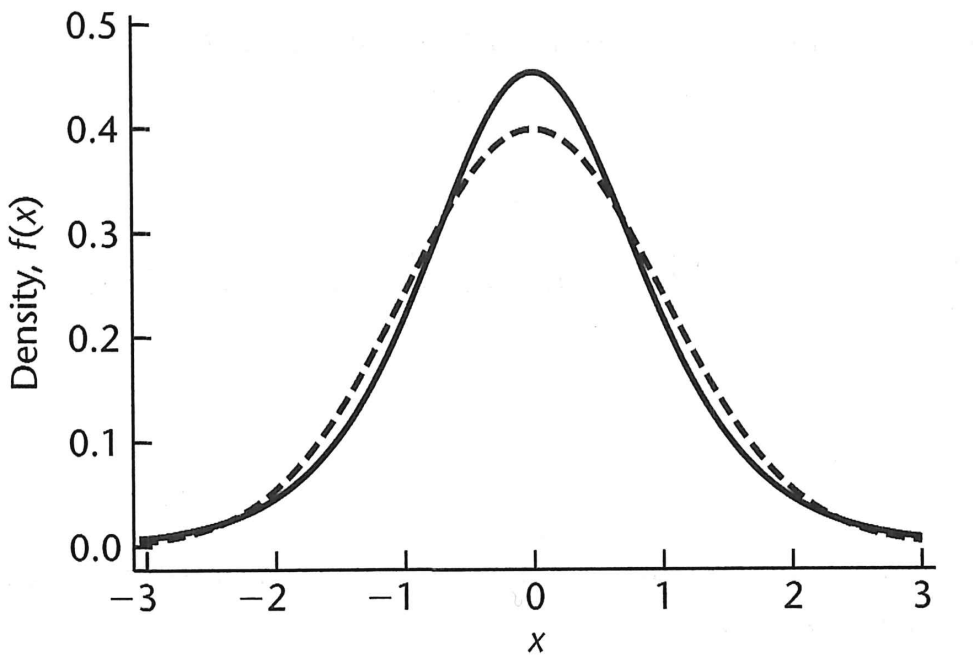
1. The probit mean response function is bounded between 0 and 1, and it approaches these limits asymptotically
2. As β_1^* increases, the mean function becomes more S-shaped, changing more rapidly in the center
3. Changing the sign of β_1^* from positive to negative changes the mean response function from a monotone increasing to a monotone decreasing function
4. Increasing or decreasing the intercept β_0^* shifts the mean response function horizontally (the direction of shift depends on the signs of both β_0^* and β_1^*)
5. As the probit response function is symmetric, a recoding of the response variables from 0/1 to 1/0, i.e., $Y_i' = 1 - Y_i$, results in the signs of all coefficients being reversed

$$\Phi(Z) = 1 - \Phi(-Z)$$

$$P(Y_i' = 1) = P(Y_i = 0) = 1 - \Phi(\beta_0^* + \beta_1^* X_i) = \Phi(-\beta_0^* - \beta_1^* X_i)$$

Logistic Mean Response Function

- The assumption of normally distributed error terms for the continuous response variable led to the use of a standard normal cumulative distribution Φ to model π_i
- An alternative error distribution that is very similar to the normal distribution is the logistic distribution



- The density of the logistic random variable ε_L having a mean of zero and standard deviation of $\sigma = \pi/\sqrt{3}$ has a simple form

$$f_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{[1 + \exp(\varepsilon_L)]^2}$$

and its cumulative distribution function is

$$F_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{1 + \exp(\varepsilon_L)}$$

- Supposing that the ε_i^c in $Y_i^c = \beta_0^c + \beta_1^c X_i + \varepsilon_i^c$ have a logistic distribution with mean zero and standard deviation σ_c, then

$$P(Y_i = 1) = P\left(\frac{\varepsilon_i^c}{\sigma_c} \leq \beta_0^* + \beta_1^* X_i\right)$$

where ε_i^c/σ_c follows a logistic distribution with mean zero and standard deviation of one

- With a little bit of algebraic manipulation, we compute and its cumulative distribution function is

$$\begin{aligned} P(Y_i = 1) = \pi_i &= P\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i^c}{\sigma_c} \leq \frac{\pi}{\sqrt{3}}\beta_0^* + \frac{\pi}{\sqrt{3}}\beta_1^* X_i\right) \\ &= P(\varepsilon_L \leq \beta_0 + \beta_1 X_i) \\ &= F_L(\beta_0 + \beta_1 X_i) \\ &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \end{aligned}$$

where $\beta_0 = (\pi/\sqrt{3})\beta_0^*$ and $\beta_1 = (\pi/\sqrt{3})\beta_1^*$ denote the logistic regression parameters

- In sum, the logistic mean response function is

$$E\{Y_i\} = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$
$$= [1 + \exp(-\beta_0 - \beta_1 X_i)]^{-1}$$

- Taking the inverse of the cumulative distribution function F_L

$$F_L^{-1}(\pi_i) = \beta_0 + \beta_1 X_i = \pi_i'$$

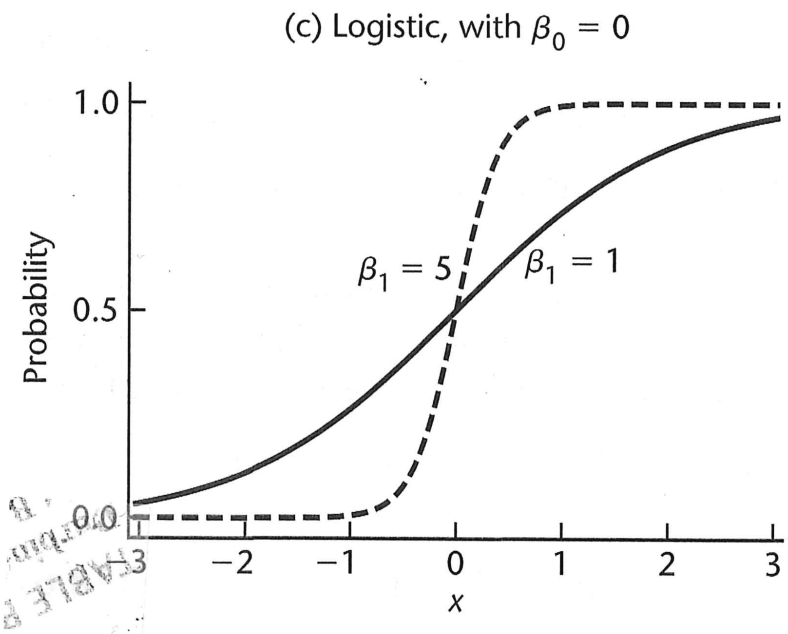
we obtain the logic response function

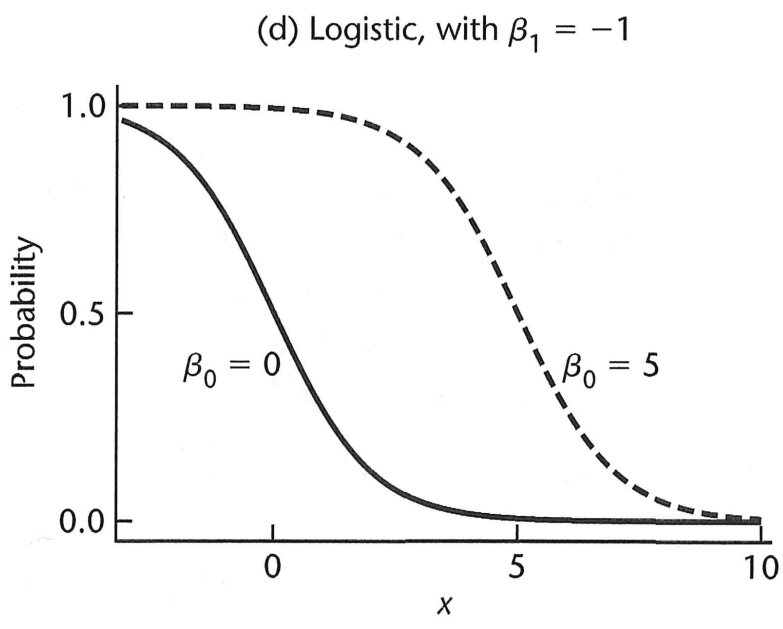
- The transformation $F_L^{-1}(\pi_i)$ is called the logit transformation of the probability π_i and is given by

$$F_L^{-1}(\pi_i) = log_e\left(\frac{\pi_i}{1 - \pi_i}\right)$$

where the ratio $\pi_i/(1 - \pi_i)$ is called the odds

Various Logistic Mean Response Functions





Subsection 2

Simple Logistic Regression

- Recall that when the response variable is binary, taking on $1/0$ values with probabilities π and $1-\pi$ respectively, Y is a Bernoulli random variable with parameter $E\{Y\} = \pi$
- The simple logistic regression model is

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- The X observations are assumed to be known constants
- There is no closed form solution to obtain estimates of β_0 and β_1, and we shall therefore defer to a statistical package (R) to compute the maximum likelihood estimates

- The fitted logistic response function is

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$$

- If we utilize the logic transformation, we can express the fitted response function as follows

$$\hat{\pi}' = b_0 + b_1 X$$

which is called the fitted logit response function, where

$$\hat{\pi}' = log_e \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right)$$

Simple Logistic Regression Model EXAMPLE

Computer Programmer Interviews

- X: months of programming experience
- Y: 0 of task not completed, 1 if completed
- Question: what is the effect of computer programming experience on the ability to complete the task?

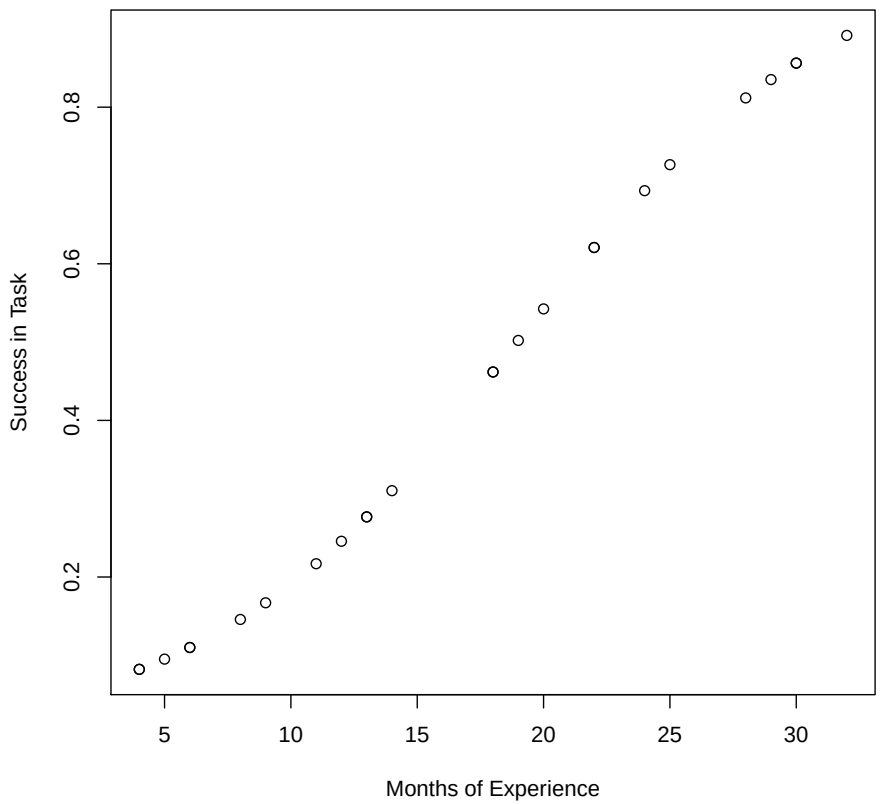
```
> compProgData <- read.table('~/Desktop/compProg.txt',sep='',header=FALSE)

> colnames(compProgData)[1] <- "monthsOfExp"
> colnames(compProgData)[2] <- "taskSuccess"

> compProgData$taskSuccess <- as.factor(compProgData$taskSuccess)

> mylogisticRegression <- glm(taskSuccess ~ monthsOfExp,
     data = compProgData, family='binomial')

Coefficients:
(Intercept)     monthsOfExp
    -3.0597       0.1615
```

Simple Logistic Regression Model EXAMPLE cont'd

- The estimated logistic regression function is

$$\hat{\pi}_i = \frac{\exp(-3.0597 + 0.1615X_i)}{1 + \exp(-3.0597 + 0.1615X_i)}$$

- Plotting the fitted values ($\hat{\pi}_i$) against the X_i, we obtain the following scatter plot



Interpretation of b_1

- The interpretation of b_1 in the fitted logistic response function is not the typical or straightforward interoperation of slope in a linear regression model
- The reason for this is that the effect of a unit increase in X varies for the logistic regression model according to the location of the starting point on the X scale
- An interpretation of b_1 is found in the property of the fitted logistic function that the estimated odds $\hat{\pi}/(1-\hat{\pi})$ are multiplied by $\exp(b_1)$ for any unit increase in X

- Consider the value of the fitted logit response function at $X = X_j$

$$\hat{\pi}'(X_j) = b_0 + b_1 X_j$$

- Now consider the value of the fitted logit response at $X = X_j + 1$

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1)$$

- The difference between the two fitted values is

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1$$

- $\hat{\pi}'(X_j)$ is the logarithm of the estimated odds when $X = X_j$, which we shall denote $\log_e(odds_1)$
- Similarly, $\hat{\pi}'(X_j + 1)$ is the logarithm of the estimated odds when $X = X_j + 1$, which we shall denote $\log_e(odds_2)$

- The difference between the two fitted logit response values can be expresses as follows:

$$\log_e(odds_2) - \log_e(odds_1) = \log_e\left(\frac{odds_2}{odds_1}\right) = b_1$$

- Exponentiating all terms above, we obtain the estimated ratio of the odds, called the odds ratio, and denoted by $\widehat{OR}$

$$\widehat{OR} = \frac{odds_2}{odds_1} = exp(b_1)$$

- Computer Programming EXAMPLE

$$\widehat{OR} = exp(b_1) = exp(0.1615) = 1.175$$

- The odds of completing a task increase by 17.5% with each additional month of experience

- In general, the estimated odds ratio when there is a difference of c units of X is $\exp(cb_1)$
- Computer Programming EXAMPLE
- A candidate who has 15 months more experience than another candidate has a $exp(15 \times 0.1615) = 11.3$-fold higher odds of completing the task

Subsection 3

Multiple Logistic Regression

Multiple Logistic Regression Model

- The simple logistic regression model is easily extended to more than one predictor variable
- In fact, several predictor variables are usually required with logistic regression to obtain an adequate description and useful predictions
- Extending the simple logistic regression model requires only replacing

$$\beta_0 + \beta_1 X_1$$

by

$$\mathbf{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$$

- We subsequently update the simple logistic response function to the multiple logistic response function as follows

$$E\{Y\} = \pi_i = \frac{\exp(\mathbf{X}'\boldsymbol{\beta})}{1 + \exp(\mathbf{X}'\boldsymbol{\beta})} = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1}$$

- Similarly, the logit transformation

$$\pi' = \log_e \left(\frac{\pi}{1-\pi} \right)$$

now leads to the logic response function or linear predictor

$$\pi' = \mathbf{X}'\boldsymbol{\beta}$$

- The multiple logistic regression model can therefore be stated as follows

Y_i are independent Bernoulli random variables with expected values $E\{Y_i\} = \pi_i$, where

$$E\{Y_i\} = \pi_i = \frac{exp(\mathbf{X}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i'\boldsymbol{\beta})}$$

- The multiple logistic regression function is sigmoidal and monotonic in shape with respect to $\mathbf{X}'\beta$ and is almost linear when $0.2 < \pi < 0.8$
- The predictor (X) variables may be different predictor variables, the may represent curvature and/or interaction effects, be quantitative or qualitative (indicator variables)
- When the logistic regression model contains only qualitative variables, it is often referred to as a log-linear model
- The fitted logistic response function and fitted values are expressed as follows

$$\hat{\pi} = \frac{exp(\mathbf{X}'\mathbf{b})}{1 + \exp(\mathbf{X}'\mathbf{b})} \qquad \hat{\pi}_i = \frac{exp(\mathbf{X}_i'\mathbf{b})}{1 + \exp(\mathbf{X}_i'\mathbf{b})}$$

U.S. Married Women's Labor-Force Participation

- Response: Labor Force Participation
  - yes/no (factor)
- Predictors

k5	Number of children aged 5 and younger (quantitative)
k618	Number of children between 6 and 18 (quantitative)
age	Wife's age in years (quantitative)
wc	Wife's college attendance (yes/no, factor)
hc	Husband's college attendance (yes/no, factor)
lwg	Log of wife's estimated wage rate (numeric)
inc	Family income excluding wife's income

```
> library('car')
> str(Mroz)

> mroz.logistic <- glm(lfp ~ k5 + k618 + age + wc + hc +lwg + inc,
   family='binomial',data=Mroz)

> summary(mroz.logistic)

#<<select output resected for concision>>

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
```

U.S. Married Women's Labor-Force Participation

- What is the estimated logistic response function?

U.S. Married Women's Labor-Force Participation

- What is the estimated logistic response function?

$$\hat{\pi} = [1 + \exp(-3.18 + 1.46\text{k5} + 0.05\text{k618} + 0.06\text{age} - 0.8\text{wc} - 0.11\text{hc} - 0.60\text{lwg} + 0.03\text{inc})]^{-1}$$

- What are the odds ratios?

U.S. Married Women's Labor-Force Participation

- What is the estimated logistic response function?

$$\hat{\pi} = [1 + \exp(-3.18 + 1.46\text{k5} + 0.05\text{k618} + 0.06\text{age} - 0.8\text{wc} - 0.11\text{hc} - 0.60\text{lwg} + 0.03\text{inc})]^{-1}$$

- What are the odds ratios?

```
> round(exp(coef(mroz.logistic)),2)
(Intercept)    k5    k618   age    wcyes   hcyes    lwg      inc
  24.10      0.23    0.94   0.94   2.24     1.12    1.83     0.97
```

- Interpret the odds ratios

- The estimated logistic regression model is given by

$$\log_e \left[\frac{\hat{\pi}}{1-\hat{\pi}} \right] = b_0 + b_1 X_1 + \ldots + b_{p-1} X_{p-1}$$

- Exponentiating both sides

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \exp(b_0) \times \exp(b_1 X_1) \times \ldots \times \exp(b_{p-1} X_{p-1})$$

- $\frac{\hat{\pi}}{1-\hat{\pi}}$ provides us with the fitted odds of success, i.e., the fitted odds of success ($\hat{\pi}$) divided by the fitted odds of failure (1-$\hat{\pi}$)
- Exponentiating the model removes the logarithms and changes it from an additive log-odds scale to one that is multiplicative in the odds scale

E.g. Increasing the age of a woman by 1 year, holding all other predictors constant, multiplies the probability of her being in the workforce by $\exp(b_{age}) = \exp(-0.06) = 0.9391$, i.e., it reduces the odds of her working by 6%

Subsection 4

Inferences about Regression Parameters

- Evaluating second-order partial derivates of the log-likelihood function results in a Hessian matrix, when evaluated at $\boldsymbol{\beta} = \mathbf{b}$, the estimated approximate variance-covariance matrix of the estimated regression coefficients for logistic regression are

$$\mathbf{s}^2\{\mathbf{b}\} = ([-g_{ij}]_{\boldsymbol{\beta}=\mathbf{b}})^{-1}$$

- Inferences about the regression coefficients for simple or multiple logistic regression models are based on the following approximate results when the sample size is large

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim z \qquad k = 0, 1, \ldots, p-1$$

Tests for a Single β_k: Wald Test

- The hypothesis test for a large-sample test of a single regression parameter is

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0$$

and the appropriate test statistic is

$$z^* = \frac{b_k}{s\{b_k\}}$$

and the decision rule is

  - If $|z^*| \leq z_{(1-\alpha/2)}$, do not reject H_0
  - If $|z^*| > z_{(1-\alpha/2)}$, reject H_0

- One sided alternatives will involve a one-sided decision rule
- A test using $(z^*)^2$ is based on the chi-square distribution with 1 degree of freedom is called the Wald test

Interval Estimation for a Single β_k

- The $1-\alpha$ confidence limits for β_k are

$$b_k \pm z_{(1-\alpha/2)} s\{b_k\}$$

- The corresponding limits for the odds ratio, $\exp(\beta_k)$ are

$$\exp[b_k \pm z_{(1-\alpha/2)} s\{b_k\}]$$

Computer Programmer Interviews

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
monthsOfExp  0.16149    0.06498   2.485   0.0129 *
#---------------------------------------------
# estimated odds ratios
(Intercept) monthsOfExp
      -3.06        0.16
```

- Is β_1 statistically significantly larger than 0?

$$H_0 : \beta_k \leq 0 \qquad H_a : \beta_k > 0$$

$$z* = 0.1615/0.065 = 2.485 > 1.645 = z_{0.95}$$

Reject H_0: as expected an increase in months of experience increases the odds of completing the project

Computer Programmer Interviews

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
monthsOfExp  0.16149    0.06498   2.485   0.0129 *
#---------------------------------------------
# estimated odds ratios
(Intercept) monthsOfExp
      -3.06        0.16
```

- Compute a 95% CI for β_1

$$0.1615 \pm 1.96(0.065) = [0.0341, 0.2889]$$

- We conclude with 95% confidence that β_1 is between 0.0341 and 0.2889
- The corresponding limits for the odd ratio are $\exp(0.0341) = 1.03$ and $\exp(0.2889) = 1.33$

Computer Programmer Interviews

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
monthsOfExp  0.16149    0.06498   2.485   0.0129 *
#---------------------------------------------------
# estimated odds ratios
(Intercept) monthsOfExp
      -3.06        0.16
```

- Compute a 95% CI for $\exp(\beta_1)$ and $\exp(5\beta_1)$

$$[\exp(0.0341) = 1.03, \exp(0.2889) = 1.33]$$

$$[\exp(5 \times 0.0341) = 1.186, \exp(5 \times 0.2889) = 4.240]$$

- With 95% confidence, we estimate the odds of success increase by between 19% and 324% with an additional 5 months of experience

Testing for Several $\beta_k = 0$: Likelihood Ratio Test

- We can use this test to identify if a given predictor variable can be dropped from the model, i.e., testing whether the associate regression coefficients $\beta_k = 0$
- The test is called the likelihood ratio test, and, like the general linear test, is based on a comparison on full and reduced models
- We begin with the full logistic model with response function

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_F)]^{-1} \qquad \text{Full model}$$

- Using the maximum likelihood estimates, we (let R) calculate the likelihood function $L(\boldsymbol{\beta})$ when $\boldsymbol{\beta}_F = \mathbf{b}_F$; we will denote this value of the likelihood function for the full model by $L(F)$

Testing for Several $\beta_k = 0$: Likelihood Ratio Test cont'd

- The hypothesis we wish to test is

$$H_0 : \beta_q = \beta_{q+1} = \ldots = \beta_{p-1}$$
$$H_a : \text{not all } \beta_k \text{ in } H_0 \text{ are equal zero}$$

- where for convenience we arrange the model so that he last $p - q$ coefficients are those tested
- The reduced logistic regression model therefor has the response function

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_R)]^{-1} \qquad \text{Reduced model}$$

- We equivalently obtain the likelihood function for the reduced model $L(R)$

n.b. It can be shown that $L(R) \leq L(F)$

Testing for Several $\beta_k = 0$: Likelihood Ratio Test cont'd

- The test statistic for the likelihood ratio test, denoted G^2, is

$$G^2 = -2 \log_e \left[\frac{L(R)}{L(F)} \right] = -2[\log_e L(R) - \log_e L(F)]$$

  where the function `logLik(<regressionObject>)` will return the log-likelihood in R
- The decision rule is
  - If $G^2 \leq \chi_{(1-\alpha, p-q)}$, do not reject H_0
  - If $G^2 > \chi_{(1-\alpha, p-q)}$, reject H_0

- For logistic regression, AIC_p or SBC_p criteria are easily adapted and are most commonly used
- The modifications are as follows

$$AIC_p = -2\log_e L(\mathbf{b}) + 2p$$

$$SBC_p = -2\log_e L(\mathbf{b}) + p\log_e(n)$$

- Promising models will yield relatively small values for these criteria
- An alternative, third criterion frequently used is simply $-2\log_e L(\mathbf{b})$, which is analogous to R^2, namely, because it can never increase as additional predictor variables are added to the model, whereas AIC_p and SBC_p include penalty functions
- Stepwise and Best Subsets approaches apply to logistic regression models