10. Applications of nonlinear estimation

- fitting distributions
- generalized linear model regression
  - log-linear model
  - probit model
  - logit model
  - Gumbel model
  - complementary log-log

Fitting distributions

fitting distribution is an example of maximum likelihood estimation:

- the user has an assumption that y obeys a certain distribution
- the goal is to estimate the pdf/pmf parameter from i.i.d. samples $\{y_i\}_{i=1}^N$

for example,

- y is Poisson(λ): $\hat{\lambda}_{\mathrm{mle}} = (1/N)\sum_{i=1}^N y_i$
- y is logistic with $f(y) = \frac{e^{-(y-\mu)/s}}{s(1+e^{-(y-\mu)/s})^2}$

$$\log f(y_1, \ldots, y_N|\mu, s) = -\sum_{i=1}^N \left(\frac{y_i - \mu}{s}\right) - N\log s - 2\sum_{i=1}^N \log\left(1 + e^{-\frac{y_i-\mu}{s}}\right)$$

$\hat{\mu}_{\mathrm{mle}}, \hat{s}_{\mathrm{mle}}$ maximize the log-likelihood function (no closed-form)

MATLAB example of fitting distribution

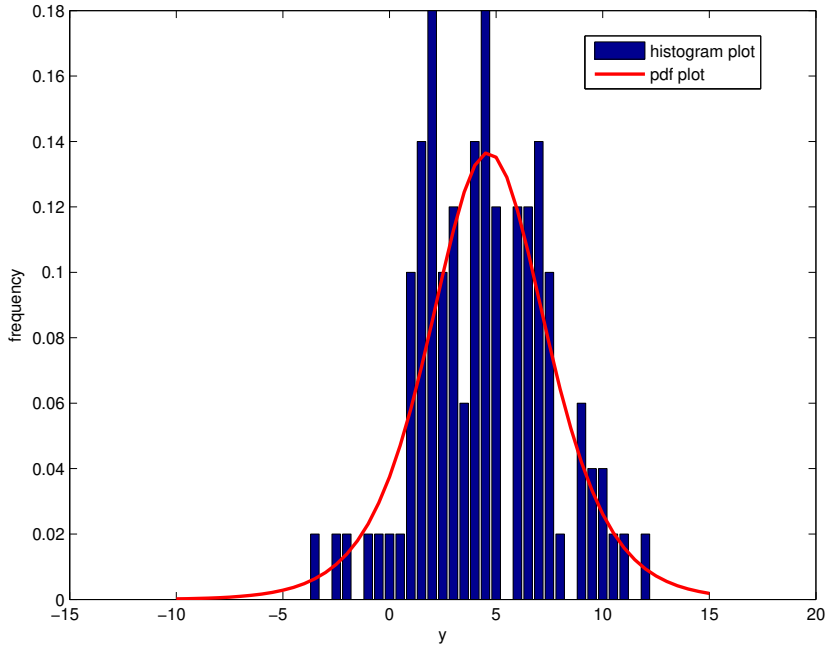
data are 100 samples from logistic distribution

```
pd = fitdist(y,'logistic')

  Logistic distribution
       mu = 4.62326    [4.00014, 5.24639]
    sigma = 1.83014    [1.55522, 2.15366]

pd.ParameterCovariance

    0.1011    0.0013
    0.0013    0.0231
```



- use `fitdist` command in MATLAB
- maximizing the loglikelihood can be solved from `mle` as well

Generalized linear model regression (GLM)

using OLS from a linear regression model $y = X\beta$ means we estimate

$$\mathbf{E}[y|X]$$

as a **linear** combination of predictors

- using linear model makes sense when y is possible in the range $(-\infty, \infty)$ (can be modelled by normal distribution)
- if y is presumably binary, integer, or non-negative, we should allow y to have an arbitrary distribution

GLM generalizes mainly two ideas of

- how y to have arbitrary distribution
- how a linear model is related to the explained variable via a *link* function

Example of nontrivial response variables

the response variable (y) can be

- binary (yes/no choice): modelled as Bernoulli
  - a diagnosis of a symptom (normal/cancer)
  - admittance to a school (admitted/rejected)
  - credit card application (accepted/rejected)
- nonnegative: exponential, Poisson, Gumbel
  - a rate of customers attending an event
  - the waiting time of the service at a bank
  - the number of credit cards that a person holds
  - a flood peak in the river (extreme value)

Genearalized linear model

in GLM, we model the expected mean of y given x

$$\mathbf{E}[y|x] = \mu = g^{-1}(\beta^T x)$$

as a function of linear combination of explanatory variables (x)

equivalence: $\beta^T x = g(\mu)$

- g is called the **link function**
- the link function gives the relation between the distribution mean y and the predictors

Binary data

setting 1: explained variable Y is binary $\{0, 1\}$

- did a subject vote last time ? (yes/no)
- subject employment status (unemployed/employed)
- credit card application result (accepted/rejected)

assumptions:

- outcomes of Y can be explained by predictors (or covariates) X
- observations of Y and X are available $\{(x_i, y_i)\}_{i=1}^{N}$

goal: develop a statistical model to explain Y when X is given

Model of binary data

setting: Y is bernoulii

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}, \qquad y = 0, 1$$

modeling: associate π as a function of a predictor x

$$\pi = F(\beta^T x)$$

since π lies in $[0, 1]$, F should ensure the properties

- F is an increasing function with x
- $\lim_{\beta^T x \to \infty} F(\beta^T x) = 1$ and $\lim_{\beta^T x \to -\infty} F(\beta^T x) = 0$

any cdf F of a continuous RV is a good candidate

three common models:

model	F	π
linear	identity	$\beta^T x$
probit	cdf of $\mathcal{N}(0,1)$	$\Phi(\beta^T x)$
logit	cdf of logistic	$\frac{e^{\beta^T x}}{1+e^{\beta^T x}}$

- linear model is simple to calculate but its value is not in $[0,1]$
- probit model involves integral calculation in evaluating Φ
- logit model has cheaper computation because of no integral

GLM format: the cdf F is the inverse of the link function

$$\mu = \pi = g^{-1}(\beta^T x)$$

goal: given data $\{(x_i, y_i)\}_{i=1}^N$, we aim to estimate β

logit model

- **the odds** is the ratio of probability to its complement

$$\text{odd} = \frac{\pi}{1-\pi}$$

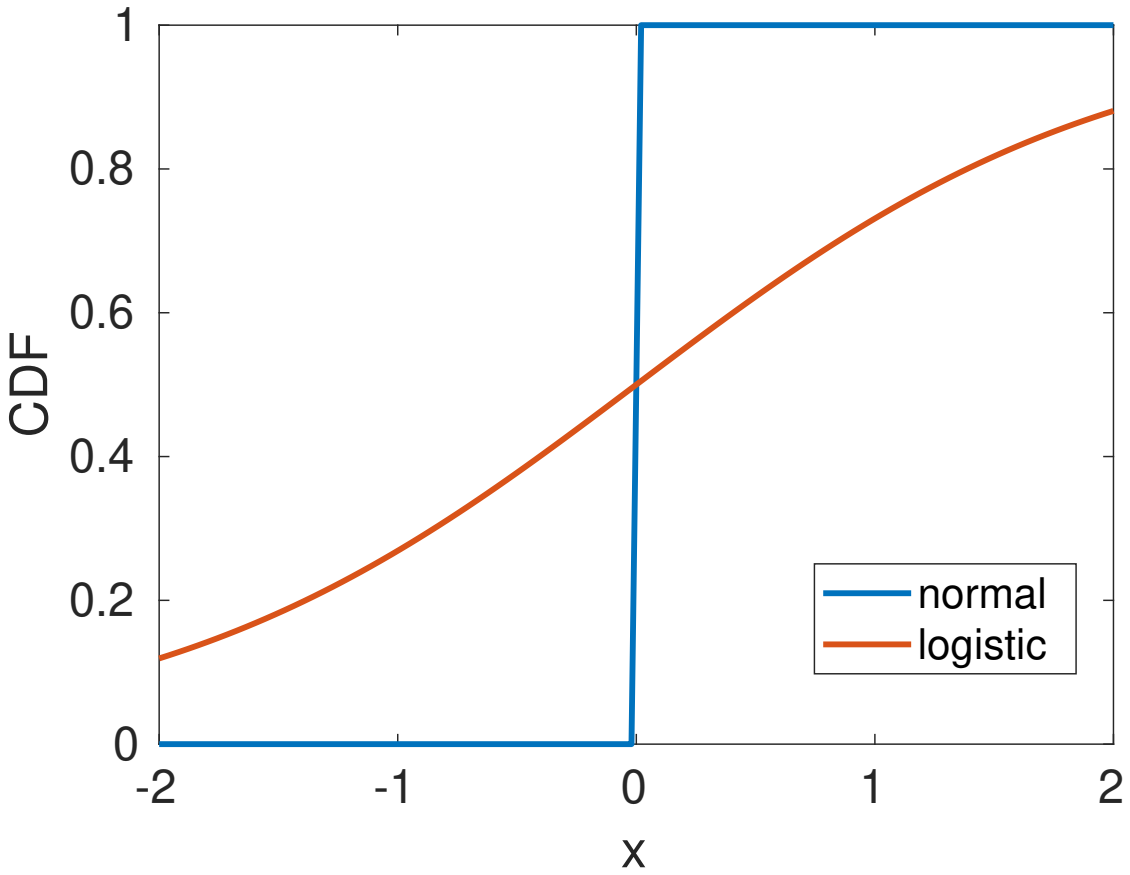
  (ratio of favorable to unfavorable cases): odd takes *any* positive value

- **the logit** is the log of odds

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

  logit maps $\mathbf{R}_+$ to $(-\infty, \infty)$ and is zero when $\pi = 1/2$

probit versus logit



the logistic cdf converges less rapidly than Φ it allows extreme values to be more frequent

Examples of GLM

distribution of y	link function	mean function	link name
normal	$\beta^T x = \mu$	$\mu = \beta^T x$	identity
exponential	$\beta^T x = \frac{1}{\mu}$	$\mu = \frac{1}{\beta^T x}$	inverse
poisson	$\beta^T x = \log(\mu)$	$\mu = e^{\beta^T x}$	log
bernoulli	$\beta^T x = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+e^{-\beta^T x}}$	logit
	$\beta^T x = \Phi^{-1}(\mu)$	$\mu = \Phi(\beta^T x)$	probit
	$\beta^T x = \log(-\log(\mu))$	$\mu = e^{-e^{\beta^T x}}$	gumbel, loglog

an estimation of GLM is to estimate β from measurements $\{(x_i, y_i)\}_{i=1}^N$

- write the log-likelihood function of y_i's based on the assumed distribution
- one way is to estimate β from maximum-likelihood estimation

notes: the related distribution functions are

distribution	support	pdf/pmf	cdf	mean
exponential	$[0, \infty)$	$\mu e^{-\mu y}$	$1 - e^{-\mu y}$	$1/\mu$
poisson	non-negative integers	$\mu^y e^{-\mu}/y!$		μ
logistic	**R**	$\frac{e^{-x}}{(1+e^{-x})^2}$	$\frac{1}{1+e^{-x}}$	0
gumbel	**R**	$e^{-(x+e^{-x})}$	$e^{-e^{-x}}$	1

MATLAB example on Probit Model

y: binary data on credit card application (accepted/rejected) x: predictors are age, monthly debt, work year, monthly income

y	AGE	DEBT (k)	YEAR	INCOME (k)
0	57	15.2694	6	23.5066
1	27	10.7196	3	53.0057
0	31	25.7597	1	32.5469
1	46	20.8113	8	45.1597
1	41	17.3692	2	38.6974

we can use `glmfit` command in MATLAB

```
b = glmfit(X,y,'binomial','link','probit','constant','off')
```

```
-0.0249
-0.1307
-0.0466
 0.1127
```

Useful MATLAB commands

command	description
`mle`	compute maximum likelihood estimates
`mlecov`	compute maximum likelihood estimates with approximated Hessian
`nlinfit`	nonlinear regression fitting
`lsqnonlin`	nonlinear least-squares problems
`fitdist`	fitting distributions by various methods (ML, GMM)
`glmfit`	estimate parameters in generalized linear model

References

Chapter 5.7 in

A.C. Cameron and P.K. Trivedi, *Microeconometircs: Methods and Applications*, Cambridge, 2005

Annette J. Dobson and Adrian G Barnett, *An Introduction to Generalized Linear Models*, Chapman, 2018