Binary choice

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Example

Ben-Akiva & Lerman (1985) Discrete Choice Analysis: Theory and Applications to Travel Demand, MIT Press (p.88)

Choice between Auto and Transit

Example

Data :

#	Time auto	Time transit	Choice	#	Time auto	Time transit	Choice
1	52.9	4.4	T	11	99.1	8.4	T
2	4.1	28.5	T	12	18.5	84.0	C
3	4.1	86.9	C	13	82.0	38.0	C
4	56.2	31.6	T	14	8.6	1.6	T
5	51.8	20.2	T	15	22.5	74.1	C
6	0.2	91.2	C	16	51.4	83.8	C
7	27.6	79.7	C	17	81.0	19.2	T
8	89.9	2.2	T	18	51.0	85.0	C
9	41.5	24.5	T	19	62.2	90.1	C
10	95.0	43.5	T	20	95.1	22.2	T
				21	41.6	91.5	C

Binary choice model

$$
\begin{array}{rcl}
U_C & = & \beta_1 T_C + \varepsilon_C \\
U_T & = & \beta_1 T_T + \varepsilon_T
\end{array}
$$

where T_C is the travel time with car (min) and T_T the travel time with transit (min).

$$
\begin{array}{rcl}
P(C|\{C,T\}) & = & P(U_C \geq U_T) \\
 & = & P(\beta_1 T_C + \varepsilon_C \geq \beta_1 T_T + \varepsilon_T) \\
 & = & P(\beta_1 T_C - \beta_1 T_T \geq \varepsilon_T - \varepsilon_C) \\
 & = & P\left(\varepsilon \leq \beta_1 (T_C - T_T)\right)
\end{array}
$$

where $\varepsilon = \varepsilon_T - \varepsilon_C$.

Error term

Three questions about the random variables ε_T and ε_C :

1. What's their mean?
2. What's their variance?
3. What's their distribution?

Error term

The mean

$$P(C|\{C,T\}) = P\left(\varepsilon \leq \beta_1(T_C - T_T)\right)$$

Assume that $E[\varepsilon] = \beta_0$ and define

$$\varepsilon' = \varepsilon - \beta_0$$

Then, $E[\varepsilon'] = 0$ and

$$\begin{aligned} P(C|\{C,T\}) &= P\left(\varepsilon' \leq \beta_1(T_C - T_T) - \beta_0\right) \\ &= P\left(\varepsilon' \leq (\beta_1 T_C - \beta_0) - \beta_1 T_T\right) \\ &= P\left(\varepsilon' \leq \beta_1 T_C - (\beta_1 T_T + \beta_0)\right) \end{aligned}$$

Error term

The mean

The mean of ε can be included as a parameter of the deterministic part.

Only the mean of the difference of the error terms is meaningful.

Alternative Specific Constant:

$$\begin{array}{rcll} U_C & = & \beta_1 T_C & +\varepsilon_C \\ U_T & = & \beta_1 T_T + \beta_0 & +\varepsilon_T \end{array} \quad \text{or} \quad \begin{array}{rcll} U_C & = & \beta_1 T_C - \beta_0 & +\varepsilon_C \\ U_T & = & \beta_1 T_T & +\varepsilon_T \end{array}$$

Error term

The mean

Note that adding the same constant to all utility functions does not affect the probability model

$$P(U_C \geq U_T) = P(U_C + K \geq U_T + K) \quad \forall K \in \mathbb{R}^n.$$

If the deterministic part of the utility functions contains an Alternative Specific Constant (ASC) for all alternatives but one, the mean of the error terms can be assumed to be zero without loss of generality.

Error term

The variance

$$P(U_C \geq U_T) = P(\alpha U_C \geq \alpha U_T) \quad \forall \alpha > 0$$

Multiplying the utility by any strictly positive number α does not affect the probability.
Moreover,

$$\begin{aligned} \text{Var}(\alpha U_C) &= \alpha^2 \text{Var}(U_C) \\ \text{Var}(\alpha U_T) &= \alpha^2 \text{Var}(U_T) \end{aligned}$$

Select α such that $\text{Var}(\alpha U_i) = a$:

$$\alpha = \sqrt{\frac{a}{\text{Var}(U_i)}}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Error term

The variance

> Imposing an arbitrary variance amounts to imposing an arbitrary scale to the utility

Error term

The distribution

Assumption 1: ε_T and ε_C are the sum of many r.v. capturing unobservable attributes (e.g. mood, experience), measurement and specification errors.

Central-limit theorem: the sum of many i.i.d. random variables approximatively follows a normal distribution

$$\varepsilon_{in} \sim N(0,1)$$

Error term

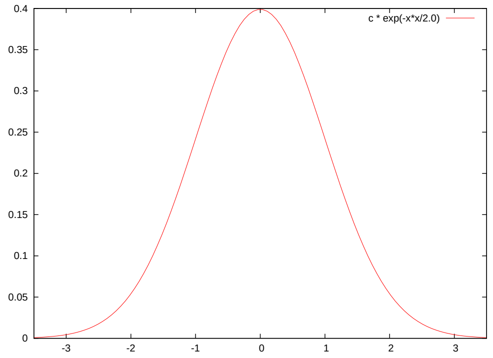
The distribution

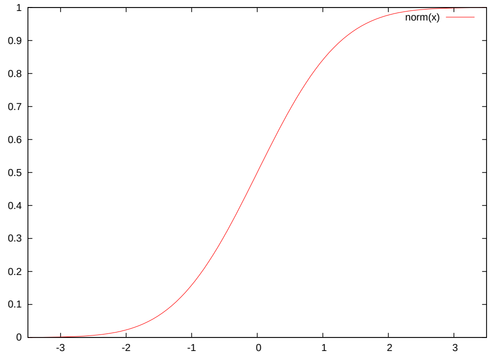
Normal distribution:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

If $\varepsilon \sim N(0,1)$, then

$$P(c \geq \varepsilon) = F(c) = \int_{-\infty}^{c} f(t)dt$$





Error term

The distribution

From the properties of the normal distribution, we have

$$\begin{aligned}\varepsilon_C &\sim N(0,1)\\ \varepsilon_T &\sim N(0,1)\\ \varepsilon = \varepsilon_T - \varepsilon_C &\sim N(0,2)\end{aligned}$$

As the variance is arbitrary, we may also assume

$$\begin{aligned}\varepsilon_C &\sim N(0,0.5)\\ \varepsilon_T &\sim N(0,0.5)\\ \varepsilon = \varepsilon_T - \varepsilon_C &\sim N(0,1)\end{aligned}$$

Error term

The distribution

$$\begin{aligned} P(C|\{C,T\}) &= P\left(\varepsilon \leq V_C - V_T\right) \\ &= F(V_C - V_T) \end{aligned}$$

$$\boxed{P(C|\{C,T\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{V_C - V_T} e^{\frac{1}{2}t^2} dt}$$

Not a closed form expression

Error term

The distribution

If the error terms are assumed to follow a normal distribution, the corresponding model is called

Probability Unit Model or Probit Model.

Error term

The distribution

Assumption 2: ε_T and ε_C are the maximum of many r.v. capturing unobservable attributes (e.g. mood, experience), measurement and specification errors.

Gumbel theorem: the maximum of many i.i.d. random variables approximately follows an Extreme Value distribution.

$$\varepsilon_C \sim \text{EV}(0, \mu)$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Error term

EV(η, μ), with $\mu > 0$:

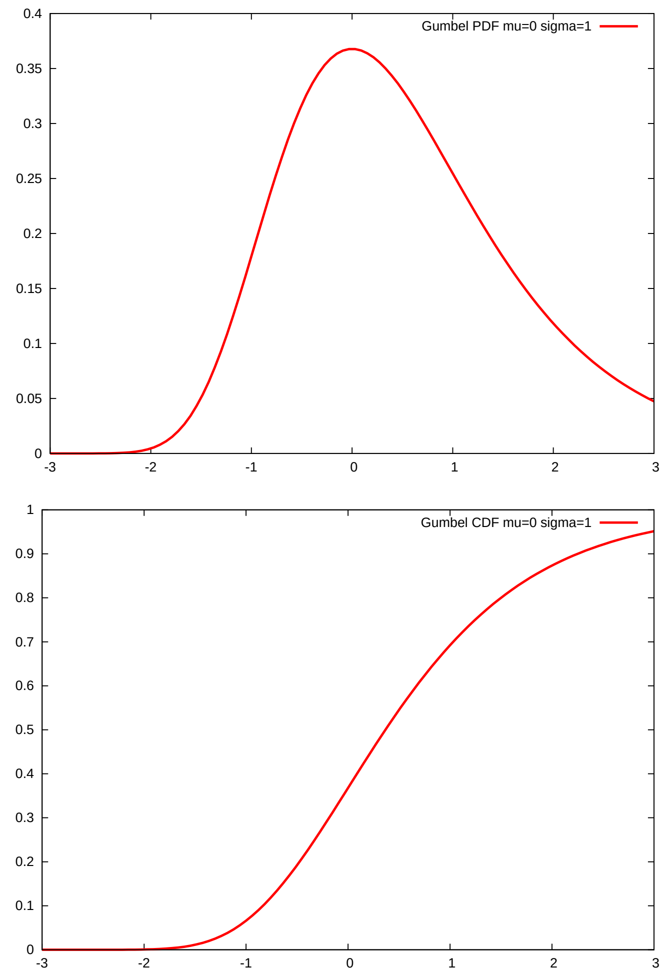
$$f(t) = \mu e^{-\mu(t-\eta)} e^{-e^{-\mu(t-\eta)}}$$

If $\varepsilon \sim$ EV(η, μ), then

The distribution

$$\begin{aligned} P(c \geq \varepsilon) = F(c) &= \int_{-\infty}^{c} f(t)dt \\ &= e^{-e^{-\mu(c-\eta)}} \end{aligned}$$

Error term


Gumbel PDF mu=0 sigma=1
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
Gumbel CDF mu=0 sigma=1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3

Error term

If

$$\varepsilon \sim \text{EV}(\eta, \mu)$$

then

$$E[\varepsilon] = \eta + \frac{\gamma}{\mu} \quad \text{and} \quad \text{Var}[\varepsilon] = \frac{\pi^2}{6\mu^2}$$

where

$$\gamma = \lim_{k \to \infty} \sum_{i=1}^{k} \frac{1}{i} - \ln k \approx 0.5772 \quad \text{Euler constant}$$

Error term

The distribution

$$P(C|\{C,T\}) = P\left(\varepsilon \leq V_C - V_T\right) = P\left(\varepsilon \leq \beta_1(T_C - T_T) - \beta_0\right)$$

where $\varepsilon = \varepsilon_T - \varepsilon_C$.

$$\begin{aligned} \varepsilon_C &\sim \text{EV}(0,\mu) \\ \varepsilon_T &\sim \text{EV}(0,\mu) \\ \varepsilon &\sim \text{Logistic}(0,\mu) \end{aligned}$$

Logit Model

Error term

The distribution

For the Logistic(0,μ), we have

$$P(c \geq \varepsilon) = F(c) = \frac{1}{1+e^{-\mu c}}$$

$$\begin{aligned} P(C|\{C,T\}) &= P\left(\varepsilon \leq V_C - V_T\right) \\ &= F(V_C - V_T) \\ &= \frac{1}{1+e^{-\mu(V_C-V_T)}} \\ &= \frac{e^{\mu V_C}}{e^{\mu V_C}+e^{\mu V_T}} \end{aligned}$$

Error term

The distribution

$$P(C|\{C,T\}) = \frac{e^{\mu V_C}}{e^{\mu V_C} + e^{\mu V_T}}$$

Binary Logistic Unit Model or Binary Logit Model

Normalize $\mu = 1$

Back to the example

Let's assume that $\beta_0 = 0.5$ and $\beta_1 = -0.1$
Let's consider the first observation:

- $T_C = 52.9$
- $T_T = 4.4$
- Choice = *transit*

What's the probability given by the model that this individual indeed chooses *transit*?

$$\begin{array}{rcccl} V_C & = & \beta_1 T_C & = & -5.29 \\ V_T & = & \beta_1 T_T + \beta_0 & = & 0.06 \end{array}$$

Back to the example

$$P(\text{transit}) = \frac{e^{V_T}}{e^{V_T} + e^{V_C}}$$

$$P(\text{transit}) = \frac{e^{0.06}}{e^{0.06} + e^{-5.29}} \cong 1$$

The model almost perfectly predicts this observation

Back to the example

Let's assume again that $\beta_0 = 0.5$ and $\beta_1 = -0.1$
Let's consider the second observation:

- $T_C = 4.1$
- $T_T = 28.5$
- Choice = *transit*

What's the probability given by the model that this individual indeed chooses *transit*?

$$
\begin{aligned}
V_C &= \beta_1 T_C &= -0.41 \\
V_T &= \beta_1 T_T + \beta_0 &= -2.35
\end{aligned}
$$

Back to the example

$$P(\text{transit}) = \frac{e^{V_T}}{e^{V_T} + e^{V_C}}$$

$$P(\text{transit}) = \frac{e^{-2.35}}{e^{-2.35} + e^{-0.41}} \cong 0.13$$

The model does not correctly predict this observation

Back to the example

The probability that the model reproduces both observations is

$$P_1(\text{transit})P_2(\text{transit}) = 0.13$$

The probability that the model reproduces all observations is

$$P_1(\text{transit})P_2(\text{transit})\ldots P_{21}(\text{auto}) = 4.62\ 10^{-4}$$

In general

$$\mathcal{L}^* = \prod_n \left(P_n(\text{auto})^{y_{\text{auto},n}} P_n(\text{transit})^{y_{\text{transit},n}}\right)$$

$$y_{\text{auto},n} = 1 - y_{\text{transit},n} = \begin{cases} 1 & \text{if individual } n \text{ chooses auto} \\ 0 & \text{otherwise} \end{cases}$$

TRANSP-OR

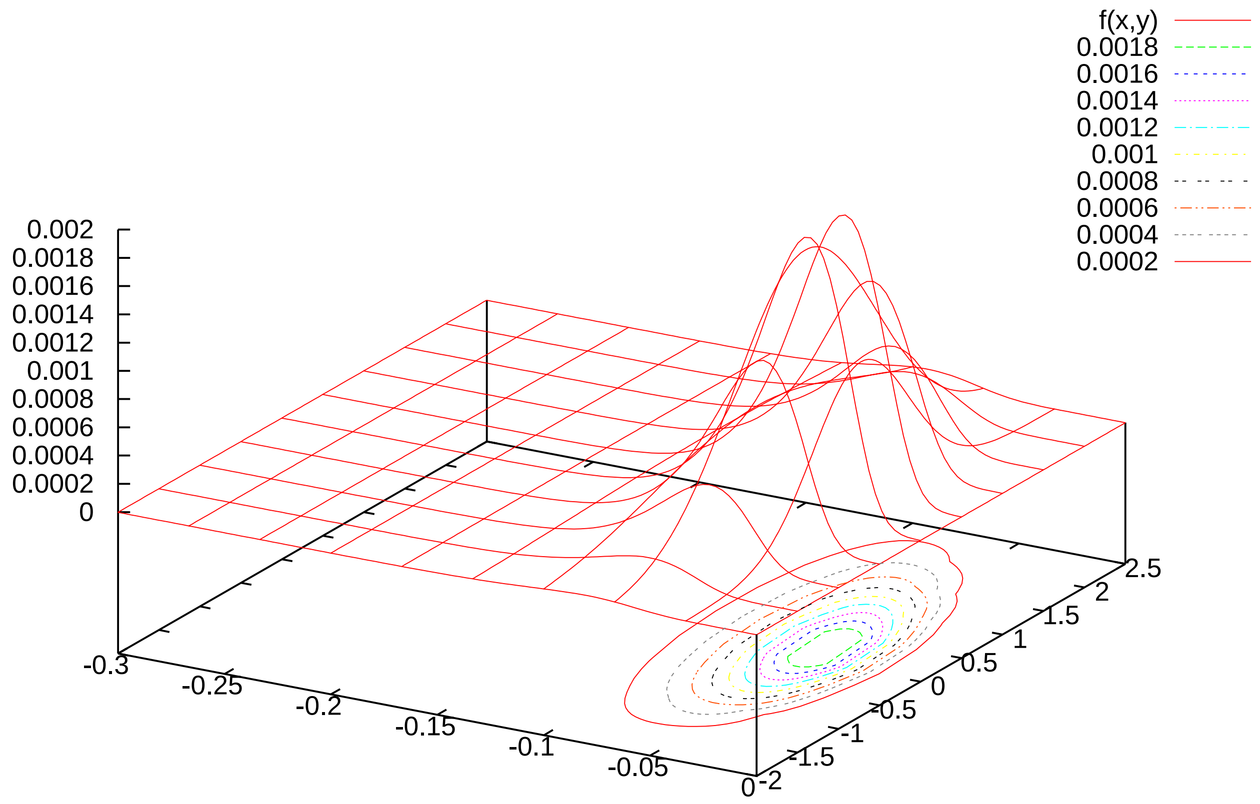
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Back to the example

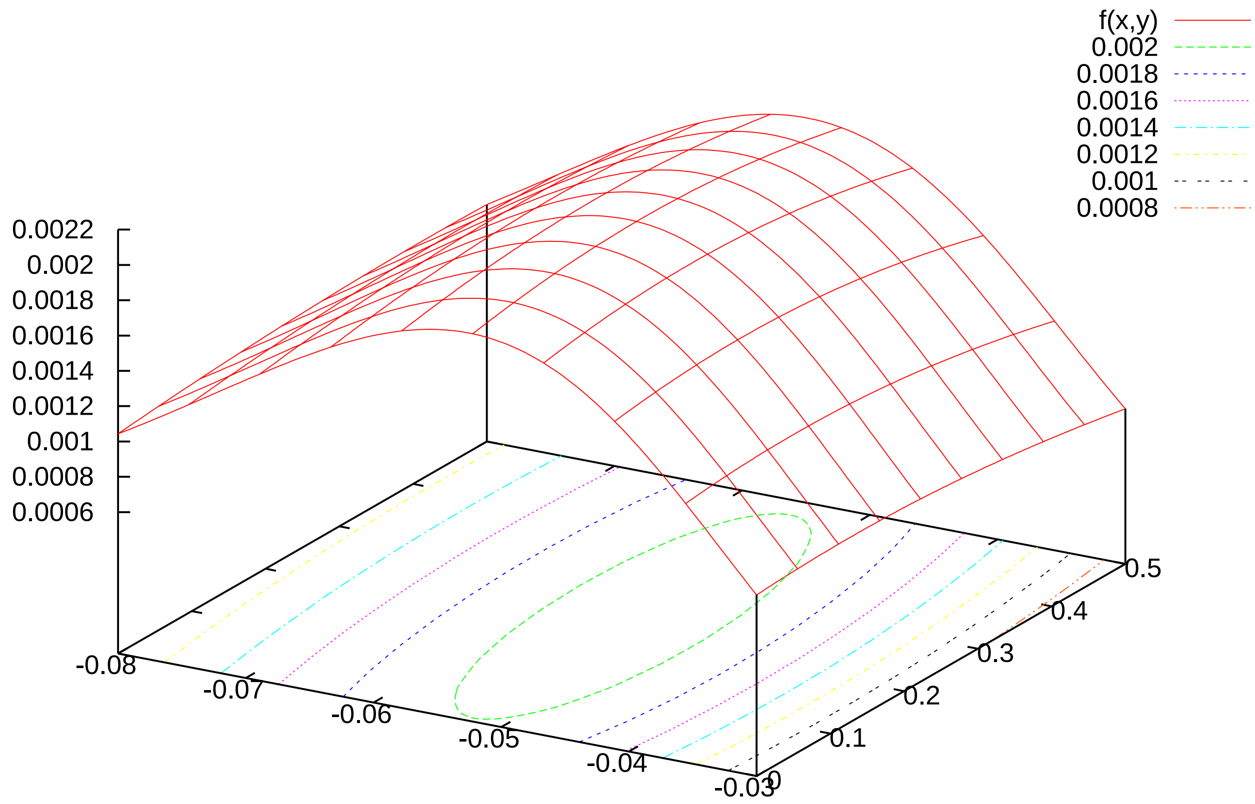
$\mathcal{L}^*$ is called the likelihood of the sample for a given model.
It is a probability.
We report this value for some values of β_0 and β_1

β_0	β_1	$\mathcal{L}^*$
0	0	4.57 10^{-07}
0	-1	1.97 10^{-30}
0	-0.1	4.1 10^{-04}
0.5	-0.1	4.62 10^{-04}

Back to the example



Back to the example



Maximum likelihood estimation

$$\max_{\beta} \prod_n \left(P_n(\text{auto})^{y_{\text{auto},n}} P_n(\text{transit})^{y_{\text{transit},n}} \right)$$

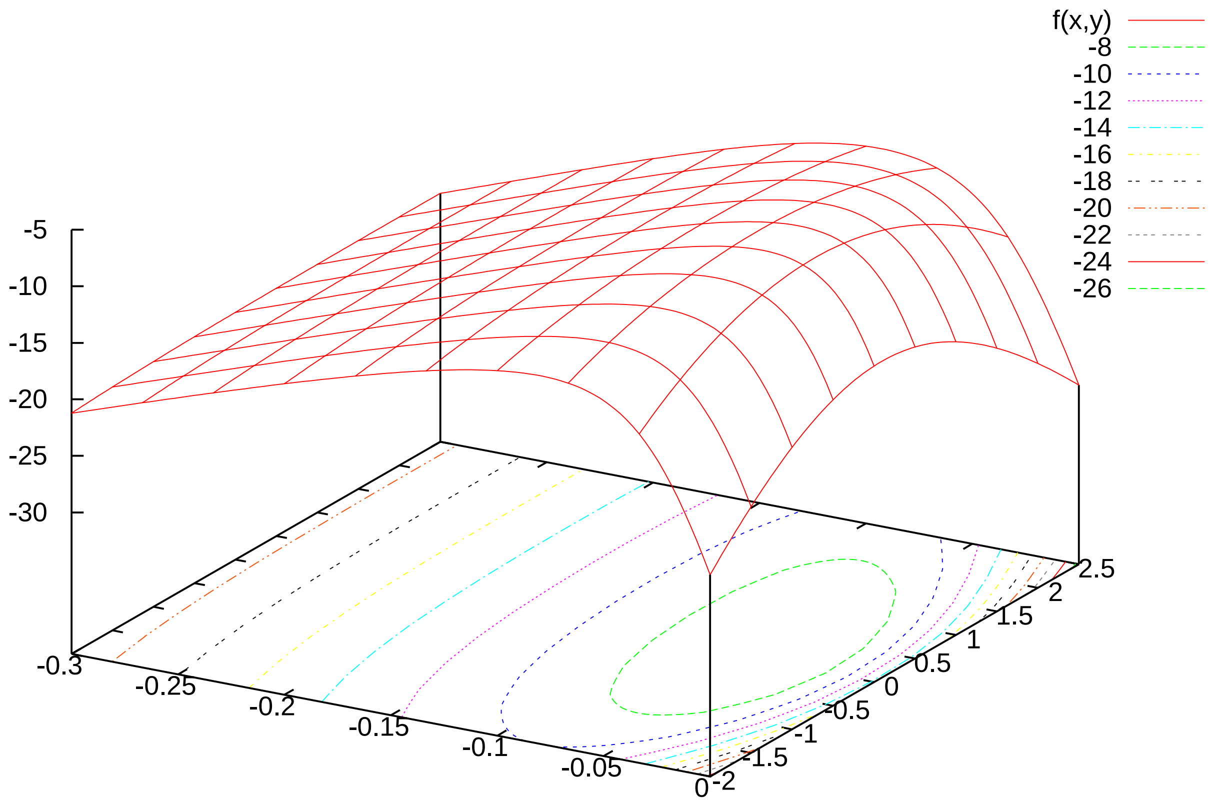
Alternatively, we prefer to maximize the log-likelihood

$$\max_{\beta} \log \prod_n \left(P_n(\text{auto})^{y_{\text{auto},n}} P_n(\text{transit})^{y_{\text{transit},n}} \right)$$

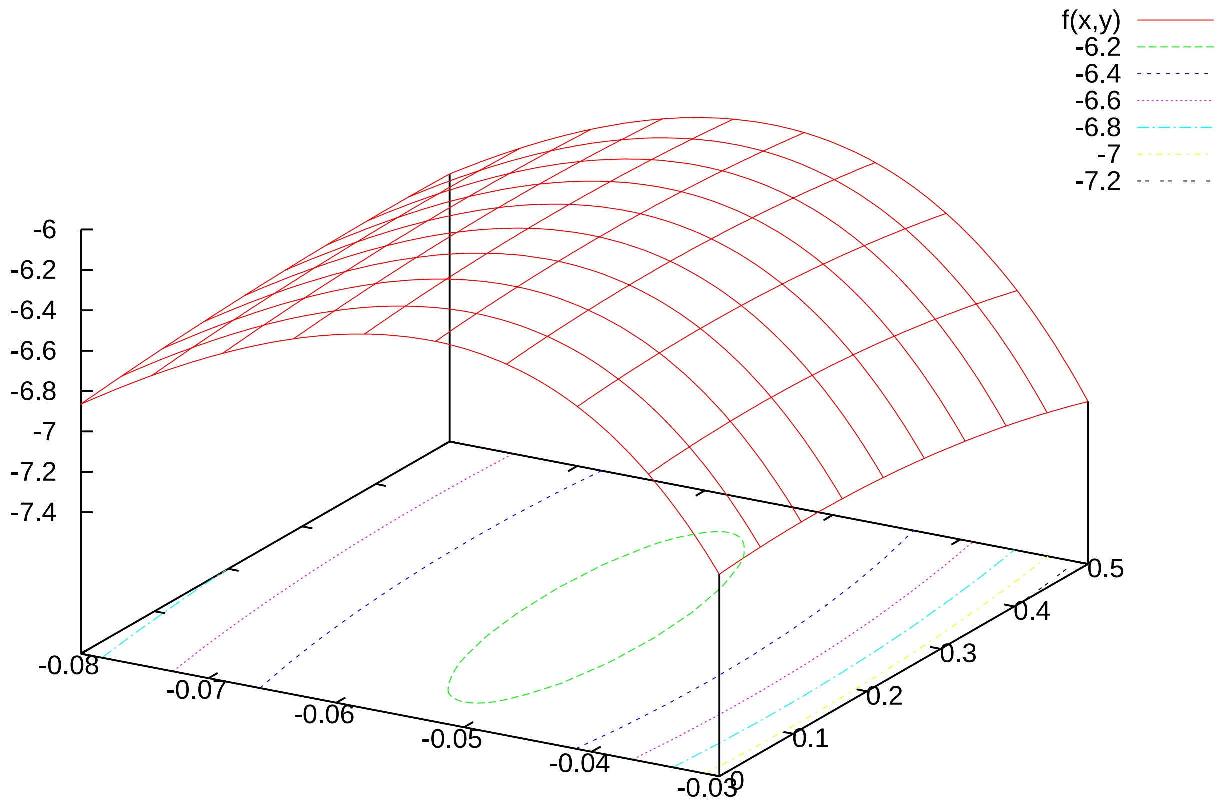
$$\max_{\beta} \sum_n \log \left(P_n(\text{auto})^{y_{\text{auto},n}} P_n(\text{transit})^{y_{\text{transit},n}} \right)$$

$$\max_{\beta} \sum_n y_{\text{auto},n} \log P_n(\text{auto}) + y_{\text{transit},n} \log P_n(\text{transit})$$

Maximum likelihood estimation



Maximum likelihood estimation

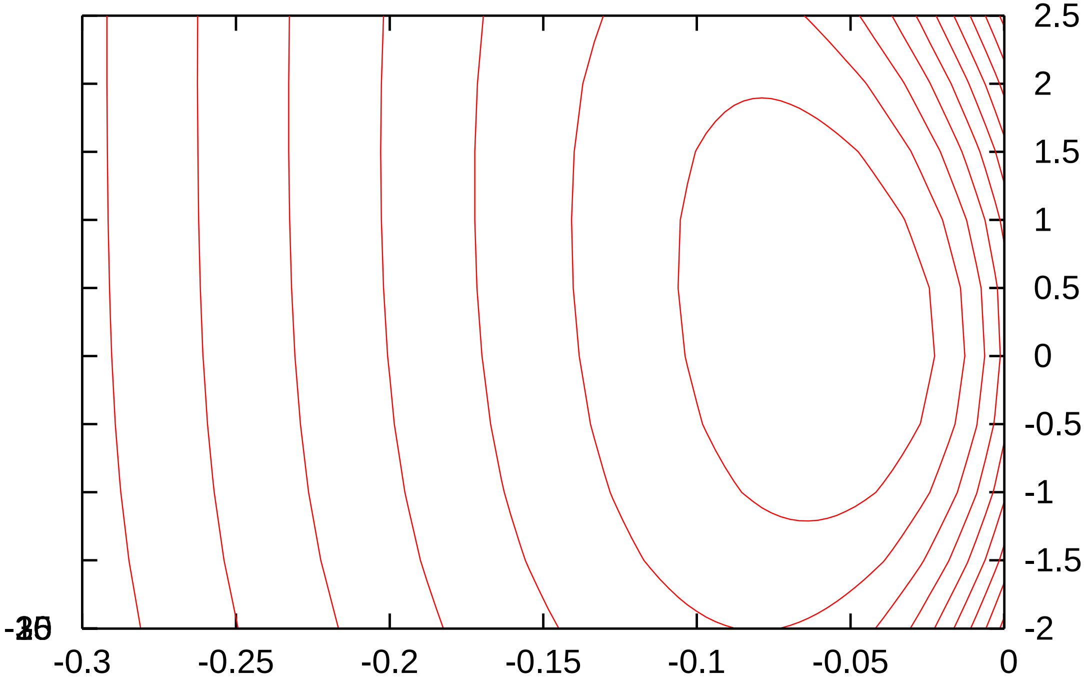


Nonlinear programming

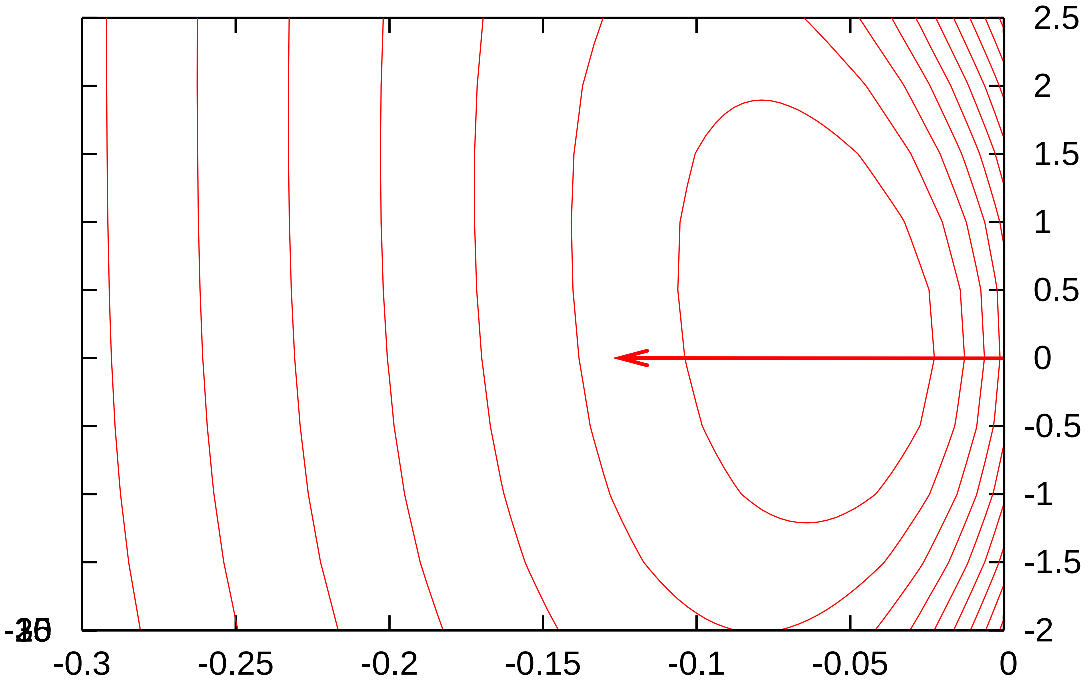
- Iterative methods
- Designed to identify a local maximum
- When the function is concave, a local maximum is also a global maximum
- For binary logit, the log-likelihood is concave
- Use the derivatives of the objective function

Example: package CFSQP used in BIOGEME

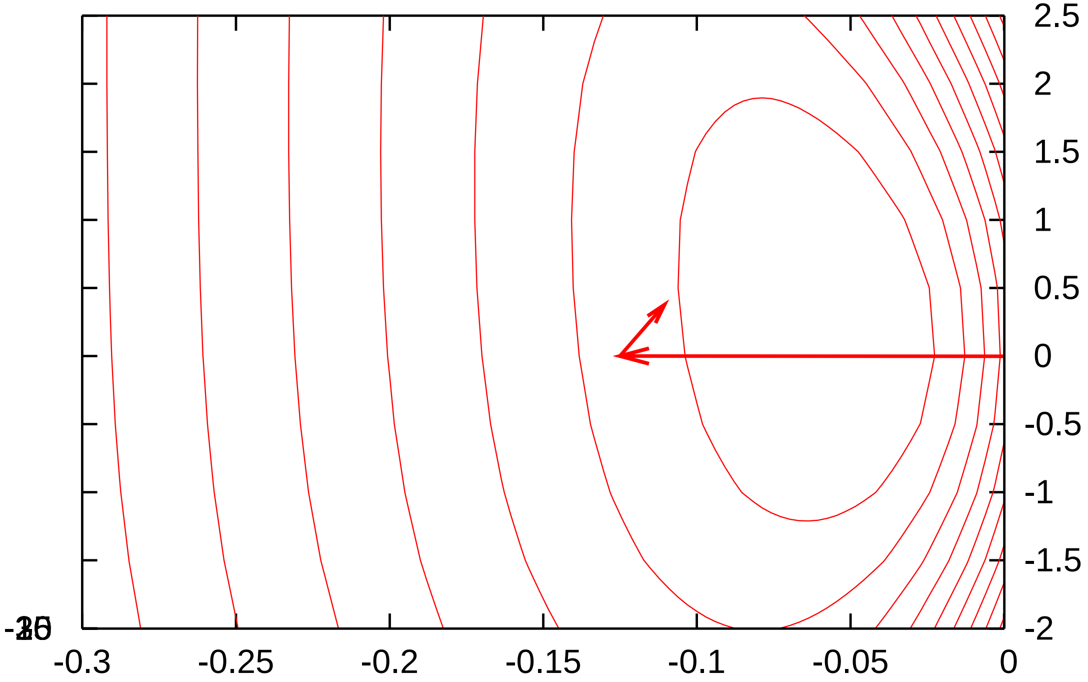
Nonlinear programming



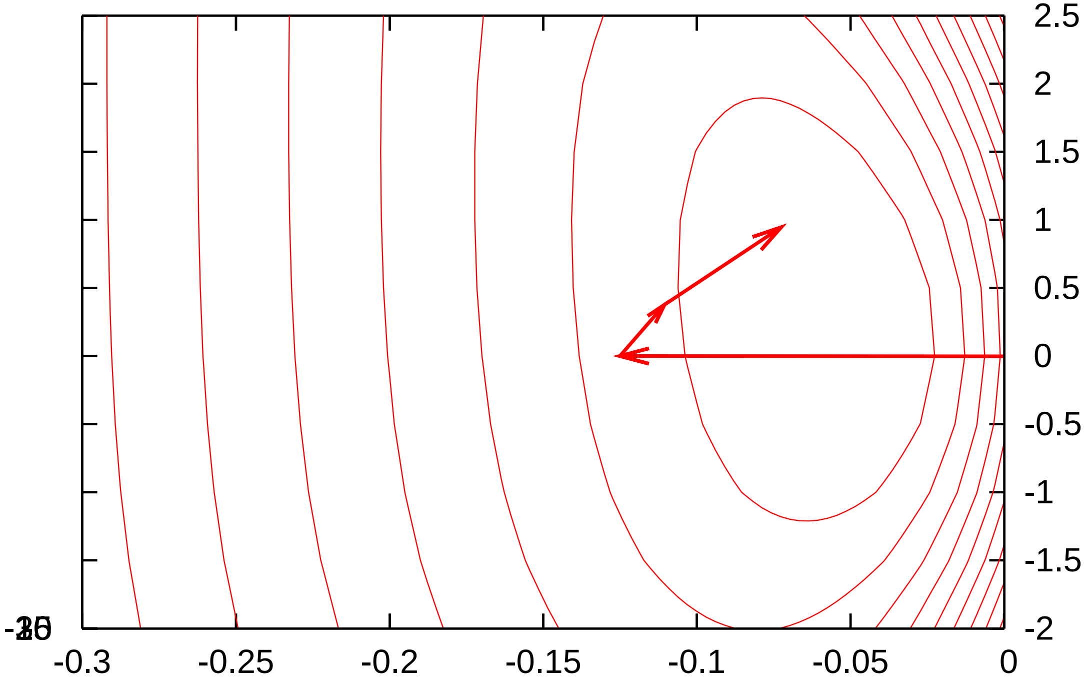
Nonlinear programming



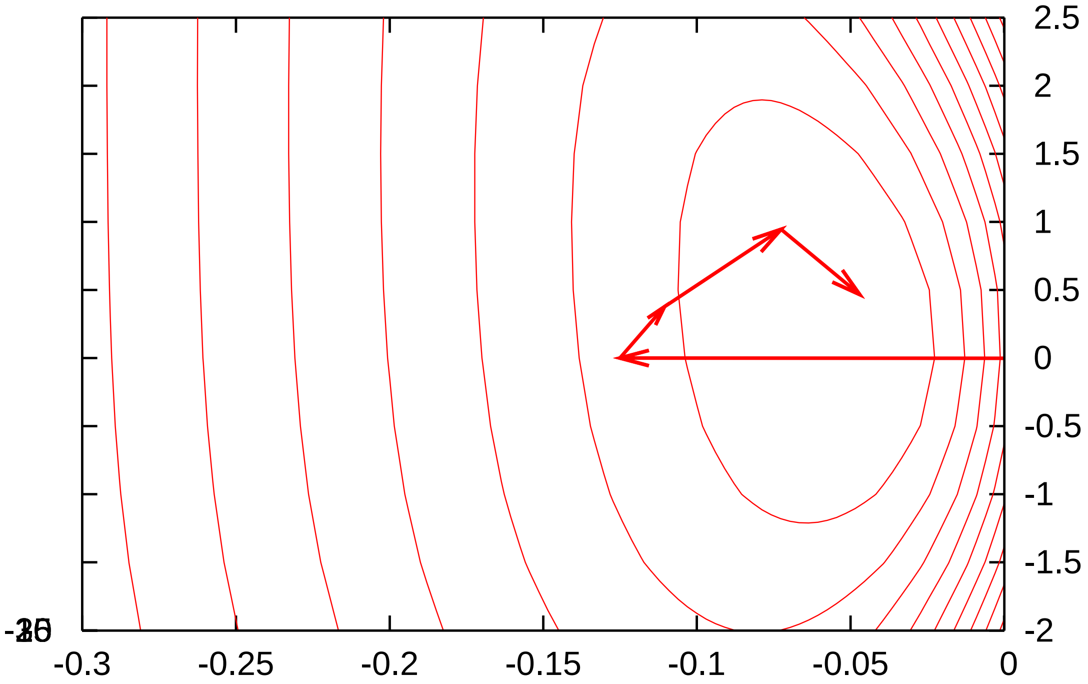
Nonlinear programming



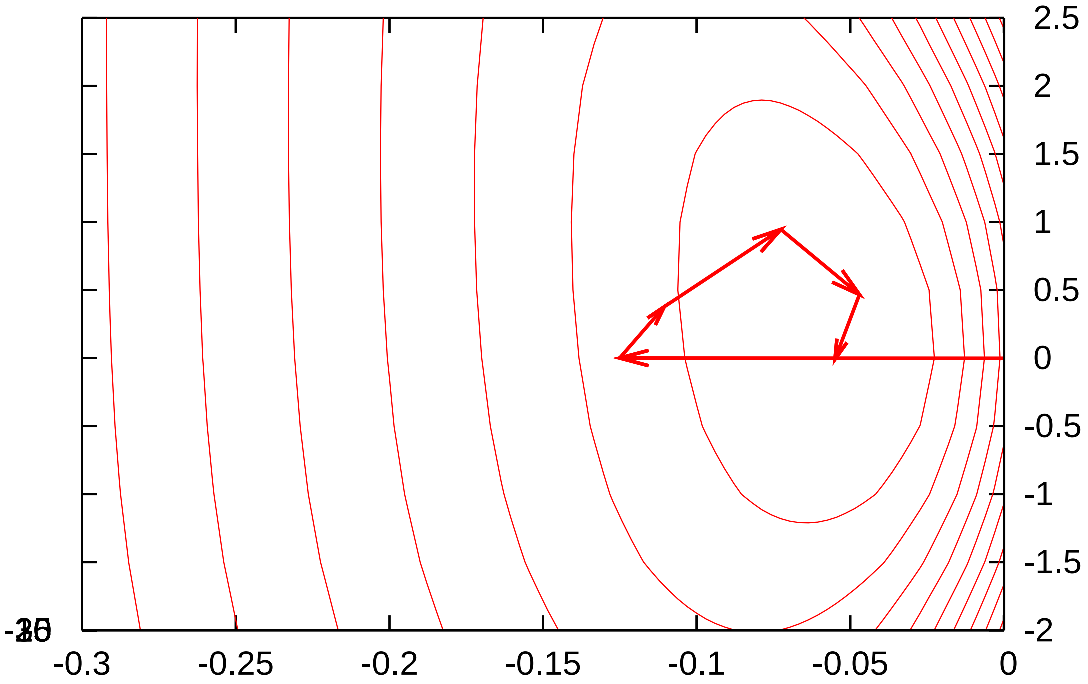
Nonlinear programming



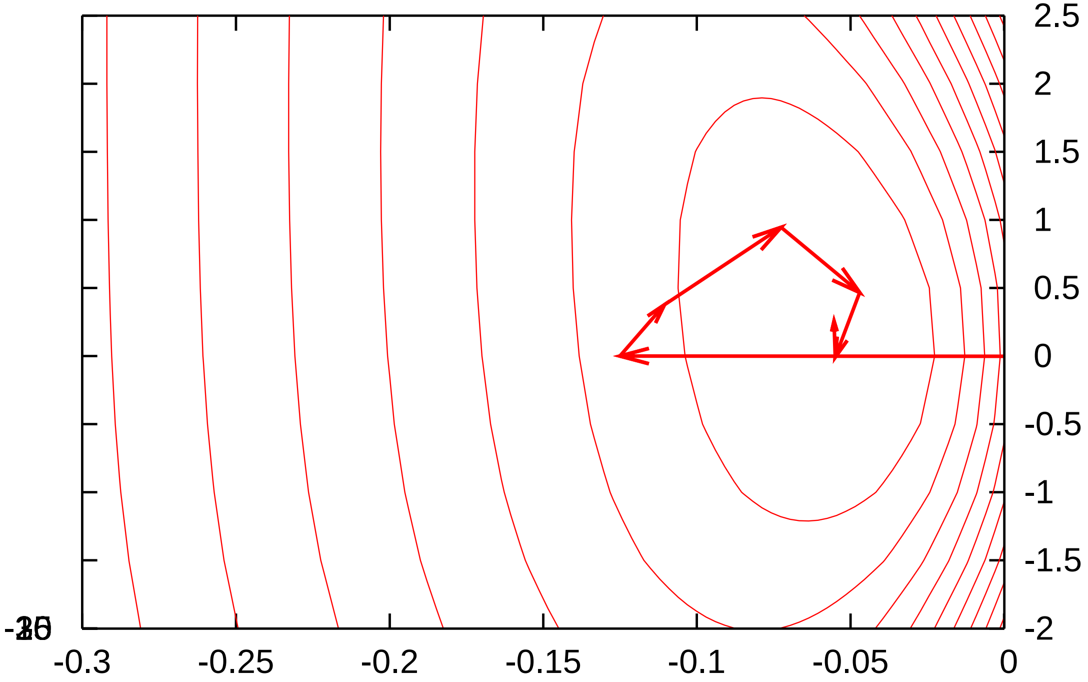
Nonlinear programming



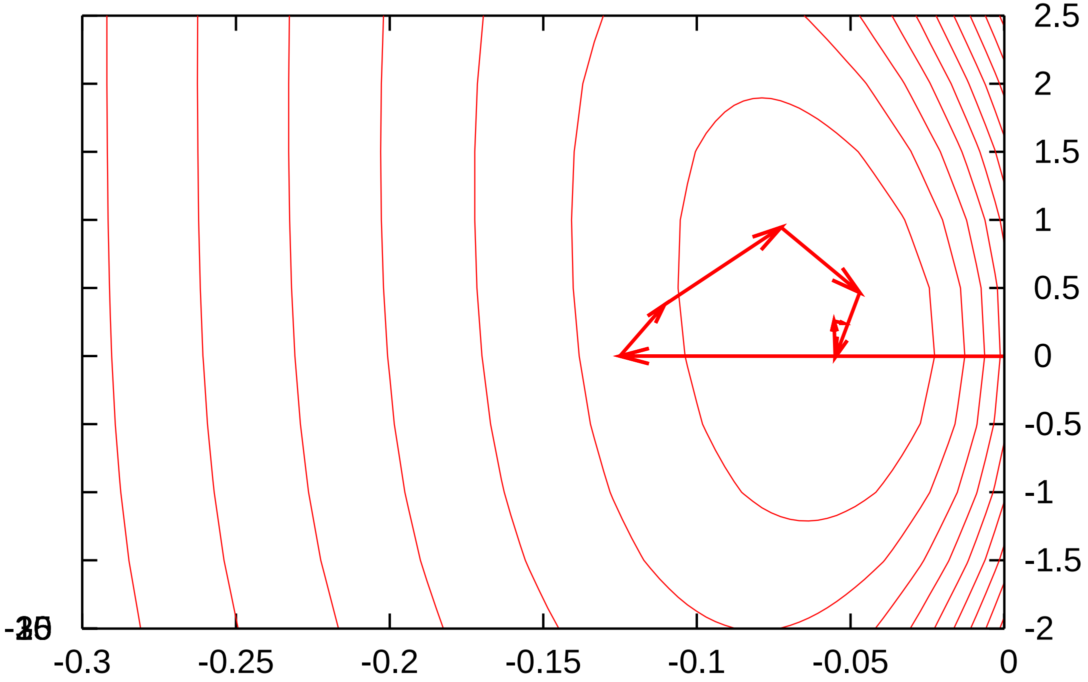
Nonlinear programming



Nonlinear programming



Nonlinear programming



Nonlinear programming

Things to be aware of

- Iterative methods terminate when a given stopping criterion is verified, based on the fact that, if β^* is the optimum,

$$\nabla\mathcal{L}(\beta^*) = 0$$

Stopping criteria usually vary across optimization packages, which may produce slightly different solutions
They are usually using a parameter defining the required precision

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Nonlinear programming

Tests with CFSQP package within BIOGEME

Prec.	β_0^*	β_1^*	$\mathcal{L}^*(\beta^*)$	$\|\nabla\mathcal{L}^*(\beta^*)\|$
1.0	+0.0000e+00	+1.4901e−08	-14.56	456.05
1.0e-01	+2.5810e−01	-5.5361e−02	-6.172	4.9646
1.0e-02	+2.4274e−01	-5.2330e−02	-6.167	1.9711
1.0e-03	+2.3732e−01	-5.3146e−02	-6.166	0.089982
1.0e-04	+2.3758e−01	-5.3110e−02	-6.166	0.0015384
1.0e-05	+2.3757e−01	-5.3110e−02	-6.166	0.0015384

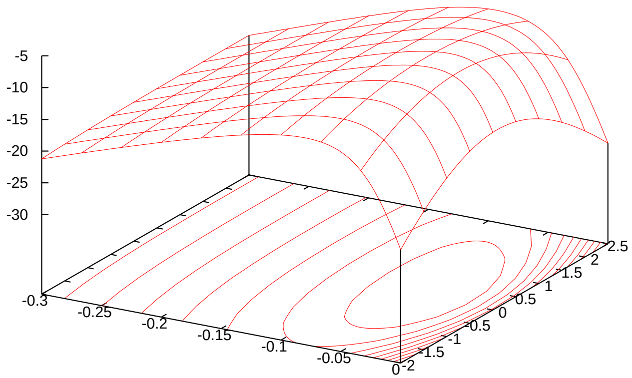
Nonlinear programming

Things to be aware of

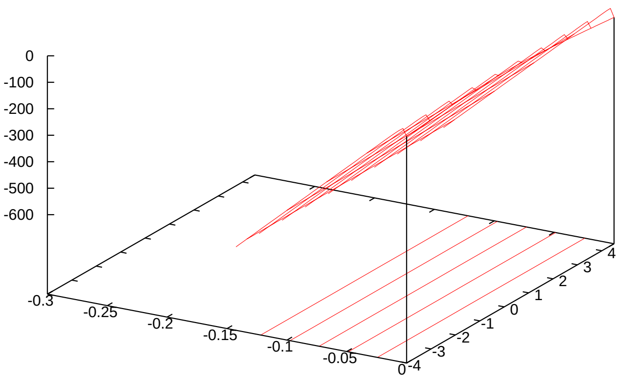
- Most methods are sensitive to the conditioning of the problem.

A well-conditioned problem is a problem for which all parameters have almost the same magnitude

Nonlinear programming

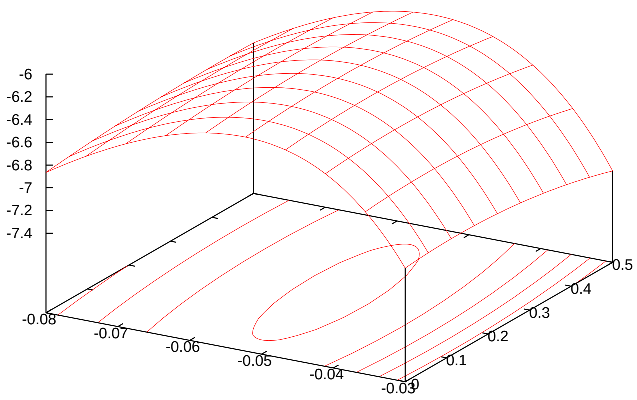


Time in min.

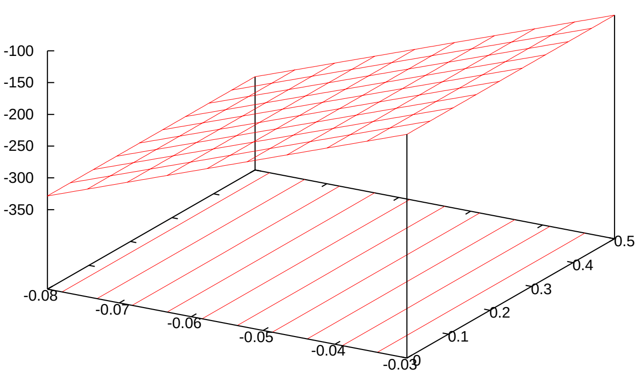


Time in sec.

Nonlinear programming



Time in min.



Time in sec.

Nonlinear programming

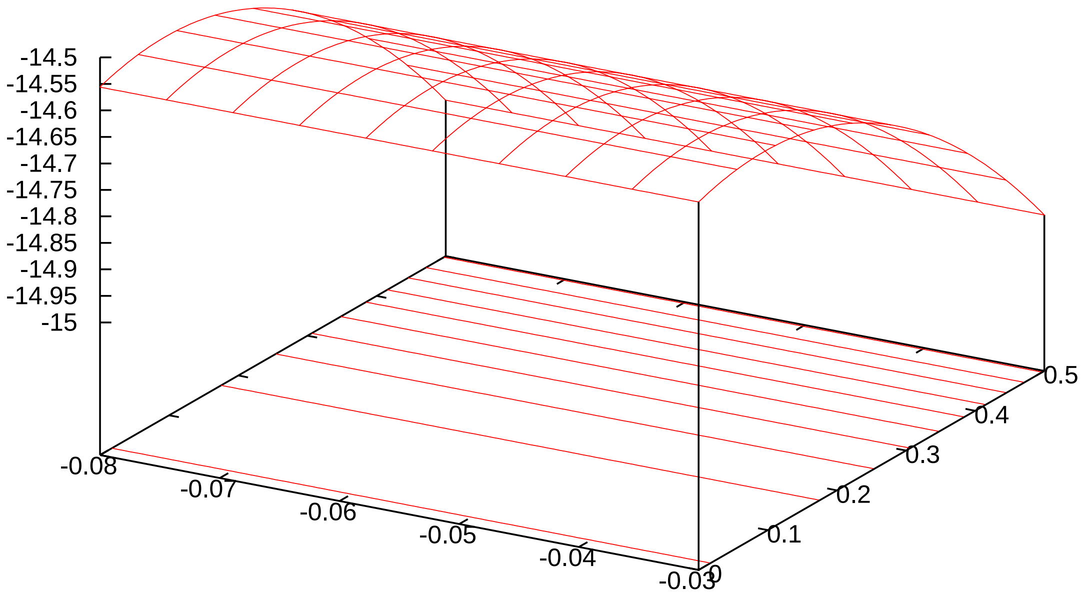
Things to be aware of

- Convergence may be very slow or even fail if the model is singular

A model is singular when some of its parameters are not identifiable

Example: all travel times are equal.

Nonlinear programming



Output of the estimation

$$\max_{\beta \in \mathbb{R}^K} \mathcal{L}(\beta)$$

Solution: β^* and $\mathcal{L}(\beta^*)$

Case study:

- $\beta_0^* = 0.2376$
- $\beta_1^* = -0.0531$
- $\mathcal{L}(\beta_0^*, \beta_1^*) = -6.166$

Output of the estimation

Information about the quality of the estimators.
Let

$$\nabla^2 \mathcal{L}(\beta^*) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \beta_1^2} & \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \partial \beta_K} \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \mathcal{L}}{\partial \beta_2^2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_2 \partial \beta_K} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_K \partial \beta_1} & \frac{\partial^2 \mathcal{L}}{\partial \beta_K \partial \beta_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_K^2} \end{pmatrix}$$

$-\nabla^2 \mathcal{L}(\beta^*)^{-1}$ is a consistent estimator of the variance-covariance matrix of the estimates

Output of the estimation

Parameter	Value	Std Err.	t-test
β_0	0.2376	0.7505	0.32
β_1	-0.0531	0.0206	-2.57

Summary statistics:

- $\mathcal{L}(\beta^*)$ = -6.166
- $\mathcal{L}(0)$ = -14.556
- $-2(\mathcal{L}(0) - \mathcal{L}(\beta^*))$ = 16.780
- ρ^2 = 0.576, $\bar{\rho}^2$ = 0.439

Output of the estimation

$\mathcal{L}(0)$ is the sample log-likelihood with a trivial model where all parameters are zero, that is a model always predicting

$$P(1|\{1,2\}) = P(2|\{1,2\}) = \frac{1}{2}$$

$$\mathcal{L}(0) = \log(\frac{1}{2^N}) = -N\log(2)$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Output of the estimation

$-2(\mathcal{L}(0) - \mathcal{L}(\beta^*))$ is the likelihood ratio.
Indeed,

$$\log\left(\frac{\bar{\mathcal{L}}(0)}{\bar{\mathcal{L}}(\beta^*)}\right) = \log(\bar{\mathcal{L}}(0)) - \log(\bar{\mathcal{L}}(\beta^*)) = \mathcal{L}(0) - \mathcal{L}(\beta^*)$$

$-2(\mathcal{L}(0) - \mathcal{L}(\beta^*))$ is asymptotically distributed as χ^2 with K degrees of freedom

Similar to the F test in regression models

Output of the estimation

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

Similar to the R^2 in regression models

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\beta^*) - K}{\mathcal{L}(0)}$$

Comparing models

- Arbitrary scale may be problematic when comparing models
- Binary probit: $\sigma^2 = \text{Var}(\varepsilon_i - \varepsilon_j) = 1$
- Binary logit: $\text{Var}(\varepsilon_i - \varepsilon_j) = \pi^2/(3\mu) = \pi^2/3$
- $\text{Var}(\mu U) = \mu^2 \text{ Var } (U)$.
- Scaled logit coeff. are $\pi/\sqrt{3}$ larger than scaled probit coeff.

Comparing models

Same example ($\pi/\sqrt{3} \approx 1.814$)

	Probit	Logit	Probit * $\pi/\sqrt{3}$
$\mathcal{L}$	-6.165	-6.166	
β_0	0.064	0.238	0.117
β_1	-0.030	-0.053	-0.054