European Doctoral School of Demography 2018-19
EDSD 220 - Statistical Demography

Logistic Regression

Giancarlo Camarda
Institut National d'Etudes Démographiques

December 2018
University of Southern Denmark, Odense

Example 1: Donner Party

- Dataset about 69 American pioneers who set out for California in a wagon train in 1846
- The group was decimated by spending a cold winter in the Sierra Nevada

Age	Sex	Survival No	Survival Yes	Total
6-14	F	0	10	10
	M	2	10	12
15-24	F	0	6	6
	M	3	3	6
25-29	F	1	1	2
	M	7	2	9
30+	F	5	3	8
	M	10	6	16
Total		28	41	69

- Our aim is to understand the effect of age and sex on the probability of surviving such harsh experience

Example 2: Fertility Data

- Dataset with aggregate info on 1607 currently married and fecund women in Fiji in 1975

Age	Education	Desires More Children?	Contraceptive Use No	Contraceptive Use Yes	Total
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25-29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30-39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40-49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

- Here current use of contraception is the response and age, education and desire for more children are the as covariates
- Contraceptive use as binary response

Bernoulli Distribution

- Binary response: Y_i with

$$\begin{aligned} P[Y_i = 1] &= \pi_i \\ P[Y_i = 0] &= 1 - \pi_i \end{aligned} \quad \Rightarrow Y_i \sim \text{Bernoulli with parameter } \pi_i$$

$$P[Y_i = y_i] = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad \text{for } y_i = 0, 1$$

- Expected values:
  $\mu_i = E[Y_i|\boldsymbol{X}_i = x_i] = P[Y_i = 1|\boldsymbol{X}_i = x_i] = \pi_i \in [0, 1]$
- Variance:
  $\sigma_i^2 = var[Y_i|\boldsymbol{X}_i = x_i] = \pi_i(1-\pi_i)$ (non constant)
- Both μ_i and σ_i^2 depend on π_i: factors affecting probability alter both mean and variance of the observations

Binomial Distribution

- We classify units according to factors into k groups: all individuals in a group have identical values of all covariates
- n_i : number of observations in group i
- y_i : number of units have the attribute of interest (e.g. use contraceptive, surviving) in group i
- y_i is a realization of a random variable Y_i that takes values $0, 1, \dots, n_i$
- If n_i are independent with the same probability π_i, then Y_i is a Binomial with parameters π_i and n_i:

$$P[Y_i = y_i] = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \quad \text{for } y_i = 0, 1, \dots, n_i$$

  - $E(Y_i) = \mu_i = n_i \pi_i$
  - $var(Y_i) = \sigma_i^2 = n_i \pi_i (1 - \pi_i)$

The Logit Transformation

- Probabilities π_i depend on observed covariates $\boldsymbol{x}_i$
- Simplest approach: $\pi_i = \boldsymbol{x}_i' \boldsymbol{\beta}$
- Problem: $\pi_i \in [0, 1]$, but $\boldsymbol{x}_i' \boldsymbol{\beta} \in [-\infty, +\infty]$
- Simple solution: transform π_i
  - move from probability π_i to the odds:

$$\text{odds}_i = \frac{\pi_i}{1 - \pi_i} \qquad \in [0, +\infty]$$

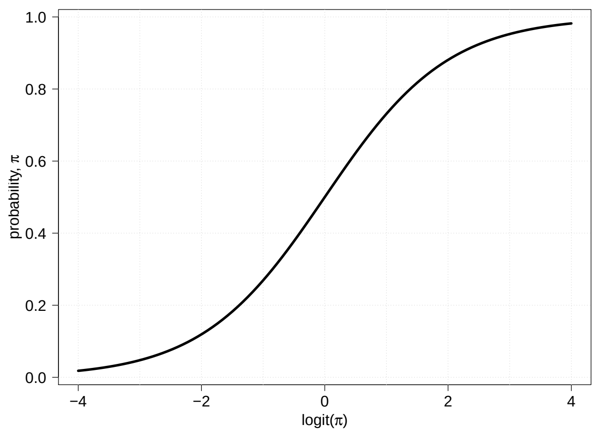
  - take the logarithm of the odds (*logit* of π_i):

$$\eta_i = \text{logit}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} \qquad \in [-\infty, +\infty]$$

- Logits map probabilities from the range (0,1) to the entire real line
- Logits may be defined in terms of the binomial mean

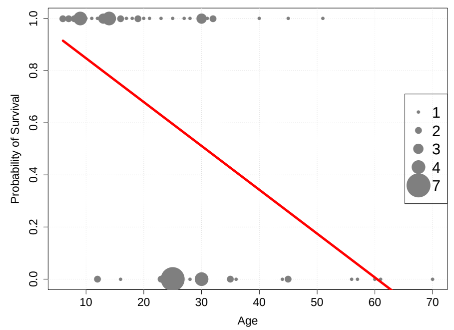
$$\eta_i = \ln \frac{\mu_i}{1 - \mu_i} = \ln \frac{n_i \pi_i}{1 - n_i \pi_i}$$

Looking at the Logit



Donner Party data using a Linear Model

- Let's estimate the *linear probability model*: $\pi_i = \boldsymbol{x}_i' \boldsymbol{\beta}$
- $\sim$ Linear Model on our Donner Party data using age as only covariate



- $\hat{\pi}_i \notin [0, 1]$

Logit on our data

- Fertility Data
  - 507 among 1607 women use contraception
  - Probability: $507/1607 = 0.316$
  - Odds: $507/1100 = 0.461$
  - Non-users outnumber users roughly two to one
  - Logit: $\ln(0.461) = -0.775$
- Donner Party Data
  - 41 among 69 pioneers survived
  - Probability: $41/69 = 0.594$
  - Odds: $41/28 = 1.464$
  - Survivors are about one and half times larger than deaths
  - Logit: $\ln(1.464) = 0.381$

The Logistic Regression Model

- We have k independent observations $y_1, \dots, y_k$
- i-th observation can be treated as a realization of a random variable Y_i
- Which distribution? (*stochastic structure*)

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

- What type of relationship? (*systematic structure*)

$$\text{logit}(\pi_i) = \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}$$

- η_i is called linear predictor

Probabilities, Odds and Log-Odds

- β_j represents the change in the logit of the probability associated with a unit change in the j-th covariate holding all other covariate constant
- Exponentiating the linear predictor:

$$\exp \eta_i = \frac{\pi_i}{1 - \pi_i} = \exp\{\boldsymbol{x}_i' \boldsymbol{\beta}\}$$

- Multiplicative model for the odds: if we were to change the j-th covariate by one unit (holding all other constant), we would multiply the odds by $\exp\{\beta_j\}$
- $\exp\{\beta_j\}$ represents an odds ratio
- Solving the logit for the probability π_i we obtain the *antilogit*:

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}}}$$

Logistic Regression - Fitting via MLE

- Likelihood:

$$L(\boldsymbol{\beta}) = \prod_i \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

- Log-likelihood:

$$l(\boldsymbol{\beta}) = \sum_i \{y_i \ln(\pi_i) + (n_i - y_i) \ln(1 - \pi_i)\}$$

  where $\text{logit}(\pi_i) = \boldsymbol{x}_i' \boldsymbol{\beta}$
- System of equation $\frac{\partial l}{\partial \beta_j} = 0 \quad \Rightarrow$ no closed-form solution
- Numerical optimization via iteratively weighted least-squares (IWLS)

Logistic Regression in R

- In R we use:

```
glm(y ~ x1 + x2 + ..., data, family=binomial(link=logit))
```

where `data` and therefore `y`, `x1`, `x2`, ... can be provided in two ways:

1. aggregate/tabular format: the response is a two-column matrix it is assumed that the first column holds the number of successes and the second holds the number of failures for each trial. Consequently, covariates are provided for each combination of covariates.
2. individual format: the response is a logical vector (or a two-level factor) and each row represent a specific individual

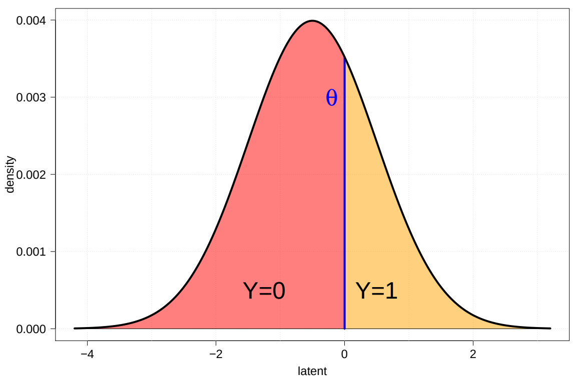
Other Choices of Link

- Any transformation that maps probabilities into the real line could be used

$$\begin{array}{ccc} \pi_i = F(\eta_i) & \Rightarrow & \eta_i = F^{-1}(\pi_i) \\ 0 < \pi_i < 1 & & -\infty < \eta_i < +\infty \end{array}$$

- We could use a *latent* variable formulation. Let's assume:
  - Y_i : (binary) manifest response
  - Y_i^* : (continuous) latent response
  - $\pi_i = P[Y_i = 1] = P[Y_i^* > \theta]$
  - $\theta = 0$
  - $\mathrm{sd}(Y_i^*) = 1$

Latent Variable and Manifest Response



- Y_i : surviving / use contraception
- Y_i^* : health condition or vitality / attitude toward contraceptive

Introducing covariates

- In a regression setting, outcomes depends on covariates

$$Y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + U_i = \eta_i + U_i$$

where U_i is an error term, note necessarily normally distributed

- Let's derive the probability of observing a positive outcome:

$$\begin{aligned} \pi_i &= P[Y_i > 0] \\ &= P[U_i > -\eta_i] \\ &= 1 - F(-\eta_i) \end{aligned}$$

- If distribution of U_i is symmetric about zero,

$$F(u) = 1 - F(-u) \quad \Rightarrow \quad \pi_i = F(\eta_i)$$

Three possible links

1. Probit:

$$U_i \sim N(0,1) \quad \Rightarrow \quad \pi_i = \Phi(\eta_i) \quad \Rightarrow \quad \eta_i = \Phi^{-1}(\pi_i)$$

Φ^{-1} have no closed form

2. Logistic

$$U_i \sim \text{Logistic} \quad \Rightarrow \quad \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \quad \Rightarrow \quad \eta_i = \ln \frac{\pi_i}{1+\pi_i}$$

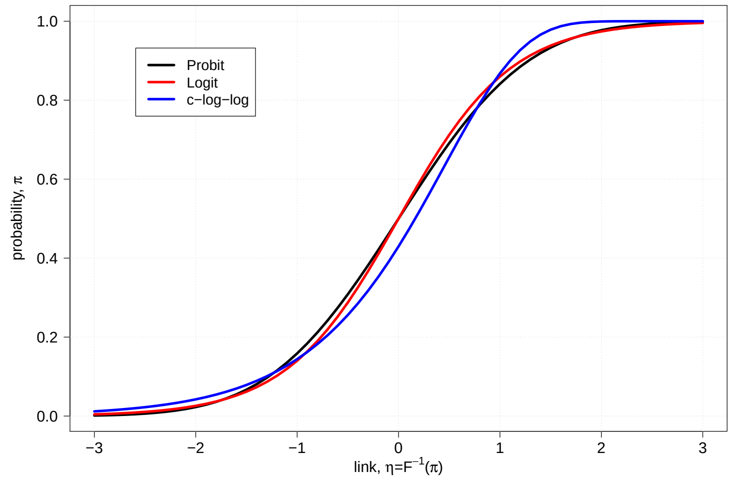
Symmetric around 0, heavier tail compared to Normal

3. Complementary Log-Log

$$U_i \sim \text{Extreme-value} \quad \Rightarrow \quad \pi_i = 1 - e^{-e^{\eta_i}} \quad \Rightarrow \quad \eta_i = \ln(-ln(1-\pi_i))$$

Useful in Discrete Time Models

Looking at the link functions



Goodness of Fit Statistics - Deviance

- A measure of discrepancy between observed and fitted values is the deviance statistic:

$$D = 2\sum_i \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right) \right\}$$

- It is twice a sum of “observed times log of observed over expected”, where the sum is over both successes and failures
- With grouped data, the distribution of the deviance statistic as the group sizes $n_i \to \infty$ for all i, converges to a chi-squared distribution with $n - p$ d.f.

Goodness of Fit Statistics - Pearson

- Alternatively, one can use Pearson’s chi-squared statistic:

$$\chi^2_p = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

- Each term in the sum is the squared difference between observed (y_i) and fitted values $(\hat{\mu}_i)$, divided by the variance
- χ^2_p Asymptotically equivalent to the deviance

Assessing the Logistic Regression: Overall effect

- Comparing nested models using deviance values
  - null hypothesis:

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0$$

  - alternative hypothesis:

$$H_A : \text{ larger model valid}$$

  - test statistics:

$$W = D_{q+1} - D_{p+1} = -2 \ln \frac{L(\boldsymbol{\mu}_{q+1})}{L(\boldsymbol{\mu}_{p+1})}, \quad \text{if } H_0 \text{ true } : W \sim \chi^2_{p-q}$$

  where where χ^2_{p-q} is the chi-squared distribution with $p - q$ degrees of freedom and D_r is the deviance for the model with r parameters

Assessing the Logistic Regression: Partial effect

- Is the covariate x_j statistically related to the response y, after controlling for the other covariates?
  - null hypothesis:

$$H_0 : \beta_j = 0$$

  - alternative hypothesis:

$$H_A : \beta_j \neq 0$$

  - test statistics:

$$z = \frac{\hat{\beta}_j}{\hat{se}[\hat{\beta}_j]}$$

  - Wald statistics z^2, if H_0 true: $z^2 \sim \chi^2_1$

Assessing the Logistic Regression: Residuals

- Discrepancy between observed y_i and fitted $\hat{y}_i = \hat{\mu}_i$
- In Linear Model: $\hat{\epsilon} = y_i - \hat{y}_i$
- More general version:
  - Pearson residuals: $r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}[\hat{\mu}_i]}}$
  - Deviance residuals: $r_{i,D} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$, with $D = \sum_i d_i$
  - standardising: $r'_{i,\cdot} = \frac{r_{i,\cdot}}{\sqrt{1-h_{ii}}}$
- What to check?
  - random noise when plotted against linear predictor
  - standardized Pearson/Deviance residuals should be approximately normal for large n_i

The Logit as GLM

	Model			
	Linear	Logit	Log-linear	General (GLM)
$Y_i \sim$	$N(\mu_i, \sigma^2)$	$B(n_i, \pi_i)$	$P(n_i\lambda_i)$	exponential family$(\boldsymbol{\theta}, \phi)$
$E(Y_i\|\boldsymbol{X}_i) =$	μ_i	$\mu_i = n_i\pi_i$	$\mu_i = n_i\lambda_i$	$b'(\boldsymbol{\theta})$
$var(Y_i\|\boldsymbol{X}_i) =$	σ^2	$n_i\pi_i(1-\pi_i)$	$n_i\lambda_i$	$b''(\boldsymbol{\theta})a(\phi)$
$\eta_i =$	$\sum \boldsymbol{x}_i\boldsymbol{\beta}$	$\sum \boldsymbol{x}_i\boldsymbol{\beta}$	$\sum \boldsymbol{x}_i\boldsymbol{\beta}$	$\sum \boldsymbol{x}_i\boldsymbol{\beta}$
$\eta_i = g(\mu_i) =$	μ_i	$\ln \frac{\mu_i}{1-\mu_i}$	$\ln \mu_i$	continuous differentiable function

- Stochastic component
- Systematic component
- Link function (canonical)