Chap. 4 Logistic Regression

For binary outcome ($Y = 0$ or 1),

$$Y \sim Bernoulli(\pi),$$

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

The logistic regression model is

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

or

$$logit(\pi(x)) = \alpha + \beta x$$

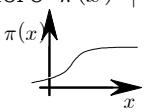
Use "logit" link for binomial Y.

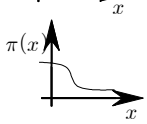
Equivalently

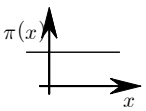
$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

Properties

- Sign of β indicates where $\pi(x) \uparrow$ or $\downarrow$ as $x \uparrow$

  $\beta > 0 : \pi(x) \uparrow 1 \ as \ x \uparrow$ 

  $\beta < 0 : \pi(x) \downarrow 0 \ as \ x \uparrow$ 

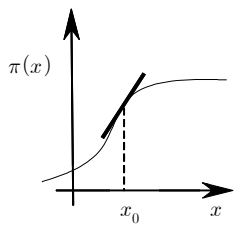
  $\beta = 0$: 

- Curve can be approximated at fixed x by straight line to describe rate of change

  eg) at x with $\pi(x) = \frac{1}{2}$, slope= $\beta(\frac{1}{2})(\frac{1}{2}) = \frac{\beta}{4}$

  at x with $\pi(x) = 0.1$ or 0.9, slope= $\beta(0.1)(0.9) = 0.09\beta$

  Steepest slope where $\pi(x) = \frac{1}{2}$



- When $\pi = \frac{1}{2}$, $\log(\frac{\pi}{1-\pi}) = \log(\frac{0.5}{0.5}) = 0 = \alpha + \beta x$

  $\Rightarrow x = -\frac{\alpha}{\beta}$ is the x value where this happens.

- $\frac{1}{\beta} \approx$ distance between x values with $\pi = 0.5$ and $\pi = 0.75$ (or 0.25)
- ML fit obtained with iterative numerical methods.

Example: Horseshoe crab Study

$n = 173$ female crabs

$Y =$ whether crab has mates (satellites) (1=Yes, 0=No)

Explanatory variables are weight, width of shell, color (light, medium light, medium, medium dark, dark) -> 1, 2, 3, 4, 5; Condition of spine.

We first consider

$$logit\ \pi(x) = \alpha + \beta x$$

where $\pi(x) \rightarrow P(Y=1)$, $x \rightarrow weight$

Crab Program

```
DATA crab;
 infile 'C:Wcrabs_SAS.dat';
 input color spine width satell weight;

  if satell>0 then y=1; if satell=0 then y=0 n=1;
  weight=weight/1000; color=color-1;
  if color=4 then dark=0; if color<4 then dark=1;

 proc genmod; /*E(Y) = β0 + β1weight */
  model y/n=weight / dist=normal link=identity;
 run;
/* Logistic regression with 'weight' (type3 option) */
❶ proc genmod;
  model y/n=weight / dist=bin link=logit type3;
 run;
❷ proc genmod;
  model y/n= /dist=bin link=logit;
 run;
/* Logistic regression with 'weight color' (type3 option) */
❸ proc genmod;
  class color;
  model y/n=weight color / dist=bin link=logit type3;
 run;
❹ proc genmod;
  model y/n=weight color / dist=bin link=logit;
 run;
❺ proc genmod;
  model y/n=weight dark / dist=bin link=logit;
 run;
❻ proc genmod;
  model y/n=weight / dist=bin link=probit;
 run;
```

❶ logistic regression with weight
❷ logistic regression without predictor
❸ logistic regression with weight and color
❹ logistic regression treating color as ordinal
❺ logistic regression treating color as whether dark(?)
❻ probit model

(Please find Crabs_sas.dat file and Input Crabs_sas.dat to SAS and running the programming.)

$$logit\ \pi(x) = \alpha + \beta x$$

where $\pi(x) \rightarrow P(Y=1)$, $x \rightarrow weight$

ML fit :

- $\hat{\text{logit}}\hat{\pi}(x) = -3.69 + 1.82x$ (or $\hat{\pi}(x) = \frac{\exp(-3.69+1.82x)}{1+\exp(-3.69+1.82x)}$)
- Estimated odds a female crab has a satellite multiplied by $e^{0.1(1.82)} = 1.2$ for each 0.1 kg increase in weight (increase by 20%).
- $\hat{\beta} > 0$ so $\hat{\pi} \uparrow$ as $x \uparrow$
- At $x = \bar{x} = 2.44$,

$$\hat{\pi}(x) = \frac{\exp(-3.69+1.82(2.44))}{1+\exp(-3.69+1.82(2.44))} = 0.676$$

- $\hat{\pi} = \frac{1}{2}$ when $x = -\frac{\hat{\alpha}}{\hat{\beta}} = -\frac{-3.69}{1.82} = 2.04$
- At $x = 2.04$, when x increases by 1-unit, $\hat{\pi}$ increases by approx. $\hat{\beta}\hat{\pi}(1-\hat{\pi}) = \frac{\hat{\beta}}{4} = 0.45$.

  However, $s = 0.58$ for weight, and 1-unit change is too large for this approximation to be good. (Actual $\hat{\pi} = 0.86$ *at* 3.04) As x increase by 0.1-unit, $\hat{\pi}$ increases by approx. $0.1\hat{\beta}\hat{\pi}(1-\hat{\pi}) = 0.045$ (Actual $\hat{\pi} = 0.547$ at 2.14)
- At $x = 5.20$ (max. value), $\hat{\pi} = 0.997$. As x increase by 0.1-unit $\hat{\pi}$ increase by approx. (0.1)(1.82)(0.997)(0.003)=0.0006. Rate of changes varies as x does.

Note

- If we assume $Y \sim$ Normal and fitted model $\mu = \alpha + \beta x$,

$$\hat{\mu} = -0.145 + 0.323x$$

  At $x = 5.2$, $\hat{\mu} = 1.53 (> 1) \rightarrow wrong$
- Alternative way to describe effect (not dependent on unit) is

$$\hat{\pi}(x_2) - \hat{\pi}(x_1)$$

such as $\hat{\pi}(UQ) - \hat{\pi}(LQ) \Rightarrow UQ : upper\ quantile$, $LQ : lower\ quantile$

ex) In logistic regression, at $x =$ weight, $LQ = 2.00$, $UQ = 2.85$

$$\hat{\pi}(2.0) = 0.48,\ \hat{\pi}(2.85) = 0.81$$

$\hat{\pi}$ increase by 0.33 (=0.81-0.48) over middle half of x value.

Inference

C.I.

95% C.I. for β is $\hat{\beta} \pm Z_{0.025}(SE)$ (Wald method)

$$1.815 \pm 1.96(0.377) = (1.08,\ 2.55)$$

95% C.I. for e^{β}, multiplication effect on odds of 1-unit increase in x, is

$$(e^{1.08},\ e^{2.55}) = (2.9,\ 12.8)$$

95% C.I. for $e^{0.1\beta}$ is

$$(e^{0.108},\ e^{0.255}) = (1.11,\ 1.29)$$

(odds increases at least 11%, at most 29%)

Note:

- For small n, safer to use likelihood-ratio C.I. than Wald C.I. (can do with LRCI option in SAS GENMOD)

  ex) LR C.I. for e^{β} is

  $$(e^{1.11},\ e^{2.60}) = (3.0,\ 13.4)$$

- For binary observation ($y = 0$ or 1), SAS(PROC GENMOD) can use model statement

  MODEL y=weight/dist=bin

  but SAS forms logit as $\log\left[\frac{P(Y=0)}{P(Y=1)}\right]$ instead of $\log\left[\frac{P(Y=1)}{P(Y=0)}\right]$ unless use "decreasing" option.

  eg) get $\log it(\hat{\pi}) = 3.69 - 1.82x$ instead of $\log it(\hat{\pi}) = -3.69 + 1.82x$

- Softeare can also construct C.I. for $\pi(x)$

  (In SAS, PROC GENMOD or PROC LOGISTIC)

  ex) At $x = 3.05$ (value for 1^{st} crab), $\hat{\pi} = 0.863$, 95% C.I. for π is

  $$(0.766,\ 0.924)$$

Significance Test

$H_0 : \beta = 0$ states that Y indep. of X (i.e., $\pi(x)$ constant)

$H_a : \beta \neq 0$

$$Z = \frac{\hat{\beta}}{SE} = \frac{1.815}{1.377} = 4.8$$

or Wald stat. $Z^2 = 23.2,\ df = 1,\ p - value < 0.0001$

Very Strang evidence that weight has positive effect on π

Likelihood-ratio test

When $\beta = 0$, $L_0 = -112.88$ (log-likelihood under H_0)

When $\beta = \hat{\beta}$, $L_1 = -97.87$

Test stat.$= -2(L_0 - L_1) = 30.0$

Under H_0, it has approx. χ^2_{df}, $df = 1 (p - value < 0.0001)$

(In PROC GENMOD, use option 'Type3')

Note: Recall for a model M,

$$\text{Deviance} = -2(L_M - L_S)$$

where L_S = log-likelihood under saturated (perfect fit) model

To compare model M_0 with a more complex model M_1,

$$\begin{aligned} LR\ stat. &= -2(L_0 - L_1) \\ &= -2(L_0 - L_S) - [-2(L_1 - L_S)] \\ &= difference\ of\ Deviances \end{aligned}$$

ex) $H_0 : \beta = 0$ in $\log it \pi(x) = \alpha + \beta x (\Rightarrow M_1)$

M_0: $logit\ \pi(x) = \alpha$

Difference of Deviance =225.76 (from ❷) -195.74 (from ❶) = 30 =LR Stat.

Multiple Logistic Regression

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

$x_1, x_2, \cdots, x_k$ can be quantitative, qualitative (using dummy variable), or both

Model form is

$$logit P(Y=1) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

or

$$\pi = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}$$

β_i =partial effect of x_i, controlling for other variables in model

e^{β_i} =conditional odds ratio between Y and x_i (1-unit change) keeping other predictors fixed

ex) Horseshoe crab data

$Y=1$ or 0 (0=no satellite)

Let x = weight, c =color (qualitative 4 categories)

$$c_1 = \begin{cases} 1 & medium\ light \\ 0 & otherwise \end{cases} \quad c_2 = \begin{cases} 1 & medium \\ 0 & otherwise \end{cases} \quad c_3 = \begin{cases} 1 & medium\ dark \\ 0 & otherwise \end{cases}$$

$c_1 = c_2 = c_3 = 0$ for dark crabs

Model

$$logit P(Y=1) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x$$

ML fit

$$logit \hat{\pi} = -4.53 + 1.27 c_1 + 1.41 c_2 + 1.08 c_3 + 1.69 x \quad (\text{from } ❸)$$

For dark crabs ($c_1 = c_2 = c_3 = 0$)

$$logit \hat{\pi} = -4.53 + 1.69 x$$

At $x = \bar{x} = 2.44$ (keep weight fixed)

$$\hat{\pi} = \frac{e^{-4.53 + 1.69(2.44)}}{1 + e^{-4.53 + 1.69(2.44)}} = 0.40$$

For medium light crabs ($c_1 = 1,\ c_2 = c_3 = 0$)

$$\log it\hat{\pi} = -4.53 + 1.27(1) + 1.69x = -3.26 + 1.69x$$

At $x = \bar{x} = 2.44,\ \hat{\pi} = 0.70$

At each weight, medium-light crabs are more likely than dark crabs to have satellites

$$\hat{\beta}_1 = 1.27,\ e^{1.27} = 3.6$$

At a given weight, estimated odds a med-light crab has satellite are 3.6 times estimated odds for dark crab ($c_1 = c_2 = c_3 = 0$)

eg) At $x = 2.44$,

$$\frac{odds\ \text{for}\ med-light}{odds\ \text{for}\ dark} = \frac{0.70/0.30}{0.40/0.60} = 3.6$$

How could you get an estimated odds ratio comparing ML to M or MD?
Compare ML ($c_1 = 1$) to M ($c_2 = 1$)

$$1.27\text{-}1.41\text{=-}0.14,\ e^{-0.14} = 0.9$$

At any given weight, estimated odds a ML crab has satellite are 0.9 times estimated odds a M crab has satellite.

Note:

- Model assumes lack of interaction between color and weight in effects on π. This implies coefficient of $x = weight$ is same for each color ($\hat{\beta}_4 = 1.69$). i.e., shape of curve for effect of x on π is same for each color.

Inference:

- Do we need color in model ?

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (Given weight (x), Y is indep. of color)

$$\begin{aligned} LRT\ statistic &= -2(l_0 - L_1) \\ &= -2[(-97.9) - (-94.3)] \\ &= difference\ of\ Deviance \\ &= 195.7 - 188.5 \\ &= 7.2 \end{aligned}$$

$df = 171 - 168 = 3,\ p - value = 0.07$

Some evidence (but not strong) of a color effect, given weight (only 22 "dark" crabs)

- There is strong evidence of weight effect ($\hat{\beta}=1.69,\ SE=0.39$). Given color, estimated odds of satellite at weight $x+1$ equal $e^{1.69}=5.4$ times estimated odds at weight x.
- Other simple models are adequate?

ex) For nominal color, color estimates

(ML, M, MD, D) ⇒ (1.27, 1.41, 1.08, 0) [from model ❸]

suggest

$$\mathrm{logit}P(Y=1)=\alpha+\beta_1 x_1+\beta_2 x_2$$

where $x_2=\begin{cases}0 & dark\\ 1 & other\ color\end{cases}$

$$\hat{\beta}_2=1.295\ (SE=0.5222)\ \Rightarrow\ Z=\frac{1.295}{0.522}=2.481,\ p-value=0.0131$$

Given weight, estimated odds of satellite for nondark crabs is $e^{1.295}=3.65$ times estimated odds for dark crabs.

Does model with 4 separate colors estimates fit better?

H_0 : Simple model (1 dummy) ⇔ $H_0:\beta_1=\beta_2=\beta_3$

H_a : More complex model (3 dummies)

$$\mathrm{logit}P(Y=1)=\alpha+\beta_1 c_1+\beta_2 c_2+\beta_3 c_3+\beta_4 x$$

$$\begin{aligned}LR\ stat. &= difference\ in\ Deviance\\ &= 189.17-188.54=0.6(df=2),\ p-value=0.741\end{aligned}$$

Simple model is adequate.

How about model allowing interaction?

$$\mathrm{logit}P(Y=1)=\alpha+\beta_1 c_1+\beta_2 c_2+\beta_3 c_3+\beta_4 x+\beta_5 c_1 x+\beta_6 c_2 x+\beta_7 c_3 x$$

color	Weight effect	
dark	β_4	$(c_1=c_2=c_3=0)$
med-light	$\beta_4+\beta_5$	$(c_1=1)$
medium	$\beta_4+\beta_6$	$(c_2=1)$
med-dark	$\beta_4+\beta_7$	$(c_3=1)$

H_0 : no interaction $(\beta_5 = \beta_6 = \beta_7 = 0)$

$$LR\ stat. = -2(L_0 - L_1) = 6.88,\ df = 3,\ p-value = 0.08$$

Weak evidence of interaction.
For easier interpretation, use simpler model.

Ordinal factors

Models with dummy variables treat color as qualitative (nominal).
To treat color as quantitative, assign scores such as (1, 2, 3, 4) and model trend

$$\mathrm{logit}\pi = \alpha + \beta_1 x_1 + \beta_2 x_2 \qquad (x_1 : weight,\ x_2 : color)$$

ML fit (model ❹)

$$\mathrm{logit}\pi = -2.03 + 1.65x_1 - 0.51x_2$$

and SE for β_1, β_2 $\Rightarrow$ (0.38), (0.22)

$\hat{\pi}\downarrow$ as color $\uparrow$ (more dark), controlling for weight $e^{-0.51} = 0.60$ which is estimated odds ratio for 1-category increase in darkness

Does model treatng color as nominal fit better?
H_0 : Simpler (ordinal) model holds
H_a : More complex (nominal) model holds

$$LR\ stat. = -2(L_0 - L_1) = diff.\ of\ Deviances$$
$$= 190.27 - 188.54 = 1.7,\ df = 2,\ p-value = 0.427$$

Do not reject H_0 (Simpler model is adequate)

Qualitative predictors

ex) FL death penalty

Victim's Race	Suspect's Race	Death Penalty Yes	Death Penalty No	n
Black	Black	4	139	143
	White	0	16	16
White	Black	11	37	48
	White	53	44	467

$\pi = P(Y = Yes)$

$\nu = \begin{cases} 1 & Victim's\ black \\ 0 & Victim's\ white \end{cases}$ $\quad d = \begin{cases} 1 & Suspect's\ black \\ 0 & Suspect's\ white \end{cases}$

Model:

$$logit\,\pi = \alpha + \beta_1 d + \beta_2 \nu$$

has ML fit

$$logit\,\hat{\pi} = -2.06 + 0.87d - 2.40\nu$$

eg) controlling for victim's race, estimated odds of death penalty for black suspect's equal $e^{0.87} = 2.38$ times estimated odds for white suspect's

95% Wald C.I.

$$e^{0.87 \pm 1.96(0.367)} = (1.16,\ 4.89)$$

95% LR C.I. (from SAS ouput)

$$(e^{0.114}, e^{1.563}) = (1.123,\ 4.773)$$

Note:

● Lack of interaction term means estimated odds ratio between Y and d same at each level of ν $(e^{0.87} = 2.38)$

ν same at each level of d $(e^{-2.40} = 0.09)(e^{2.4} = \frac{1}{0.09} = 11.1)$

i.e., Controlling for d, estimated odds of death penalty when $\nu =$ white were 11.1 times estimated odds when $\nu =$ black

(homogeneous association) means same odds ratio at each level of the other variable.

● $H_0 : \beta_1 = 0$ (Y conditional independence of d given ν

$H_a : \beta_1 \neq 0$

Wald Test

$$Z = \frac{\hat{\beta}}{SE} = \frac{0.868}{0.367} = 2.36$$

Wald stat.= $Z^2 = 5.59,\ df = 1,\ p-value = 0.018$

Evidence that death penalty was more likely for black suspect controlling for ν

Likelihood-ratio test

Test of $H_0 : \beta_1 = 0$. Compares model

$$H_0 : \log it(\pi) = \alpha + \beta_2 \nu$$
$$H_a : \log it(\pi) = \alpha + \beta_1 d + \beta_2 \nu$$

$$LR\ stat. = -2(L_0 - L_1) = -2(-211.99 - (-209.48)) = 5.0$$
$$= difference\ of\ Deviances$$
$$= 5.39 - 0.38 = 5.01,\ df = 1 (p-value = 0.025)$$

```
data death;
 input v $ d $ p total @@;
cards;
b b 4 143  b w  0  16
w b 11 48  w w 53 467
run;

❶ proc genmod data=death;
 class v d;
 model p/total = d v / dist=bin link=logit lrci type3;
run;

❷ proc genmod data=death;
 class v d;
 model p/total = v / dist=bin link=logit lrci;
run;
```

OUTPUT:

❶

The GENMOD Procedure
Model Information

Data Set	WORK.DEATH
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	p
Response Variable (Trials)	total
Number of Observations Read	4
Number of Observations Used	4
Number of Events	68
Number of Trials	674

Class Level Information

Class	Levels	Values
v	2	b w
d	2	b w

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.3798	0.3798
Scaled Deviance	1	0.3798	0.3798
Pearson Chi-Square	1	0.1978	0.1978

Scaled Pearson X2	1	0.1978	0.1978
Log Likelihood		-209.4783	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.0595	0.1458	-2.3565	-1.7836	199.40	<.0001
d	b	1	0.8678	0.3671	0.1140	1.5633	5.59	0.0181
d	w	0	0.0000	0.0000	0.0000	0.0000	.	.
v	b	1	-2.4044	0.6006	-3.7175	-1.3068	16.03	<.0001
v	w	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

The GENMOD Procedure
LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
d	1	5.01	0.0251
v	1	20.35	<.0001

❷

The GENMOD Procedure
Model Information

Data Set	WORK.DEATH
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	p
Response Variable (Trials)	total
Number of Observations Read	4
Number of Observations Used	4
Number of Events	68
Number of Trials	674

Class Level Information

Class	Levels	Values
v	2	b w
d	2	b w

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	5.3940	2.6970
Scaled Deviance	2	5.3940	2.6970
Pearson Chi-Square	2	5.8109	2.9054
Scaled Pearson X2	2	5.8109	2.9054
Log Likelihood		-211.9854	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.9526	0.1336	-2.2234	-1.6989	213.68	<.0001
v	b	1	-1.7045	0.5237	-2.9072	-0.7995	10.59	0.0011
v	w	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Exercise

- Conduct Wald, LR tests of $H_0 : \beta_2 = 0$
- Get point and interval estimate of odds ratio for effect of victim's race.

controlling for d

Note:

● A common application for logistic regression having multiple 2×2 tables is multi-center clinical trials

Center	Treatment	Response S	Response F
1	1		
	2		
2	1		
	2		
⋮	⋮		
K	1		
	2		

$$logitP(Y=1)=\alpha+\beta_1c_1+\beta_2c_2+\beta_{K-1}c_{K-1}+\beta x$$

where c_i: clinical center

Assumes odds ratio = e^{β} in each table

A model like this with several dummy variables for a factor is often expressed as

$$logitP(Y=1)=\alpha+\beta_i^c+\beta x \quad (\beta_K^c=0),$$

where β_i^c is effect for center i (relative to last center)

To test $H_0:\beta=0$ about treatment effect for several 2×2 tables, could use (a) likelihood-ratio test, (b) Wald test, (c) Cochran-Mantel-Haenszel test (p114), and (d) small-sample generalization of Fisher's exact test (p158-159)

Example: Exercise 4.19 in 2nd Edition

A sample of subjects were asked their opinion about current laws legalization abortion (support, oppose). For the explanatory variables gender (female, male), religious affiliation (Protestant, Catholic, Jewish), and political party affiliation (Democrat, Republican, Independent), the model for the probability π of supporting legalized abortion is given by:

Y= support current abortion law (1=yes, 0=no)

$$\log itP(Y=1)=\alpha+\beta_h^G+\beta_i^R+\beta_j^P+\beta x$$

where β_h^G is for gender, β_i^R is for religion, and β_j^P is for party affiliation.

For religion (Protestant, Catholic, Jewish),

$$\hat{\beta}_1^R = -0.57,\ \hat{\beta}_2^R = -0.66,\ \hat{\beta}_3^R = 0.00$$

β_i^R represents terms

$$\hat{\beta}_1^R r_1 + \hat{\beta}_2^R r_2 = -0.57 r_1 - 0.66 r_2$$

where $r_1 = \begin{cases} 1, & \text{Protestant} \\ 0, & otherwise \end{cases}$, $r_2 = \begin{cases} 1, & Catholic \\ 0, & otherwise \end{cases}$

Probit model

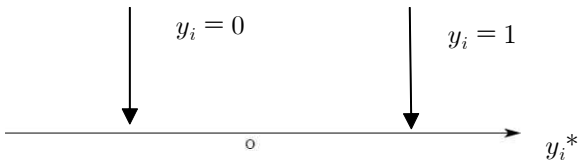
● model : $\Phi^{-1}[P(Y_i=1)]=\alpha+\beta x_i$, where Φ^{-1} is the inverse function of standard normal cdf.

● motivation

- Latent variable model with threshold.

Suppose there is underlying normal response y_i^* and we observe

$\begin{cases} y_i=0 & \text{if } y_i^* \le 0 \\ y_i=1 & \text{if } y_i^* > 0 \end{cases}$ (threshold).



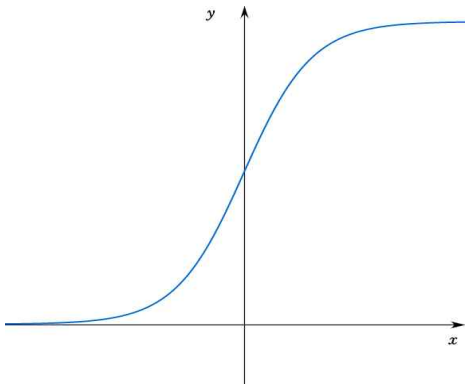
Suppose $y_i^* = \alpha+\beta x_i+\epsilon_i,\ \epsilon_i \sim N(0,1)$

Then

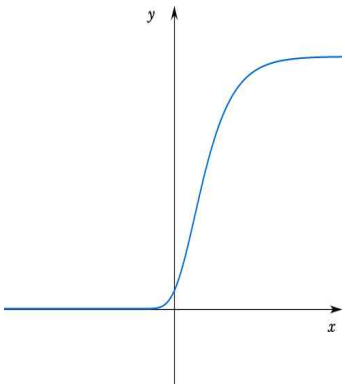
$$\begin{aligned} P(Y_i=1) &= P(Y_i^*>0) \\ &= P(\alpha+\beta x_i+\epsilon_i>0) \\ &= P(\epsilon_i > -(\alpha+\beta x_i)) \\ &= 1-\Phi(-\alpha-\beta x_i)=\Phi(\alpha+\beta x_i) \end{aligned}$$

Thus β = change in $E(y_i^*)$ for 1 unit increase in x.

Complementary Log-Log models



logit



toxicology
(complementary log-log)

- Logit and probit models have shape of cdf's for symmetric pdf's.
- If instead of $\pi = P(Y=1)$ approaches 1 at a different rate than it approaches 0, alternative link functions may be better.
- For example, if π approaches 1 sharply, useful models uses complementary log-log link,

$$\log[-\log(1-\pi(x))] = \alpha + \beta x$$

$$\Leftrightarrow \pi(x) = 1 - \exp[-\exp(\alpha + \beta x)]$$

To interpret, note

$$\log[-\log(1-\pi(x+1))] - \log[-\log(1-\pi(x))] = \beta$$

or

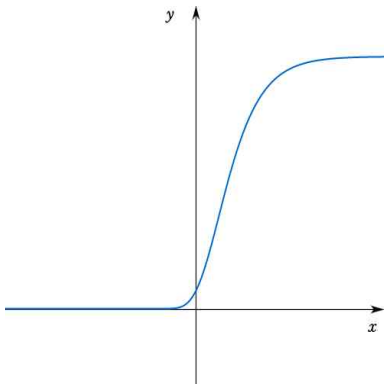
$$\frac{\log[(1-\pi(x+1)]}{\log(1-\pi(x))} = e^{\beta}$$

or

$$1-\pi(x+1) = (1-\pi(x))^{e^{\beta}}$$

i.e. $P(failure)$ at $x+1$ is power e^{β} of $P(failure)$ at x

Log-Log link



The log-log link give model $\log[-\log(\pi(x))] = \alpha + \beta x_i$

and $\pi(x) = \exp[-\exp(\alpha + \beta x)]$: cdf of extreme value (Gumbel) distribution.

Gumbel distribution is

$$F(x) = \exp[-\exp(-\frac{x-a}{b})],\ b > 0$$

● Note

log-log model for $P(success)$=complementary log-log model for $P(failure)$

● Example Bliss(1935)

# of beetles killed after exposure to gaseous carbon disulfide at various concentrations.

$x = \log(dose)$	# of beetles	# of killed	proportion killed
1.69	59	6	.10
1.72	60	13	.22
1.75	62	18	.29
1.78	56	28	.50
1.81	63	52	.83
1.84	59	53	.90
1.86	62	61	.98
1.88	60	60	1.00

See beetle.sas.

● Model

$link(\pi(x)) = \alpha + \beta x$, where $\pi(x)$ = prob. of death at $\log(dose) = x$

Link	Deviance $(df = 6)$
Logit	11.1
Probit	9.98
$c-\log-\log$	3.51
$\log-\log$	27.57

For $c-\log-\log$ link,

$$\log[-\log(1-\pi_i)] = \alpha + \beta x_i$$

$$1-\hat{\pi}_i = \exp[-\exp(\hat{\alpha}+\hat{\beta}x_i)]$$

$$= \exp[-\exp(-39.52+22.01x_i)] \quad \cdots \text{from 3}$$

$$(1-\pi(x+0.1)) = (1-\pi(x))^{e^{22.01\times 0.1}} = (1-\pi(x))^{9.03}$$

Estimated $prob(survival)$ at $x+0.1$ is the power $e^{22.01\times 0.1} = 9.03$ of $prob(survival)$ at x

$\log(dose)$	$1-\hat{\pi}_i$
1.7	$\exp[-\exp(-39.52+22.01(1.7))] = 0.8851$
1.8	$(0.8851)^{9.03} = 0.3319$
1.9	$(0.3319)^{9.03} = 0.00005$

Probit model

$$\Phi^{-1}(\hat{\pi}_i) = -34.96 + 19.74x_i$$