Estimation of the parameters

Maximum likelihood estimation

Michel Bierlaire

Mathematical Modeling of Behavior

Outline

Maximum likelihood estimation

Motivation

- The logit model involves unknown parameters: μ, β_k.
- Their value must be inferred from a sample of observations.
- We use maximum likelihood to estimate their value.

Note

Heterogeneity

- In the theoretical derivation, we assume wlog that the parameters μ, β_k are constant across individuals n.
- Indeed, if it were not the case, we can always write

$$\mu_n = \mu_n(s_n; \mu), \text{ and } \beta_n = \beta_n(s_n; \beta)$$

where s_n is a vector of socio-economic characteristics, and μ and β are vectors of parameters that are constant across individuals n.

Example: specification table of the model

	Alternative i	Alternative j
β_c	cost of trip (CHF)	cost of trip (CHF)
β_1	car (0/1)	car (0/1)
β_2	travel time (hours)	travel time (hours)
β_3	headway if train (min.)	headway if train (min.)

Observed variables

1. An indicator variable defined as

$$y_{in} = \begin{cases} 1 & \text{if individual } n \text{ chose alternative } i, \\ 0 & \text{if individual } n \text{ chose alternative } j. \end{cases}$$

   For notational convenience, we also define $y_{jn} = 1 - y_{in}$.
2. Two vectors of explanatory variables z_{in} and z_{jn}, each containing $K = 4$ values.

Example: raw data

	Individual 1	Individual 2	Individual 3
Train cost	40.00	7.80	40.00
Car cost	5.00	8.33	3.20
Train travel time	2.50	1.75	2.67
Car travel time	1.17	2.00	2.55
Headway	60	60	30
Choice	Car	Train	Train

Example: formatted data

Unlabeled alternatives

n	cost_{in}	car_{in}	time_{in}	headway_{in}	cost_{jn}	car_{jn}	time_{jn}	headway_{jn}
1	5	1	1.17	0	40	0	2.5	60
2	7.8	0	1.75	60	8.33	1	2	0
3	40	0	2.67	30	3.2	1	2.55	0

Chosen alternative: i.

Example: observed variables

$$y_{i1} = y_{i2} = y_{i3} = 1, y_{j1} = y_{j2} = y_{j3} = 0.$$

$$\begin{aligned}
z_{i1} &= (5 \quad 1 \quad 1.17 \quad 0)^T \\
z_{j1} &= (40 \quad 0 \quad 2.5 \quad 60)^T \\
z_{i2} &= (7.8 \quad 0 \quad 1.75 \quad 60)^T \\
z_{j2} &= (8.33 \quad 1 \quad 2 \quad 0)^T \\
z_{i3} &= (40 \quad 0 \quad 2.67 \quad 30)^T \\
z_{j3} &= (3.2 \quad 1 \quad 2.55 \quad 0)^T
\end{aligned}$$

Choice model

$$\beta = \begin{pmatrix} -1 \\ \beta_{\text{car}} \\ \beta_{\text{time}} \\ \beta_{\text{headway}} \end{pmatrix}$$

$$P_n(i; \beta, \mu) = \frac{e^{\mu \beta^T z_{in}}}{e^{\mu \beta^T z_{in}} + e^{\mu \beta^T z_{jn}}}.$$

Likelihood

Probability that the model replicates all the observations.

Example: likelihood

Individuals

- Each individual n has chosen alternative i.
- This choice is predicted by the model with probability $P_n(i; \beta, \mu)$.
- Independence of the error terms across n.

Likelihood

$$\mathcal{L}^*(\beta, \mu) = P_1(i; \beta, \mu) P_2(i; \beta, \mu) P_3(i; \beta, \mu).$$

where $\beta \in \mathbb{R}^{K=4}$ and $\mu \in \mathbb{R}$.

Example: likelihood

Assume that

$$\beta = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ \mu = 10^{-8},$$

we have

n	V_{in}	V_{jn}	$P_n(i)$	$P_n(j)$
1	-5.0	-40.00	0.5	0.5
2	-7.8	-8.33	0.5	0.5
3	-40.0	-3.20	0.5	0.5

$$\mathcal{L}^* = 0.5 \ \cdot \ 0.5 \ \cdot \ 0.5 = 0.125. \quad (1)$$

Example: likelihood

Assume that

$$\beta = \begin{pmatrix} -1 \\ -1 \\ -15 \\ -0.3 \end{pmatrix}, \ \mu = 0.1$$

we have

n	V_{in}	V_{jn}	$P_n(i)$	$P_n(j)$
1	-23.55	-95.50	0.999	0.001
2	-52.05	-39.33	0.219	0.781
3	-89.05	-42.45	0.009	0.991

$$\mathcal{L}^* = 0.999 \ \cdot \ 0.219 \ \cdot \ 0.009 = 0.00197.$$

Example: likelihood

Assume that

$$\beta = \begin{pmatrix} -1 \\ -20 \\ -20 \\ -0.1 \end{pmatrix}, \ \mu = 0.04$$

we have

n	V_{in}	V_{jn}	$P_n(i)$	$P_n(j)$
1	-48.40	-96.00	0.807	0.193
2	-48.80	-68.33	0.642	0.358
3	-96.40	-74.20	0.339	0.661

$$\mathcal{L}^* = 0.807 \cdot 0.642 \cdot 0.339 = 0.176.$$

Definitions

Likelihood

$$\mathcal{L}^*(\beta, \mu) = \prod_{n=1}^{N} P_n(i; \beta, \mu)^{y_{in}} P_n(j; \beta, \mu)^{y_{jn}},$$

where $\beta \in \mathbb{R}^K$ and $\mu \in \mathbb{R}^N$.

Log likelihood

$$\mathcal{L}(\beta, \mu) = \sum_{n=1}^{N} (y_{in} \ln P_n(i; \beta, \mu) + y_{jn} \ln P_n(j; \beta, \mu)).$$

Maximum likelihood estimation

Optimization problem

$$\widehat{\beta}, \widehat{\mu} = \operatorname{argmax}_{\beta \in \mathbb{R}^K, \mu \in \mathbb{R}} \mathcal{L}(\beta, \mu) = \mathcal{L}(\beta_1, \beta_2, \ldots, \beta_K, \mu).$$

Software

`biogeme.epfl.ch`

Estimation of the parameters

Unknown parameters

$$\mu, \beta_k, k = 1, \ldots, K.$$

Contribution to the likelihood of observation n

$$P_n(i|\mathcal{C}) = \frac{e^{-\mu \mathrm{cost}_{in} + \sum_k \mu \beta_k z_{ink}}}{e^{-\mu \mathrm{cost}_{in} + \sum_k \mu \beta_k z_{ink}} + e^{-\mu \mathrm{cost}_{jn} + \sum_k \mu \beta_k z_{jnk}}}.$$

Issue: non linearity

- Non-concave formulation.
- Algorithms may converge to local maxima.
- A concave formulation is desirable.

Estimation of the parameters

Rename the parameters

$$\beta'_c = -\mu \text{ and } \beta'_k = \mu\beta_k, \; \forall k.$$

$$P_n(i|\mathcal{C}) = \frac{e^{\beta'_c \text{cost}_{in} + \sum_k \beta'_k z_{ink}}}{e^{\beta'_c \text{cost}_{in} + \sum_k \beta'_k z_{ink}} + e^{\beta'_c \text{cost}_{jn} + \sum_k \beta'_k z_{ijk}}}.$$

Notes

- It is equivalent to the original specification, if μ is normalized to 1.
- Logit with this specification has a concave log-likelihood function.
- Once the parameters are estimated, the inverse transform must be applied to obtain the willingness to pay parameters

$$\beta_t = \frac{\beta'_t}{\mu} = -\frac{\beta'_t}{\beta'_c}.$$

Moneymetric specification

Unnormalized version: includes all β's and μ

$$\mu V_{in} = \mu\beta_c \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}.$$

Normalization: $\beta_c = -1$

$$\mu V_{in} = -\mu \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}.$$

Moneymetric specification

Normalization: $\beta_c = -1$

$$\mu V_{in} = -\mu \text{cost}_{in} + \sum_k \mu \beta_k z_{ink}.$$

Advantages

- Convenient unit.
- Easy interpretation.
- Explicit representation of μ.
- Scale can vary with n: $\mu_n(s_n; \mu)$

Drawbacks

- Not linear in the parameters.
- More complicated to estimate.
- Possibility to be caught in local maxima.

Linear-in-parameters specification

Unnormalized version: includes all β's and μ

$$\mu V_{in} = \mu\beta_c \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}.$$

Normalization: $\mu = 1$

$$\mu V_{in} = \beta_c \text{cost}_{in} + \sum_k \beta_k z_{ink}.$$

Linear-in-parameters specification

Normalization: $\mu = 1$

$$\mu V_{in} = V_{in} = \beta_c \text{cost}_{in} + \sum_k \beta_k z_{ink}.$$

Advantages

- Linear in the parameters.
- Simple to estimate.
- With logit, concave log-likelihood function.

Drawbacks

- Unitless.
- Coefficients difficult to interpret.
- No explicit representation of μ.
- Scale μ the same for all individuals.

Linear-in-parameters normalization

Heteroscedasticity

- What if μ_n varies across individuals?
- In practice, there is a scale parameter for each segment.
- Normalization: select one segment, and normalize the corresponding scale parameter to 1.
- Estimate the scale parameters for the other segments.
- Consequence: not linear-in-parameters anymore.

Normalization

Notes

- The choice of a specific normalization is arbitrary, as both lead to the exact same choice model.
- The linear-in-parameters normalization has been widely adopted in the literature, for historical reasons.
- The moneymetric normalization provides a better interpretation.
- Warning: if some parameters are assumed to be distributed (see the lecture on mixtures), the choice of the distribution is conditional on the type of normalization.

Comparison with linear regression

Linear regression

$$y_n = \sum_k \beta_k z_{nk} + \varepsilon_n$$

- $\varepsilon_n \sim N(\eta, \sigma^2)$.
- ε_n independent from x.
- y_n is observable.
- All parameters are identified.

Choice model

$$U_{in} = \sum_k \beta_k z_{ink} + \varepsilon_{in}$$

- $\varepsilon_{in} \sim EV(\eta, \mu)$.
- ε_{in} independent from x.
- U_{in} is latent, not observable.
- Location: η does not play any role.
- Units: normalization is needed.

Outline

Example: interurban mode choice in Switzerland

Sample

- Revealed preference data.
- Survey conducted between 2009 and 2010 for PostBus.
- Questionnaires sent to people living in rural areas.
- Each observation corresponds to a sequence of trips from home to home.
- Sample size: 1899.

Model: 3 alternatives

- Car,
- public transportation (PT),
- slow modes.

Attributes

Travel cost

- Alternatives: car and public transportation.
- Units: CHF.
- Coefficient: generic.

Travel time

- Alternatives: car and public transportation.
- Units: minutes.
- Coefficient: alternative specific, interacted with trip purpose.

Attributes

Waiting time

- Alternative: public transportation.
- Units: minutes
- Coefficient: interacted with socio-professional category.

Distance

- Alternative: slow mode.
- Units: kilometers.
- Coefficient: interaction with age.

Parameters of the error terms

Alternative specific constants

- Alternatives: car and public transportation.
- Units: none.
- Coefficient: interacted with public transportation subscription (GA) and language region.

Scale parameter

Interacted with public transportation subscription (GA).

Utility functions

Public transportation

$$\mu_n V_{\text{pt},n} = \mu_n(\text{asc}_{\text{pt},n} + \beta_c \text{tc}_{\text{pt},n} + \beta_{t,\text{pt},n}\text{tt}_{\text{pt},n} + \beta_{w,n}\text{wt}_n),$$

where

$$\begin{aligned}
\text{asc}_{\text{pt},n} =& \text{asc}_{\text{pt}} + \text{asc}_{\text{pt,GA}} \, \mathbb{1}(n \in \text{GA}) + \text{asc}_{\text{pt,german}} \, \mathbb{1}(n \in \text{german}), \\
\mu_n =& \mu_{\text{GA}} \, \mathbb{1}(n \in \text{GA}) + \mu_{\text{no GA}} \, \mathbb{1}(n \notin \text{GA}), \\
\beta_{t,\text{pt},n} =& \beta_{t,\text{pt}} + \beta_{t,\text{pt,non work}} \, \mathbb{1}(n \in \text{non work}), \\
\beta_{w,n} =& \beta_w + \beta_{w,\text{craftman}} \, \mathbb{1}(n \in \text{craftman}) + \beta_{w,\text{manager}} \, \mathbb{1}(n \in \text{manager}) + \\
& \beta_{w,\text{intellectual}} \, \mathbb{1}(n \in \text{intellectual}).
\end{aligned}$$

Utility functions

Car

$$\mu_n V_{\text{car},n} = \mu_n(\text{asc}_{\text{car},n} + \beta_c \text{tc}_{\text{car},n} + \beta_{t,\text{car},n} \text{tt}_{\text{car},n}),$$

where

$$\begin{aligned}
\text{asc}_{\text{car},n} =& \text{asc}_{\text{car}} + \text{asc}_{\text{car,GA}} \ \mathbb{1}(n \in \text{GA}) + \text{asc}_{\text{car,german}} \ \mathbb{1}(n \in \text{german}), \\
\mu_n =& \mu_{\text{GA}} \ \mathbb{1}(n \in \text{GA}) + \mu_{\text{no GA}} \ \mathbb{1}(n \notin \text{GA}), \\
\beta_{t,\text{car},n} =& \beta_{t,\text{car}} + \beta_{t,\text{car,non work}} \ \mathbb{1}(n \in \text{non work}),
\end{aligned}$$

Utility functions

Slow modes

$$\mu_n V_{\text{sm},n} = \mu_n(\beta_{d,n}\text{distance}_n),$$

where

$$\begin{aligned}
\mu_n =& \mu_{\text{GA}}\, \mathbb{1}(n \in \text{GA}) + \mu_{\text{no GA}}\, \mathbb{1}(n \notin \text{GA}), \\
\beta_{d,n} =& \beta_d + \beta_{d,\text{young}}\, \mathbb{1}(n \in \text{age } \leqslant 65).
\end{aligned}$$

Parameters

19 parameters

asc_{pt}	$\beta_{w,\text{craftman}}$
$\text{asc}_{\text{pt,GA}}$	$\beta_{w,\text{manager}}$
$\text{asc}_{\text{pt,german}}$	$\beta_{w,\text{intellectual}}$
asc_{car}	$\beta_{t,\text{car}}$
$\text{asc}_{\text{car,GA}}$	$\beta_{t,\text{car,non work}}$
$\text{asc}_{\text{car,german}}$	β_{d}
β_{c}	$\beta_{d,\text{young}}$
$\beta_{t,\text{pt}}$	μ_{GA}
$\beta_{t,\text{pt,non work}}$	μ_{noGA}
β_{w}	

Normalizations: moneymetric: $\beta_c = -1$, linear-in-parameters: $\mu_{\text{no GA}} = 1$.

Estimation results

General statistics

	Moneymetric	Linear
Number of estimated parameters	18	18
Sample size	1899	1899
Null log likelihood	-2046.529	-2046.529
Final log likelihood	-1066.542	-1066.542
Likelihood ratio test for the null model	1959.975	1959.975
Rho-square for the null model	0.479	0.479
Rho-square-bar for the null model	0.47	0.47
Akaike Information Criterion	2169.083	2169.083
Bayesian Information Criterion	2268.967	2268.967
Number of iterations	110	37

Estimation results: moneymetric

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Cte reference (car)	41.4	17.3	2.4	0.0166
2	Cte diff. yearly subscription (car)	-35.3	16.6	-2.13	0.0335
3	Cte diff. German speaking (car)	-29.5	11.8	-2.5	0.0126
4	Cte. reference (PT)	-21.9	14.4	-1.51	0.13
5	Cte diff. yearly subscription (PT)	6.57	13.1	0.5	0.617
6	Cte diff. German speaking (PT)	6.24	9.61	0.649	0.516
7	Cost [CHF]	-1.0			
8	Travel time reference (car) [min.]	-1.28	0.382	-3.34	0.000841
9	Travel time diff. other trip purposes (car) [min.]	0.653	0.35	1.87	0.0618
10	Travel time reference (PT) [min.]	-0.328	0.136	-2.41	0.016
11	Travel time diff. other trip purposes (PT) [min.]	0.148	0.124	1.19	0.235
12	Waiting time reference [min.]	-0.837	0.35	-2.39	0.0168
13	Waiting time diff. craftman [min.]	0.638	0.392	1.62	0.104
14	Waiting time diff. intellectual [min.]	0.601	0.457	1.32	0.188
15	Waiting time diff. manager [min.]	-1.25	0.55	-2.26	0.0236
16	Distance reference [km]	-11.3	4.19	-2.7	0.00684
17	Distance diff. age $\leqslant$ 65 [km]	4.68	2.91	1.61	0.108
18	Scale (no yearly subscription)	0.0319	0.00894	3.56	0.000367
19	Scale (yearly subscription)	0.0498	0.0201	2.47	0.0134

Estimation results: linear-in-parameters

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Cte reference (car)	1.32	0.382	3.45	0.000561
2	Cte diff. yearly subscription (car)	-1.13	0.375	-3.0	0.00268
3	Cte diff. German speaking (car)	-0.94	0.272	-3.46	0.00054
4	Cte. reference (PT)	-0.697	0.406	-1.72	0.0862
5	Cte diff. yearly subscription (PT)	0.209	0.411	0.509	0.611
6	Cte diff. German speaking (PT)	0.199	0.3	0.662	0.508
7	Cost [CHF]	-0.0319	0.00894	-3.56	0.000369
8	Travel time reference (car) [min.]	-0.0407	0.00777	-5.24	1.63e-07
9	Travel time diff. other trip purposes (car) [min.]	0.0208	0.00844	2.46	0.0137
10	Travel time reference (PT) [min.]	-0.0104	0.00379	-2.76	0.00583
11	Travel time diff. other trip purposes (PT) [min.]	0.0047	0.00352	1.34	0.181
12	Waiting time reference [min.]	-0.0267	0.00843	-3.16	0.00156
13	Waiting time diff. craftman [min.]	0.0203	0.0114	1.78	0.0744
14	Waiting time diff. intellectual [min.]	0.0192	0.014	1.37	0.171
15	Waiting time diff. manager [min.]	-0.0397	0.0146	-2.72	0.00656
16	Distance reference [km]	-0.361	0.1	-3.6	0.000317
17	Distance diff. age $\leqslant$ 65 [km]	0.149	0.0841	1.77	0.0764
18	Scale (no yearly subscription)	1.0			
19	Scale (yearly subscription)	1.56	0.525	2.98	0.0029

Estimation results: comparison

Parameter	Description	Moneymetric	Linear in parameters
1	Cte reference (car)	41.4	1.32
2	Cte diff. yearly subscription (car)	-35.3	-1.13
3	Cte diff. German speaking (car)	-29.5	-0.94
4	Cte. reference (PT)	-21.9	-0.697
5	Cte diff. yearly subscription (PT)	6.57	0.209
6	Cte diff. German speaking (PT)	6.24	0.199
7	Cost [CHF]	-1.0	-0.0319
8	Travel time reference (car) [min.]	-1.28	-0.0407
9	Travel time diff. other trip purposes (car) [min.]	0.653	0.0208
10	Travel time reference (PT) [min.]	-0.328	-0.0104
11	Travel time diff. other trip purposes (PT) [min.]	0.148	0.0047
12	Waiting time reference [min.]	-0.837	-0.0267
13	Waiting time diff. craftman [min.]	0.638	0.0203
14	Waiting time diff. intellectual [min.]	0.601	0.0192
15	Waiting time diff. manager [min.]	-1.25	-0.0397
16	Distance reference [km]	-11.3	-0.361
17	Distance diff. age $\leqslant$ 65 [km]	4.68	0.149
18	Scale (no yearly subscription)	0.0319	1.0
19	Scale (yearly subscription)	0.0498	1.56

Estimation results: moneymetric

$$\beta_c = -0.0319$$

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Cte reference (car)	1.32	0.551	2.4	0.0166
2	Cte diff. yearly subscription (car)	-1.13	0.53	-2.13	0.0335
3	Cte diff. German speaking (car)	-0.941	0.377	-2.5	0.0126
4	Cte. reference (PT)	-0.697	0.461	-1.51	0.13
5	Cte diff. yearly subscription (PT)	0.209	0.419	0.499	0.617
6	Cte diff. German speaking (PT)	0.199	0.306	0.649	0.516
7	Cost [CHF]	-0.0319			
8	Travel time reference (car) [min.]	-0.0407	0.0122	-3.34	0.000841
9	Travel time diff. other trip purposes (car) [min.]	0.0208	0.0112	1.87	0.0618
10	Travel time reference (PT) [min.]	-0.0105	0.00434	-2.41	0.016
11	Travel time diff. other trip purposes (PT) [min.]	0.00471	0.00396	1.19	0.235
12	Waiting time reference [min.]	-0.0267	0.0112	-2.39	0.0168
13	Waiting time diff. craftman [min.]	0.0203	0.0125	1.62	0.104
14	Waiting time diff. intellectual [min.]	0.0192	0.0146	1.32	0.188
15	Waiting time diff. manager [min.]	-0.0398	0.0176	-2.26	0.0236
16	Distance reference [km]	-0.362	0.134	-2.7	0.00684
17	Distance diff. age $\leqslant$ 65 [km]	0.149	0.093	1.61	0.108
18	Scale (no yearly subscription)	0.999	0.28	3.56	0.000368
19	Scale (yearly subscription)	1.56	0.631	2.47	0.0134

Comments

Moneymetric vs linear-in-parameters

- The results are identical, up to numerical precision.
- Moneymetric is more challenging for the optimization algorithm.

Alternative specific constants: moneymetric model

Car vs. Slow modes

GA	German	Value	
No	No	41.4.	
No	Yes	11.9	= 41.4 - 29.5.
Yes	No	6.1	= 41.4 - 35.3.
Yes	Yes	-23.4	= 41.4 - 29.5 - 35.3.

Public transp. vs. Slow modes

GA	German	Value	
No	No	-21.9.	
No	Yes	-15.66	= -21.9 + 6.24.
Yes	No	-15.33	= -21.9 + 6.57.
Yes	Yes	-9.09	= -21.9 + 6.24 + 6.57.

Travel time coefficient

Car

- Trip purpose work: -1.28.
- Other trip purpose: -1.28 + 0.653 = -0.627.

Public transportation

- Trip purpose work: -0.328.
- Other trip purpose: + 0.148 = -0.18.

Waiting time coefficient

- Craftman: -0.837+0.638=-0.199.
- Intellectual: -0.837+0.601=-0.236.
- Manager: -0.837-1.25=-2.807.
- Others: -0.837.

Distance coefficient

- 65 or less: -11.3 $+$ 4.68 $=$ -6.62.
- More than 65: -11.3.

Outline

Output of the estimation

Main outputs

- the parameter estimates $\widehat{\beta}$,
- the value of the log likelihood function at the parameter estimates $\mathcal{L}(\widehat{\beta})$.

Other output

- variance-covariance matrix of the estimates,
- standard errors,
- *t*-statistics,
- *p*-values,
- goodness of fit.

Variance-covariance: Cramer-Rao bound

Definition

$$-\,\mathrm{E}\left[\nabla^2\mathcal{L}(\beta)\right]^{-1} = \left\{-\,\mathrm{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta\partial\beta^T}\right]\right\}^{-1}.$$

Estimator

$$A = \mathrm{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_k\partial\beta_m}\right] \approx \sum_{n=1}^{N}\left[\frac{\partial^2\left(y_{in}\ln P_n(i) + y_{jn}\ln P_n(j)\right)}{\partial\beta_k\partial\beta_m}\right]_{\beta=\widehat{\beta}},$$

$$\widehat{\Sigma}_{\beta}^{\mathrm{CR}} = -\widehat{A}^{-1}.$$

Variance-covariance: robust estimator

BHHH matrix

$$-E\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta\partial\beta^T}\right] \approx \sum_{n=1}^{N} \nabla L_n(\widehat{\beta})\nabla L_n(\widehat{\beta})^T = \widehat{B},$$

where

$$\nabla L_n(\widehat{\beta}) = \nabla(y_{in}\ln P_n(i) + y_{jn}\ln P_n(j)).$$

Robust or sandwich estimator

$$\widehat{\Sigma}^{\text{R}}_{\beta} = (-\widehat{A})^{-1}\ \widehat{B}\ (-\widehat{A})^{-1} = \widehat{\Sigma}^{\text{CR}}_{\beta}\ (\widehat{\Sigma}^{\text{BHHH}}_{\beta})^{-1}\ \widehat{\Sigma}^{\text{CR}}_{\beta}.$$

Variance-covariance matrix

Notes

- When the true likelihood function is maximized, these estimators are asymptotically equivalent.
- When other consistent estimators are used, different from the maximum likelihood, the robust estimator must be used.

Standard errors

Definition

$$\sigma_k = \sqrt{\widehat{\Sigma}_\beta(k, k)},$$

where $\widehat{\Sigma}_\beta(k, k)$ is the kth entry of the diagonal of the matrix $\widehat{\Sigma}_\beta$.

t statistics

Definition

$$t_k = \frac{\widehat{\beta}_k - \beta_0}{\sigma_k},$$

where β_0 is the value associated wth the null hypothesis (usually 0).

Role

Typically used to test the null hypothesis that the true value of a coefficient is zero. This hypothesis can be rejected with 95% of confidence if

$$|t_k| \geqslant 1.96. \tag{2}$$

p values

Definition

- It is the probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis that $\beta_k = 0$.
- Consider an estimate $\widehat{\beta}_k$ of the parameter β_k, and t_k its t statistic. The p value is calculated as

$$p_k = 2(1 - \Phi(t_k)),$$

where $\Phi(\cdot)$ is the cumulative density function of the univariate standard normal distribution.

Role

- Exact same role as the t statistics.
- The null hypothesis can be rejected at the confidence level p_k.

Goodness of fit

Preliminary remarks

- There are several measures of goodness of fit.
- None of them can be used in an absolute way.
- They can only be used to compare two models, estimated on the same data set, with the same dependent variable.

Goodness of fit

Log likelihood

$$\mathcal{L}(\widehat{\beta}).$$

Normalized log likelihood

$$\rho^2 = 1 - \frac{\mathcal{L}(\widehat{\beta})}{\mathcal{L}(0)}.$$

Comments on ρ^2

- It is not the square of anything. It mimics R^2 in linear regression.
- In general, value strictly between 0 (null model) and 1 (perfect fit).
- But the value is meaningless as such.

Goodness of fit: accounting for the number of parameters

Akaike Information Criterion (AIC)

$$2K - 2\mathcal{L}(\widehat{\beta}).$$

Note: the lower, the better.

Normalized AIC

$$\bar{\rho}^2 = 1 + \frac{\text{AIC}}{2\mathcal{L}(0)} = 1 - \frac{\mathcal{L}(\widehat{\beta}) - K}{\mathcal{L}(0)}.$$

Note: the higher, the better.

Goodness of fit: accounting for sample size

Bayesian Information Criterion (BIC)

$$K \ln(N) - 2\mathcal{L}(\widehat{\beta}).$$

Note: the lower, the better.

Goodness of fit: benchmark models

Benchmark model with 0 parameter

$$P_n(i) = \frac{1}{J_n}.$$

$$\mathcal{L}(0) = -\sum_{n=1}^{N} \log(J_n),$$

where N is the number of observations.

Goodness of fit: benchmark models

Benchmark model with $J - 1$ parameters

We assume that $J_n = J$, $\forall n$:

$$P_n(i) = p_i = \frac{N_i}{N}.$$

There are J parameters $p_1, \ldots, p_J$. They must sum up to one, removing one degree of freedom.

$$\mathcal{L}(c) = \sum_{i=1}^{J} N_i (\ln N_i - \ln N) = \sum_{i=1}^{J} N_i \ln N_i - N \ln N.$$

where N_i is the number of observations choosing alternative i.

Likelihood ratio test

Null hypothesis

Two models are equivalent.

Statistic

$$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta}))$$

is asymptotically distributed as χ^2 with K degrees of freedom.

Statistic

$$-2(\mathcal{L}(c) - \mathcal{L}(\widehat{\beta}))$$

is asymptotically distributed as χ^2 with $K - (J - 1)$ degrees of freedom.

Summary

- Maximum likelihood estimation.
- Alternative normalization: $\mu = 1$.
- Output of the estimation.