Generalized Linear Models

Example 1

- We would like to predict *$admit$* given *gre* and *gpa*

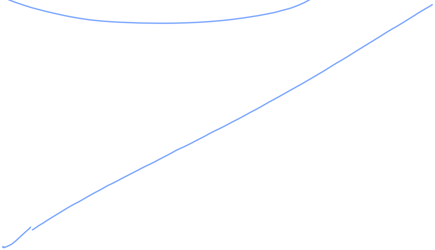
response

predictors

admit	gre	gpa
0	380	3.61
1	660	3.67
1	800	4.00
1	640	3.19
0	520	2.93
1	760	3.00
1	560	2.98
0	400	3.08
1	540	3.39
0	700	3.92

- Can we use linear model here?

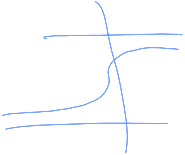
$$\underline{admit} = \boxed{\beta_0 + \beta_1 \cdot \underline{gre} + \beta_2 \cdot \underline{gpa}}$$



- Multiple Linear Regression cannot handle binary/categorical response

- Linear Model

$$admit = \beta_0 + \beta_1 \cdot gre + \beta_2 \cdot gpa$$



- Logistic Regression

$$P(admit = 1) = \frac{1}{1 + exp(-\beta_0 - \beta_1 \cdot gre - \beta_2 \cdot gpa)}$$

- When fit the logistic regression on the data we obtain:

$$P(admit = 1) = \frac{1}{1 + exp(-\underset{=}{0.73} - \underset{=}{0.02} \cdot \underset{\underset{600}{\downarrow}}{gre} + \underset{=}{3.57} \cdot \underset{\underset{4.0}{\downarrow}}{gpa})}$$

GRE = 600

GPA = 4.0

- For example, with a student having 380 GRE and 3.61 gpa, the model will predict

$$P(admit = 1) = \frac{1}{1 + exp(-0.73 - 0.02 \cdot 600 + 3.57 \cdot 4.0)} \boxed{= 0.01}$$

- This means that the chance of the student being admitted is 0.01 or the student will not be admitted by the model prediction.

Logistic Regression

$$E(y) = 0 \cdot P(y=0) + 1 \cdot P(y=1)$$

$$= \boxed{P(y=1)}$$

- Suppose the response y can only takes two values $\underline{0}$ and $\underline{1}$. The logistic regression models the probability of $y = 1$ as follows.

$$E(y) = \pi = \underline{P(y = 1)} = \frac{1}{1 + exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2)}$$

or, equivalently $\quad = g_1(\pi)$

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$g_2(\pi) = \pi = E(y) = \underline{\beta_0 + \beta_1 X_1 + \beta_2 X_2} \quad (MLR)$$

Generalized Linear Model

- The GLM models $\mu = E(y)$ as follows.

$$g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x'\beta$$

  where y is assumed to follow an exponential distribution family.

- Exponential distribution family includes all the basic distribution such as normal distribution, binomial distribution, Poisson distribution...
- $g(\mu)$ is called the canonical link function
- For logistic regression, the link function is a logit function

$$g(x) = \ln\left(\frac{x}{1-x}\right)$$

Some GLMs

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = x'\beta$$

Distribution	Canonical Link Function	Mathematical Form
Normal	Identity	$g(\mu) = \mu$
Binomial	Logit	$g(\pi) = \ln[\pi/(1-\pi)]$
Poisson	Natural log	$g(\mu) = \ln \mu$
Gamma	Inverse	$g(\mu) = 1/\mu$
Inverse Gaussian	Squared inverse	$g(\mu) = 1/\mu^2$

Example 2

A statistician uses logistic regression to model a probability of success of a random variable. The estimated parameters for the intercepts and two predictors are $\hat{\beta}_0 = 0.02$, $\hat{\beta}_1 = -0.4$, and $\hat{\beta}_2 = 0.3$. Calculate the predicted probability of success at $x_1 = 1$ and $x_2 = 1$.

$$p\left(\frac{\pi}{1-\pi}\right) = .02 - .4x_1 + .3x_2$$

$$\pi = \frac{1}{1 + e^{-.02 + .4x_1 - .3x_2}} = \boxed{.48}$$

Example 3

$$\pi = \frac{1}{1+e^{-\beta_0-\beta_1 x_1}}$$

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0+\beta_1 x_1$$

A statistician uses logistic regression to model a probability of success of a random variable. You are given

- There is one predictors and an intercept in the model
- The estimates of success at $x = 4$ and $x = 6$ are 0.8 and 0.9, respectively.

Calculate $\hat{\beta}_1$ the estimated slope of the model.

$x = 4, \Rightarrow \pi = .8 \Rightarrow$

$x = 6 \Rightarrow \pi = .9 \Rightarrow$

$$\ln\left(\frac{.8}{.2}\right) = \beta_0 + \beta_1 \cdot 4$$

$$\ln\left(\frac{.9}{.1}\right) = \beta_0 + \beta_1 \cdot 6$$

$$\ln 4 = \beta_0 + \beta_1 4$$

$$\ln 9 = \beta_0 + \beta_1 \cdot 6$$

$$\ln 4 - \ln 9 = -2\beta_1$$

$$\Rightarrow \beta_1 = \frac{\ln 4 - \ln 9}{-2} = \boxed{.405}$$

Example 4

You are given the following for a fitted GLM. Calculate the modeled probability of an Urban driver having an accident.

Response variable	Occurrence of Accidents
Response distribution	Binomial → ①
Link	Logit → ②

Parameter	df	$\hat{\beta}$		se
Intercept	1	−2.358	$\hat{\beta}_0$	0.048
x_1 Area	2			
Suburban	0	0.000		
Urban	1	0.905	$\hat{\beta}_1$	0.062
Rural	1	−1.129		0.151

$$\ln\left(\frac{\pi}{1-\pi}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$$ ①

$$\pi = \frac{1}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1}}$$ ②

$$= \frac{1}{1 + e^{2.358 - .905}} = \boxed{.1895}$$

Odd of an event

$$\ln\left(\boxed{\frac{\pi}{1-\pi}}\right) = \beta_0 + \beta_1 x_1 + \cdots$$

$\ln($odd of $-1 = 1)$

- The odds of an event A is the ratio of the probability that A occurs to the probability that A does not occur.
- The odd of tossing an coin and see Tail is 1:1
- The odd of rolling a die and see $\underline{\text{number}}$ 6 is $\frac{1/6}{5/6} = \underline{1:5}$

- Logistic regression in terms of Odd

$$\ln(\text{Odd of Success}) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Possion Regression and Other link functions

- Poisson Regression are used when the response are count variables, for example: the number of claims of a customer...
- The response is assume to follow a Poisson distribution and the link function used is a log link function, ln.

$$\underset{\overset{\|}{g(\mu)}}{\ln(\mu)} = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}_{}$$

Example 5

You are given the following when fitted a GLM model. Calculate the predicted Y value for a female with age of 22.

Response variable	Y
Response distribution	Poisson
Link	log
AIC	221.254

Parameter	$\hat{\beta}$	s.e.$(\hat{\beta})$
Intercept	5.421	0.228
x_1 Gender		
Male	0.000	0.000
Female	−0.557	0.217
x_2 Age	0.107	0.002

$$\log(y) = 5.421 - .557 + .107 \times 22$$

$$= 7.218$$

$$\Rightarrow y = e^{7.218}$$

$$= 1363.759$$

Example 6

You are given the following GLM output. Calculate the predicted premium for an insured in Risk Group 2 with Vehicle Symbol 2.

	Response variable	Pure Premium		
	Response distribution	Gamma		
	Link	log		
	Parameter	df	$\hat{\beta}$	
	Intercept	1	4.78	$\hat{\beta}_0$
+1	Risk Group	2		
	Group 1	0	0.00	
	Group 2	1	−0.20	$\hat{\beta}_1$
+2	Group 3	1	−0.35	
	Vehicle Symbol	1		
	Symbol 1	0	0.00	
	Symbol 2	1	0.42	$\hat{\beta}_2$

$$\log(y) = 4.78 - .2 + .42$$

$$= 5$$

$$\Rightarrow y = e^{5}$$

$$= 148.41$$

Example 7

You are given the following output of an GLM. Calculate the probability of a policy with 5 years of tenure that experienced at a 10% prior rate increase and has 100,000 in amount of insurance will retain into the next policy term.

Response variable		retention	
Response distribution		binomial	
Link		square root	
Pseudo R^2		0.6521	
Parameter	df	$\hat{\beta}$	
Intercept	1	0.6102	$\hat{\beta}_0$
x_1 Tenure			
< 5 years	0	0.0000	
≥ 5 years	1	0.1320	$\hat{\beta}_1$
x_2 Prior Rate Change			
$< 0\%$	1	0.0160	
[0%,10%]	0	0.0000	$\hat{\beta}_2$
$> 10\%$	1	-0.0920	
x_3 Amount of Insurance (000's)	1	0.0015	$\hat{\beta}_3$

$$\sqrt{y} = .6102 + .1320 + 0 + .0015(100)$$

$$\sqrt{y} = .8922$$

$$y = .8922^2 = \boxed{.796}$$