# 9: Inferential Statistics I of Discrete-Choice Models: Maximum-Likelihood Estimation

- 9.1. Maximum-Likelihood Estimation
- 9.2 Estimation Errors: Variance-Covariance Matrix
  - 9.2.1 Example 1: SP Survey in the Audience
  - 9.2.2 Example 2: RP Survey in the Audience

# 9.1. Maximum-Likelihood Estimation: the likelihood function

- The **maximum-likelihood (ML)** estimation is applicable for general stochastic models where the probabilities depend on a parameter vector $\boldsymbol{\beta}$
- The goal is to maximize the **likelihood function $L(\boldsymbol{\beta})$**, i.e., the probability that the model predicts *all* data points $(\boldsymbol{y}_n, \boldsymbol{x}_n)$, $n = 1, ..., N$:

$$L(\boldsymbol{\beta}) = P\left( \hat{\boldsymbol{y}}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \quad ..., \quad \hat{\boldsymbol{y}}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N \right)$$

  where $\hat{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}(\boldsymbol{x}_n)$ gives the model estimate for $\boldsymbol{x}_n$
- For continuous endogenous variables, the likelihood function is given by the multi-dimensional probability density at the data points:

$$L(\boldsymbol{\beta}) = f_{\hat{\boldsymbol{y}}_1(\boldsymbol{\beta}), ..., \hat{\boldsymbol{y}}_N(\boldsymbol{\beta})}(\boldsymbol{y}_1, ..., \boldsymbol{y}_N)$$

? Verify that the density formulation is equivalent to the probability definition by requiring the model estimations to be in small intervals around the data instead of hitting the data exactly.

! The multi-dimensional probability density $f(.)$ is defined such that $dP = f_{\boldsymbol{y}_1, ..., \boldsymbol{y}_N}(\boldsymbol{y}) d^N \boldsymbol{y}$. Keeping $d^N \boldsymbol{y}$ small and constant, $dP$ and thus $P$ is maximized if and only if $f(.)$ is maximized.

# 9.1. Maximum-Likelihood Estimation: the likelihood function

- The **maximum-likelihood (ML)** estimation is applicable for general stochastic models where the probabilities depend on a parameter vector $\boldsymbol{\beta}$
- The goal is to maximize the **likelihood function $L(\boldsymbol{\beta})$**, i.e., the probability that the model predicts *all* data points $(\boldsymbol{y}_n, \boldsymbol{x}_n)$, $n = 1, ..., N$:

$$\boxed{L(\boldsymbol{\beta}) = P\big(\ \hat{\boldsymbol{y}}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \quad ..., \quad \hat{\boldsymbol{y}}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N\ \big)}$$

where $\hat{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}(\boldsymbol{x}_n)$ gives the model estimate for $\boldsymbol{x}_n$

- For continuous endogenous variables, the likelihood function is given by the multi-dimensional probability density at the data points:

$$L(\boldsymbol{\beta}) = f_{\hat{\boldsymbol{y}}_1(\boldsymbol{\beta}), ..., \hat{\boldsymbol{y}}_N(\boldsymbol{\beta})}(\boldsymbol{y}_1, ..., \boldsymbol{y}_N)$$

? Verify that the density formulation is equivalent to the probability definition by requiring the model estimations to be in small intervals around the data instead of hitting the data exactly.

! The multi-dimensional probability density $f(.)$ is defined such that $\mathrm{d}P = f_{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N}(\boldsymbol{y})\mathrm{d}^N\boldsymbol{y}$. Keeping $\mathrm{d}^N\boldsymbol{y}$ small and constant, $\mathrm{d}P$ and thus $P$ is maximized if and only if $f(.)$ is maximized.

# 9.1. Maximum-Likelihood Estimation: the likelihood function

- The **maximum-likelihood (ML)** estimation is applicable for general stochastic models where the probabilities depend on a parameter vector $\boldsymbol{\beta}$
- The goal is to maximize the **likelihood function $L(\boldsymbol{\beta})$**, i.e., the probability that the model predicts *all* data points $(\boldsymbol{y}_n, \boldsymbol{x}_n)$, $n = 1, ..., N$:

$$\boxed{L(\boldsymbol{\beta}) = P\left( \hat{\boldsymbol{y}}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \quad ..., \quad \hat{\boldsymbol{y}}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N \right)}$$

  where $\hat{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}(\boldsymbol{x}_n)$ gives the model estimate for $\boldsymbol{x}_n$
- For continuous endogenous variables, the likelihood function is given by the multi-dimensional probability density at the data points:

$$L(\boldsymbol{\beta}) = f_{\hat{\boldsymbol{y}}_1(\boldsymbol{\beta}),...,\hat{\boldsymbol{y}}_N(\boldsymbol{\beta})}(\boldsymbol{y}_1, ..., \boldsymbol{y}_N)$$

? Verify that the density formulation is equivalent to the probability definition by requiring the model estimations to be in small intervals around the data instead of hitting the data exactly.

! The multi-dimensional probability density $f(.)$ is defined such that $\mathrm{d}P = f_{\hat{\boldsymbol{y}}_1,...,\hat{\boldsymbol{y}}_N}(\boldsymbol{y})\mathrm{d}^N\boldsymbol{y}$. Keeping $\mathrm{d}^N\boldsymbol{y}$ small and constant, $\mathrm{d}P$ and thus $P$ is maximized if and only if $f(.)$ is maximized.

# 9.1. Maximum-Likelihood Estimation: the likelihood function

- The **maximum-likelihood (ML)** estimation is applicable for general stochastic models where the probabilities depend on a parameter vector $\boldsymbol{\beta}$
- The goal is to maximize the **likelihood function $L(\boldsymbol{\beta})$**, i.e., the probability that the model predicts *all* data points $(\boldsymbol{y}_n, \boldsymbol{x}_n)$, $n = 1, \ldots, N$:

$$L(\boldsymbol{\beta}) = P\left( \hat{\boldsymbol{y}}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \quad \ldots, \quad \hat{\boldsymbol{y}}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N \right)$$

  where $\hat{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}(\boldsymbol{x}_n)$ gives the model estimate for $\boldsymbol{x}_n$
- For continuous endogenous variables, the likelihood function is given by the multi-dimensional probability density at the data points:

$$L(\boldsymbol{\beta}) = f_{\hat{\boldsymbol{y}}_1(\boldsymbol{\beta}), \ldots, \hat{\boldsymbol{y}}_N(\boldsymbol{\beta})}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$$

? Verify that the density formulation is equivalent to the probability definition by requiring the model estimations to be in small intervals around the data instead of hitting the data exactly.

! The multi-dimensional probability density $f(.)$ is defined such that $\mathrm{d}P = f_{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_N}(\boldsymbol{y})\mathrm{d}^N\boldsymbol{y}$. Keeping $\mathrm{d}^N\boldsymbol{y}$ small and constant, $\mathrm{d}P$ and thus $P$ is maximized if and only if $f(.)$ is maximized.

# 9.1. Maximum-Likelihood Estimation: the likelihood function

- The **maximum-likelihood (ML)** estimation is applicable for general stochastic models where the probabilities depend on a parameter vector $\boldsymbol{\beta}$
- The goal is to maximize the **likelihood function $L(\boldsymbol{\beta})$**, i.e., the probability that the model predicts *all* data points $(\boldsymbol{y}_n, \boldsymbol{x}_n)$, $n = 1, ..., N$:

$$\boxed{L(\boldsymbol{\beta}) = P\left( \hat{\boldsymbol{y}}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \quad ..., \quad \hat{\boldsymbol{y}}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N \right)}$$

  where $\hat{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}(\boldsymbol{x}_n)$ gives the model estimate for $\boldsymbol{x}_n$
- For continuous endogenous variables, the likelihood function is given by the multi-dimensional probability density at the data points:

$$L(\boldsymbol{\beta}) = f_{\hat{\boldsymbol{y}}_1(\boldsymbol{\beta}),...,\hat{\boldsymbol{y}}_N(\boldsymbol{\beta})}(\boldsymbol{y}_1, ..., \boldsymbol{y}_N)$$

? Verify that the density formulation is equivalent to the probability definition by requiring the model estimations to be in small intervals around the data instead of hitting the data exactly.

! The multi-dimensional probability density $f(.)$ is defined such that $\mathrm{d}P = f_{\hat{\boldsymbol{y}}_1,...,\hat{\boldsymbol{y}}_N}(\boldsymbol{y})\mathrm{d}^N\boldsymbol{y}$. Keeping $\mathrm{d}^N\boldsymbol{y}$ small and constant, $\mathrm{d}P$ and thus $P$ is maximized if and only if $f(.)$ is maximized.

# Maximum-likelihood estimation

- The ML method maximizes the likelihood function:

$$\boxed{\hat{\beta} = \arg\max_{\beta} L(\beta)}$$

- Equivalently, and often better, one maximizes the **log-likelihood**:

$$\hat{\beta} = \arg\max_{\beta} \tilde{L}(\beta), \quad \tilde{L}(\beta) = \ln L(\beta)$$

? Why it does not matter whether to maximize the likelihood or the log-likelihood?

! Since, as a probability or probability density, $L > 0$ and the log function is defined and strictly monotonously increasing in this range. Since (i) in this case

$$x > y \Leftrightarrow f(x) > f(y)$$

(ii) the maximum function is based on this inequality relation, the *argument* of the maximum remains unchanged.

# Maximum-likelihood estimation

- The ML method maximizes the likelihood function:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$$

- Equivalently, and often better, one maximizes the **log-likelihood**:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}(\boldsymbol{\beta}), \quad \tilde{L}(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$$

? Why it does not matter whether to maximize the likelihood or the log-likelihood?

! Since, as a probability or probability density, $L > 0$ and the log function is defined and strictly monotonously increasing in this range. Since (i) in this case

$$x > y \Leftrightarrow f(x) > f(y)$$

(ii) the maximum function is based on this inequality relation, the *argument* of the maximum remains unchanged.

## Maximum-likelihood estimation

- The ML method maximizes the likelihood function:

$$\boxed{\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})}$$

- Equivalently, and often better, one maximizes the **log-likelihood**:

$$\boxed{\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}(\boldsymbol{\beta}), \quad \tilde{L}(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})}$$

? Why it does not matter whether to maximize the likelihood or the log-likelihood?

! Since, as a probability or probability density, $L > 0$ and the log function is defined and strictly monotonously increasing in this range. Since (i) in this case

$$x > y \Leftrightarrow f(x) > f(y)$$

(ii) the maximum function is based on this inequality relation, the *argument* of the maximum remains unchanged.

# Maximum-likelihood estimation

- The ML method maximizes the likelihood function:

$$\boxed{\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})}$$

- Equivalently, and often better, one maximizes the **log-likelihood**:

$$\boxed{\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}(\boldsymbol{\beta}), \quad \tilde{L}(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})}$$

? Why it does not matter whether to maximize the likelihood or the log-likelihood?

! Since, as a probability or probability density, $L > 0$ and the log function is defined and strictly monotonously increasing in this range. Since (i) in this case

$$x > y \Leftrightarrow f(x) > f(y)$$

(ii) the maximum function is based on this inequality relation, the *argument* of the maximum remains unchanged.

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$L(\boldsymbol{\beta}) \overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n)$$

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$L(\boldsymbol{\beta}) \overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d.N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right],$$

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d.N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta x}_n)^2}{2\sigma^2}\right], \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln f_n(y_n) = \sum_{n=1}^{N} \left\{-\frac{1}{2}(\ln 2\pi + \ln \sigma^2) - \left[\frac{(y_n - \boldsymbol{\beta x}_n)^2}{2\sigma^2}\right]\right\}
\end{aligned}
$$

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d.N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right], \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln f_n(y_n) = \sum_{n=1}^{N} \left\{ -\frac{1}{2}(\ln 2\pi + \ln\sigma^2) - \left[\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\,\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\,\boldsymbol{\beta})
\end{aligned}
$$

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d.N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right], \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln f_n(y_n) = \sum_{n=1}^{N} \left\{-\frac{1}{2}(\ln 2\pi + \ln\sigma^2) - \left[\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

Except for the irrelevant additive and multiplicative constants, this is the SSE function of the OLS method and therefore leads to the same estimator!

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d.N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right], \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln f_n(y_n) = \sum_{n=1}^{N} \left\{-\frac{1}{2}(\ln 2\pi + \ln\sigma^2) - \left[\frac{(y_n - \boldsymbol{\beta}\boldsymbol{x}_n)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

Except for the irrelevant additive and multiplicative constants, this is the SSE function of the OLS method and therefore leads to the same estimator!

**?** Why it is possible to express $L(\beta)$ as a product?

# Application 1: Regression models

Besides OLS, the ML can also be used to estimate regression models. Does it give the same result, at least if the statistical Gauß-Markow conditions are satisfied?

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\overset{\epsilon_n \text{ independent}}{=} \prod_{n=1}^{N} f_n(y_n) \overset{\epsilon_n \sim i.d. N(0,\sigma^2)}{=} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \boldsymbol{\beta x}_n)^2}{2\sigma^2}\right], \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{n=1}^{N} \ln f_n(y_n) = \sum_{n=1}^{N} \left\{-\frac{1}{2}(\ln 2\pi + \ln\sigma^2) - \left[\frac{(y_n - \boldsymbol{\beta x}_n)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

Except for the irrelevant additive and multiplicative constants, this is the SSE function of the OLS method and therefore leads to the same estimator!

**?** Why it is possible to express $L(\boldsymbol{\beta})$ as a product?

**!** Since the random terms $\epsilon_n \sim i.i.dN(0, \sigma^2)$, particularly, they are *independent* from each other

## Application 2: Discrete-choice models

- Probability to predict the chosen alternative $i_n$ for a *single* decision $n$:

$$P\left(\hat{\boldsymbol{Y}}_n = \boldsymbol{y}_n\right) = P\left(\hat{Y}_{n1} = y_{n1},\ \ldots,\ \hat{Y}_{nI} = y_{nI}\right)$$

# Application 2: Discrete-choice models

- Probability to predict the chosen alternative $i_n$ for a *single* decision $n$:

$$
\begin{aligned}
P\left(\hat{\boldsymbol{Y}}_n = \boldsymbol{y}_n\right) &= P\left(\hat{Y}_{n1} = y_{n1},\ \ldots,\ \hat{Y}_{nI} = y_{nI}\right) \\
&= \prod_{i=1}^{I} \left[P_{ni}(\boldsymbol{\beta})\right]^{y_{ni}} = P_{ni_n}(\boldsymbol{\beta})
\end{aligned}
$$

(this relies on the exclusivity/completeness of $\mathcal{A}_n$ and of independent RUs)

# Application 2: Discrete-choice models

- Probability to predict the chosen alternative $i_n$ for a *single* decision $n$:

$$\begin{aligned} P\left(\hat{\boldsymbol{Y}}_n = \boldsymbol{y}_n\right) &= P\left(\hat{Y}_{n1} = y_{n1}, \ \dots, \ \hat{Y}_{nI} = y_{nI}\right) \\ &= \prod_{i=1}^{I} \left[P_{ni}(\boldsymbol{\beta})\right]^{y_{ni}} = P_{ni_n}(\boldsymbol{\beta}) \end{aligned}$$

(this relies on the exclusivity/completeness of $\mathcal{A}_n$ and of independent RUs)

- Probability to predict *all* the decisions correctly assuming independent decisions:

$$L(\boldsymbol{\beta}) = P\left(\boldsymbol{Y}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \dots, \boldsymbol{Y}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N\right)$$

# Application 2: Discrete-choice models

- Probability to predict the chosen alternative $i_n$ for a *single* decision $n$:

$$\begin{aligned} P\left(\hat{\boldsymbol{Y}}_n = \boldsymbol{y}_n\right) &= P\left(\hat{Y}_{n1} = y_{n1},\ \ldots,\ \hat{Y}_{nI} = y_{nI}\right) \\ &= \prod_{i=1}^{I} [P_{ni}(\boldsymbol{\beta})]^{y_{ni}} = P_{ni_n}(\boldsymbol{\beta}) \end{aligned}$$

(this relies on the exclusivity/completeness of $\mathcal{A}_n$ and of independent RUs)

- Probability to predict *all* the decisions correctly assuming independent decisions:

$$\begin{aligned} L(\boldsymbol{\beta}) &= P\left(\boldsymbol{Y}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N\right) \\ &= \prod_{n=1}^{N} \prod_{i=1}^{I} [P_{ni}(\boldsymbol{\beta})]^{y_{ni}} \end{aligned}$$

# Application 2: Discrete-choice models

- Probability to predict the chosen alternative $i_n$ for a *single* decision $n$:

$$\begin{aligned} P\left(\hat{\boldsymbol{Y}}_n = \boldsymbol{y}_n\right) &= P\left(\hat{Y}_{n1} = y_{n1},\ \ldots,\ \hat{Y}_{nI} = y_{nI}\right) \\ &= \prod_{i=1}^{I} [P_{ni}(\boldsymbol{\beta})]^{y_{ni}} = P_{ni_n}(\boldsymbol{\beta}) \end{aligned}$$

(this relies on the exclusivity/completeness of $\mathcal{A}_n$ and of independent RUs)

- Probability to predict *all* the decisions correctly assuming independent decisions:

$$\begin{aligned} L(\boldsymbol{\beta}) &= P\left(\boldsymbol{Y}_1(\boldsymbol{\beta}) = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_N(\boldsymbol{\beta}) = \boldsymbol{y}_N\right) \\ &= \prod_{n=1}^{N} \prod_{i=1}^{I} [P_{ni}(\boldsymbol{\beta})]^{y_{ni}} \end{aligned}$$

ML estimation:

$$\hat{\boldsymbol{\beta}} = \text{arg}\max_{\boldsymbol{\beta}} \tilde{L}(\boldsymbol{\beta}), \quad \tilde{L}(\boldsymbol{\beta}) = \sum_{n=1}^{N} \sum_{i=1}^{I} y_{ni} \ln P_{ni}(\boldsymbol{\beta}) = \sum_{n=1}^{N} \ln P_{ni_n}(\boldsymbol{\beta})$$

## Question

**?** Show that, in deriving the main ML result $\tilde{L} = \sum_n \sum_i y_{ni} \ln P_{ni}$, the random utilities need not to be uncorrelated between alternatives, only between choices

**!** Because of the exclusivity/completeness requirement for the alternatives, exactly one alternative can be chosen per decision so it is enough to maximize the corresponding probability (which, of course, depends on possible correlations)

# Question

? Show that, in deriving the main ML result $\tilde{L} = \sum_n \sum_i y_{ni} \ln P_{ni}$, the random utilities need not to be uncorrelated between alternatives, only between choices

! Because of the exclusivity/completeness requirement for the alternatives, exactly one alternative can be chosen per decision so it is enough to maximize the corresponding probability (which, of course, depends on possible correlations)

## Estimating models with only ACs

If there are no exogenous variables, we are left with just the ACs reflecting that people prefer certain alternatives over others for unknown reasons:

$$V_{ni} = \sum_{m=1}^{I-1} \beta_m \delta_{mi} \quad \text{or} \quad V_{ni} = \beta_i \text{ if } i \neq I, \; V_{nI} = 0$$

This **AC-only model** will be the "reference case" when estimating the model quality, e.g., by the **likelihood-ratio index**.

? Show that the estimated models gives probabilities $P_{ni} = P_i$ that are equal to the observed choice fractions $N_i/N$. (*Hint:* Lagrange multiplicators to satisfy $\sum_i P_i = 1$)

! we have $\tilde{L}(\boldsymbol{P}) = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i$; maximize under the constraint $\sum_i P_i = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}P_i}\left(\tilde{L}(\boldsymbol{P}) - \lambda(\sum_i P_i - 1)\right) \stackrel{!}{=} 0 \;\Rightarrow\; \frac{N_i}{P_i} = \lambda \Rightarrow\; P_i \propto N_i$$

? Based on this result $P_i = N_i/N$, give the parameters for the AC-only MNL and for the binary i.i.d. Probit model Logit: $P_i/P_I = N_i/N_I = \exp(\beta_i)$ (notice that $I$ is the reference w/o AC)

# Estimating models with only ACs

If there are no exogenous variables, we are left with just the ACs reflecting that people prefer certain alternatives over others for unknown reasons:

$$V_{ni} = \sum_{m=1}^{I-1} \beta_m \delta_{mi} \quad \text{or} \quad V_{ni} = \beta_i \text{ if } i \neq I, \; V_{nI} = 0$$

This **AC-only model** will be the "reference case" when estimating the model quality, e.g., by the **likelihood-ratio index**.

**?** Show that the estimated models gives probabilities $P_{ni} = P_i$ that are equal to the observed choice fractions $N_i/N$. (*Hint:* Lagrange multiplicators to satisfy $\sum_i P_i = 1$)

**!** we have $\tilde{L}(\boldsymbol{P}) = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i$; maximize under the constraint $\sum_i P_i = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}P_i}\left(\tilde{L}(\boldsymbol{P}) - \lambda\Big(\sum_i P_i - 1\Big)\right) \stackrel{!}{=} 0 \;\Rightarrow\; \frac{N_i}{P_i} = \lambda \Rightarrow \; P_i \propto N_i$$

**?** Based on this result $P_i = N_i/N$, give the parameters for the AC-only MNL and for the binary i.i.d. Probit model Logit: $P_i/P_I = N_i/N_I = \exp(\beta_i)$ (notice that $I$ is the reference w/o AC)

# Estimating models with only ACs

If there are no exogenous variables, we are left with just the ACs reflecting that people prefer certain alternatives over others for unknown reasons:

$$V_{ni} = \sum_{m=1}^{I-1} \beta_m \delta_{mi} \quad \text{or} \quad V_{ni} = \beta_i \text{ if } i \neq I, \; V_{nI} = 0$$

This **AC-only model** will be the "reference case" when estimating the model quality, e.g., by the **likelihood-ratio index**.

? Show that the estimated models gives probabilities $P_{ni} = P_i$ that are equal to the observed choice fractions $N_i/N$. (*Hint:* Lagrange multiplicators to satisfy $\sum_i P_i = 1$)

! we have $\tilde{L}(\boldsymbol{P}) = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i$; maximize under the constraint $\sum_i P_i = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}P_i}\left(\tilde{L}(\boldsymbol{P}) - \lambda(\sum_i P_i - 1)\right) \stackrel{!}{=} 0 \;\Rightarrow\; \frac{N_i}{P_i} = \lambda \Rightarrow \; P_i \propto N_i$$

? Based on this result $P_i = N_i/N$, give the parameters for the AC-only MNL and for the binary i.i.d. Probit model Logit: $P_i/P_I = N_i/N_I = \exp(\beta_i)$ (notice that $I$ is the reference w/o AC)

## Estimating models with only ACs

If there are no exogenous variables, we are left with just the ACs reflecting that people prefer certain alternatives over others for unknown reasons:

$$V_{ni} = \sum_{m=1}^{I-1} \beta_m \delta_{mi} \quad \text{or} \quad V_{ni} = \beta_i \text{ if } i \neq I, \; V_{nI} = 0$$

This **AC-only model** will be the "reference case" when estimating the model quality, e.g., by the **likelihood-ratio index**.

? Show that the estimated models gives probabilities $P_{ni} = P_i$ that are equal to the observed choice fractions $N_i/N$. (*Hint:* Lagrange multiplicators to satisfy $\sum_i P_i = 1$)

! we have $\tilde{L}(\boldsymbol{P}) = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i$; maximize under the constraint $\sum_i P_i = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}P_i}\left(\tilde{L}(\boldsymbol{P}) - \lambda(\sum_i P_i - 1)\right) \stackrel{!}{=} 0 \;\Rightarrow\; \frac{N_i}{P_i} = \lambda \Rightarrow\; P_i \propto N_i$$

? Based on this result $P_i = N_i/N$, give the parameters for the AC-only MNL and for the binary i.i.d. Probit model Logit: $P_i/P_I = N_i/N_I = \exp(\beta_i)$ (notice that $I$ is the reference w/o AC)

# Estimating models with only ACs

If there are no exogenous variables, we are left with just the ACs reflecting that people prefer certain alternatives over others for unknown reasons:

$$V_{ni} = \sum_{m=1}^{I-1} \beta_m \delta_{mi} \quad \text{or} \quad V_{ni} = \beta_i \text{ if } i \neq I, \ V_{nI} = 0$$

This **AC-only model** will be the "reference case" when estimating the model quality, e.g., by the **likelihood-ratio index**.

- **?** Show that the estimated models gives probabilities $P_{ni} = P_i$ that are equal to the observed choice fractions $N_i/N$. (*Hint:* Lagrange multiplicators to satisfy $\sum_i P_i = 1$)
- **!** we have $\tilde{L}(\boldsymbol{P}) = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i$; maximize under the constraint $\sum_i P_i = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}P_i}\left(\tilde{L}(\boldsymbol{P}) - \lambda(\sum_i P_i - 1)\right) \stackrel{!}{=} 0 \ \Rightarrow \ \frac{N_i}{P_i} = \lambda \Rightarrow \ P_i \propto N_i$$

- **?** Based on this result $P_i = N_i/N$, give the parameters for the AC-only MNL and for the binary i.i.d. Probit model Logit: $P_i/P_I = N_i/N_I = \exp(\beta_i)$ (notice that $I$ is the reference w/o AC)

# Exercise: simple binomial model with an AC and travel time

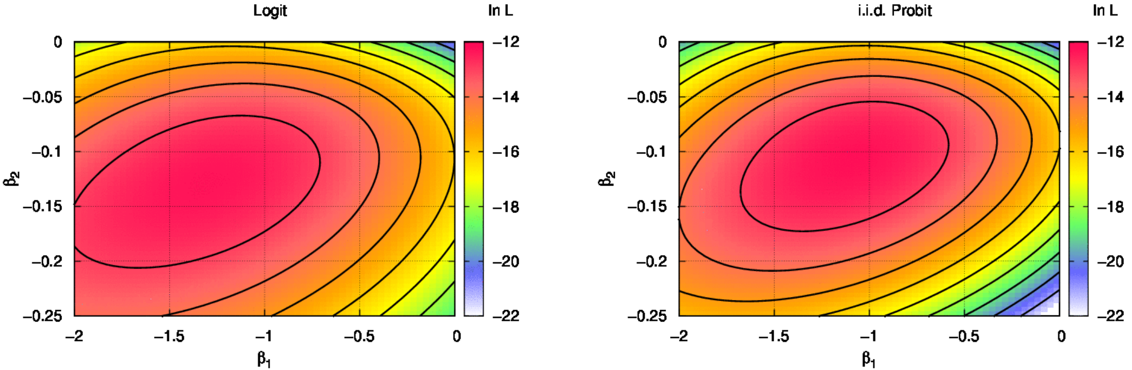
$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$

| Choice set | $T_{\text{ped}} = T_1$ [min] | $T_{\text{bike}} = T_2$ [min] | # chosen 1 | # chosen 2 |
|---|---|---|---|---|
| 1 | 15 | 30 | 3 | 2 |
| 2 | 10 | 15 | 2 | 3 |
| 3 | 20 | 20 | 1 | 4 |
| 4 | 30 | 25 | 1 | 4 |
| 5 | 30 | 20 | 0 | 5 |
| 6 | 60 | 30 | 0 | 5 |

# Exercise: simple binomial model with an AC and travel time

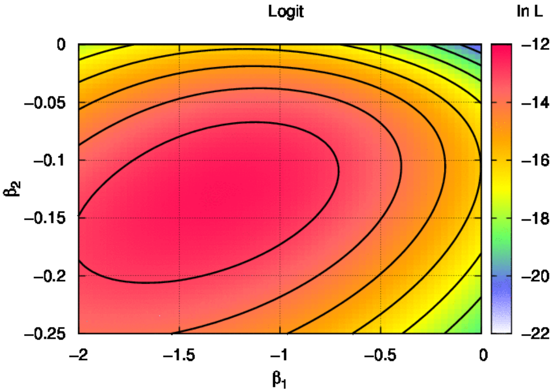
$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$

| Choice set | $T_{\text{ped}} = T_1$ [min] | $T_{\text{bike}} = T_2$ [min] | # chosen 1 | # chosen 2 |
|---|---|---|---|---|
| 1 | 15 | 30 | 3 | 2 |
| 2 | 10 | 15 | 2 | 3 |
| 3 | 20 | 20 | 1 | 4 |
| 4 | 30 | 25 | 1 | 4 |
| 5 | 30 | 20 | 0 | 5 |
| 6 | 60 | 30 | 0 | 5 |



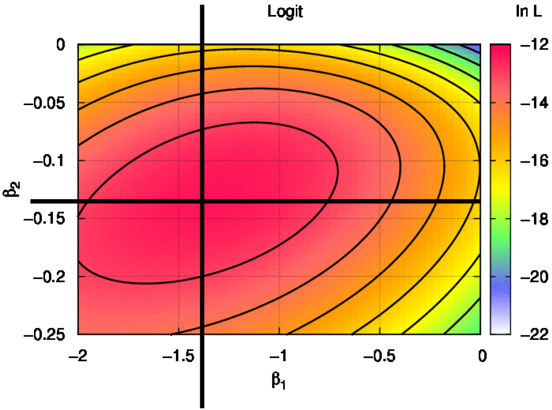
# I: Graphical solution

$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$



# I: Graphical solution

$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$




**Logit**:

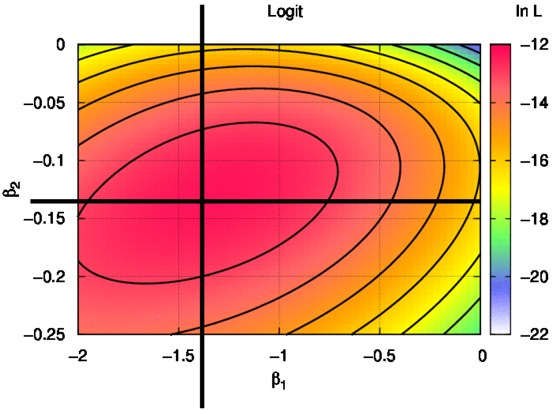
$\tilde{L} = -12$

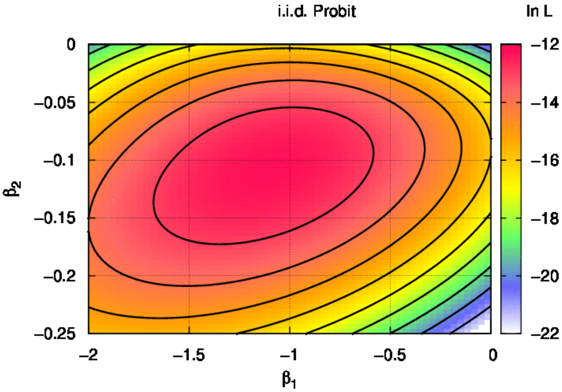
$\hat{\beta}_1 = -1.3, \ \hat{\beta}_2 = -0.14,$

AC in minutes: $-\dfrac{\hat{\beta}_1}{\hat{\beta}_2} = -9\,\mathrm{min}$

# I: Graphical solution

$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$




**Logit**:

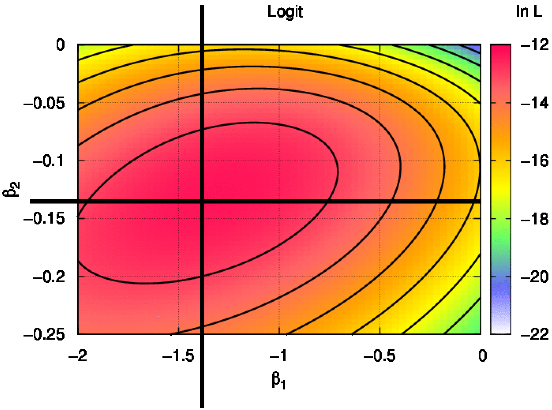
$\tilde{L} = -12$

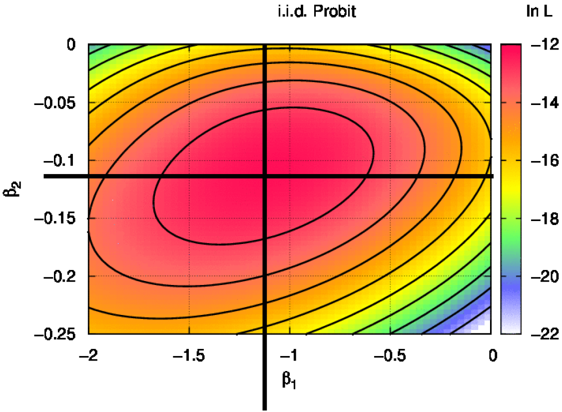
$\hat{\beta}_1 = -1.3, \ \hat{\beta}_2 = -0.14,$

AC in minutes: $-\frac{\hat{\beta}_1}{\hat{\beta}_2} = -9\,\mathrm{min}$

# I: Graphical solution

$$V_{ni} = \beta_1 \delta_{i1} + \beta_2 T_{ni}$$




**Logit**:

$\tilde{L} = -12$
$\hat{\beta}_1 = -1.3, \ \hat{\beta}_2 = -0.14,$
AC in minutes: $-\frac{\hat{\beta}_1}{\hat{\beta}_2} = -9\,\mathrm{min}$

**Probit**:

$\tilde{L} = -12$
$\hat{\beta}_1 = -1.1, \ \hat{\beta}_2 = -0.12,$
AC in minutes: $-\frac{\hat{\beta}_1}{\hat{\beta}_2} = -9\,\mathrm{min}$

# II. Numerical solution

- Generally, we have a nonlinear optimization problem.
- For parameter-linear utilities, we know for the MNL that a maximum exists and is unique.
- Standard methods of nonlinear optimization are possible:
  - **Newton's and quasi-Newton method**: Fast but may be unstable
  - **Gradient/steepest descent methods**: slow but reliable
  - **Broyden-Fletcher-Goldfarb-Shanno (BFGS)** or **Levenberg-Marquardt algorithm** combining gradient and Newton methods. Such methods are used in many software packages
  - **genetic algorithms** if the *objective function landscape* is complicated (nonlinear utilities).

# II. Numerical solution

- Generally, we have a nonlinear optimization problem.
- For parameter-linear utilities, we know for the MNL that a maximum exists and is unique.
- Standard methods of nonlinear optimization are possible:
  - **Newton's and quasi-Newton method**: Fast but may be unstable
  - **Gradient/steepest descent methods**: slow but reliable
  - **Broyden-Fletcher-Goldfarb-Shanno (BFGS)** or **Levenberg-Marquardt algorithm** combining gradient and Newton methods. Such methods are used in many software packages
  - **genetic algorithms** if the *objective function landscape* is complicated (nonlinear utilities).

# Special case: estimating the MNL

The special structure of the MNL with parameter-linear utilities, $V_{ni} = \sum_m \beta_m X_{mni}$ allows for an intuitive formulation of the estimation problem:

The observed and modeled **property sums** sums of the factors $X$ for a given parameter $m$ should be the same

$$
\begin{aligned}
X_m^{\text{MNL}} &= X_m^{\text{data}}, \\
\sum_{n,i} x_{mni} \, P_{ni}(\hat{\beta}) &= \sum_{n,i} x_{mni} \, y_{ni} = \sum_n x_{mni_n}
\end{aligned}
$$

# Example: four factors, two alternatives

MNL model, $V_{ni} = \beta_1 T_{ni} + \beta_2 C_{ni} + \beta_3 g_i \delta_{i1} + \beta_4 \delta_{i1}$, $g_{♂} = 0$, $g_{♀} = 1$:

- $X_1 = T$: Total travel time for the chosen alternatives:

$$T^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) T_{ni}, \quad T^{\text{data}} = \sum_{n,i} y_{ni} T_{ni} = \sum_n T_{n i_n}$$

- $X_2 = C$: Total money spent by the decision makers:

$$C^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) C_{ni}, \quad C^{\text{data}} = \sum_{n,i} y_{ni} C_{ni} = \sum_n C_{n i_n}$$

- $X_3 = N_{1,♀}$: number of woman choosing alternative 1:

$$N_{1,♀}^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}) g_n, \quad N_{1,♀}^{\text{data}} = \sum_n y_{n1} g_n$$

- $X_4 = N_1$: total number of persons choosing alternative 1:

$$N_1^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}), \quad N_1^{\text{data}} = \sum_n y_{n1}$$

## Example: four factors, two alternatives

MNL model, $V_{ni} = \beta_1 T_{ni} + \beta_2 C_{ni} + \beta_3 g_i \delta_{i1} + \beta_4 \delta_{i1}$, $g_{♂} = 0$, $g_{♀} = 1$:

- $X_1 = T$: Total travel time for the chosen alternatives:

$$T^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) T_{ni}, \quad T^{\text{data}} = \sum_{n,i} y_{ni} T_{ni} = \sum_n T_{ni_n}$$

- $X_2 = C$: Total money spent by the decision makers:

$$C^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) C_{ni}, \quad C^{\text{data}} = \sum_{n,i} y_{ni} C_{ni} = \sum_n C_{ni_n}$$

- $X_3 = N_{1,♀}$: number of woman choosing alternative 1:

$$N_{1,♀}^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}) g_n, \quad N_{1,♀}^{\text{data}} = \sum_n y_{n1} g_n$$

- $X_4 = N_1$: total number of persons choosing alternative 1:

$$N_1^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}), \quad N_1^{\text{data}} = \sum_n y_{n1}$$

# Example: four factors, two alternatives

MNL model, $V_{ni} = \beta_1 T_{ni} + \beta_2 C_{ni} + \beta_3 g_i \delta_{i1} + \beta_4 \delta_{i1}$, $g_{♂} = 0$, $g_{♀} = 1$:

- $X_1 = T$: Total travel time for the chosen alternatives:

$$T^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) T_{ni}, \quad T^{\text{data}} = \sum_{n,i} y_{ni} T_{ni} = \sum_n T_{ni_n}$$

- $X_2 = C$: Total money spent by the decision makers:

$$C^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) C_{ni}, \quad C^{\text{data}} = \sum_{n,i} y_{ni} C_{ni} = \sum_n C_{ni_n}$$

- $X_3 = N_{1,♀}$: number of woman choosing alternative 1:

$$N^{\text{MNL}}_{1,♀} = \sum_n P_{n1}(\boldsymbol{\beta}) g_n, \quad N^{\text{data}}_{1,♀} = \sum_n y_{n1} g_n$$

- $X_4 = N_1$: total number of persons choosing alternative 1:

$$N^{\text{MNL}}_1 = \sum_n P_{n1}(\boldsymbol{\beta}), \quad N^{\text{data}}_1 = \sum_n y_{n1}$$

## Example: four factors, two alternatives

MNL model, $V_{ni} = \beta_1 T_{ni} + \beta_2 C_{ni} + \beta_3 g_i \delta_{i1} + \beta_4 \delta_{i1}$, $g_{♂} = 0$, $g_{♀} = 1$:

- $X_1 = T$: Total travel time for the chosen alternatives:

$$T^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) T_{ni}, \quad T^{\text{data}} = \sum_{n,i} y_{ni} T_{ni} = \sum_n T_{ni_n}$$

- $X_2 = C$: Total money spent by the decision makers:

$$C^{\text{MNL}} = \sum_{n,i} P_{ni}(\boldsymbol{\beta}) C_{ni}, \quad C^{\text{data}} = \sum_{n,i} y_{ni} C_{ni} = \sum_n C_{ni_n}$$

- $X_3 = N_{1,♀}$: number of woman choosing alternative 1:

$$N_{1,♀}^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}) g_n, \quad N_{1,♀}^{\text{data}} = \sum_n y_{n1} g_n$$

- $X_4 = N_1$: total number of persons choosing alternative 1:

$$N_1^{\text{MNL}} = \sum_n P_{n1}(\boldsymbol{\beta}), \quad N_1^{\text{data}} = \sum_n y_{n1}$$

# 9.2 Estimation Errors: Variance-Covariance Matrix

Since the log-likelihood is maximized at $\hat{\boldsymbol{\beta}}$, we have

$$\frac{\partial \tilde{L}}{\partial \boldsymbol{\beta}} = 0 \ \Rightarrow \ \tilde{L}(\boldsymbol{\beta}) \approx \tilde{L}_{\text{max}} + \frac{1}{2} \Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}, \quad \Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$$

with the (negative definite) Hessian $H_{lm} = \left.\frac{\partial^2 \tilde{L}(\boldsymbol{\beta})}{\partial \beta_l \ \partial \beta_m}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$

Compare $L(\boldsymbol{\beta})$ near its maximum with the density $f(\boldsymbol{x})$ of the general multivariate normal distribution with variance-covariance matrix $\Sigma$:

$$\begin{aligned} L(\boldsymbol{\beta}) &= L_{\text{max}} \exp\left(\frac{1}{2} \Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}\right), \\ f(\boldsymbol{x}) &= \left((2\pi)^M \text{Det}\boldsymbol{\Sigma}\right)^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}\right) \end{aligned}$$

Identify $\Delta\boldsymbol{\beta}$ with $\boldsymbol{x}$, the sought-after variance-covariance matrix $\mathbf{V}$ with $\Sigma$, and assume the asymptotic limit (higher than quadratic terms in $\tilde{L}(\hat{\boldsymbol{\beta}})$ negligible): $\Rightarrow$

$$\boxed{\mathbf{V} = \text{Cov}(\hat{\boldsymbol{\beta}}) = E\left[\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)'\right] \approx -\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}})}$$

# 9.2 Estimation Errors: Variance-Covariance Matrix

Since the log-likelihood is maximized at $\hat{\boldsymbol{\beta}}$, we have

$$\frac{\partial \tilde{L}}{\partial \boldsymbol{\beta}} = 0 \ \Rightarrow \ \tilde{L}(\boldsymbol{\beta}) \approx \tilde{L}_{\text{max}} + \frac{1}{2}\Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}, \quad \Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$$

with the (negative definite) Hessian $H_{lm} = \left.\frac{\partial^2 \tilde{L}(\boldsymbol{\beta})}{\partial \beta_l \ \partial \beta_m}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$

Compare $L(\boldsymbol{\beta})$ near its maximum with the density $f(\boldsymbol{x})$ of the general multivariate normal distribution with variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\begin{aligned} L(\boldsymbol{\beta}) &= L_{\text{max}} \exp\left(\frac{1}{2}\Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}\right), \\ f(\boldsymbol{x}) &= \left((2\pi)^M \text{Det}\boldsymbol{\Sigma}\right)^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \end{aligned}$$

Identify $\Delta\beta$ with $x$, the sought-after variance-covariance matrix $\mathbf{V}$ with $\Sigma$, and assume the asymptotic limit (higher than quadratic terms in $\tilde{L}(\hat{\beta})$ negligible): $\Rightarrow$

$$\mathbf{V} = \text{Cov}(\hat{\beta}) = E\left[\left(\beta - \hat{\beta}\right)\left(\beta - \hat{\beta}\right)'\right] \approx -\mathbf{H}^{-1}(\hat{\beta})$$

# 9.2 Estimation Errors: Variance-Covariance Matrix

Since the log-likelihood is maximized at $\hat{\boldsymbol{\beta}}$, we have

$$\frac{\partial \tilde{L}}{\partial \boldsymbol{\beta}} = 0 \;\Rightarrow\; \tilde{L}(\boldsymbol{\beta}) \approx \tilde{L}_{\text{max}} + \frac{1}{2}\Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}, \quad \Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$$

with the (negative definite) Hessian $H_{lm} = \left.\frac{\partial^2 \tilde{L}(\boldsymbol{\beta})}{\partial\beta_l\;\partial\beta_m}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$

Compare $L(\boldsymbol{\beta})$ near its maximum with the density $f(\boldsymbol{x})$ of the general multivariate normal distribution with variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\begin{aligned} L(\boldsymbol{\beta}) &= L_{\text{max}} \exp\left(\frac{1}{2}\Delta\boldsymbol{\beta}^T \cdot \mathbf{H} \cdot \Delta\boldsymbol{\beta}\right), \\ f(\boldsymbol{x}) &= \left((2\pi)^M \text{Det}\boldsymbol{\Sigma}\right)^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \end{aligned}$$

Identify $\Delta\boldsymbol{\beta}$ with $\boldsymbol{x}$, the sought-after variance-covariance matrix $\mathbf{V}$ with $\boldsymbol{\Sigma}$, and assume the asymptotic limit (higher than quadratic terms in $\tilde{L}(\hat{\boldsymbol{\beta}})$ negligible): $\Rightarrow$

$$\boxed{\mathbf{V} = \text{Cov}(\hat{\boldsymbol{\beta}}) = E\left[\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)'\right] \approx -\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}})}$$

# Fisher's information matrix

The variance-covariance matrix is related to **Fisher's information matrix $\mathcal{I}$**:

$$\mathcal{I} = \mathbf{V}^{-1} = -\mathbf{H}, \quad I_{lm} = -\frac{\partial^2 \tilde{L}(\hat{\boldsymbol{\beta}})}{\partial \beta_l \, \partial \beta_m}$$

- Roughly speaking, information is missing uncertainty, so the higher the main components of $\mathcal{I}$, the lower the main components of $\mathbf{V}$
- **Cramér-Rao inequality:** A lower bound for the variance-covariance matrix is the inverse of Fisher's information matrix $\Rightarrow$ The ML estimator is **asymptotically efficient**
- Comparison with the OLS estimator $\mathbf{V}_{\text{OLS}} = 2\sigma^2 \mathbf{H}_{\text{SSE}}^{-1}$ of regression models:

$$\mathcal{I} = -\mathbf{H} = \mathbf{H}_{\text{SSE}}/(2\sigma^2) = \mathbf{X}'\mathbf{X}/\sigma^2$$

The negative Hesse matrix of $\tilde{L}(\beta)$ is proportional to the Hesse matrix of the regression SSE $S(\beta)$.

# Fisher's information matrix

The variance-covariance matrix is related to **Fisher's information matrix $\mathcal{I}$**:

$$\mathcal{I} = \mathbf{V}^{-1} = -\mathbf{H}, \quad I_{lm} = -\frac{\partial^2 \tilde{L}(\hat{\boldsymbol{\beta}})}{\partial \beta_l \, \partial \beta_m}$$

- Roughly speaking, information is missing uncertainty, so the higher the main components of $\mathcal{I}$, the lower the main components of $\mathbf{V}$
- **Cramér-Rao inequality:** A lower bound for the variance-covariance matrix is the inverse of Fisher's information matrix $\Rightarrow$ The ML estimator is **asymptotically efficient**
- Comparison with the OLS estimator $\mathbf{V}_{\text{OLS}} = 2\sigma^2 \mathbf{H}_{\text{SSE}}^{-1}$ of regression models:

$$\mathcal{I} = -\mathbf{H} = \mathbf{H}_{\text{SSE}}/(2\sigma^2) = \mathbf{X}'\mathbf{X}/\sigma^2$$

The negative Hesse matrix of $\tilde{L}(\beta)$ is proportional to the Hesse matrix of the regression SSE $S(\beta)$.

# Fisher's information matrix

The variance-covariance matrix is related to **Fisher's information matrix $\mathcal{I}$**:

$$\mathcal{I} = \mathbf{V}^{-1} = -\mathbf{H}, \quad I_{lm} = -\frac{\partial^2 \tilde{L}(\hat{\boldsymbol{\beta}})}{\partial \beta_l \, \partial \beta_m}$$

- Roughly speaking, information is missing uncertainty, so the higher the main components of $\mathcal{I}$, the lower the main components of $\mathbf{V}$
- **Cramér-Rao inequality:** A lower bound for the variance-covariance matrix is the inverse of Fisher's information matrix $\Rightarrow$ The ML estimator is **asymptotically efficient**
- Comparison with the OLS estimator $\mathbf{V}_{\text{OLS}} = 2\sigma^2 \mathbf{H}_{\text{SSE}}^{-1}$ of regression models:

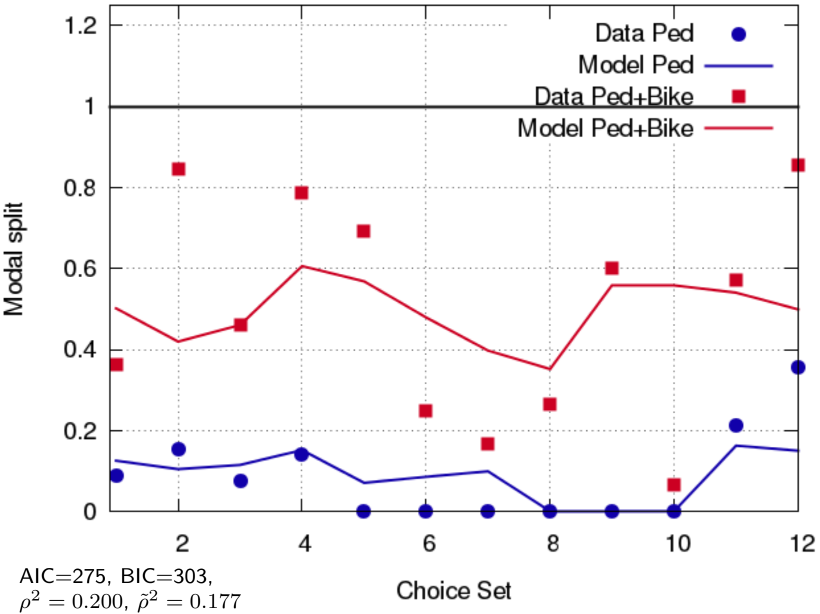
  $$\mathcal{I} = -\mathbf{H} = \mathbf{H}_{\text{SSE}}/(2\sigma^2) = \mathbf{X}'\mathbf{X}/\sigma^2$$

  The negative Hesse matrix of $\tilde{L}(\boldsymbol{\beta})$ is proportional to the Hesse matrix of the regression SSE $S(\boldsymbol{\beta})$.

## 9.2.1 Example 1 from past lecture: SP Survey in the Audience WS18/19 (red: bad weather, $W = 1$)

| Choice Set | Alt. 1: Ped | Alt. 2: Bike | Alt. 3: PT/Car | Alt 1 | Alt 2 | Alt 3 |
|---|---|---|---|---|---|---|
| 1 | 30 min | 20 min | 20 min+0€ | 1 | 3 | 7 |
| 2 | 30 min | 20 min | 20 min+2€ | 2 | 9 | 2 |
| 3 | 30 min | 20 min | 20 min+1€ | 1 | 5 | 7 |
| 4 | 30 min | 20 min | 30 min+0€ | 2 | 9 | 3 |
| 5 | 50 min | 20 min | 30 min+0€ | 0 | 9 | 4 |
| 6 | 50 min | 30 min | 30 min+0€ | 0 | 3 | 9 |
| 7 | 50 min | 40 min | 30 min+0€ | 0 | 2 | 10 |
| 8 | 180 min | 60 min | 60 min+2€ | 0 | 4 | 11 |
| 9 | 180 min | 40 min | 60 min+2€ | 0 | 9 | 6 |
| 10 | 180 min | 40 min | 60 min+2€ | 0 | 1 | 14 |
| 11 | 12 min | 8 min | 10 min+0€ | 3 | 5 | 6 |
| 12 | 12 min | 8 min | 10 min+1€ | 5 | 7 | 2 |

# Model specification for Model 1 of the past lecture




$$V_i = \beta_0\delta_{i1} + \beta_1\delta_{i2} + \beta_2 K_i + \beta_3 T_i$$

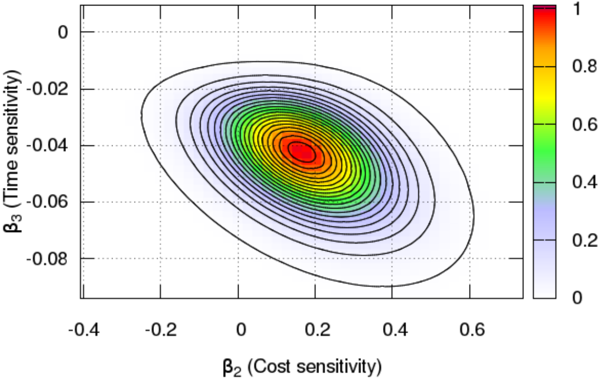
$$\boxed{\begin{aligned}\beta_0 &= -0.95 \pm 0.37,\\ \beta_1 &= -0.28 \pm 0.24,\\ \beta_2 &= +0.17 \pm 0.19,\\ \beta_3 &= -0.04 \pm 0.02\end{aligned}}$$

$$\frac{\beta_0}{-\beta_3} = -22.4\,\mathrm{min},\quad \frac{\beta_1}{-\beta_3} = -6.6\,\mathrm{min},\quad \frac{60\beta_3}{\beta_2} = -15\,€/\mathrm{h}$$

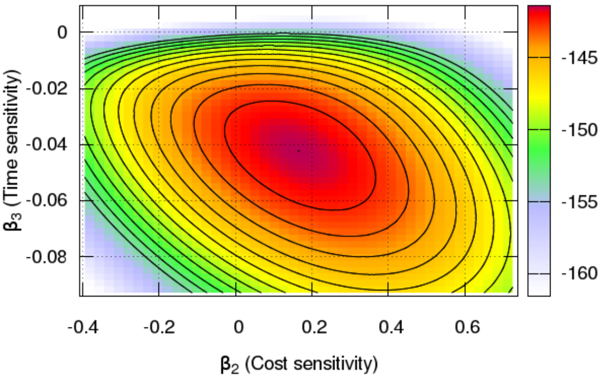
AIC=275, BIC=303,
$\rho^2 = 0.200$, $\tilde{\rho}^2 = 0.177$

# Likelihood and log-likelihood function for varying cost ($\beta_2$) and time ($\beta_3$) sensitivities

$$V_i = \beta_0 \delta_{i1} + \beta_1 \delta_{i2} + \beta_2 K + \beta_3 T$$



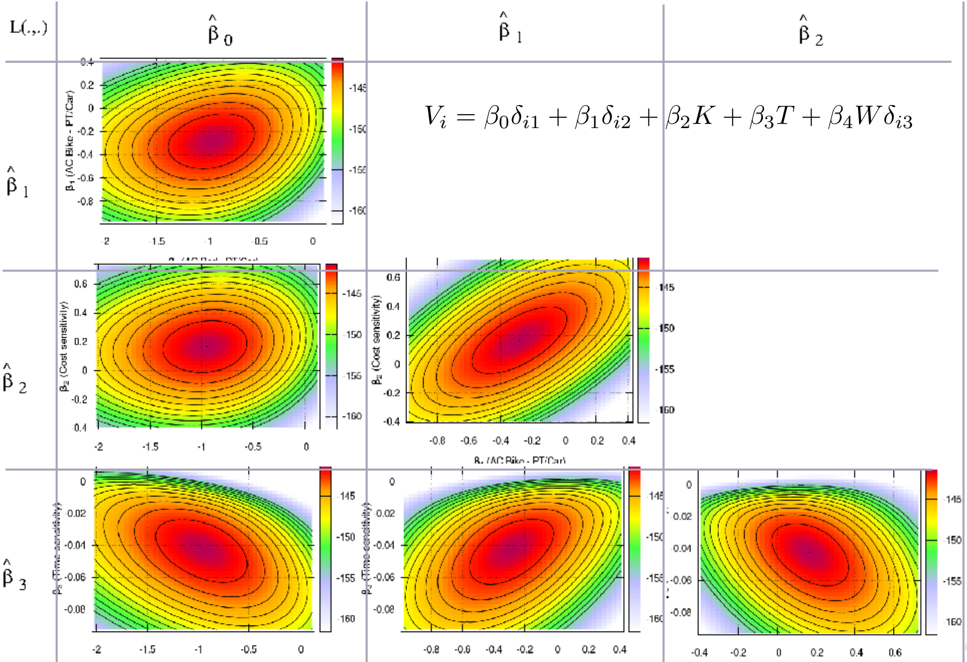

Likelihood function
$L(\beta_2, \beta_3 | \hat{\beta}_0, \hat{\beta}_0)$




Log-likelihood function
$\tilde{L}(\beta_2, \beta_3 | \hat{\beta}_0, \hat{\beta}_1)$

# Log-likelihood function in parameter space

$$V_i = \beta_0\delta_{i1} + \beta_1\delta_{i2} + \beta_2 K + \beta_3 T + \beta_4 W\delta_{i3}$$



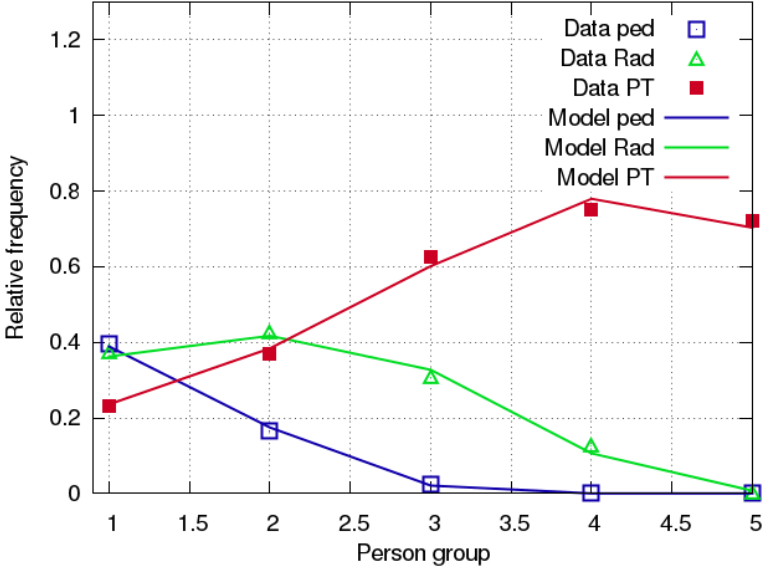
# 9.2.2 Example 2: RP Survey in the Audience

Distance classes for the trip home to university (cumulated till 2018)

Weather: good

| Distance | Class-center | Choice Alt. 1: ped | Choice Alt. 2: bike | Choice Alt. 2: PT | Choice Alt. 3: car |
|---|---|---|---|---|---|
| 0-1 km | 0.5 km | 17 | 16 | 10 | 0 |
| 1-2 km | 1.5 km | 9 | 23 | 20 | 2 |
| 2-5 km | 3.5 km | 2 | 27 | 55 | 4 |
| 5-10 km | 7.5 km | 0 | 7 | 42 | 7 |
| 10-20 km | 12.5 km | 0 | 0 | 18 | 7 |

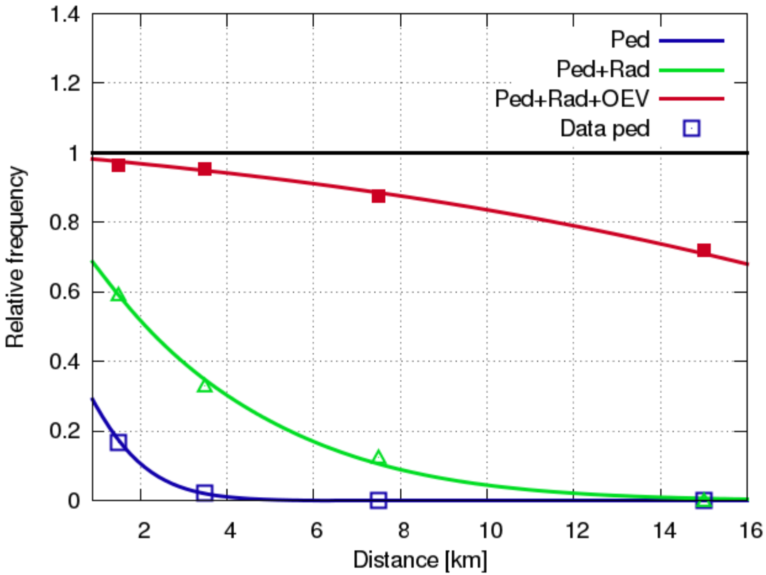
# Revealed Choice: fit quality




$$V_1 = \beta_1 + \beta_4 r,$$
$$V_2 = \beta_2 + \beta_5 r,$$
$$V_3 = \beta_3 + \beta_6 r,$$
$$V_4 = 0$$

$$\beta_1 = 4.1 \pm 0.6,$$
$$\beta_2 = 3.6 \pm 0.5,$$
$$\beta_3 = 3.0 \pm 0.5,$$
$$\beta_4 = -1.43 \pm 0.26,$$
$$\beta_5 = -0.48 \pm 0.08,$$
$$\beta_6 = -0.14 \pm 0.05$$

# Revealed Choice: Modal split as a function of distance

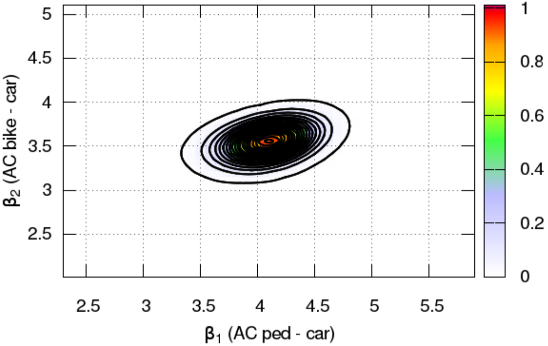



$$V_1 = \beta_1 + \beta_4 r,$$
$$V_2 = \beta_2 + \beta_5 r,$$
$$V_3 = \beta_3 + \beta_6 r,$$
$$V_4 = 0$$

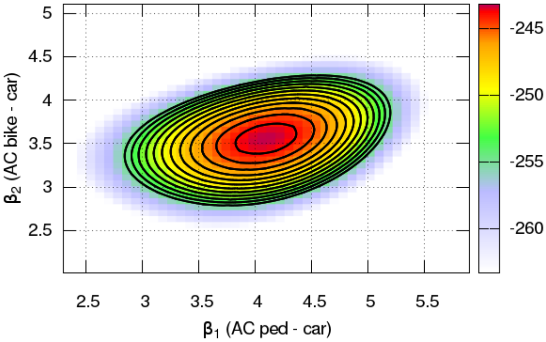
$$\beta_1 = 4.1 \pm 0.6,$$
$$\beta_2 = 3.6 \pm 0.5,$$
$$\beta_3 = 3.0 \pm 0.5,$$
$$\beta_4 = -1.43 \pm 0.26,$$
$$\beta_5 = -0.48 \pm 0.08,$$
$$\beta_6 = -0.14 \pm 0.05$$

# Likelihood and Log-Likelihood as $f(\beta_1, \beta_2)$

$$V_i = \sum_{m=1}^{3} \beta_m \delta_{m,i} + \sum_{m=1}^{3} \beta_{m+3}\ r\delta_{m,i}$$




Likelihoodfunktion
$L(\beta_1, \beta_2, \hat{\beta}_3, ...)$




Log-Likelihoodfunktion
$\tilde{L}(\beta_1, \beta_2, \hat{\beta}_3, ...)$

# Log-Likelihood: Sections through parameter space