Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 14 – Logistic Regression, Poisson Regression, and Generalized Linear Models

Regression Models with Binary Response Variable

- the response variable of interest has only two possible qualitative outcomes
- be represented by a binary indicator variables: taking values on 0 and 1
- A binary response variable is said to involve *binary responses* or *Dichotomous responses*

Meaning of Response Function when Outcome Variable is Binary

1. Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad Y_i = 0, 1$$
$$(E\{\varepsilon_i\} = 0)$$
$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

2. Consider Y_i to be a Bernoulli random variable:

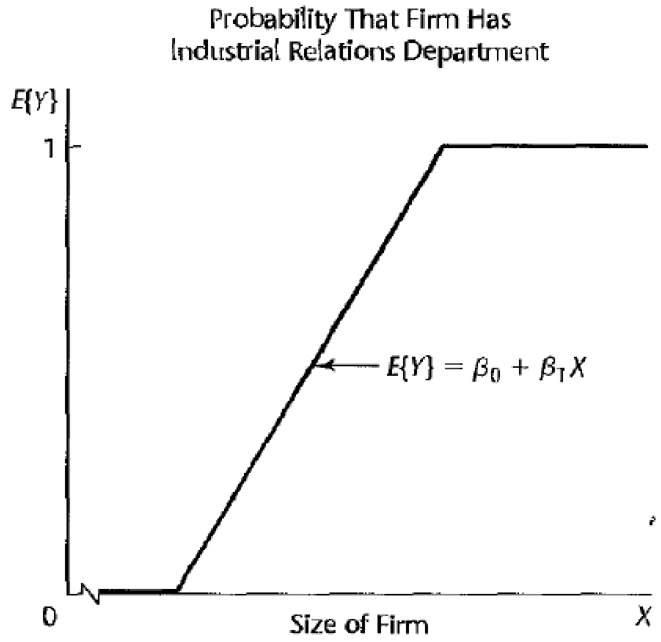
Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

$$\Rightarrow E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = P(Y_i = 1) = \beta_0 + \beta_1 X_i$$

Meaning of Response Function when Outcome Variable is Binary, cont'd

- The mean response $E\{Y_i\} = \beta_0 + \beta_1 X_i$ is simply the probability that $Y_i = 1$ when the level of the predictor variable is X_i

FIGURE 14.1 Illustration of Response Function when Response Variable Is Binary—Industrial Relations Department Example.



Special Problems when Response Variable is Binary

1. Nonnormal Error Terms: $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$

$$\text{When } Y_i = 1: \quad \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$
$$\text{When } Y_i = 0: \quad \varepsilon_i = -\beta_0 - \beta_1 X_i$$

2. Nonconstant Error Variance: $\varepsilon_i = Y_i - \pi_i$ (π_i: constant)

$$\Rightarrow \sigma^2\{Y_i\} = \sigma^2\{\varepsilon_i\} = \pi_i(1-\pi_i) = (E\{Y_i\})(1 - E\{Y_i\})$$

3. Constraints on Response Function:

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

Many response function do not automatically posses this constraint.

Sigmoidal Response Functions for Binary Response

Ex: health researcher studying the effect of a mother's use of alcohol

- X: an index of degree of alcohol use during pregnancy
- Y^c : the duration of her pregnancy (continuous response)
- simple linear regression model:

$$Y_i^c = \beta_0^c + \beta_1^c X_i + \varepsilon_i^c, \quad N(0, \sigma_c^2)$$

$\Rightarrow$ the usual simple linear regression analysis

Sigmoidal Response Functions for Binary Response, cont'd

- Coded each pregnancy:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^c \le T \text{ weeks} \\ 0 & \text{if } Y_i^c > T \text{ weeks} \end{cases}$$

$$\begin{aligned} \Rightarrow P(Y_i = 1) = \pi_i = P(Y_i^c \le T) &= P(\beta_0^c + \beta_1^c X_i + \varepsilon_i^c \le T) \\ &= P\left(\underbrace{\frac{\varepsilon_i^c}{\sigma_c}}_{(Z)} \le \underbrace{\frac{T - \beta_0^c}{\sigma_c}}_{(\beta_0^*)} - \underbrace{\frac{\beta_1^c}{\sigma_c}}_{(-\beta_1^*)} X_i \right) \\ &= P(Z \le \beta_0^* + \beta_1^* X_i) \\ &= \Phi(\beta_0^* + \beta_1^* X_i) \end{aligned}$$

$$(\Phi(z) = P(Z \le z), \quad Z \sim N(0,1)\ )$$

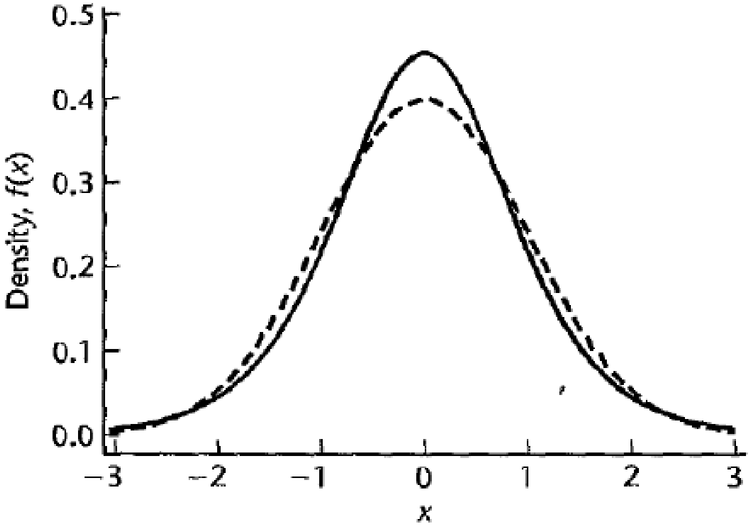
Sigmoidal Response Functions for Binary Response, cont'd

1. probit mean response function:

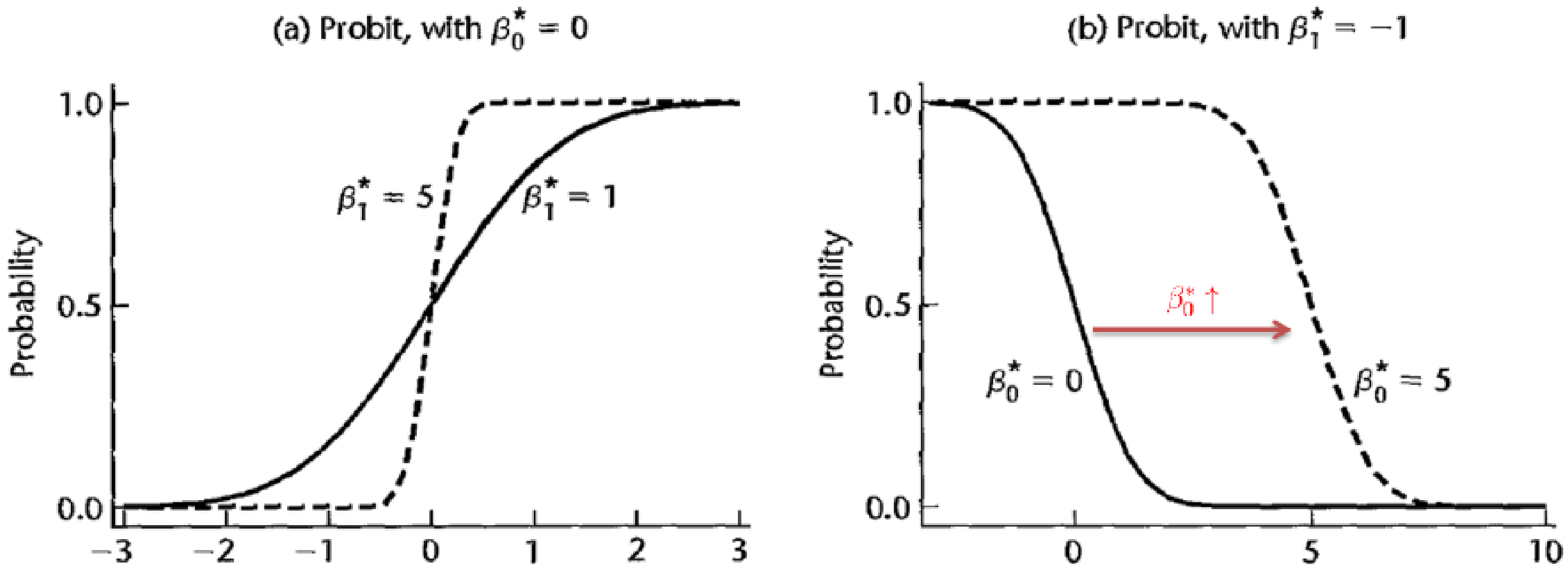
$$E\{Y_i\} = \pi_i = \Phi(\beta_0^* + \beta_1^* X_i)$$
$$\Rightarrow \Phi^{-1}(\pi_i) = \pi_i' = \beta_0^* + \beta_1^* X_i$$

- Φ^{-1}: sometimes called the probit transformation
- $\pi_i' = \beta_0^* + \beta_1^* X_i$: probit response function or linear predictor

FIGURE 14.3 Plots of Normal Density (dashed line) and Logistic Density (solid line), Each Having Mean 0 and Variance 1.



Sigmoidal Response Functions for Binary Response, cont'd

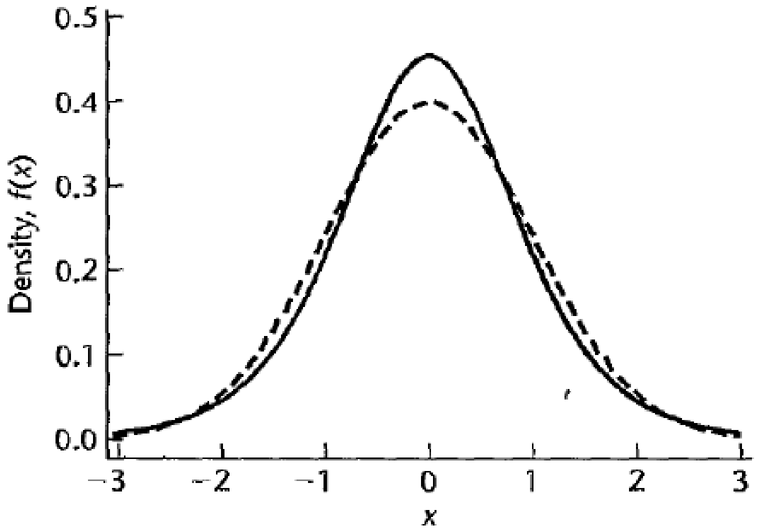


- bounded between 0 and 1
- $\beta_1^* \uparrow$ $(\beta_0^* = 0)$ more S-shaped, changing rapidly in the center
- changing the sign of β_1^*
- Symmetric function: $Y_i' = 1 - Y_i \quad (\Phi(Z) = 1 - \Phi(-Z))$

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

Sigmoidal Response Functions for Binary Response, cont'd

FIGURE 14.3
Plots of Normal Density (dashed line) and Logistic Density (solid line), Each Having Mean 0 and Variance 1.



Logistic function

- Similar to the normal distribution: mean=0, variance=1
- slightly heavier tails

Sigmoidal Response Functions for Binary Response, cont'd

- $\varepsilon_L \sim$ logistic r.v. $\mu = 0, \sigma = \pi/\sqrt{3}$

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

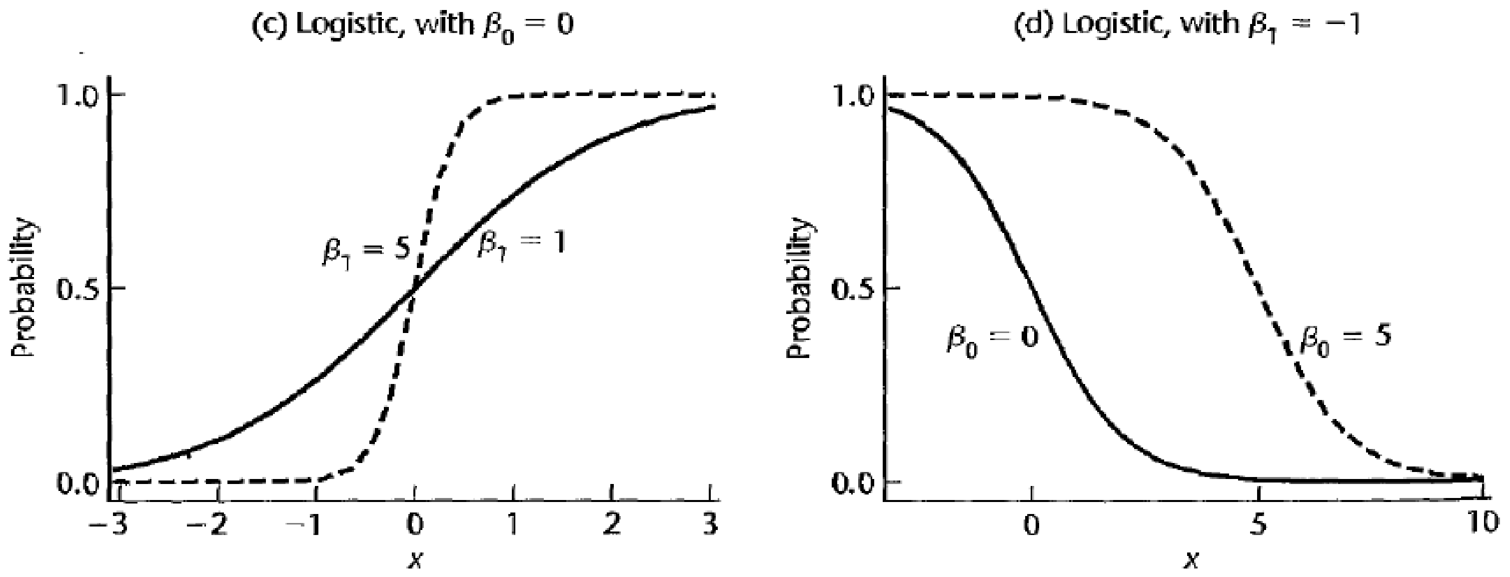
- If $\varepsilon_i^c \sim$ Logistic distribution with $\mu = 0, \sigma_c$

$$P(Y_i = 1) = \pi_i = P\left(\frac{\varepsilon_i^c}{\sigma_c} \le \beta_0^* + \beta_1^* X_i\right), \quad (E(\frac{\varepsilon_i^c}{\sigma_c}) = 0,\ \sigma\{\frac{\varepsilon_i^c}{\sigma_c}\} = 1)$$

$$= P\left(\underbrace{\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i^c}{\sigma_c}}_{(\varepsilon_L)} \le \underbrace{\frac{\pi}{\sqrt{3}}\beta_0^*}_{(\beta_0)} + \underbrace{\frac{\pi}{\sqrt{3}}\beta_1^*}_{(\beta_1)} X_i\right)$$

$$= P(\varepsilon_L \le \beta_0 + \beta_1 X_i) = F_L(\beta_0 + \beta_1 X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

Sigmoidal Response Functions for Binary Response, cont'd



2. logistic mean response function:

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

$$= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)}$$

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

Sigmoidal Response Functions for Binary Response, cont'd

- $F^{-1}(\pi_i) = \beta_0 + \beta_1 X_i = \log_e\left(\dfrac{\pi_i}{1-\pi_i}\right)$:
  called the logit transformation of the probability π_i

- $\dfrac{\pi_i}{1-\pi_i}$: called the odds

Sigmoidal Response Functions for Binary Response, cont'd

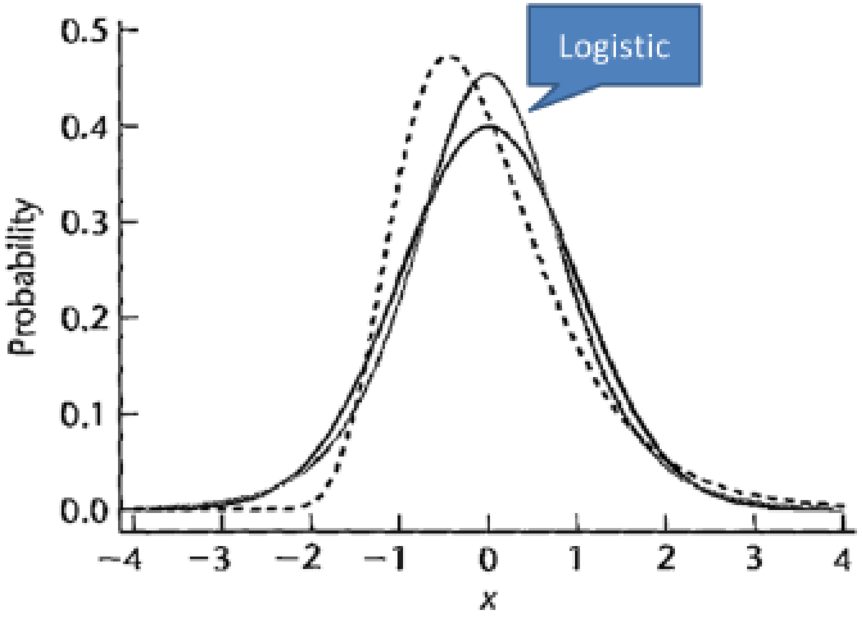
3. complementary log-log response function: extreme value of Gumbel probability distribution

$$E\{Y_i\} = \pi_i = 1 - \exp(-\exp(\beta_0^G + \beta_1^G X_i))$$
$$\Rightarrow \pi' = \log[-\log(1 - \pi(X_i))] = \beta_0^G + \beta_1^G X_i$$

- $F^{-1}(\pi) = \log_e\left(\frac{\pi_i}{1-\pi_i}\right)$: called the logit transformation of the probability π
- $\frac{\pi_i}{1-\pi_i}$: called the odds

Sigmoidal Response Functions for Binary Response, cont'd

FIGURE 14.4 Plots of Gumbel (dashed line), Normal (black line), and Logistic (gray line) Density Functions, Each Having Mean 0 and Variance 1.



Simple Logistic Regression

The most widely used:

1. The regression parameters have relatively simple and useful interpretations
2. statistical software is widely available for analysis of logistic regression models

Estimation parameters:

- Estimation: MLE to estimate the parameters of the logistic response function
- Utilize the Bernoulli distribution for a binary random variable

Simple Logistic Regression, cont'd

Simple Logistic Regression Model

- $Y \sim \mathcal{B}er(\pi)$: $E\{Y\} = \pi$

$$\Rightarrow Y_i = E\{Y_i\} + \varepsilon_i$$

- the distribution of ε_i depends on the Bernoulli distribution of the response Y_i

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

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

Simple Logistic Regression, cont'd

Likelihood function:

- Each Y_i:

$$P(Y_i = 1) = \pi_i; \quad P(Y_i = 0) = 1 - \pi_i$$

$$\Rightarrow f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \quad Y_i = 0, 1;\ i = 1, \ldots, n$$

- The joint probability function:

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

Simple Logistic Regression, cont'd

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\Rightarrow \ln g(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} \left[Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

$$(\because 1-\pi_i = [1+\exp(\beta_0+\beta_1 X_i)]^{-1})$$

$$(\Rightarrow \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i)$$

$$\Rightarrow \ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 X_i) - \sum_{i=1}^{n} \ln[1+\exp(\beta_0+\beta_1 X_i)]$$

No closed-form solution for β_0, β_1 that $\max \ln L(\beta_0, \beta_1)$

Simple Logistic Regression, cont'd

Maximum Likelihood Estimation:

- To find the MLE b_0, b_1: require computer-intensive numerical search procedures
- Once b_0, b_1 are found:

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

- the logit transformation:

$$\hat{\pi}' = \ln\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = b_0 + b_1 X$$

Example:

TABLE 14.1 Data and Maximum Likelihood Estimates—Programming Task Example.

(a) Data

Person i	(1) Months of Experience X_i	(2) Task Success Y_i	(3) Fitted Value $\hat{\pi}_i$
1	14	0	.310
2	29	0	.835
3	6	0	.110
...	...	...	...
23	28	1	.812
24	22	1	.621
25	8	1	.146

(b) Maximum Likelihood Estimates

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation
β_0	−3.0597	1.259
β_1	.1615	.0650

Example, cont'd

```
mylogit <- glm(Y ~ X, data =
Dataset_14TA01,family = "binomial")
> summary(mylogit)

Call:
glm(formula = Y ~ X, family = "binomial", data = Dataset_14TA01)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.8991601  -0.7508920  -0.4140037   0.7992195   1.9623537

Coefficients:
              Estimate  Std. Error  z value Pr(>|z|)
(Intercept) -3.05969586  1.25934986 -2.42958 0.015116 *
X            0.16148592  0.06498001  2.48516 0.012949 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296490  on 24  degrees of freedom
Residual deviance: 25.424574  on 23  degrees of freedom
AIC: 29.424574

Number of Fisher Scoring iterations: 4
```

```
> as.matrix(mylogit$fitted.values)
            [,1]
1  0.31026237072
2  0.83526292179
3  0.10999615830
4  0.72660237188
5  0.46183704246
6  0.08213001754
7  0.46183704246
8  0.24566554235
9  0.62081157675
10 0.10999615830
11 0.85629861504
12 0.21698039329
13 0.85629861504
14 0.09515416130
15 0.54240353449
16 0.27680233903
17 0.16709980122
18 0.89166416440
19 0.69337940941
20 0.27680233903
21 0.50213414135
22 0.08213001754
23 0.81182461437
24 0.62081157675
25 0.14581507520
```

Example, cont'd

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = -3.0597 + 0.1615X_i$$

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

For X=14,

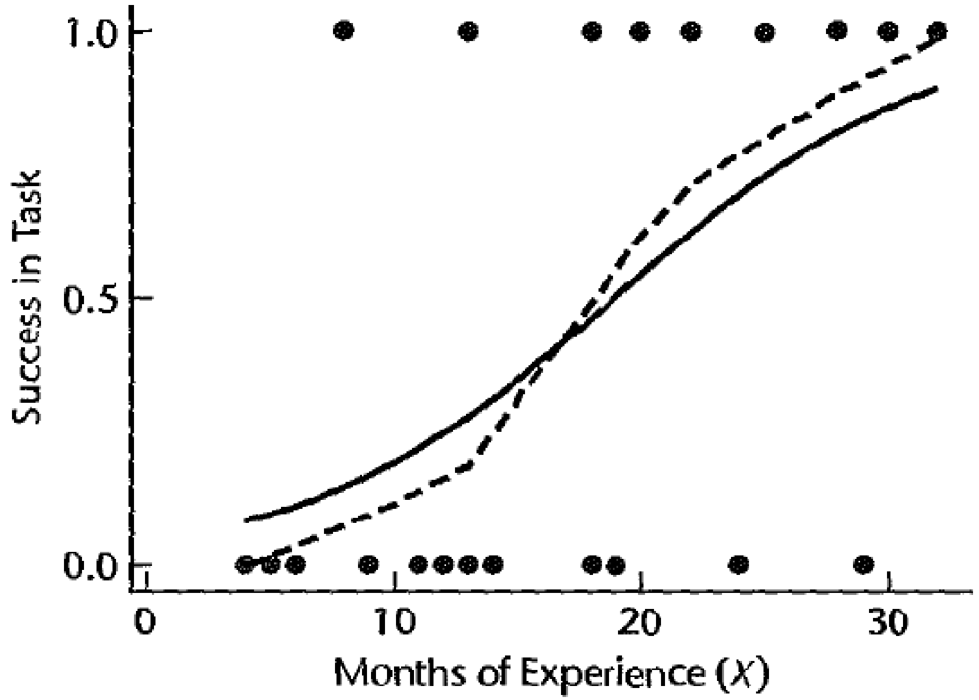
$$\hat{\pi} = \frac{\exp(-3.0597 + 0.1615 * 14)}{1 + \exp(-3.0597 + 0.1615 * 14)} = 0.31$$

```
> cbind(Dataset_14TA01$X,mylogit$fitted.values)
   [,1]          [,2]
1    14 0.31026237072
2    29 0.83526292179
3     6 0.10999615830
4    25 0.72660237188
5    18 0.46183704246
6     4 0.08213001754
7    18 0.46183704246
8    12 0.24566554235
9    22 0.62081157675
10    6 0.10999615830
11   30 0.85629861504
12   11 0.21698039329
13   30 0.85629861504
14    5 0.09515416130
15   20 0.54240353449
16   13 0.27680233903
17    9 0.16709980122
18   32 0.89166416440
19   24 0.69337940941
20   13 0.27680233903
21   19 0.50213414135
22    4 0.08213001754
23   28 0.81182461437
24   22 0.62081157675
25    8 0.14581507520
```

Simple Logistic Regression, cont'd

1. examine the appropriateness of the fitted response function
2. if the fit is good $\Rightarrow$ make a variety of inferences and predictions

FIGURE 14.5 Scatter Plot, Lowess Curve (dashed line), and Estimated Logistic Mean Response Function (solid line)—Programming Task Example.



Simple Logistic Regression, cont'd

Interpretation of b_1:

- The interpretation of the estimated regression coefficient b_1 in the fitted logistic response function is not the straightforward interpretation of the slope in a linear regression model.
- An interpretation of b_1:
  the estimated odds $\hat{\pi}/(1-\hat{\pi})$ are multiplied by $\exp(b_1)$ for any unit increase in X
- $\hat{\pi}'(X_j)$: the logarithm of the estimated odds when $X = X_j$ (denoted as $\ln(odds_1)$)

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

Simple Logistic Regression, cont'd

- Similarly, $\ln(odds_2) = \hat{\pi}'(X_j + 1)$

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

- odds ratio $\widehat{OR}$

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

Example, cont'd

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = -3.0597 + 0.1615X_i$$

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

- the odds of completing the task increase by 17.5% with each additional month of experience.
- Compare month 10 vs 25 $\Rightarrow \widehat{OR} = \exp\big((25-10) * 0.1615\big)$ =11.3
  - the odds of completing the task increase over 11 fold for experienced persons compared to relatively inexperienced persons

Multiple Logistic Regression Model

Similar to the multiple linear regression, matrix notation will be used.

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{X}} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{X}_i} = \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

We then have:

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

$$\mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

Multiple Logistic Regression Model, cont'd

$$E(Y) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}$$

Or

$$E(Y) = \left[1 + e^{-X'\beta}\right]^{-1}$$

$$\Rightarrow E(Y) = \frac{1}{\left[1+e^{-X'\beta}\right]} = \frac{1}{\left[1+\frac{1}{e^{X'\beta}}\right]} = \frac{1}{\left[\frac{1+e^{X'\beta}}{e^{X'\beta}}\right]} = \frac{e^{X'\beta}}{1+e^{X'\beta}}$$

- When the logistic regression model contains only qualitative variables, it is often referred to as a log-linear model. See Reference 14.2 for an in-depth discussion of the analysis of log-linear models.

Fitting of Model

The log-likelihood function for simple logistic regression in extends directly for multiple logistic regression:

$$log_e L(\beta) = \sum Y_i(X_i'\beta) - \sum log_e[1 + \exp(X_i'\beta)]$$

- Numerical search procedures will be used to find values of $\beta_0, \beta_1, \cdots, \beta_{p-1}$ that maximize $log_e L(\beta)$
- MLEs are denoted by $b_0, b_1, \cdots, b_{p-1}$

$$\underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

Fitting of Model, cont'd

- The fitted logistic response function and fitted values can then be expressed as follows:

$$\hat{\pi} = \frac{\exp(\mathbf{X}'\mathbf{b})}{1+\exp(\mathbf{X}'\mathbf{b})} = [1+\exp(-\mathbf{X}'\mathbf{b})]^{-1}$$

$$\hat{\pi}_i = \frac{\exp(\mathbf{X}_i'\mathbf{b})}{1+\exp(\mathbf{X}_i'\mathbf{b})} = [1+\exp(-\mathbf{X}_i'\mathbf{b})]^{-1}$$

where:

$$\mathbf{X}'\mathbf{b} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1}$$

$$\mathbf{X}_i'\mathbf{b} = b_0 + b_1 X_{i1} + \cdots + b_{p-1} X_{i,p-1}$$

Example

	Case i	(1) Age X_{i1}	(2) Socioeconomic Status X_{i2}	(3) X_{i3}	(4) City Sector X_{i4}	(5) Disease Status Y_i	(6) Fitted Value $\hat{\pi}_i$
	1	33	0	0	0	0	.209
(Coded)	2	35	0	0	0	0	.219
	3	6	0	0	0	0	.106
	4	60	0	0	0	0	.371
	5	18	0	1	0	1	.111
	6	26	0	1	0	0	.136
	...	...	...	...	...	...	...
	98	35	0	1	0	0	.171

(a) Estimated Coefficients, Standard Deviations, and Odds Ratios

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	Estimated Odds Ratio
β_0	−3.8877	.9955	—
β_1	.02975	.01350	1.030
β_2	.4088	.5990	1.505
β_3	−.30525	.6041	.737
β_4	1.5747	.5016	4.829

(b) Estimated Approximate Variance-Covariance Matrix

$$\mathbf{s}^2\{\mathbf{b}\} = \begin{array}{c} \begin{array}{ccccc} b_0 & b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{bmatrix} .4129 & -.0057 & -.1836 & -.2010 & -.1632 \\ -.0057 & .00018 & .00115 & .00073 & .00034 \\ -.1836 & .00115 & .3588 & .1482 & .0129 \\ -.2010 & .00073 & .1482 & .3650 & .0623 \\ -.1632 & .00034 & .0129 & .0623 & .2516 \end{bmatrix} \end{array}$$

$$\hat{\pi} = [1 + \exp(3.8877 - .02975X_1 - .4088X_2 + .30525X_3 - 1.5747X_4]^{-1}$$

the estimated mean response for case $i = 1$, where $X_{11} = 33$, $X_{12} = 0$, $X_{13} = 0$, $X_{14} = 0$, is:

$$\hat{\pi}_1 = \{1 + \exp[2.3129 - .02975(33) - .4088(0) + .30525(0) - 1.5747(0)]\}^{-1} = .209$$

Example

```
f.1403 <- glm(Y ~ X1+X2+X3+X4, data = Dataset_14TA03,family = "binomial")
summary(f.1403)
Call:
glm(formula = Y ~ X1 + X2 + X3 + X4, family = "binomial", data =
Dataset_14TA03)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.6551788  -0.7529130  -0.4787573   0.8558046   2.0976704

Coefficients:
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept) -2.31293482  0.64258788 -3.59941 0.00031894 ***
X1           0.02975009  0.01350281  2.20325 0.02757702 *
X2           0.40879024  0.59900377  0.68245 0.49495432
X3          -0.30525456  0.60412836 -0.50528 0.61336152
X4           1.57474923  0.50162060  3.13932 0.00169339 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.31761  on 97  degrees of freedom
Residual deviance: 101.05415  on 93  degrees of freedom
AIC: 111.05415

Number of Fisher Scoring iterations: 4
```

```
> round(exp(f.1403$coefficients),3)
(Intercept)         X1         X2         X3         X4
      0.099      1.030      1.505      0.737      4.830
```

the odds of a person having contracted the disease increase by about 3.0 percent with
each additional year of age (X1), for given socioeconomic status and city sector location

Polynomial Logistic Regression

Occasionally, the first-order logistic model may not provide an adequate fit to the data and a more complicated model may be needed. One such model is the kth-order polynomial logistic regression model, with logit response function:

$$\pi'(x) = \beta_0 + \beta_{11}x + \beta_{22}x^2 + \cdots + \beta_{kk}x^k$$

The second order polynomial:

$$\pi'(x) = \beta_0 + \beta_{11}x + \beta_{22}x^2$$

Example: The IPO data set is listed in Appendix C.11

Predictor	Estimated Coefficient	Estimated Standard Error	z^*	P-value
Constant	$b_0 = 0.3005$	0.1240	2.42	0.015
x	$b_{11} = 0.5516$	0.1385	3.98	0.000
x^2	$b_{22} = -0.8615$	0.1404	-6.14	0.000

$$\hat{\pi}' = .3005 + .5516x - .8615x^2$$

Inferences about Regression Parameters

The inference procedures rely on large sample sizes.

- N$\uparrow \infty \;\rightarrow\;$ $Normally\ Distributed$
- Approximate variances and covariances that are functions of the second-order partial derivatives of the logarithm of the likelihood function.

Specifically, let G denote the matrix of second-order partial derivatives of the loglikelihood function in (14.42), the derivatives being taken with regard to the parameters $\beta_0, \beta_1, \cdots, \beta_{p-1}$:

$$\underset{p\times p}{\mathbf{G}} = [g_{ij}] \qquad i = 0, 1, \ldots, p-1; \; j = 0, 1, \ldots, p-1$$

where:

$$g_{00} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0^2}$$

$$g_{01} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1}$$

etc.

Inferences about Regression Parameters, cont'd

G is called the Hessian matrix.

- When the second-order partial derivatives in the Hessian matrix are evaluated at β = b, that is, at the maximum likelihood estimates, the estimated approximate variance-covariance matrix of the estimated regression coefficients for logistic regression can be obtained as follows:

$$s^2\{b\} = \left(\left| 1 - g_{ij} \right|_{\beta=b} \right)^{-1}$$

- With the large sample theory:

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

where z is N(0,1) and $s^2\{b_k\} = \left(\left| 1 - g_{ij} \right|_{\beta_k=b_k} \right)^{-1}$

Test Concerning a Single β_k: Wald Test

A large-sample test of a single regression parameter can be constructed based on (14.52). For the alternatives:

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

an appropriate test statistic is:

- With the large sample theory:

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

and the decision rule is:

If $|z^*| \leq z(1 - \alpha/2)$, conclude H_o
If $|z^*| > z(1 - \alpha/2)$, conclude H_a

This test is called Wald test.

Example:

In the programming task example, β_1 was expected to be positive. The alternatives of interest therefore are:

$$H_o: \beta_1 \leq 0$$
$$H_a: \beta_1 > 0$$

$$z^* = \frac{0.1615}{0.0650} = 2.485$$

For α = 0.05, we require *z(.95)* = 1.645. The decision rule therefore is:

If $|z^*| \leq 1.645$, conclude H_o
If $|z^*| > 1.645$, conclude H_a

Since z^* = 2.485 > 1.645, we conclude H_a, that β_1 is positive, as expected. The one-sided P-value of this test is 0.0065.

Interval Estimation of a Single β_k

The approximate 1 - α confidence limits for β_k:

$$b_k \pm z\left(1 - \frac{\alpha}{2}\right) s\{b_k\}$$

The corresponding confidence limits for the odds ratio exp(β_k) are:

$$exp\left[b_k \pm z\left(1 - \frac{\alpha}{2}\right) s\{b_k\}\right]$$

Example

For the programming task example, it is desired to estimate β_1 with an approximate 95 percent confidence interval. We require $z(.975) = 1.960$, as well as the estimates $b_1 = .1615$ and $s\{b_1\} = .0650$ which are given in Table 14.1b. Hence, the confidence limits are $.1615 \pm 1.960(.0650)$, and the approximate 95 percent confidence interval for β_1 is:

$$.0341 \leq \beta_1 \leq .2889$$

Thus, we can conclude with approximately 95 percent confidence that β_1 is between .0341 and .2889. The corresponding 95 percent confidence limits for the odds ratio are $\exp(.0341) = 1.03$ and $\exp(.2889) = 1.33$.

Test whether Several $\beta_k = 0$: Likelihood Ratio Test

Deviance goodness of fit test:

- completely analogous to the F test for lack of fit for simple and multiple linear regression models
- Assume:
  - c unique combinations of the predictors: $X_1, \ldots, X_c$
  - n_j: $\#\{$repeat binary observations at $X_j\}$
  - Y_{ij}: the ith binary response ar X_j
- Deviance goodness of fit test: Likelihood ratio test of the reduced model:

$$\text{Reduced model: } E\{Y_i j\} = [1 + \exp(-\boldsymbol{X}_j'\beta)]^{-1}$$

$$\text{Full model: } E\{Y_i j\} = \pi_j,\ j = 1, \ldots, c$$

π_j: are parameters

Test whether Several $\beta_k = 0$: Likelihood Ratio Test, cont'd

- $p_j = \frac{Y_{\cdot j}}{n_j}, j = 1, \ldots, c$
- $\hat{\pi}$: the reduced model estimate of π
- The likelihood ratio test statistic:

$$
\begin{aligned}
\text{Deviance: } G^2 &= -2\left[\ln L(R) - \ln L(F)\right] \\
&= -2\sum_{j=1}^{c}\left[Y_{\cdot j}\ln\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j})\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right] \\
&= DEV(X_0, X_1, \ldots, X_{p-1})
\end{aligned}
$$

Test whether Several $\beta_k = 0$: Likelihood Ratio Test, cont'd

- Test the alternative:

$$H_0 : E\{Y_i j\} = [1 + \exp(-\boldsymbol{X}_j'\beta)]^{-1}$$

$$H_a : E\{Y_i j\} \neq [1 + \exp(-\boldsymbol{X}_j'\beta)]^{-1}$$

Decision rule:

If $DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1-\alpha; c-p) \Rightarrow$ conclude H_0

If $DEV(X_0, X_1, \ldots, X_{p-1}) > \chi^2(1-\alpha; c-p) \Rightarrow$ conclude H_a

Example:

$$H_o: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

L(F)=L$(b_0, b_1, b_2, b_3, b_4) = -50.27$
L(R)=L$(b_0, b_2, b_3, b_4) = -53.102$

Hence the required test statistic is:

$G^2 = -2[\log_e L(R) - \log_e L(F)] = -2\,[-50.27 + 53.102] = 5.15$

For $\alpha = 0.05$, $\chi^2(0.95; 1) = 3.84$. The decision rule is:

If $G^2 \leq 3.84$, conclude H_o
If $G^2 > 3.84$, conclude H_a

- Since $G^2 = 5.15 > 3.84$, we conclude H_a, that X_1 should not be dropped from the model. The *P* -value of this test is .023.

Example, cont'd

- Are two-factor interaction terms are required in the model?

$X'\beta_F = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \beta_8 X_2 X_4 + \beta_9 X_3 X_4$
$X'\beta_R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

L(F)=L$(b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9) = -46.998$

L(R)=L$(b_0, b_2, b_3, b_4) = -50.527$

Hence the required test statistic is:

G^2=-2[$\log_e$L(R) - $\log_e$L(F)] = -2 [-50.527 +46.998] = 7.058

For α = 0.05, $\chi^2(0.95; 5) = 11.07$. The decision rule is:

If $G^2 \leq 11.07$, conclude H_o

If $G^2 > 11.07$, conclude H_a

- Since G^2 = 7.058 ≤ 11.07, we conclude H_o ,that we conclude Ho, that the two-factor interactions are not needed in the logistic regression model. The *P* -value of this test is .22.

Automatic Model Selection Methods:

For logistic regression modeling, the AIC_p and SBC_p criteria are easily adapted and are modified.

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

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

Best Subsets Procedures:

These procedures are applicable to the logistic regression models. AIC_p and SBC_p can be used to identify the best subset.

Stepwise Procedures:

May not be feasible but it can be applied for logistic regression models

Example:

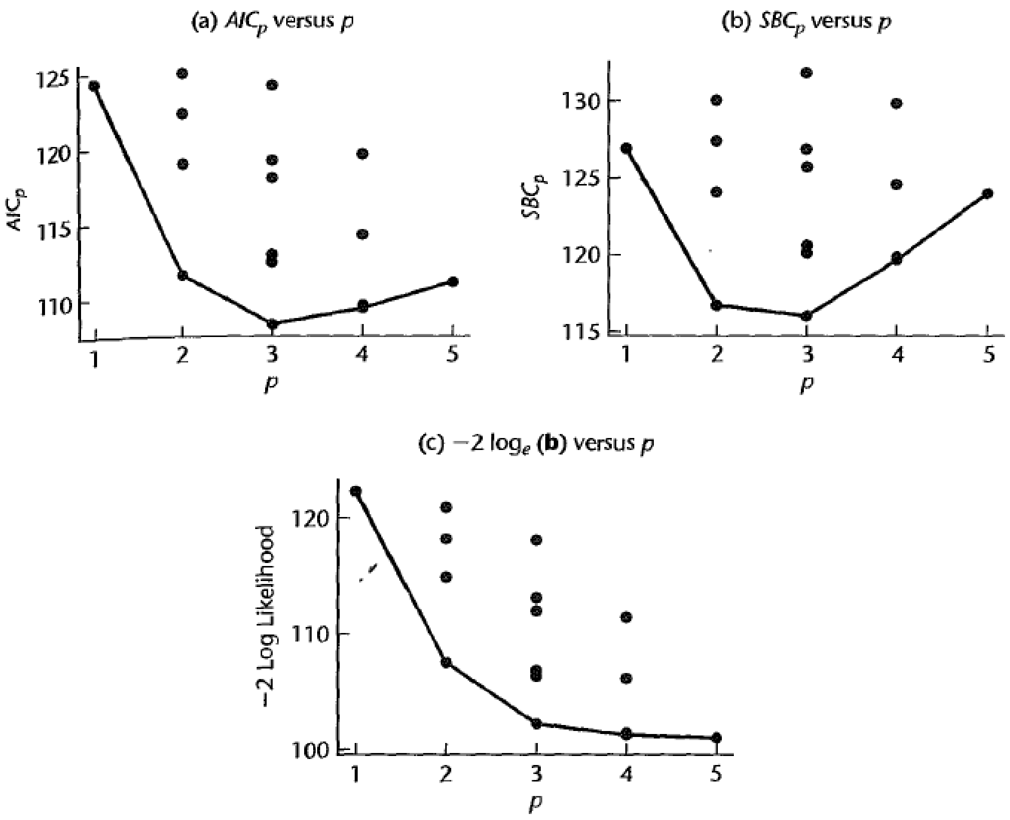
TABLE 14.6 Best Subsets Results—Disease Outbreak Example.

(a) Results for All Possible Models ($X_{ij} = 1$ if X_j in model i; $X_{ij} = 0$ otherwise)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Socioeconomic Status		City Sector			
Model i	Parameters p	Age X_{i1}	X_{i2}	X_{i3}	X_{i4}	AIC_p	SBC_p	$-2\log_e L(\mathbf{b})$
1	1	0	0	0	0	124.318	126.903	122.318
2	2	1	0	0	0	118.913	124.083	114.913
3	2	0	1	0	0	124.882	130.052	120.882
4	2	0	0	1	0	122.229	127.399	118.229
5	2	0	0	0	1	111.534	116.704	107.534
6	3	1	1	0	0	119.109	126.864	113.109
7	3	1	0	1	0	117.968	125.723	111.968
8	3	1	0	0	1	108.259	116.014	102.259
9	3	0	1	1	0	124.085	131.840	118.085
10	3	0	1	0	1	112.881	120.636	106.881
11	3	0	0	1	1	112.371	120.126	106.371
12	4	1	1	1	0	119.502	129.842	111.502
13	4	1	1	0	1	109.310	119.650	101.310
14	4	1	0	1	1	109.521	119.861	101.521
15	4	0	1	1	1	114.204	124.543	106.204
16	5	1	1	1	1	111.054	123.979	101.054

(b) Best Four Models for Each Criterion

	AIC_p Criterion		SBC_p Criterion	
Rank	Predictors	AIC_p	Predictors	SBC_p
1	X_1, X_4	108.259	X_1, X_4	116.014
2	X_1, X_2, X_4	109.310	X_4	116.704
3	X_1, X_3, X_4	109.521	X_1, X_2, X_4	119.650
4	X_1, X_2, X_3, X_4	111.054	X_1, X_3, X_4	119.861



Tests for Goodness of Fit Test

The appropriateness of the fitted logistic regression model needs to be examined before it is accepted for use, as is the case for all regression models. In particular, we need to examine whether the estimated response function for the data is monotonic and sigmoidal in shape key properties of the logistic response function.

Pearson Chi-Square Goodness of Fit Test

- The Pearson chi-square goodness of fit test assumes only that the Y_{ij} observations are independent and that replicated data of reasonable sample size are available.
- The test can detect major departures from a logistic response function, but is not sensitive to small departures from a logistic response function.
- The alternatives of interest are:

$$H_o: E(Y) = \left[1 + e^{-X'\beta}\right]^{-1}$$

$$H_a: E(Y) \neq \left[1 + e^{-X'\beta}\right]^{-1}$$

If the logistic response function is appropriate, the expected value of Y_{ij} is given by:

$$\hat{\pi}_j = \left[1 + e^{-X_j'\beta}\right]^{-1}$$

Tests for Goodness of Fit Test, cont'd

If the logistic response function is appropriate, the expected value of Y_{ij} is given by:

$$\hat{\pi}_j = \left[1 + e^{-X_j'\beta}\right]^{-1}$$

Expected number of cases for the jth class are estimated to be:

$$E_{j1} = n_j \times \hat{\pi}_j$$

$$E_{j0} = n_j \times (1 - \hat{\pi}_j) = n - E_{j1}$$

Actual observed estimates are denoted by $O_{j1}\ and O_{jo}$. Now we have actual and predicted frequencies and we can use the chi-squared test:

$$X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

The decision rule is

$$\text{If } X^2 \leq \chi^2(1-\alpha; c-p), \text{ conclude } H_0$$
$$\text{If } X^2 > \chi^2(1-\alpha; c-p), \text{ conclude } H_a$$

Example

Class j	n_j	$\hat{\pi}_j$	p_j	Number of Coupons Not Redeemed: Observed O_{j0}	Number of Coupons Not Redeemed: Expected E_{j0}	Number of Coupons Redeemed: Observed O_{j1}	Number of Coupons Redeemed: Expected E_{j1}
1	200	.1736	.150	170	165.3	30	34.7
2	200	.2543	.275	145	149.1	55	50.9
3	200	.3562	.350	130	128.8	70	71.2
4	200	.4731	.500	100	105.4	100	94.6
5	200	.7028	.685	63	59.4	137	140.6

$$X^2 = \frac{(170-165.3)^2}{165.3} + \frac{(30-34.7)^2}{34.7} + \cdots + \frac{(137-140.6)^2}{140.6}$$
$$= 2.15$$

For $\alpha = 0.05$ and $c - p = 5 - 2 = 3$, we require $\chi^2(.95; 3) = 7.81$. Since $X^2 = 2.15 \leq 7.81$, we conclude H_0, that the logistic response function is appropriate. The P-value of the test is .54.

Deviance Goodness of Fit Test

- $p_j = \frac{Y_{\cdot j}}{n_j}, j = 1, \ldots, c$
- $\hat{\pi}$: the reduced model estimate of π
- The likelihood ratio test statistic:

$$
\begin{aligned}
\text{Deviance: } G^2 &= -2\left[\ln L(R) - \ln L(F)\right] \\
&= -2\sum_{j=1}^{c}\left[Y_{\cdot j}\ln\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j})\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right] \\
&= DEV(X_0, X_1, \ldots, X_{p-1})
\end{aligned}
$$

Deviance Goodness of Fit Test, cont'd

- Test the alternative:

$$H_0 : E\{Y_ij\} = [1 + \exp(-\boldsymbol{X}'_j \beta)]^{-1}$$

$$H_a : E\{Y_ij\} \neq [1 + \exp(-\boldsymbol{X}'_j \beta)]^{-1}$$

Decision rule:

If $DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1 - \alpha; c - p) \Rightarrow$ conclude H_0

If $DEV(X_0, X_1, \ldots, X_{p-1}) > \chi^2(1 - \alpha; c - p) \Rightarrow$ conclude H_a

Example

Level j	(1) Price Reduction X_j	(2) Number of Households n_j	(3) Number of Coupons Redeemed $Y_{\cdot j}$	(4) Proportion of Coupons Redeemed p_j	(5) Model-Based Estimate $\hat{\pi}_j$
1	5	200	30	.150	.1736
2	10	200	55	.275	.2543
3	15	200	70	.350	.3562
4	20	200	100	.500	.4731
5	30	200	137	.685	.7028

$$
\begin{aligned}
DEV(X_0, X_1) &= -2\left[30 \log_e\left(\frac{.1736}{.150}\right) + (200-30)\log_e\left(\frac{.8264}{.850}\right)\right. \\
&\quad \left. + \cdots + 137\log_e\left(\frac{.7028}{.685}\right) + (200-137)\log_e\left(\frac{.2972}{.315}\right)\right] \\
&= 2.16
\end{aligned}
$$

For $\alpha = .05$ and $c - p = 3$, we require $\chi^2(.95; 3) = 7.81$. Since $DEV(X_0, X_1) = 2.16 \leq 7.81$, we conclude H_0, that the logistic model is a satisfactory fit. The P-value of this test is approximately .54, the same as that obtained earlier for the Pearson chi-square goodness of fit test.

Hosmer-Lemeshow Goodness of Fit Test

Extension of the Chi-Square tests and are widely used for logistic regression models. The data are put into groups/bins based on the predicted values or independent values. For example, FICO bins.

Once the data is put onto bins, the chi-square test can be applied.

TABLE 14.8 Hosmer-Lemeshow Goodness of Fit Test for Logistic Regression Function—Disease Outbreak Example.

Class j	$\hat{\pi}'_j$ Interval	n_j	Number of Persons without Disease: Observed O_{j0}	Number of Persons without Disease: Expected E_{j0}	Number of Persons with Disease: Observed O_{j1}	Number of Persons with Disease: Expected E_{j1}
1	−2.60—under −2.08	20	19	18.196	1	1.804
2	−2.08—under −1.43	20	17	17.093	3	2.907
3	−1.43—under −.70	20	14	14.707	6	5.293
4	−.70—under .16	19	9	10.887	10	8.113
5	.16—under 1.70	19	8	6.297	11	12.703
	Total	98	67	67.180	31	30.820

$$X^2 = \frac{(19-18.196)^2}{18.196} + \frac{(1-1.804)^2}{1.804} + \cdots + \frac{(8-6.297)^2}{6.297} + \frac{(11-12.703)^2}{12.703}$$
$$= 1.98$$

$\chi^2(.95; 3) = 7.81.$

Since $X^2 = 1.98 \leq 7.81$, we conclude H_0,

Logistic Regression Diagnostic

various residuals

- ordinary residuals: e_i

$$e_i = \begin{cases} 1 - \hat{\pi}_i & \text{if } Y_i = 1 \\ -\hat{\pi}_i & \text{if } Y_i = 0 \end{cases}$$

  - not be normally distributed
  - Plots of e_i against $\hat{Y}_i$ or X_j will be uninformative

- Pearson Residuals: r_{p_i}

$$r_{p_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

  - related to Pearson chi-square goodness of fit statistic

Logistic Regression Diagnostic

- Studentized Pearson Residuals: r_{SP_i}

$$r_{SP_i} = \frac{r_{p_i}}{1 - h_{ii}}, \quad H = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2}$$

$\widehat{\boldsymbol{W}}$: the $n \times n$ diagonal matrix with elements $\hat{\pi}_i(1 - \hat{\pi}_i)$

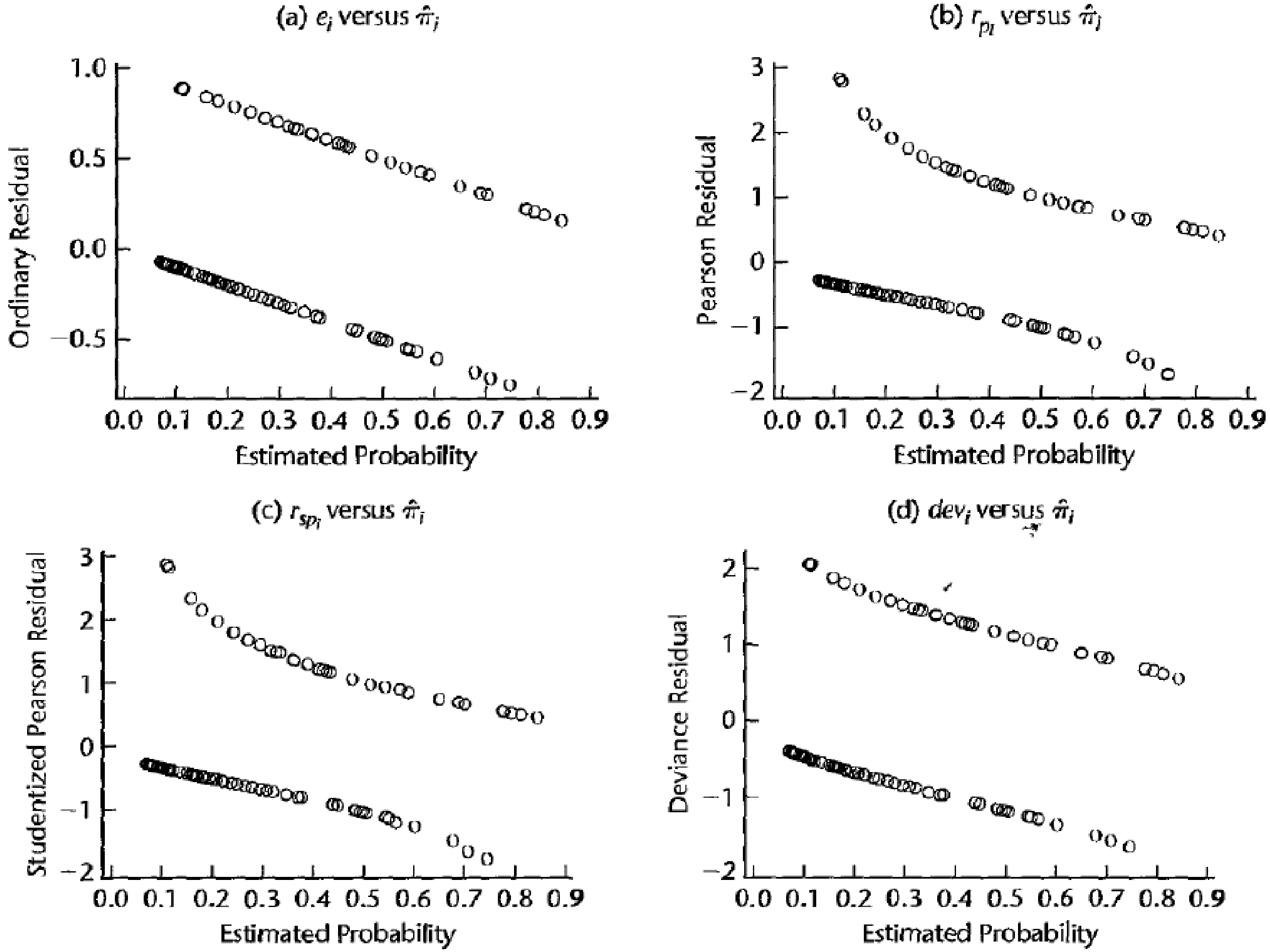
- Deviance Residuals: dev_i

$$DEV(X_0, X_1, \dots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i \ln(\hat{\pi}_i + (1 - Y_i)\ln(1 - \hat{\pi}_i))]$$

$$dev_i = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2\sum_{i=1}^{n}[Y_i \ln(\hat{\pi}_i + (1 - Y_i)\ln(1 - \hat{\pi}_i))]}$$

$$(\sum_{i=1}^{n} dev_i^2 = DEV(X_0, X_1, \dots, X_{p-1}))$$

Logistic Regression Diagnostic



Inferences about Mean Response

The mean response of interest by π_h

$$\pi_h = \left[1 + e^{-X_h'\beta}\right]^{-1}$$

The point estimator of π_h will be denoted by $\hat{\pi}_h$ and is as follows:

$$\hat{\pi}_h = \left[1 + e^{-X_h'b}\right]^{-1}$$

Interval Estimation

The expression by using the fact that $E\{Y_h\} = \pi_h$ and $X_h'\beta = \pi_h'$:

$$\pi_h = \left[1 + e^{-\pi_h'}\right]^{-1} \qquad s^2\{\hat{\pi}_h'\} = s^2\{\mathbf{X}_h'\mathbf{b}\} = \mathbf{X}_h'\mathbf{s}^2\{\mathbf{b}\}\mathbf{X}_h$$

Approximate I - α; large-sample confidence limits for the logit mean response π_h' are then obtained in the usual fashion:

$$L = \hat{\pi}_h' - z(1 - \alpha/2)s\{\hat{\pi}_h'\} \qquad L^* = [1 + \exp(-L)]^{-1}$$

$$U = \hat{\pi}_h' + z(1 - \alpha/2)s\{\hat{\pi}_h'\} \qquad U^* = [1 + \exp(-U)]^{-1}$$

Simultaneous confidence intervals can be applied by changing the critical value by z(1 – α / (2g)).

Example

Want to make the prediction for X_h: [1 10 0 1 0]

$$\hat{\pi}'_h = \mathbf{X}'_h \mathbf{b} = -2.3129(1) + .02975(10) + .4088(0) - .30525(1) + 1.5747(0)$$
$$= -2.32065$$

$$s^2\{\hat{\pi}'_h\} = .2945$$

$$L = -2.32065 - 1.960(.54268) = -3.38430$$
$$U = -2.32065 + 1.960(.54268) = -1.25700$$

Finally, we use (14.94) to obtain the confidence limits for the mean response π_h:

$$L^* = [1 + \exp(3.38430)]^{-1} = .033$$
$$U^* = [1 + \exp(1.25700)]^{-1} = .22$$

Thus, the approximate 95 percent confidence interval for the mean response π_h is:

$$.033 \le \pi_h \le .22$$

Prediction of a New Observation

Choice of Prediction Rule:

1. *Use.5 as the cutoff.* With this approach. the prediction rule is:

If $\hat{\pi}_h$ exceed 0.5. predict 1; otherwise predict 0

2. *Find the best cutoff based on the numerical search:*

3. *Use prior probabilities, expert judgment, and costs of incorrect predictions in determining the cutoff*

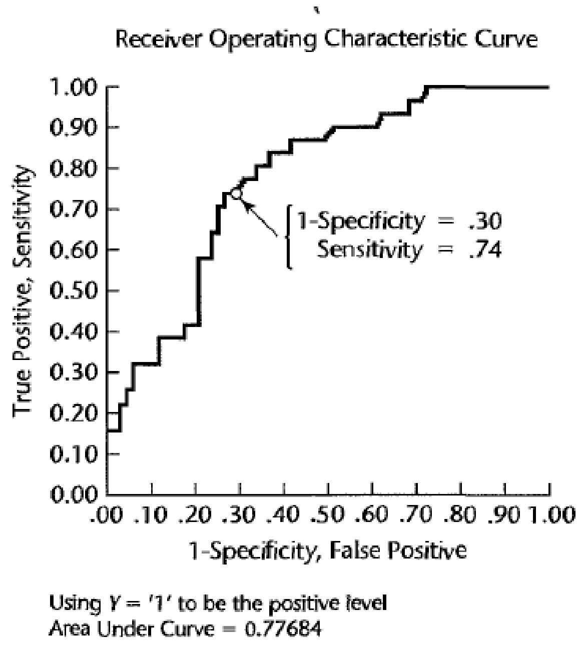
TABLE 14.12 Classification Based on Logistic Response Function (14.46) and Prediction Rules (14.95) and (14.96)—Disease Outbreak Example.

True Classification	(a) Rule (14.95)			(b) Rule (14.96)		
	$\hat{Y}=0$	$\hat{Y}=1$	Total	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=0$	47	20	67	50	17	67
$Y=1$	8	23	31	9	22	31
Total	55	43	98	59	39	98

Prediction of a New Observation

An effective way to display this information graphically is through the *receiver operating characteristic* (ROC) *curve,* which plots $P(\hat{Y} = 1|Y = 1)$ (also called *sensitivity)* as a function of 1 - $P(\hat{Y} = 0|Y = 0)$(also called *specificity)* for the possible cutpoints $\hat{\pi}_h$.

$P(\hat{Y} = 1|Y = 1)$=23/31=0.74



The Confusion Matrix

		Actual Class	
		Good Payer	Defaulter
Predicted Class	Good payer	True positive (TP)	False positive (FP)
	Defaulter	False Negative (FN)	True Negative (TN)

continued...

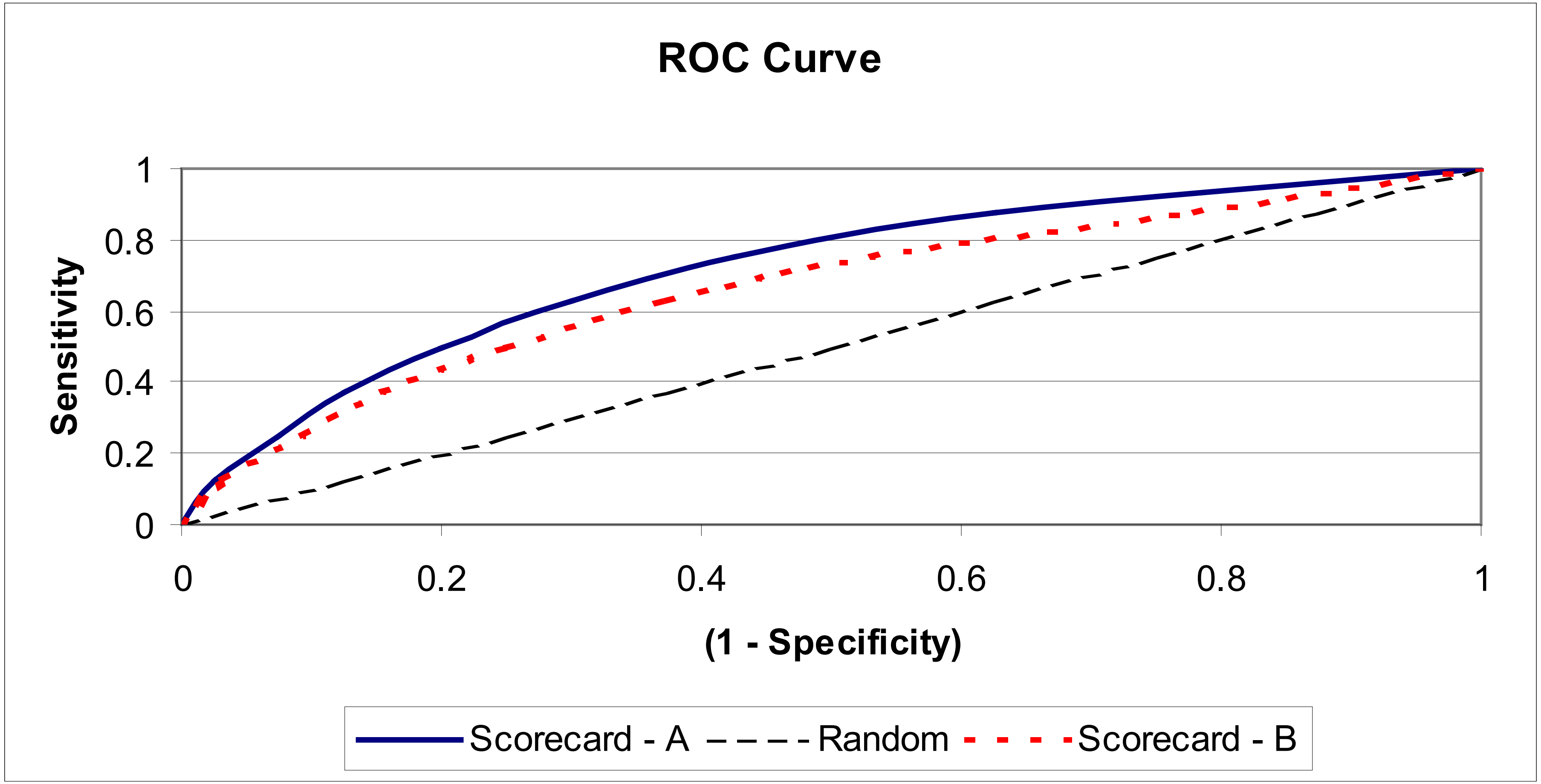
The Confusion Matrix

- Classification accuracy
  = (TP+TN) / (TP+FP+TN+FN)
- Error rate = (FP+FN) / (TP+FP+TN+FN)
- Sensitivity = TP / (TP+FN)
- Specificity = TN / (TN+FP)
- All these measures vary when the classification cut-off is varied.
- Extremes
  - Predict all customers as good:
    - Sensitivity=100, specificity=0
  - Predict all customers as bad:
    - Sensitivity=0, specificity=100

The Receiver Operating Characteristic (ROC) Curve

- The ROC curve is a two-dimensional graphical illustration of the sensitivity on the Y-axis versus (1-specificity) on the X-axis for various values of the classification threshold.
- In a credit scoring context, the sensitivity is the percentage of goods predicted to be good, and 1-specificity is the percentage of bads predicted to be good.
- It basically illustrates the behaviour of a classifier without regard to class distribution or error cost, so it effectively decouples classification performance from these factors.

The Receiver Operating Characteristic Curve

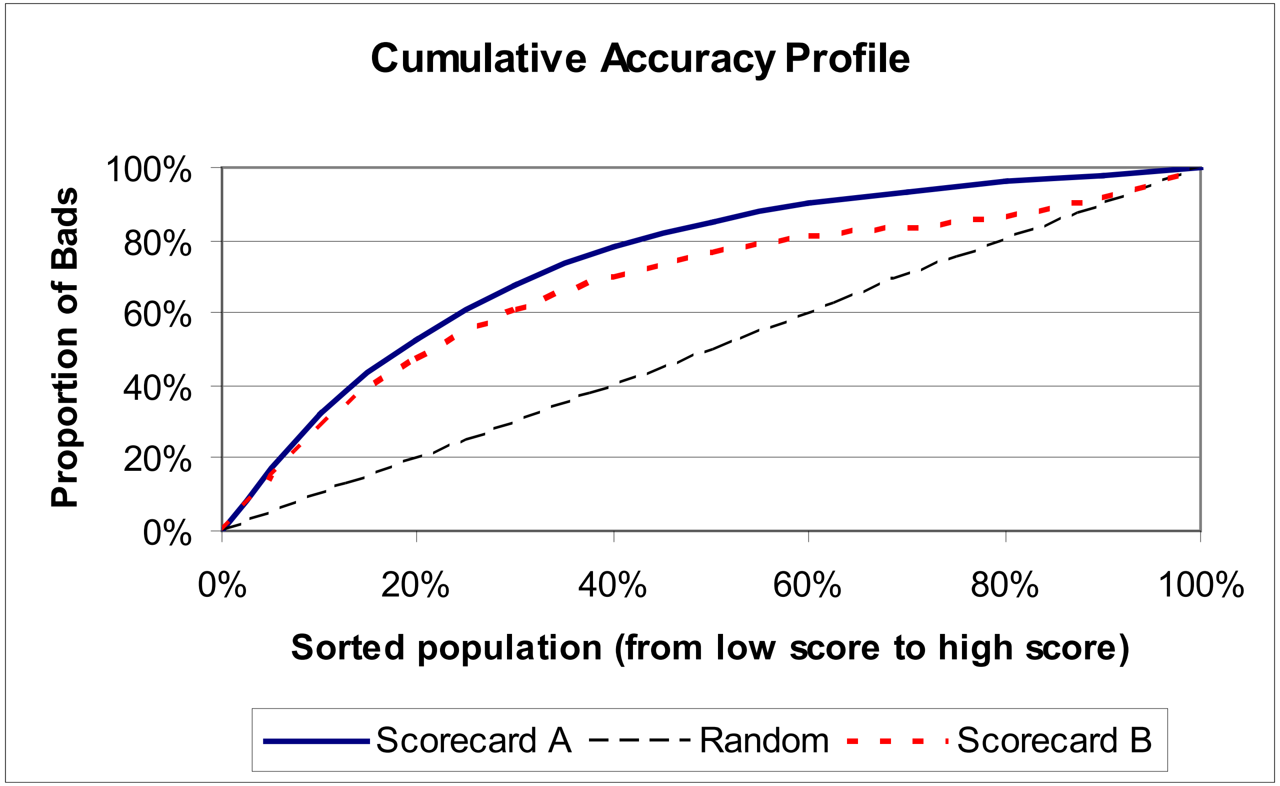


The Area Under the ROC Curve

- How to compare intersecting ROC curves?
- The area under the ROC curve (AUC).
- The AUC provides a simple figure-of-merit for the performance of the constructed classifier.
- An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen instance of class 1 is correctly ranked higher than a randomly chosen instance of class 0 (Hanley and McNeil, 1983) (Wilcoxon or Mann-Whitney or U statistic).
- The higher the better.
- A good classifier should have an AUC larger than 0.5.

The Cumulative Accuracy Profile (CAP)

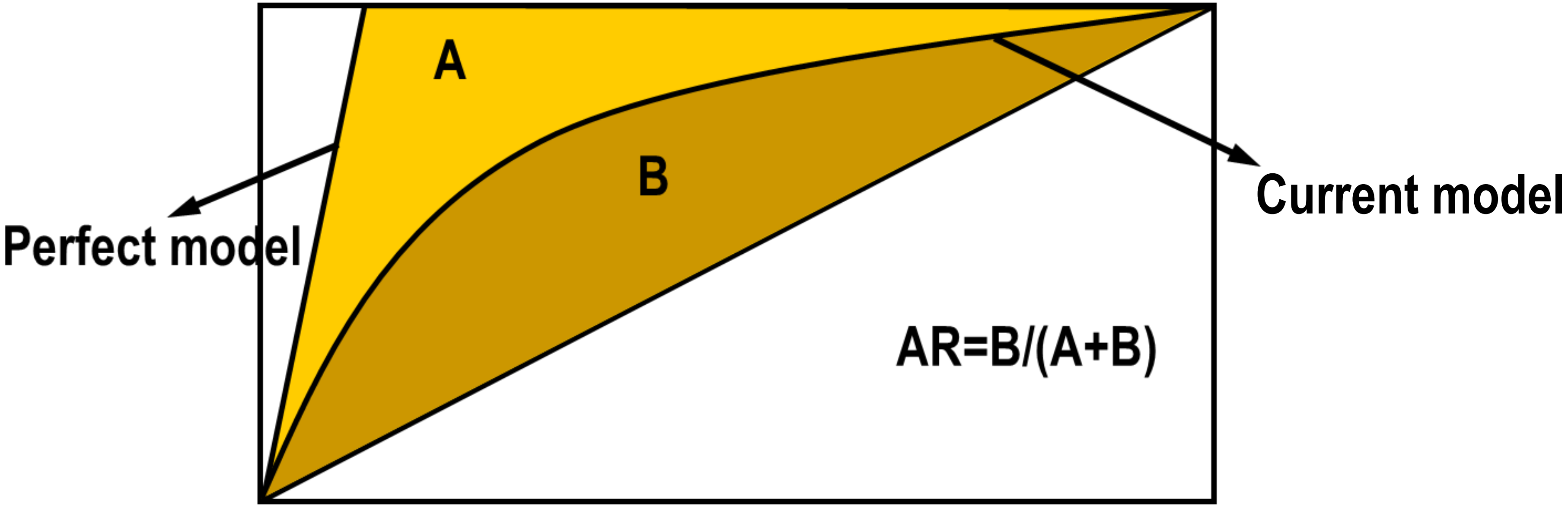
- Distribution of “bad” cases and total cases by deciles across all score ranges



Also called the Lorenz or Power curve

...

Accuracy Ratio



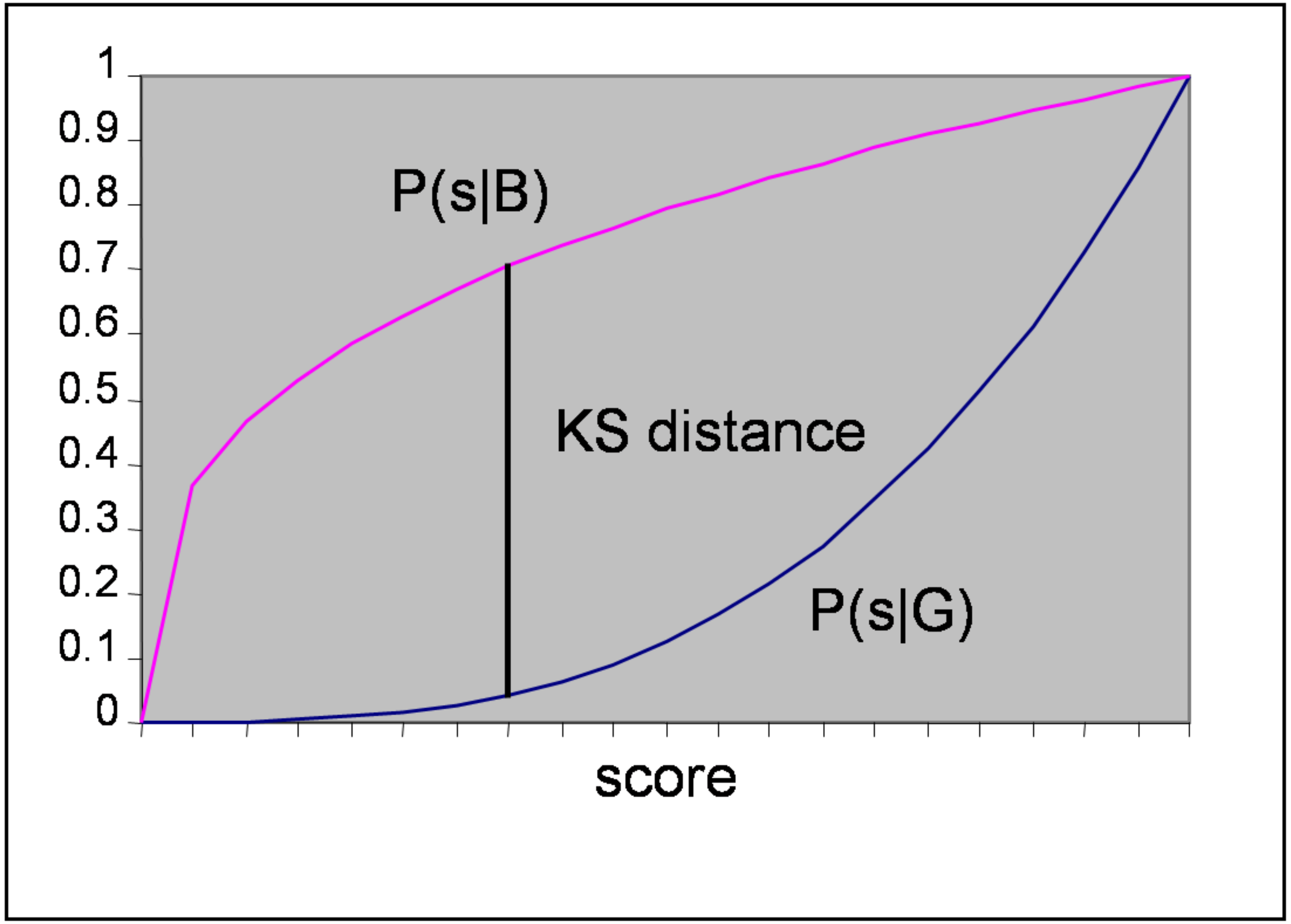
- The accuracy ratio (AR) is defined as follows:
  (Area below power curve for current model-Area below power curve for random model) /
  (Area below power curve for perfect model-Area below power curve for random model)
- Perfect model has an AR of 1.
- Random model has an AR of 0.
- AR is sometimes also called the *Gini coefficient*.
- **AR=2*AUC-1**.

The Kolmogorov-Smirnov (KS) Distance

- Separation measure.
- Measures the distance between the cumulative score distributions P(s|B) and P(s|G).
- KS = $\max_s |P(s|G)-P(s|B)|$, where:
  - $P(s|G) = \sum_{x \leq s} p(x|G)$ (equals 1- sensitivity)
  - $P(s|B) = \sum_{x \leq s} p(x|B)$ (equals the specificity)
- KS distance metric is the maximum vertical distance between both curves.
- KS distance can also be measured on the ROC graph:
  - Maximum vertical distance between ROC graph and diagonal

continued...

The Kolmogorov-Smirnov Distance



The Mahalanobis Distance

- Better than Euclidean distance because it takes the distribution (standard deviation) of the scores into account
- Measure the Mahalanobis distance between the two mean scores of the scorecards

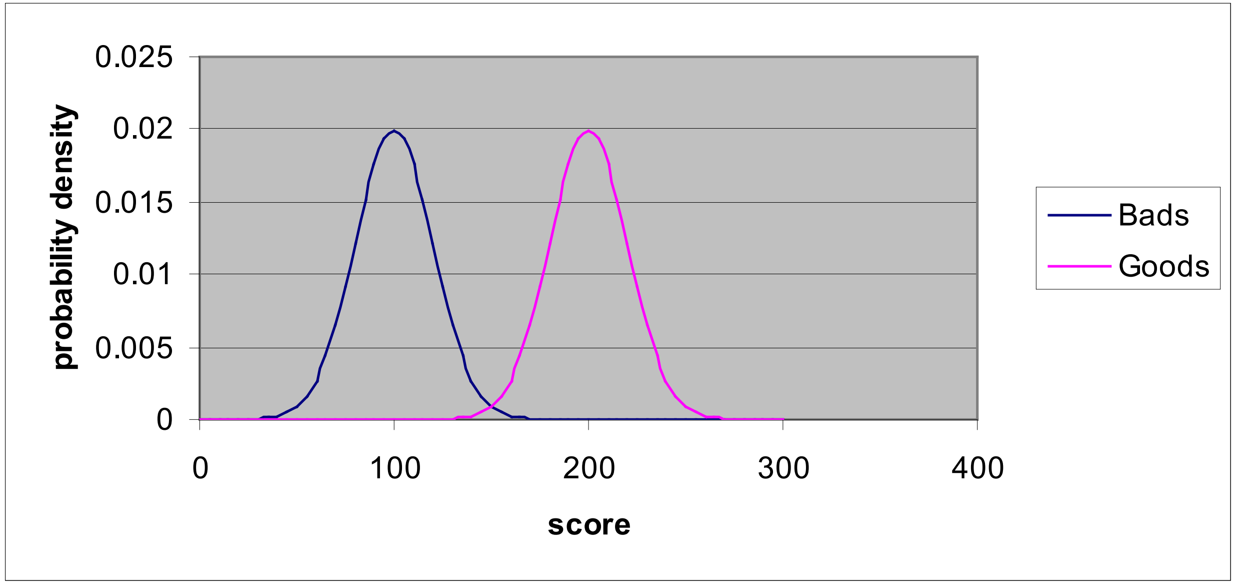
$$M = \frac{\left|\mu_G - \mu_B\right|}{\sigma}$$

with σ the (pooled) standard deviation of the scores of the goods and the bads from their respective means

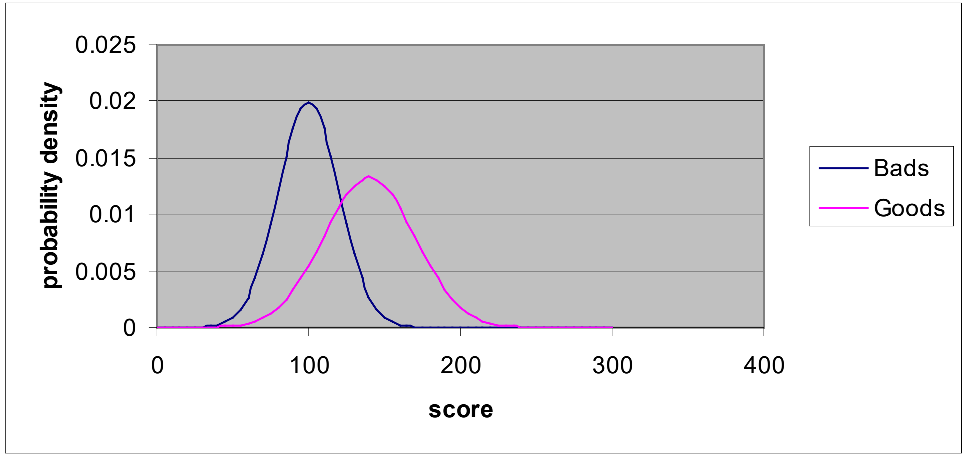
continued...

The Mahalanobis Distance

Good separation



Bad separation



Poisson Regression

Poisson regression is useful when the outcome is a count, with large-count outcomes being rare events. The Poisson probability distribution is as follows:

$$f(Y) = \frac{\mu^Y \exp(-\mu)}{Y!} \qquad Y = 0, 1, 2, \ldots$$

where $f(Y)$ denotes the probability that the outcome is Y and $Y! = Y(Y-1)\cdots 3 \cdot 2 \cdot 1$.

The mean and variance of the Poisson probability distribution are:

$$E\{Y\} = \mu$$

$$\sigma^2\{Y\} = \mu$$

Poisson Regression, cont'd

The Poisson regression model, like any nonlinear regression model, can be stated as follows:

$$Y_i = E\{Y_i\} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

The mean response for the ith case, to be denoted now by μ_i for simplicity, is assumed as always to be a function of the set of predictor variables, $X_1, \ldots, X_{p-1}$. We use the notation $\mu(\mathbf{X}_i, \boldsymbol{\beta})$ to denote the function that relates the mean response μ_i to $\mathbf{X}_i$, the values of the predictor variables for case i, and $\boldsymbol{\beta}$, the values of the regression coefficients. Some commonly used functions for Poisson regression are:

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i'\boldsymbol{\beta}$$

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i'\boldsymbol{\beta})$$

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \log_e(\mathbf{X}_i'\boldsymbol{\beta})$$

The most commonly used response function is $\mu_i = \exp(\mathbf{X}'\boldsymbol{\beta})$.

Maximum Likelihood Estimation

For Poisson regression model (14.113), the likelihood function is as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \frac{[\mu(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i} \exp[-\mu(\mathbf{X}_i, \boldsymbol{\beta})]}{Y_i!}$$

$$= \frac{\left\{\prod_{i=1}^{n} [\mu(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i}\right\} \exp\left[-\sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta})\right]}{\prod_{i=1}^{n} Y_i!}$$

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log_e[\mu(\mathbf{X}_i, \boldsymbol{\beta})] - \sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta}) - \sum_{i=1}^{n} \log_e(Y_i!)$$

Numerical search procedures are used to find the maximum likelihood estimates $b_0, b_1, \ldots, b_{p-1}$. Iteratively reweighted least squares can again be used to obtain these estimates. We shall rely on standard statistical software packages specifically designed to handle Poisson regression to obtain the maximum likelihood estimates.

Model Development

Model development for a Poisson regression model is carried out in a similar fashion to that for logistic regression, conducting tests for individual coefficients or groups of coefficients based on the likelihood ratio test statistic G^2. For Poisson regression, the model deviance is as follows:

$$DEV(X_0, X_1, \ldots, X_{p-1}) = -2\left[\sum_{i=1}^{n} Y_i \log_e\left(\frac{\hat{\mu}_i}{Y_i}\right) + \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)\right]$$

Deviance residual for the ith case is:

$$dev_i = \pm\left[-2Y_i \log_e\left(\frac{\hat{\mu}_i}{Y_i}\right) - 2(Y_i - \hat{\mu}_i)\right]^{1/2}$$

Inferences for a Poisson regression model are carried out in the same way as for logistic regression.

Example

TABLE 14.14 Data—Miller Lumber Company Example.

Census Tract i	Housing Units X_1	Average Income X_2	Average Age X_3	Competitor Distance X_4	Store Distance X_5	Number of Customers Y
1	606	41,393	3	3.04	6.32	9
2	641	23,635	18	1.95	8.89	6
3	505	55,475	27	6.54	2.05	28
...	...	...	...	...	...	...
108	817	54,429	47	1.90	9.90	6
109	268	34,022	54	1.20	9.51	4
110	519	52,850	43	2.92	8.62	6

(a) Fitted Poisson Response Function

$$\hat{\mu} = \exp[2.942 + .000606X_1 - .0000117X_2 - .00373X_3 + .168X_4 - .129X_5]$$

$$DEV(X_0, X_1, X_2, X_3, X_4, X_5) = 114.985$$

(b) Estimated Coefficients, Standard Deviations, and G^2 Test Statistics

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	G^2	P-value
β_0	2.9424	.207		
β_1	.0006058	.00014	18.21	.000
β_2	−.00001169	.0000021	31.80	.000
β_3	−.003726	.0018	4.38	.036
β_4	.1684	.026	41.66	.000
β_5	−.1288	.016	67.50	.000

Census Tract i	Y_i	$\hat{\mu}_i$	dev_i
1	9	12.3	−.999
2	6	8.8	−.992
3	28	28.1	−.024
...	...	...	...
108	6	5.3	.289
109	4	4.4	−.197
110	6	6.4	−.171

Generalized Linear Models

General class of linear models that are made up of 3 components: Random, Systematic, and Link Function

1. Random component: $Y_1, \ldots, Y_n$ *are n* independent responses that follow a probability distribution belonging to the *exponential family* of probability distributions, with expected value $E\{Y_i\} = \mu_i$

2. Systematic component: A *linear predictor* based on the predictor variables $X_{i,1}, \ldots, X_{i,p-1}$ is utilized, denoted by $X_i'\beta$

$$\mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

3. The *link function: g* relates the linear predictor to the mean response:

$$X_i'\beta = \mathrm{g}(\mu_i)$$

Common Link Functions

- Identity link (form used in *normal* and *gamma* regression models):

$$g(\mu) = \mu$$

- Log link (used when μ cannot be negative as when data are *Poisson* counts):

$$g(\mu) = \log(\mu)$$

- Logit link (used when μ is bounded between 0 and 1 as when data are binary):

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$