# Binary Choice

Matthieu de Lapparent
matthieu.delapparent@epfl.ch

Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne

# Outline

# Example

## Data

- Unit of analysis: travelers (simulated observations)
- Choice set: choice of car (C) or transit (T)
- Independent variable: travel time

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

# Example

## Data from 21 decision makers

| # | 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

## Specification of utility functions

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

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

## Choice model

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

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

# Error term

Three questions about the random variables $\varepsilon_T$ and $\varepsilon_C$

1. What's their distribution?
2. What's their moments:
   1. Mean?
   2. Variance?

Note

For binary choice it is sufficient to make assumptions about $\varepsilon = \varepsilon_T - \varepsilon_C$

# First-order moment: mean

**Note**

Adding the same constant $\mu$ to all utility functions does not affect the choice model

$$\Pr(U_C \geq U_T) = \Pr(U_C + \mu \geq U_T + \mu) \quad \forall \mu \in \mathbb{R}.$$

**Why?**

An utility function is defined up to a monotone increasing transformation.

# First-order moment: mean, cont.

## Change of variables

- Assume that $\mathrm{E}[\varepsilon_C] = \beta_C$ and $\mathrm{E}[\varepsilon_T] = \beta_T$.
- Define $\varepsilon'_C = \varepsilon_C - \beta_C$ and $\varepsilon'_T = \varepsilon_T - \beta_T$,
- so that $\mathrm{E}[\varepsilon'_C] = \mathrm{E}[\varepsilon'_T] = 0$.

## Choice model

$P(C|\{C, T\}) =$

$$\begin{array}{lll}
\Pr(\beta_1(T_C - T_T) & \geq & \varepsilon_T - \varepsilon_C) = \\
\Pr(\beta_1(T_C - T_T) & \geq & \varepsilon'_T + \beta_T - \varepsilon'_C - \beta_C) = \\
\Pr(\beta_1(T_C - T_T) + (\beta_C - \beta_T) & \geq & \varepsilon'_T - \varepsilon'_C) = \\
\Pr(\beta_1(T_C - T_T) + \beta_0 & \geq & \varepsilon')
\end{array}$$

where $\beta_0 = \beta_C - \beta_T$ and $\varepsilon' = \varepsilon'_T - \varepsilon'_C$.

# First-order moment: mean, cont.

## Mean

- The mean of $\varepsilon$ can be included as a parameter of the deterministic part of utility
- Only the mean of the difference of the error terms is meaningful

## Alternative Specific Constant (ASC)

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

In practice, one needs to associate an ASC with all alternatives but one: exclusion constraint to define a one-to-one mapping between vector of parameters and choice probabilities

# Second-order moment: the variance

### Utility is ordinal

Utilities can be scaled up or down without changing the choice probability

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

### Repeat once more!

A utility function is defined up to a monotone increasing transformation.

### Link with the variance

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

### Variance is not identified

- As any $\alpha$ can be selected arbitrarily, any variance can be assumed.
- No way to identify the variance of the error terms from data.

# Practical summary

### Only difference in levels of utility matters

It is not possible to estimate all ASC but only their differences. Choose arbitrarily one of the ASCs as reference and fix it to 0: estimated differences of ASCs are wrt to this reference

### Scale is arbitrary

It means for a linear utility function that the values of the parameters are not sensible.

# The normal distribution

### Assumption 1

$\varepsilon_T$ and $\varepsilon_C$ 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 approximately follows a normal distribution: $N(\mu, \sigma^2)$.

### Assumed distribution

$$\varepsilon_C \sim N(0,1), \quad \varepsilon_T \sim N(0,1), \varepsilon_C \perp \varepsilon_T$$

# The normal distribution, cont.

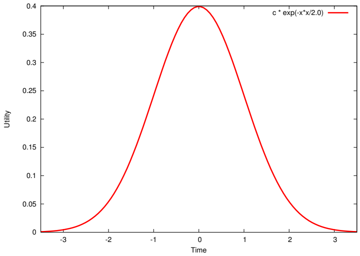
Probability density function (pdf):

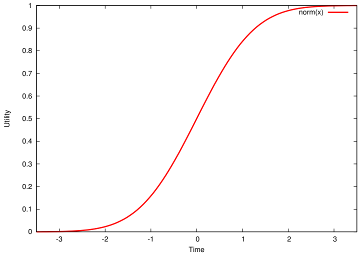
$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Cumulative distribution function (CDF)

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

No closed form



# The normal distribution, cont.

$\varepsilon = \varepsilon_T - \varepsilon_C$

- From the properties of the normal distribution, we have

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

- As the variance is arbitrary, we may also assume

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

# The binary probit model

## Choice model

$$P(C|\{C,T\}) = \Pr(\beta_1(T_C - T_T) + \beta_0 \geq \varepsilon) = F_\varepsilon(\beta_1(T_C - T_T) + \beta_0)$$

## The binary probit model

$$P(C|\{C,T\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1(T_C - T_T) - \beta_0} e^{-\frac{1}{2}t^2} dt$$

Not a closed form expression

# The binary probit model

### The distribution

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

Probability Unit Model or Probit Model.

# The Gumbel distribution

### Assumption 2

$\varepsilon_T$ and $\varepsilon_C$ 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: $\text{EV}(\eta, \mu)$.

### Assumed distribution

$$\varepsilon_C \sim \text{EV}(0,1), \quad \varepsilon_T \sim \text{EV}(0,1), \varepsilon_C \perp \varepsilon_T$$

# The type 1 Extreme Value distribution EV1$(\eta, \mu)$

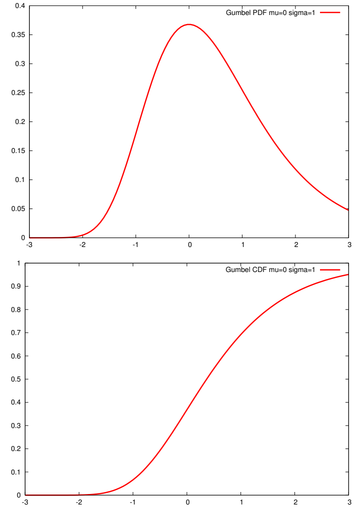
## Probability density function (pdf)

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

## Cumulative distribution function (CDF)

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

# The type 1 Extreme Value distribution



# The type 1 Extreme Value distribution

## Properties

If

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

then

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

where $\gamma$ is Euler's constant.

## Euler's constant

$$\gamma = \lim_{k \to \infty} \sum_{i=1}^{k} \frac{1}{i} - \ln k = -\int_0^{\infty} e^{-x} \ln x dx \approx 0.5772$$

# Difference of independent type 1 Extreme Value distributions

$\varepsilon = \varepsilon_T - \varepsilon_C$

From the properties of the extreme value distribution, we have

$$
\begin{aligned}
\varepsilon_C &\sim \text{EV}(0, 1) \\
\varepsilon_T &\sim \text{EV}(0, 1) \\
\varepsilon &\sim \text{Logistic}(0, 1)
\end{aligned}
$$

# The Logistic distribution: Logistic($\eta$,$\mu$)

**Probability density function (pdf)**

$$f(t) = \frac{\mu e^{-\mu(t-\eta)}}{(1 + e^{-\mu(t-\eta)})^2}$$

**Cumulative distribution function (CDF)**

$$P(c \geq \varepsilon) = F(c) = \int_{-\infty}^{c} f(t)dt = \frac{1}{1 + e^{-\mu(c-\eta)}}$$

with $\mu > 0$.

# The binary logit model

## Choice model

$$P(C|\{C,T\}) = \Pr(\beta_1(T_C - T_T) + \beta_0 \geq \varepsilon) = F_\varepsilon(\beta_1(T_C - T_T) + \beta_0)$$

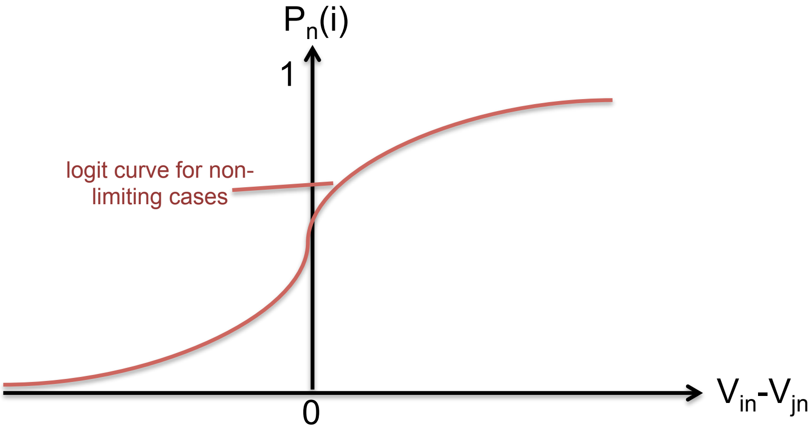
## The binary logit model

$$P(C|\{C,T\}) = \frac{1}{1 + e^{-(\beta_1(T_C - T_T) + \beta_0)}} = \frac{e^{\beta_1 T_C + \beta_0}}{e^{\beta_1 T_C + \beta_0} + e^{\beta_1 T_T}}$$

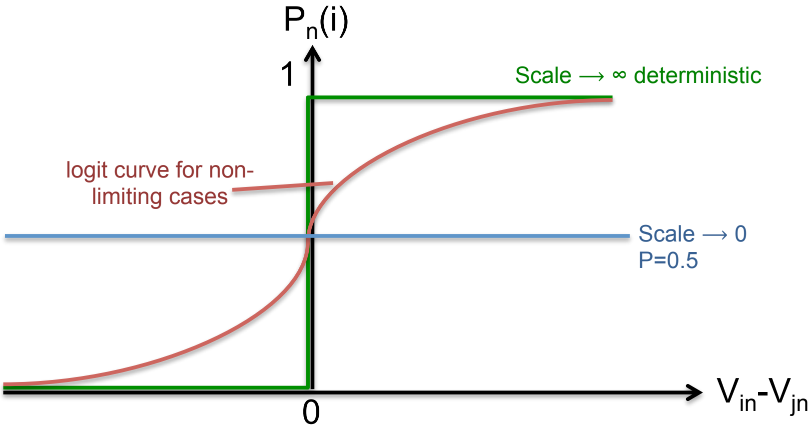
## The binary logit model

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

# Logit curve



# Logit curve: limiting cases



# Back to the example

## Remember the data from our 21 decision makers?

| # | 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 |

# First individual

### Parameters

Let's assume that $\beta_0 = 0.5$ and $\beta_1 = -0.1$

### Variables

Let's consider the first observation:

- $T_{C1} = 52.9$
- $T_{T1} = 4.4$
- Choice $=$ *transit*: $y_{\text{auto},1} = 0$, $y_{\text{transit},1} = 1$

### Choice

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

# First individual

## Utility functions

$$\begin{array}{rcl} V_{C1} & = & \beta_1 T_{C1} & = & -5.29 \\ V_{T1} & = & \beta_1 T_{T1} + \beta_0 & = & 0.06 \end{array}$$

## Choice model

$$P_1(\text{transit}) = \frac{e^{V_{T1}}}{e^{V_{T1}} + e^{V_{C1}}} = \frac{e^{0.06}}{e^{0.06} + e^{-5.29}} \cong 1$$

## Comments

- The model fits the observation very well.
- Consistent with the assumption that travel time is the only explanatory variable.

# Second individual

## Parameters

Let's assume that $\beta_0 = 0.5$ and $\beta_1 = -0.1$

## Variables

- $T_{C2} = 4.1$
- $T_{T2} = 28.5$
- Choice $=$ *transit*: $y_{\text{auto},2} = 0$, $y_{\text{transit},2} = 1$

## Choice

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

# Second individual

### Utility functions

$$\begin{array}{rcl} V_{C2} & = & \beta_1 T_{C2} & = & -0.41 \\ V_{T2} & = & \beta_1 T_{T2} + \beta_0 & = & -2.35 \end{array}$$

### Choice model

$$P_2(\text{transit}) = \frac{e^{V_{T2}}}{e^{V_{T2}} + e^{V_{C2}}} = \frac{e^{-2.35}}{e^{-2.35} + e^{-0.41}} \cong 0.13$$

### Comment

- The model fits the observation poorly.
- But the assumption is that travel time is the only explanatory variable.
- Still, the probability is not small.

# Back to the example

### Two observations

The probability that the model reproduces both observations is

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

### All observations

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

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

where $y_{j,n}$ is 1 if individual $n$ has chosen alternative $j$, 0 otherwise

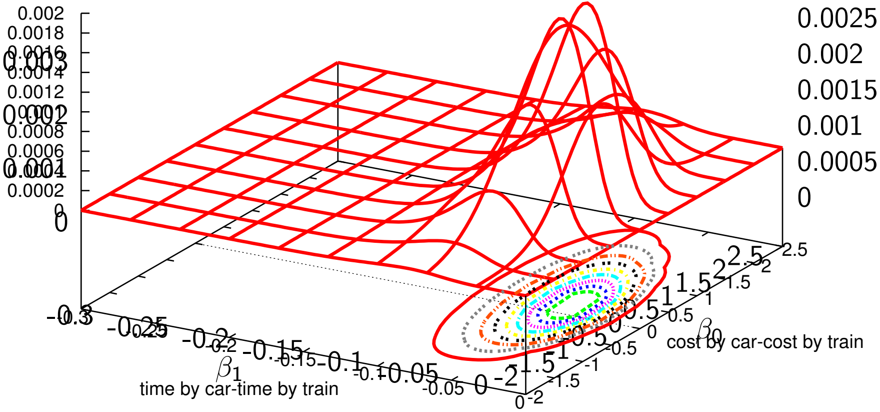
# Back to the example

- $\mathcal{L}^*$ is called the likelihood of the sample for a given model.
- Probability that the model fits all observations
- It is a function of the parameters

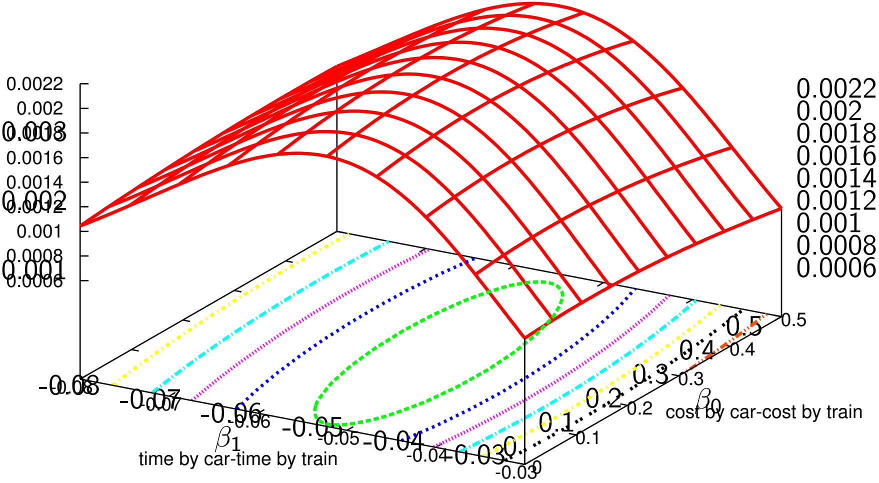
Examples for some values of $\beta_0$ and $\beta_1$

| $\beta_0$ | $\beta_1$ | $\mathcal{L}^*$ |
|---|---|---|
| 0 | 0 | $4.57\ 10^{-07}$ |
| 0 | -1 | $1.97\ 10^{-30}$ |
| 0 | -0.1 | $4.1\ 10^{-04}$ |
| 0.5 | -0.1 | $4.62\ 10^{-04}$ |

# Likelihood function



# Likelihood function (zoom)



# Maximum likelihood estimation

## Estimators for the parameters

Parameters that achieve the maximum likelihood

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

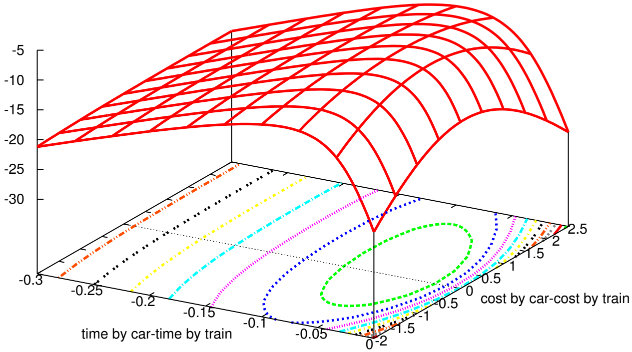
## Log likelihood

Alternatively, we prefer to maximize the log likelihood

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

$$\max_{\beta} \sum_{n} \ln \left( y_{\text{auto},n} P_n(\text{auto}) + y_{\text{transit},n} P_n(\text{transit}) \right)$$

# Maximum likelihood estimation



# Solving the optimization problem

## Unconstrained nonlinear optimization

- 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

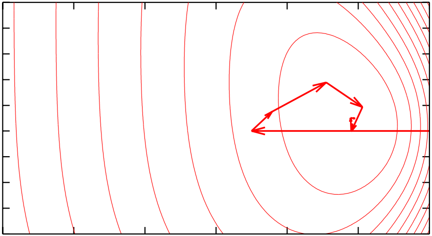
Example: package CFSQP used in BIOGEME

# Example of algorithm

Tests with CFSQP package within BIOGEME

| Prec. | $\beta_0^*$ | $\beta_1^*$ | $\mathcal{L}^*(\beta^*)$ | $\|\nabla\mathcal{L}^*(\beta^*)\|$ |
|---|---|---|---|---|
| 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 |

# Example of algorithm: CFSQP



# Nonlinear optimization

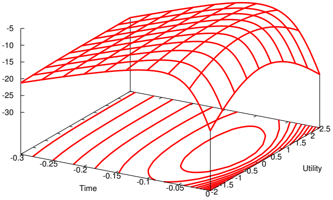
Things to be aware of...

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

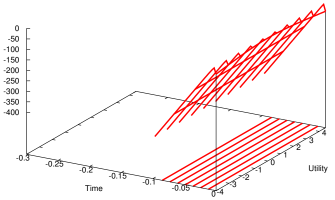
$$\nabla \ln \mathcal{L}(\beta^*) = 0$$

- Stopping criteria vary across optimization packages (based on required precision) $\rightarrow$ slightly different solutions
- Most methods are sensitive to the conditioning of the problem
- A well-conditioned problem $\rightarrow$ all parameters have almost the same magnitude

# Nonlinear optimization




Time in min.



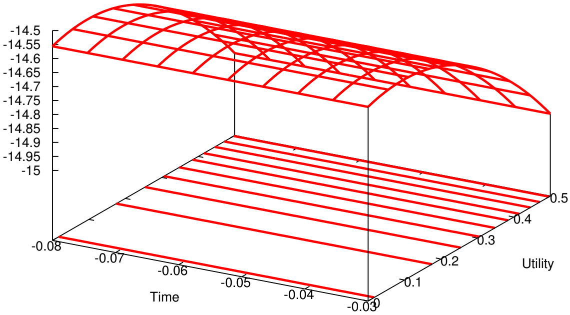

Time in sec.

# Nonlinear optimization

Things to be aware of...

- Convergence may be very slow or even fail if likelihood function is flat
- It happens when the model is not identifiable
- Structural flaw in the model (e.g. full set of alternative specific constants)
- Lack of variability in the data (all prices are the same across the sample)

# Nonlinear programming



# Output of the estimation

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

- $\beta^*$
- $\ln \mathcal{L}(\beta^*)$

### Case study

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

# Second derivatives

Information about the quality of the estimators.

Let

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

$-\nabla^2 \ln \mathcal{L}(\beta^*)^{-1}$ is a consistent estimator of the variance-covariance matrix of the estimates... if the assumed distribution is "the true one"!

# Statistics

## Statistics on the parameters

| Parameter | Value | Std Err. | $t$-test |
|---|---|---|---|
| $\beta_0$ | 0.2376 | 0.7505 | 0.32 |
| $\beta_1$ | -0.0531 | 0.0206 | -2.57 |

## Summary statistics

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

# Null log likelihood

**$\ln \mathcal{L}(0)$**

sample log likelihood with a trivial model where all parameters are zero, that is a model always predicting

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

**Purely a function of sample size**

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

# Likelihood ratio

$-2(\ln \mathcal{L}(0) - \ln \mathcal{L}(\beta^*))$

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

## Likelihood ratio test

- $H_0$: the two models are equivalent
- Under $H_0$, $-2(\ln \mathcal{L}(0) - \ln \mathcal{L}(\beta^*))$ is asymptotically distributed as $\chi^2$ with $K$ degrees of freedom (K is the difference between the number of parameters in the full model and the number of parameters in the restricted model. The 2 models needs to be nested).
- Similar to the $F$ test in regression models

# Rho (bar) squared

$\rho^2$

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

Similar to the $R^2$ in regression models

$\bar{\rho}^2$

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

# Comparing models

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

# Comparing models

## Estimation results

| | Probit | Logit | Probit * $\pi/\sqrt{3}$ |
|---|---|---|---|
| $\mathcal{L}$ | -6.165 | -6.166 | |
| $\beta_0$ | 0.064 | 0.238 | 0.117 |
| $\beta_1$ | -0.030 | -0.053 | -0.054 |

Note: $\pi/\sqrt{3} \approx 1.814$

# Appendix

# Maximum likelihood for binary logit

- Let $\mathcal{C}_n = \{i, j\}$
- Let $y_{in} = 1$ if $i$ is chosen by $n$, 0 otherwise
- Let $y_{jn} = 1$ if $j$ is chosen by $n$, 0 otherwise
- Obviously, $y_{in} = 1 - y_{jn}$
- Log-likelihood of the sample

$$\sum_{n=1}^{N} \left( y_{in} \ln \frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}} + y_{jn} \ln \frac{e^{V_{jn}}}{e^{V_{in}} + e^{V_{jn}}} \right)$$