Chapter 14

Logistic Regression with Binary Response

Instructor: Li C.X.

Outline

- Binary variables
- Odds and odds ratio
- Modeling binary outcome variables
- The logistic model
- Parameter estimation
- Inferences about regression parameters

Binary response variables

- A binary response variable Y which takes on the values 0 or 1.
- In these situations, a parameter which is usually of interest is

$$\pi = P(Y = 1)$$

- *Odds:* $$Odds(\pi) = \frac{\pi}{1-\pi} = \frac{P(Y=1)}{1-P(Y=1)}.$$

$$Odds(\pi) < 1 \Leftrightarrow \pi < 0.5,$$
$$Odds(\pi) = 1 \Leftrightarrow \pi = 0.5,$$
$$Odds(\pi) > 1 \Leftrightarrow \pi > 0.5.$$

- When two fair coins are flipped, P(two heads)=1/4 , P(not two heads)=3/4.The odds in favor of getting two heads is: 1/3, or sometimes referred to as 1 to 3 odds.

Odds ratio

- Often in applied statistics, we are interested in comparing the probability of $Y = 1$, across two groups.

$$\pi_1 = P(Y = 1 \mid \text{group } 1),$$
$$\pi_2 = P(Y = 1 \mid \text{group } 2).$$

- The odds ratio (OR) is simply defined as the ratio of the odds in favor of $Y=1$ in the two groups:

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} \quad .$$

- OR take on values from 0 to ∞ .

$$OR < 1 \Leftrightarrow \pi_1 < \pi_2,$$
$$OR = 1 \Leftrightarrow \pi_1 = \pi_2,$$
$$OR > 1 \Leftrightarrow \pi_1 > \pi_2.$$

- $Y \sim B(1, \pi)$

$$E(Y) = \sum yP(Y = y)$$

$$= 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0)$$

$$= 1 \cdot \pi + 0 \cdot (1 - \pi)$$

$$= \pi \ .$$

- Based on an i.i.d. random sample Y_1 , $Y_2, \ldots,$ $Y_n \sim B(1, \pi)$, the MLE

$$\hat{\pi} = \frac{\sum_{i=1}^{n} Y_i}{n} .$$

Modeling binary outcome variables

- Until this point our dependent variable of interest of regression has been (assumed) continuous.
- Consider the simple linear regression model:

Independent $Y_1, Y_2, \ldots, Y_n$, $Y_i \sim N(\mu_i, \sigma^2)$

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}$$

or $$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad \varepsilon_i \sim N\left(0, \sigma_i^2\right), i.i.d.$$

- In applied statistics, we often encounter the situation where the dependent variable is binary.
  - Response to treatment, presence/absence of a certain genetic trait.

- Sometimes this binary response variable is dependent on some other continuous background variable, i.e. $Y \sim B(1, \pi(X))$.

- A sample: independent $Y_1, Y_2, \ldots, Y_n$.

$$Y_i \sim B(1, \pi_i)$$

$$\pi_i = E(Y_i) = P(Y_i = 1) = g(X_{i1}, X_{i2}, \cdots, X_{i,p-1})$$

- Function g?

Probit mean response function

- Assume that underlying the binary outcome *Y* is a possibly unobservable continuous variable *Y'*

$$Y = 1 \Leftrightarrow Y' < \tau \, , \qquad \pi_i = P(Y_i = 1) = P(Y_i' < \tau) \, .$$
$$Y = 0 \Leftrightarrow Y' > \tau \, .$$

- Assume a linear relationship between *Y'* and the predictors.

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

- Probit response function

$$\pi_i = P(Y_i' < \tau) = P(\beta_0' + \beta_1' X_{i1} + \ldots + \beta_{p-1}' X_{i,p-1} + \varepsilon_i < \tau)$$
$$= P\left[\frac{\varepsilon_i}{\sigma} < \frac{\tau - \beta_0' + \beta_1' X_{i1} + \ldots + \beta_{p-1}' X_{i,p-1}}{\sigma} \right]$$
$$= \Phi(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1})$$

- Then $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}$

Logistic mean response function

- Sample: independent $Y_1, Y_2, \ldots, Y_n$. $\quad Y_i \sim B(1, \pi_i)$

$$\pi_i = E(Y_i) = P(Y_i = 1) = g(X_{i1}, X_{i2}, \cdots, X_{i,p-1})$$

- Logistic mean response function

$$\pi_i = E(Y_i) = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)}$$

$$= \left[1 + \exp\left(-\beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)\right]^{-1}$$

- This model can be linearized, using the transformation, known as the logit transformation.

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}$$

Simple Logistic model

- $\{Y_i\}$ are independent Bernoulli random variables with mean π_i

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

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

- **Parameter interpretation**
  - β_1 represents the change in the logit, or log odds, for a unit increase in the predictor.

$$\log\left(\frac{\pi_{X=x}}{1-\pi_{X=x}}\right) = \beta_0 + \beta_1 x\,, \qquad \log\left(\frac{\pi_{X=x+1}}{1-\pi_{X=x+1}}\right) = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1\,,$$

$$\beta_1 = \log\left(\frac{\pi_{X=x+1}}{1-\pi_{X=x+1}}\right) - \log\left(\frac{\pi_{X=x}}{1-\pi_{X=x}}\right) = \log\left(\frac{\pi_{X=x+1}}{1-\pi_{X=x+1}} \Big/ \frac{\pi_{X=x}}{1-\pi_{X=x}}\right).$$

$$OR = \exp(\beta_1) = \frac{\pi_{X=x+1}}{1-\pi_{X=x+1}} \Big/ \frac{\pi_{X=x}}{1-\pi_{X=x}}\,.$$

Simple Logistic model

- Log-Likelihood

$$
\begin{aligned}
\log(L) &= \log\left\{\prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}\right\} \\
&= \sum Y_i \log(\pi_i) + \sum (1-Y_i)\log(1-\pi_i) \\
&= \sum Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum \log(1-\pi_i) \\
&= \sum Y_i(\beta_0+\beta_1 X_i) - \sum \log(1+\exp(\beta_0+\beta_1 X_i))
\end{aligned}
$$

- Maximum likelihood estimators (MLEs) b0 and b1 of parameters do not have analytical closed formulas.
- Computer packages use iterative numerical procedures to find MLEs.
- These estimates are used to calculate

$$
\hat{\pi}_i = \frac{\exp(b_0+b_1X_i)}{1+\exp(b_0+b_1X_i)} \qquad \operatorname{logit}(\hat{\pi}_i) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = b_0 + b_1 X_i\,. \qquad \widehat{OR} = \exp(b_1)\,.
$$

Example

- Dataset

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

- Estimates

(b) Maximum Likelihood Estimates

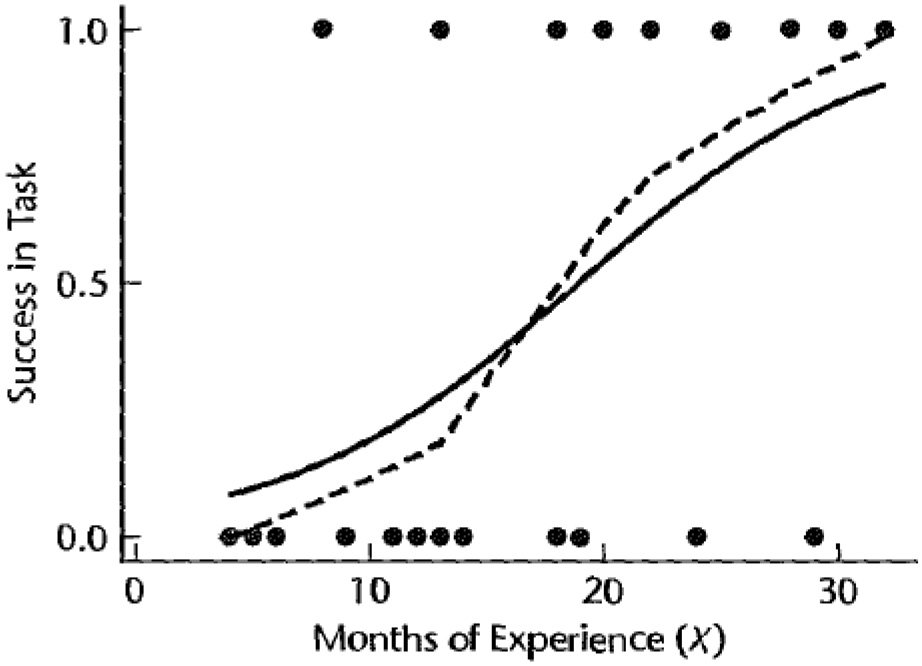
Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation
β_0	−3.0597	1.259
β_1	.1615	.0650

$$\operatorname{logit}(\hat{\pi}_i) = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.0597 + 0.1615X_i$$

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

$$\widehat{OR} = \exp(b_1) = \exp(.1615) = 1.175$$

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



Multiple Logistic model

$$\underset{n\times p}{\boldsymbol{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} \qquad \boldsymbol{X}_i = \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i,p-1} \end{bmatrix}. \qquad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)}{1+\exp\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)} = \frac{\exp\left(\mathbf{X}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}_i'\boldsymbol{\beta}\right)} = \frac{1}{1+\exp\left(-\mathbf{X}_i'\boldsymbol{\beta}\right)}$$

The log likelihood

$$\begin{aligned} \log L &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1}) \\ &\quad - \sum_{i=1}^{n} \log\left(1+\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1})\right) \\ &= \sum_{i=1}^{n} y_i(\boldsymbol{X}_i'\boldsymbol{\beta}) - \sum_{i=1}^{n} \log\left(1+\exp(\boldsymbol{X}_i'\boldsymbol{\beta})\right). \end{aligned}$$

Categorical predictors

Disease outbreak example:

- Socioeconomic status (3 levels) and city sectors (2 levels)

Class	X_2	X_3
Upper	0	0
Middle	1	0
Lower	0	1

$X_4 = 0$ for sector 1 and $X_4 = 1$ for sector 2.

	(1)	(2)	(3)	(4)	(5)	(6)
		Socioeconomic Status				
Case i	Age X_{i1}	X_{i2}	X_{i3}	City Sector X_{i4}	Disease Status Y_i	Fitted Value $\hat{\pi}_i$
1	33	0	0	0	0	.209
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

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

For case i=1,

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

The meaning of OR values?

Inferences about Regression Parameters

- Maximum likelihood estimators for logistic regression are approximately normally distributed, with little or no bias.

$$\frac{b_k - \beta_k}{s(b_k)} \sim N(0,1), \text{ approximately}$$

- **Wald Z test for a single β_k** : $H_0: \beta_k = 0 \quad H_a: \beta_k \neq 0$

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

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

- Disease outbreak example: test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$$z^* = \frac{b_1}{s(b_1)} = \frac{0.02975}{0.01350} = 2.204 > 1.96, \quad p = 0.0275 < \alpha = 0.05$$

- The approximate $1-\alpha$ confidence interval for β_k:

$$b_k \pm z(1-\alpha/2)s(b_k)$$

- The approximate $1-\alpha$ confidence interval for odds ratio $\exp(\beta_k)$

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

- Disease outbreak example: Find 95% confidence intervals for β_2 and for the odds ratio $\exp(\beta_2)$

- Remark: Approximate joint CIs for g logistic paramters can be developed by Bonferroni procedure.

$$b_k \pm z(1-\alpha/(2g))s(b_k)$$

Testing a subset of parameters

- Testing a subset of parameters

  H_0: $\beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

  H_a: not all of the β_k in H_0 equal zero

- **Review: Likelihood ratio test (LRT)**

$$H_0 : \theta \in \Theta_0 \ \textit{versus} \ \ H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta} \mid H_0)}{L(\hat{\theta})}$$

$$-2\ln\Lambda = -2\left[\ln L(\hat{\theta} \mid H_0) - \ln L(\hat{\theta})\right] \dot{\sim} \ \chi^2(k) \ \text{ under } H_0,$$

where $k = \dim(\Theta) - \dim(\Theta_0)$

Large values support H1

LRT for testing a subset of parameters

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

$$H_a: \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero}$$

- Based on Full and Reduced Model
- Original model (Full model)

$$\pi_i = \left[1 + \exp\left(-\beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)\right]^{-1} = \left[1 + \exp\left(-\mathbf{X}_i' \boldsymbol{\beta}_F\right)\right]^{-1}$$

- Under H0, the model is reduced to be:

$$\pi_i = \left[1 + \exp\left(-\beta_0 - \beta_1 X_{i1} - \ldots - \beta_{q-1} X_{i,q-1}\right)\right]^{-1} = \left[1 + \exp\left(-\mathbf{X}_i' \boldsymbol{\beta}_R\right)\right]^{-1}$$

  - It is nested within the full model.
- LR test $\chi_L^2 = -2\left[\ln L(\hat{\boldsymbol{\beta}} \mid H_0) - \ln L(\hat{\boldsymbol{\beta}})\right]$

$$= -2\left[\ln L(\hat{\beta}_0^* .., \hat{\beta}_{q-1}^*, 0, .., 0) - \ln L(\hat{\beta}_0, ..., \hat{\beta}_{p-1})\right]$$

$$= -2\left[\ln L(\hat{\boldsymbol{\beta}}_R) - \ln L(\hat{\boldsymbol{\beta}}_F)\right] \dot{\sim} \chi^2(p-q) \text{ under } H_0$$

$$\boxed{\chi_L^2 = -2\left[\ln L(\text{Reduced model}) - \ln L(\text{Full model})\right]}$$

Testing a subset of parameters

- Testing a subset of parameters

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

H_a: not all of the β_k in H_0 equal zero

- LRT: $$\boxed{\chi_L^2 = -2\left[\ln L(\text{Reduced model}) - \ln L(\text{Full model})\right]}$$

Reject H0 if $\chi_L^2 > \chi^2(1-\alpha; p-q)$

- Disease outbreak example: test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$$\ln L(\text{F}) = \ln(b_0, b_1, b_2, b_3, b_4) = -50.527$$

$$\ln L(\text{R}) = \ln(b_0^*, b_2^*, b_3^*, b_4^*) = -53.502$$

$$\chi_L^2 = -2\left[\ln L(\text{R}) - \ln L(\text{F})\right] = 5.15 > 3.84, \quad p = 0.023 < 0.05$$

- How to test $H_0: \beta_2 = \beta_3 = 0$ (Socioeconomic status has no effect) vs $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$?

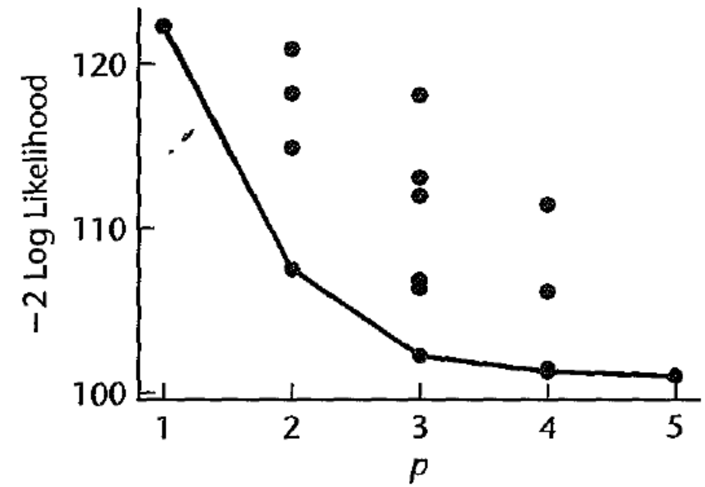
Model selection criteria

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

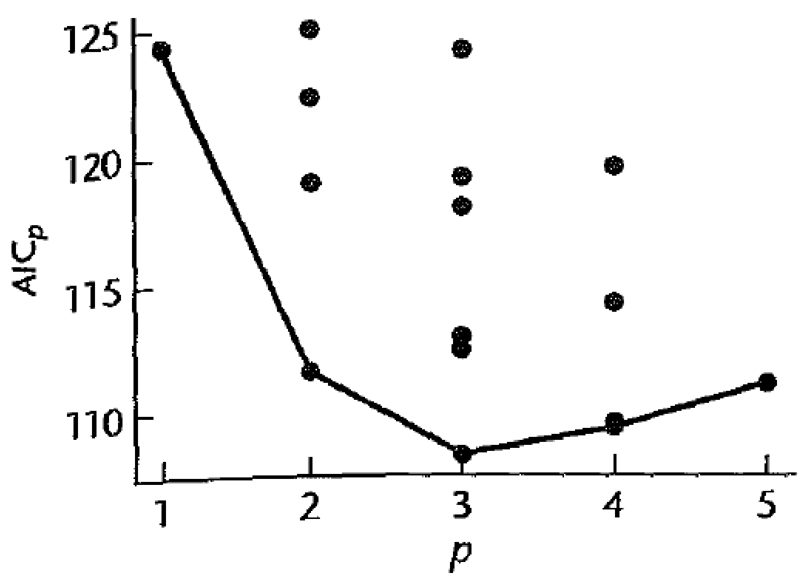
$$SBC_p = -2\log_e L(\mathbf{b}) + p\log_e(n)$$

Remark: for nested models can use LRT.

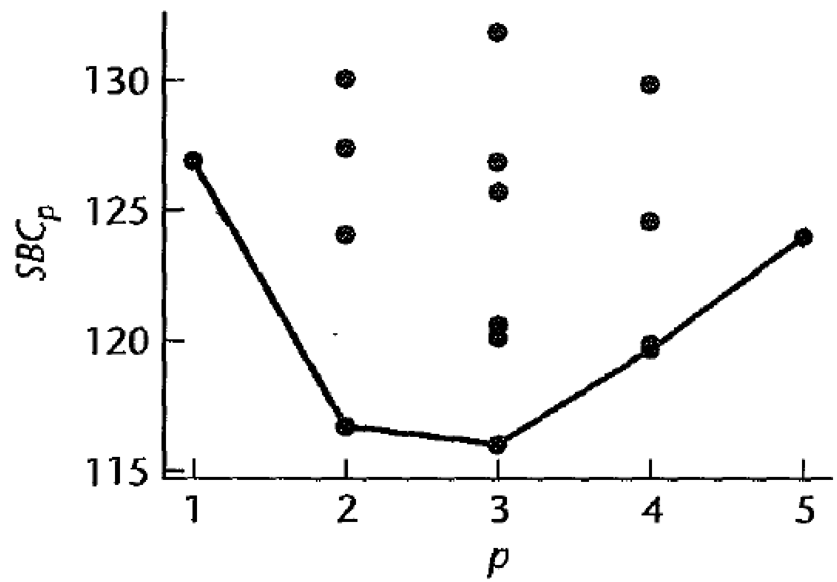
(c) $-2\log_e(\mathbf{b})$ versus p



(a) AIC_p versus p



(b) SBC_p versus p



Model selection: stepwise logistic

Logistic Regression
Block 1: Method = Forward Stepwise (Wald)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SECTOR	1.743	.473	13.593	1	.000	5.716
	Constant	−3.332	.765	18.990	1	.000	.036
Step 2[b]	AGE	.029	.013	4.946	1	.026	1.030
	SECTOR	1.673	.487	11.791	1	.001	5.331
	Constant	−4.009	.873	21.060	1	.000	.018

a. Variable(s) entered on step 1: SECTOR.

b. Variable(s) entered on step 2: AGE.

Prediction

- For a given $\mathbf{X}_h$, $\hat{Y}_h = ?$

$$\hat{Y}_h = 1, \text{ if } \hat{\pi}_h > p_c; \ \hat{Y}_h = 0, \ otherwise$$

$$\Leftrightarrow \hat{Y}_h = 1, \text{ if } \mathbf{X}_h'\hat{\boldsymbol{\beta}} > c; \ \hat{Y}_h = 0, \ otherwise$$

- Choice of prediction rule:

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

   If $\hat{\pi}_h$ exceeds .5, predict 1; otherwise predict 0. $\hat{Y}_h = 1,$

2. *Find the best cutoff for the data set*

   This approach involves evaluating different cutoffs.

   The cutoff for which the proportion of incorrect predictions is lowest

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

$$\text{Predict 1 if } \hat{\pi}_h \geq .316\text{; predict 0 if } \hat{\pi}_h < .316 \quad (14.95)$$

$$\text{Predict 1 if } \hat{\pi}_h \geq .325\text{; predict 0 if } \hat{\pi}_h < .325 \quad (14.96)$$

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

For rule (14.95):

- Sensitivity (true positive rate, TPR):

$$P(\hat{Y}=1|Y=1)=\frac{23}{31}=.74$$

- 1−Specificity(false positive rate, FPR)

$$1-P(\hat{Y}=0|Y=0)=1-\frac{47}{67}=.30$$

  - Specificity(true negative rate, TNR)

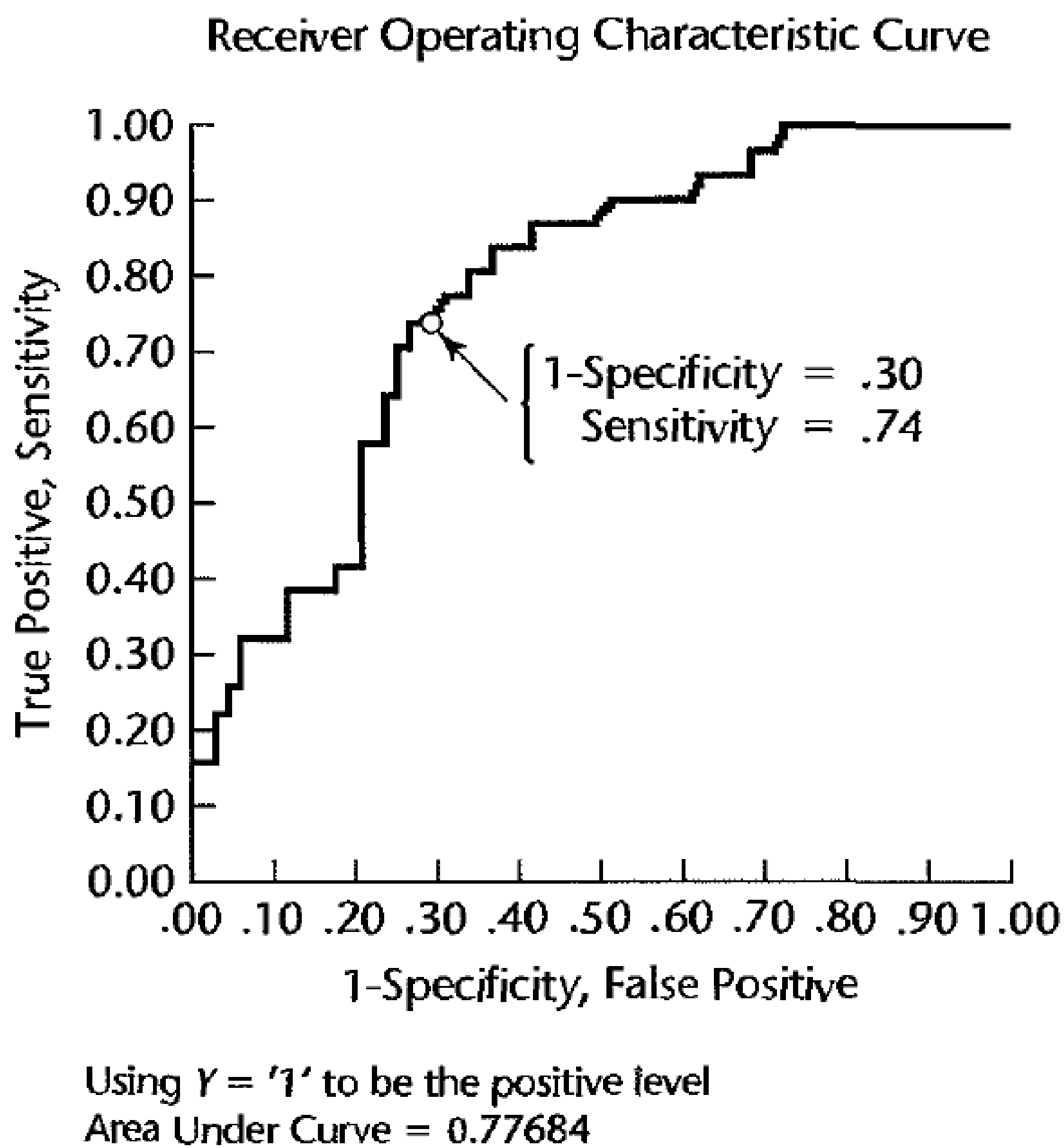
ROC curve

- Youden Index

J=Sensitivity+Specificity–1

=TPR–FPR

- Can use Youden Index for choice of cutoff value



Exponential Family and Generalized Linear Models

- Exponential family
  - Discrete distributions: Multinomial, Bernoulli, Binomial, Poisson,…
  - Continuous distributions: Gaussian, Exponential, Laplace, Gamma, Beta, Weibull,…

$$p(x|\eta) \ = \ h(x) \ \exp\left(\eta^\top \, t(x) \ - \ a(\eta)\right)$$

where, η is called "natural parameter", t(x) is related "sufficient statistic", h(x) is the "underlying measure" and a(η) is called "log normalizer", which ensures that the distribution integrates to one. Hence,

$$a(\eta) = \log \int h(x) \ \exp\left(\eta^\top t(x)\right) dx.$$

Exponential Family

- Bernoulli: let $\pi = \Pr(x = 1)$.

$$p(x|\pi) = \pi^x(1-\pi)^{1-x}.$$
$$= \exp\left\{x \log \frac{\pi}{1-\pi} + \log(1-\pi)\right\}$$

  - $\eta = \log \frac{\pi}{1-\pi}$,
  - $t(x) = x$,
  - $a(\eta) = -\log(1-\pi) = \log(1+e^\eta)$,
  - and $h(x) = 1$.

- Poisson: $p(x|\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!} = \dfrac{1}{x!}\exp\{x \log \lambda - \lambda\}$,

  - $\eta = \lambda$,
  - $t(x) = x$,
  - $a(\eta) = \lambda = e^\eta$,
  - and $h(x) = \frac{1}{x!}$.

Moments of Exponential Family

$$
\begin{aligned}
\frac{d\,a(\eta)}{d\eta} &= \frac{d}{d\eta}\left\{\log\left(\int \exp\{\eta^\top t(x)\}h(x)dx\right)\right. \\
&= \frac{\frac{d}{d\eta}\int \exp\left\{\eta^\top t(x)\right\} h(x)dx}{\int \exp\{\eta^\top t(x)\}h(x)dx} \\
&= \frac{\int t(x)h(x)\exp\{\eta^\top t(x)\}dx}{\int \exp\{\eta^\top t(x)\}h(x)dx} \\
&= \frac{\int t(x)\exp\{\eta^\top t(x)\}h(x)dx}{\exp\{-a(\eta)\}} \\
&= \int t(x)\exp\{\eta^\top t(x) - a(\eta)\}h(x)dx \\
&= \mathbb{E}\left[t(x)\right].
\end{aligned}
$$

- Likewise, it can be shown that:

$$
\frac{d^2\,a(\eta)}{d\eta^2} = \operatorname{Var}\left(t(x)\right) = \mathbb{E}\left[t(x)^2\right] - \mathbb{E}\left[t(x)\right]^2.
$$

Exponential Family

- Overdispersed exponential families
- The pdf or pmf can be written in the form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

where ϕ is the dispersion parameter and θ is the canonical parameter.

- It can be shown that

$$E(Y) = b'(\theta) = \mu$$
$$\text{var}(Y) = \phi b''(\theta) = \phi V(\mu)$$

Examples

- Example 1: $Y \sim N(\mu, \sigma^2)$.

$$\theta = \mu,\ \phi = \sigma^2,\ b(\theta) = \mu^2 / 2 = \theta^2 / 2$$

$$\Rightarrow \quad E(Y) = \mu = b^{'}(\theta) = \theta,\ Var(Y) = b^{''}(\theta)\phi = \phi = \sigma^2$$

- Example 2 (Poisson): $f(y,\theta,\phi) = \dfrac{\mu^y e^{-\mu}}{y!} = e^{y\log(\mu)-\mu-\log(y!)}$

$$\theta = \log(\mu),\ \phi = 1,\quad b(\theta) = \mu = e^\theta$$

$$\Rightarrow \qquad E(Y) = b^{'}(\theta) = e^\theta = \mu,\ Var(Y) = b^{''}(\theta)\phi = e^\theta = \mu$$

- Example 3 (Binormial) $Y \sim \dfrac{B(m,p)}{m}$ $\qquad f(y,\theta,\phi) = \begin{pmatrix} m \\ my \end{pmatrix} p^{my}(1-p)^{m-my}$

$$\theta = \log\left(\frac{p}{1-p}\right),\ \phi = \frac{1}{m},\ b(\theta) = \log\left(\frac{1}{1-p}\right) = \log\left(1 + e^\theta\right)$$

$$\Rightarrow \quad E(Y) = \mu = b^{'}(\theta) = \frac{e^\theta}{1+e^\theta} = p,\quad Var(Y) = b^{''}(\theta)\phi = \frac{1}{m}p(1-p)$$

Generalized Linear Models

- Canonical Links
- For a glm where the response follows an exponential distribution, we have

$$g(\mu_i) = g(b'(\theta_i)) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

- The canonical link is defined as

$$g = (b')^{-1}$$

$$\Rightarrow g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

- The maximum likelihood estimates are obtained

The log-likelihood for the sample $y_1, \dots, y_n$ is

$$l = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)$$

	连接函数	回归模型	分布
恒等	$x^T\beta = E(y)$	$E(y) = x^T\beta$	正态分布
对数	$x^T\beta = \ln E(y)$	$E(y) = \exp(x^T\beta)$	Poisson分布
Logit	$x^T\beta = Logit E(y)$	$E(y) = \dfrac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$	二项分布
逆	$x^T\beta = \dfrac{1}{E(y)}$	$E(y) = \dfrac{1}{x^T\beta}$	Gamma分布

The glm Function in R

- Generalized linear models can be fitted in R using the glm function, which is similar to the lm function for fitting linear models.
- The arguments to a glm call are as follows

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = glm.control(...), model = TRUE,
    method = "glm.fit", x = FALSE, y = TRUE,
    contrasts = NULL, ...)
```

- The formula is specified to glm as, e.g. `y ~ x1 + x2`

Family Argument

- The family argument takes (the name of) a family function which specfies
  - the link function
  - the variance function
- The exponential family functions available in R are

  - `binomial(link = "logit")`
  - `gaussian(link = "identity")`
  - `Gamma(link = "inverse")`
  - `inverse.gaussian(link = "1/mu^2")`
  - `poisson(link = "log")`

Extractor Functions

- The glm function returns an object of class c("glm", "lm").
- There are several glm or lm methods available for accessing/displaying components of the glm object, including:
  - `residuals()`
  - `fitted()`
  - `predict()`
  - `coef()`
  - `deviance()`
  - `formula()`
  - `summary()`