

1. Introduction to multinomial discrete choice models

1.1 Background

Multiple linear regression model as basic econometric approach:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i$$

y_i : Dependent variable for individual i

x_{i1}, \dots, x_{ik} : k explanatory variables for individual i

β_0 : Intercept

β_1 : This parameter measures the effect of an increase of x_{i1} on y_i , holding all other observed and unobserved factors fixed

⋮

β_k : This parameter measures the effect of an increase of x_{ik} on y_i , holding all other observed and unobserved factors fixed

ε_i : Error term

- The parameters are commonly estimated by the ordinary method of least squares (OLS method) due to its attractive properties (e.g. unbiasedness, efficiency, consistency, asymptotic efficiency, asymptotic normality) under several conditions (on this basis, t-tests or F-tests can be conducted)
- However, the useful application of linear regression models requires quantitative continuous dependent variables with an unrestricted range

1.2 General model structure

Multinomial dependent variables in a microeconomic analysis:

These qualitative variables have more than two possible mutually exclusive categories which are not ordered

Examples for microeconomic analyses with multinomial response models:

- Analysis of the employment status of a person (e.g. blue-collar worker, white-collar employee, self-employed person)
- Analysis of the choice of a voter among several parties (e.g. CDU/CSU, SPD, AfD, Bündnis 90/Die Grünen, Die Linke)
- Analysis of the choice of a person among several means of transportation (e.g. car, bus, train, plane)
- Analysis of the choice of a pupil among several secondary school types (e.g. Hauptschule, Realschule, Gymnasium)
- Analysis of the (hypothetically stated) choice of a car buyer among several energy sources (e.g. gasoline, diesel, hybrid, gas, biofuel, hydrogen, electric)
- Analysis of the (hypothetically stated) choice of a financial decision maker among several investments that are characterized by specific attributes

Utility function of multinomial discrete choice models (“random utility models”):
 The basis of the microeconomic motivation is that an individual i can choose among J mutually exclusive alternatives of a qualitative variable. The hypothetical (linear) utility function of i for alternative j is as follows:

$$u_{ij} = \beta_j' x_i + \gamma' z_{ij} + \varepsilon_{ij} \quad \text{for } i = 1, \dots, n; j = 1, \dots, J$$

The deterministic component of the utility function comprises the k_1 -dimensional vector $x_i = (x_{i1}, \dots, x_{ik_1})'$ of individual characteristics, the k_2 -dimensional vector $z_{ij} = (z_{ij1}, \dots, z_{ijk_2})'$ of alternative specific attributes, and corresponding parameter vectors $\beta_j = (\beta_{j1}, \dots, \beta_{jk_1})'$ and $\gamma = (\gamma_1, \dots, \gamma_{k_2})'$. The stochastic component of the utility function refers to the error term ε_{ij} that comprises all unobservable factors. The z_{ij} are summarized in the $J \cdot k_2$ -dimensional vector $z_i = (z'_{i1}, \dots, z'_{iJ})'$ and then the x_i and z_i are summarized in the $(k_1 + J \cdot k_2)$ -dimensional vector $X_i = (x'_i, z'_i)'$. The β_j are summarized in the $J \cdot k_1$ -dimensional vector $\beta = (\beta'_1, \dots, \beta'_J)'$.

While the utilities u_{ij} are unobservable, the realizations of the following dummy variables can be observed ($i = 1, \dots, n; j = 1, \dots, J$):

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ chooses alternative } j \\ 0 & \text{otherwise} \end{cases}$$

According to the random utility maximization hypothesis, individual i chooses category j if the utility of alternative j is the largest of all utilities, i.e. $u_{ij} > u_{ij'}$ ($i = 1, \dots, n; j, j' = 1, \dots, J, j \neq j'$).

Choice probabilities (i.e. probabilities that i chooses j) in multinomial discrete choice models:

$$\begin{aligned}
 p_{ij}(X_i, \beta, \gamma) &= P(y_{ij}=1|X_i, \beta, \gamma) = P(u_{ij} > u_{ij'}; \forall j \neq j'|X_i, \beta, \gamma) \\
 &= P\left(\beta'_j x_i + \gamma'z_{ij} + \varepsilon_{ij} > \beta'_1 x_i + \gamma'z_{i1} + \varepsilon_{i1}; \dots; \right. \\
 &\quad \beta'_j x_i + \gamma'z_{ij} + \varepsilon_{ij} > \beta'_{j-1} x_i + \gamma'z_{i,j-1} + \varepsilon_{i,j-1}; \\
 &\quad \beta'_j x_i + \gamma'z_{ij} + \varepsilon_{ij} > \beta'_{j+1} x_i + \gamma'z_{i,j+1} + \varepsilon_{i,j+1}; \dots; \\
 &\quad \left. \beta'_j x_i + \gamma'z_{ij} + \varepsilon_{ij} > \beta'_J x_i + \gamma'z_{iJ} + \varepsilon_{iJ} \right) \\
 &= P\left(\varepsilon_{i1} - \varepsilon_{ij} < (\beta'_j x_i + \gamma'z_{ij}) - (\beta'_1 x_i + \gamma'z_{i1}); \dots; \right. \\
 &\quad \varepsilon_{i,j-1} - \varepsilon_{ij} < (\beta'_j x_i + \gamma'z_{ij}) - (\beta'_{j-1} x_i + \gamma'z_{i,j-1}); \\
 &\quad \varepsilon_{i,j+1} - \varepsilon_{ij} < (\beta'_j x_i + \gamma'z_{ij}) - (\beta'_{j+1} x_i + \gamma'z_{i,j+1}); \dots; \\
 &\quad \left. \varepsilon_{iJ} - \varepsilon_{ij} < (\beta'_j x_i + \gamma'z_{ij}) - (\beta'_J x_i + \gamma'z_{iJ}) \right) \\
 &= P\left(\varepsilon_{ij'} - \varepsilon_{ij} < (\beta'_j x_i + \gamma'z_{ij}) - (\beta'_{j'} x_i + \gamma'z_{ij'}); \forall j \neq j'\right)
 \end{aligned}$$

These choice probabilities are the basis for the discrete choice analysis. Different distribution assumptions for the stochastic component ε_{ij} lead to different choice probabilities and thus to different multinomial discrete choice models. The special case of $J = 2$ leads to binary discrete choice models.

Willingness to pay:

If one alternative specific attribute is a price or cost variable (alternatively also income as individual characteristic could be considered), the willingness to pay (WTP) for other alternative specific attributes can be derived. The basis for the calculation is the total derivative of the utility function with respect to the price or cost variable z_{ij1} and the other attribute z_{ij2} , assuming that all other explanatory variables are held fixed. This total derivative is then set to zero so that the utility does not change:

$$du_{ij} = \gamma_1 dz_{ij1} + \gamma_2 dz_{ij2} = 0$$

$$\gamma_1 dz_{ij1} = -\gamma_2 dz_{ij2}$$

For the change of z_{ij1} that keeps the utility constant in the case of a (marginal) change of z_{ij2} , it follows for the WTP:

$$\text{WTP} = \frac{dz_{ij1}}{dz_{ij2}} = -\frac{\gamma_2}{\gamma_1}$$

The negative sign of the WTP reveals that the two changes are in the opposite direction, i.e. to keep the utility constant, z_{ij1} increases (decreases) when z_{ij2} decreases (increases). If γ_2 is negative (e.g. for a transport time attribute in the choice among several means of transportation) and γ_1 (for prices or costs) is commonly negative, the WTP is negative (which means that the prices or costs increase for the same utility value when transport time decreases).

1.3 Maximum likelihood estimation

Basis:

While the distribution and thus the probability or density function $f(y; \theta)$ of a random variable y (e.g. Bernoulli distribution, Poisson distribution, normal distribution) is known, some parameters of this distribution that are summarized in the vector $\theta = (\theta_1, \theta_2, \dots, \theta_m)'$ are unknown and have to be estimated

→ For this purpose a random sample y_1, \dots, y_n of n individuals is drawn from the corresponding distribution where $f_i(y_i; \theta)$ is the probability or density function of y_i for individual i

Due to the independence of the y_i , the joint probability function (for discrete random variables) or density function (for continuous random variables) of the y_1, \dots, y_n is the product of the individual probability or density functions:

$$f(y_1, \dots, y_n; \theta) = f_1(y_1; \theta) \cdots f_n(y_n; \theta) = \prod_{i=1}^n f_i(y_i; \theta)$$

If this function is not considered as a function of the random sample y_1, \dots, y_n given the parameters in θ , but as a function of θ for a given random sample y_1, \dots, y_n , it can be interpreted as a likelihood function:

$$L(\theta) = \prod_{i=1}^n f_i(y_i; \theta)$$

It should be noted that the likelihood function as well as all following derived functions are random variables (or random vectors or random matrixes) before the drawing of a random sample.

The idea of the maximum likelihood method (ML) is to find the value $\hat{\theta}$ of θ that maximizes the likelihood function on the basis of the random sample y_1, \dots, y_n . However, the maximization procedure is generally not based on the likelihood function, but on the log-likelihood function, i.e. the following natural logarithm:

$$\log L(\theta) = \log f_1(y_1; \theta) + \dots + \log f_n(y_n; \theta) = \sum_{i=1}^n \log f_i(y_i; \theta)$$

The maximization of $\log L(\theta)$ leads to the same estimator $\hat{\theta}$ as the maximization of $L(\theta)$. Advantages of the use of the log-likelihood function:

- The use of $\log L(\theta)$ avoids extremely small values in the case of discrete random variables (due to the multiplication of probabilities and thus values that are smaller than one) and also high values in the case of continuous random variables (if values that are higher than one in the density functions are multiplied)
- Generally, the maximization process to receive $\hat{\theta}$ is much simpler for the log-likelihood function than for the likelihood function

Maximization approach:

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)' = \arg \max_{\theta} [\log L(\theta)] = \arg \max_{\theta} \left[\sum_{i=1}^n \log f_i(y_i; \theta) \right]$$

The ML estimator $\hat{\theta}$ is therefore the value of θ that maximizes the log-likelihood function.

The first derivative of the log-likelihood function is called score (function):

$$s(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \log L(\theta)}{\partial \theta_1} \\ \frac{\partial \log L(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial \log L(\theta)}{\partial \theta_m} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum_{i=1}^n \log f_i(y_i; \theta)}{\partial \theta_1} \\ \frac{\partial \sum_{i=1}^n \log f_i(y_i; \theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial \sum_{i=1}^n \log f_i(y_i; \theta)}{\partial \theta_m} \end{pmatrix}$$

Due to the sum of the terms in the log-likelihood function, the score is also an additive function:

$$s(\theta) = \sum_{i=1}^n s_i(\theta) = \sum_{i=1}^n \frac{\partial \log f_i(y_i; \theta)}{\partial \theta}$$

Expectation of the score at the true, but unknown parameter vector θ :

$$E[s(\theta)] = 0$$

The second derivative of the log-likelihood function is called Hessian matrix:

$$H(\theta) = \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} = \begin{pmatrix} \frac{\partial^2 \log L(\theta)}{(\partial \theta_1)^2} & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_m} \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\theta)}{(\partial \theta_2)^2} & \dots & \frac{\partial^2 \log L(\theta)}{\partial \theta_2 \partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_m \partial \theta_1} & \frac{\partial^2 \log L(\theta)}{\partial \theta_m \partial \theta_2} & \dots & \frac{\partial^2 \log L(\theta)}{(\partial \theta_m)^2} \end{pmatrix}$$

Due to the sum of the terms in the log-likelihood function, the Hessian matrix is also an additive function:

$$H(\theta) = \sum_{i=1}^n H_i(\theta) = \sum_{i=1}^n \frac{\partial^2 \log f_i(y_i; \theta)}{\partial \theta \partial \theta'}$$

A necessary condition for the ML estimator $\hat{\theta}$ is that the score for this value of θ is zero:

$$\left. \frac{\partial \log L(\theta)}{\partial \theta} \right|_{\hat{\theta}} = s(\hat{\theta}) = \sum_{i=1}^n s_i(\hat{\theta}) = 0$$

As a consequence, the maximization process of the ML estimator can be characterized as follows:

$$\hat{\theta} = \underset{\theta}{\text{arg solves}} \left[s(\theta) = \sum_{i=1}^n s_i(\theta) = \sum_{i=1}^n \frac{\partial \log f_i(y_i; \theta)}{\partial \theta} = 0 \right]$$

Additional necessary and sufficient condition for the ML estimator $\hat{\theta}$:

The Hessian matrix for this value of θ must be negative definite (if there is a solution at an inner point of the parameter space). The maximum can be local or global. In many simple cases the log-likelihood function is globally concave so that the solution of the first order condition leads to a unique and global maximum of the log-likelihood function.

All previous concepts for unconditional models with a random variable y_i can be simply transferred to conditional models and thus microeconomic models with a dependent variable y_i and k explanatory variables (as in the case of linear regression models) which are summarized in the vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})'$ of explanatory variables for individual i .

With the conditional probability or density function $f_i(y_i; \mathbf{x}_i, \theta)$ of y and a random sample (y_i, \mathbf{x}_i) ($i = 1, \dots, n$) it follows for the conditional joint probability or density function:

$$f(y_1, \dots, y_n; \mathbf{x}_1, \dots, \mathbf{x}_n, \theta) = f_1(y_1; \mathbf{x}_1, \theta) \cdots f_n(y_n; \mathbf{x}_n, \theta) = \prod_{i=1}^n f_i(y_i; \mathbf{x}_i, \theta)$$

It follows for the log-likelihood function, the score, the Hessian matrix, and the maximization approach:

$$\log L(\theta) = \sum_{i=1}^n \log f_i(y_i; \mathbf{x}_i, \theta)$$

$$s(\theta) = \sum_{i=1}^n \frac{\partial \log f_i(y_i; \mathbf{x}_i, \theta)}{\partial \theta}$$

$$H(\theta) = \sum_{i=1}^n \frac{\partial^2 \log f_i(y_i; \mathbf{x}_i, \theta)}{\partial \theta \partial \theta'}$$

$$\hat{\theta} = \arg \max_{\theta} \left[\sum_{i=1}^n \log f_i(y_i; \mathbf{x}_i, \theta) \right]$$

Application to multinomial discrete choice models:

In the following, the J-dimensional vector $y_i = (y_{i1}, \dots, y_{iJ})'$ comprises the observable dependent variables as discussed above and X_i comprises all explanatory variables. Furthermore, all (free) parameters (particularly in β and γ , but possibly also variance covariance parameters, see later) are summarized in the vector θ . In the case of multinomial discrete choice models, the y_i are multinomially distributed with the parameters 1 and the choice probabilities $p_{ij}(X_i, \beta, \gamma)$. Based on a random sample (X_i, y_i) for $i = 1, \dots, n$ individuals, the likelihood and log-likelihood functions are:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f_i(y_i; X_i, \theta) = \prod_{i=1}^n p_{i1}(X_i, \theta)^{y_{i1}} p_{i2}(X_i, \theta)^{y_{i2}} \dots p_{iJ}(X_i, \theta)^{y_{iJ}} \\ &= \prod_{i=1}^n \prod_{j=1}^J p_{ij}(X_i, \theta)^{y_{ij}} \end{aligned}$$

$$\log L(\theta) = \sum_{i=1}^n \sum_{j=1}^J y_{ij} \log p_{ij}(X_i, \theta)$$

It follows for the ML estimator: :

$$\hat{\theta} = \arg \max_{\theta} \left[\sum_{i=1}^n \sum_{j=1}^J y_{ij} \log p_{ij}(X_i, \theta) \right] = \arg \text{solves}_{\theta} \left[\sum_{i=1}^n \sum_{j=1}^J y_{ij} \frac{\partial \log p_{ij}(X_i, \theta)}{\partial \theta} = 0 \right]$$

Finite sample properties of an ML estimator $\hat{\theta}$ of θ :

- $\hat{\theta}$ is often a biased estimator of θ (the unbiased ML estimator of parameters in the classical linear regression model is an exception)
- $\hat{\theta}$ is generally not normally distributed (the normal distribution of ML estimators in the classical linear regression model is also an exception)
- The generally unknown small sample properties of ML estimators can be examined by Monte Carlo experiments for specific (microeconomic) models and specific parameter values

Asymptotic properties of the ML estimator $\hat{\theta}$ of θ (under several regularity conditions and if the underlying model is correctly specified):

- Consistency, i.e. $P(|\hat{\theta} - \theta| > \xi)$ converges (for $\xi > 0$) to zero for $n \rightarrow \infty$ or $\text{plim}(\hat{\theta}) = \theta$. This means that the asymptotic distribution of $\hat{\theta}$ is centered at θ and its variance goes to zero. An alternative notation for consistency is:

$$\hat{\theta} \xrightarrow{p} \theta$$

- Asymptotic normality
- Asymptotic efficiency

The asymptotic normality does not directly refer to the ML estimator $\hat{\theta}$, but to $\sqrt{n}(\hat{\theta}-\theta)$:

$$\sqrt{n}(\hat{\theta} - \theta) \stackrel{a}{\sim} N(0; I(\theta)^{-1}) \quad \text{or} \quad \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0; I(\theta)^{-1})$$

This means that $\sqrt{n}(\hat{\theta}-\theta)$ converges in distribution to the normal distribution, i.e. is asymptotically normally distributed with an expectation vector zero and the variance covariance matrix $I(\theta)^{-1}$. The matrix $I(\theta)$ is called information matrix and has the following form:

$$I(\theta) = -E\left(\frac{\partial^2 \log f_i(y_i; \theta)}{\partial \theta \partial \theta'}\right) = -E[H_i(\theta)]$$

The inverse of the information matrix is the Cramer Rao lower bound which implies that the difference between this matrix and the corresponding variance covariance matrix for any other consistent estimator of θ , for which $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotically normally distributed, is negative definite. Since $\sqrt{n}(\hat{\theta}-\theta)$ reaches this lower bound, the ML estimator is asymptotically efficient.

Information matrix equality at the true, but unknown parameter vector θ :

$$I(\theta) = -E[H_i(\theta)] = E[s_i(\theta)s_i(\theta)'] = \text{Var}[s_i(\theta)]$$

This equality means that the variance covariance matrix of the score for observation i is identical to the negative expectation of the Hessian matrix for i and thus the information matrix.

From the asymptotic normality of $\sqrt{n}(\hat{\theta}-\theta)$ it follows that the ML estimator $\hat{\theta}$ is approximately normally distributed for large but finite samples of n observations:

$$\hat{\theta} \stackrel{\text{appr}}{\sim} \mathbf{N}\left(\theta; [\mathbf{nI}(\theta)]^{-1}\right)$$

The variance covariance matrix of $\hat{\theta}$ thus has the following form:

$$\text{Var}(\hat{\theta}) = \begin{pmatrix} \text{Var}(\hat{\theta}_1) & \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \vdots & \text{Cov}(\hat{\theta}_1, \hat{\theta}_m) \\ \text{Cov}(\hat{\theta}_2, \hat{\theta}_1) & \text{Var}(\hat{\theta}_2) & \vdots & \text{Cov}(\hat{\theta}_2, \hat{\theta}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\theta}_m, \hat{\theta}_1) & \text{Cov}(\hat{\theta}_m, \hat{\theta}_2) & \cdots & \text{Var}(\hat{\theta}_m) \end{pmatrix} = [\mathbf{nI}(\theta)]^{-1} = -\mathbf{E}[\mathbf{nH}_i(\theta)]^{-1}$$

However, the information matrix (and this symmetric and positive definite variance covariance matrix) is unknown in practice since it depends on the unknown θ and thus has to be (consistently) estimated, e.g. for statistical tests and the construction of confidence intervals. $\text{Var}(\hat{\theta})$ can be estimated by including the ML estimator $\hat{\theta}$ instead of the true parameter vector:

$$\hat{\text{Var}}(\hat{\theta}) = -\mathbf{E}[\mathbf{nH}_i(\hat{\theta})]^{-1}$$

However, it is often not possible to obtain an exact expression for the expectation.

In practice the following three estimators for the variance covariance matrix (including consistent estimators of the information matrix) are commonly used:

$$\begin{aligned} \text{Vâr}(\hat{\theta})_1 &= \left[n\hat{I}(\hat{\theta})_1 \right]^{-1} = - \left[n \frac{1}{n} \sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} = - \left[\sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} \\ \text{Vâr}(\hat{\theta})_2 &= \left[n\hat{I}(\hat{\theta})_2 \right]^{-1} = \left[n \frac{1}{n} \sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})' \right]^{-1} = \left[\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})' \right]^{-1} \\ \text{Vâr}(\hat{\theta})_3 &= \left[n\hat{I}(\hat{\theta})_3 \right]^{-1} = \left[\sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} \left[\sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})' \right] \left[\sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} \\ &= \text{Vâr}(\hat{\theta})_1 \text{Vâr}(\hat{\theta})_2^{-1} \text{Vâr}(\hat{\theta})_1 \end{aligned}$$

All estimators of the information matrix are asymptotically equivalent, but can be very different in small samples.

Problem of maximization of the log-likelihood function:

A very efficient way to find the ML estimator $\hat{\theta}$ in the maximization process is the analytical optimization by equating the score to zero and solving for the maximizing parameters. However, in many cases the analytical optimization is not available due to the non-linearity of the log-likelihood function so that iterative numerical maximization algorithms have to be applied such as the Newton Raphson, BHHH, DFP, or BFGS algorithms.

1.4 Statistical testing

Problem:

The ML estimation of (e.g. multinomial discrete choice) models leads to a point estimate, but does not account for the sampling variability, which is the basis for statistical tests

Testable hypotheses refer to restrictions on the parameter space. The following general null and alternative hypotheses are based on q such restrictions:

$$H_0: c(\theta) = 0 \quad \Leftrightarrow \quad \begin{cases} c_1(\theta) = 0 \\ \vdots \\ c_q(\theta) = 0 \end{cases}$$

$$H_1: c(\theta) \neq 0$$

In the simplest case, the dimension of the function $c(\theta)$ is $q = 1$ and $c(\theta)$ refers to specific values of one parameter θ_1 of the vector θ which leads to the following null hypothesis (with an arbitrary constant a):

$$H_0: \theta_1 = a$$

On the basis of the ML estimation of (e.g. multinomial discrete choice) models, these null hypotheses can be statistically tested with the Wald or likelihood ratio test (or also the score test), which are asymptotically equivalent.

Wald test procedure:

This test is based on the unrestricted ML estimator $\hat{\theta}$. Under the null hypothesis $H_0: c(\theta) = 0$ it follows that $c(\hat{\theta})$ converges stochastically to zero since it is a consistent estimator of $c(\theta)$. Therefore, the alternative hypothesis implies that $c(\hat{\theta})$ strongly differs from the null vector.

Wald test statistic:

$$\text{WT} = n c(\hat{\theta})' \left[\frac{\partial c(\hat{\theta})}{\partial \theta'} \hat{I}(\hat{\theta})^{-1} \frac{\partial c(\hat{\theta})'}{\partial \theta} \right]^{-1} c(\hat{\theta})$$

$\hat{I}(\hat{\theta})$ is a consistent estimator of the information matrix and can e.g. be estimated on the basis of the three versions as discussed above. If $H_0: c(\theta) = 0$ is true, it follows that the Wald test statistic is asymptotically χ^2 distributed with q degrees of freedom, i.e.:

$$\text{WT} \xrightarrow{d} \chi_q^2$$

Thus, the null hypothesis is (for a large sample size n) rejected in favor of the alternative hypothesis at the significance level α if:

$$\text{WT} > \chi_{q;1-\alpha}^2$$

In the case of the specific simple null hypothesis $H_0: \theta_1 = a$ it follows that the corresponding Wald test statistic is asymptotically χ^2 distributed with one degree of freedom. Therefore, it follows the simplest and most important version of a Wald test statistic, namely the z-statistic (or t-statistic), which is asymptotically standard normally distributed:

$$\text{WT} = z = \frac{\hat{\theta}_1 - a}{\sqrt{\text{Var}(\hat{\theta}_1)}} \xrightarrow{d} N(0; 1)$$

The estimated variance of $\hat{\theta}_1$ is the l-th diagonal element of the estimated variance covariance matrix of $\hat{\theta}$ (e.g. on the basis of the three versions as discussed above). The null hypothesis is thus (for a large sample size n) rejected at the significance level α if:

$$|z| > z_{1-\alpha/2}$$

Restricted ML estimation:

An ML estimator $\hat{\theta}_r$ is called restricted ML estimator if the underlying ML estimation is based on specific restrictions for the unknown parameters. In the simplest case the restriction refers to specific values for unknown parameters. For example, if $\theta = (\alpha, \beta)'$, a possible restriction is $\alpha = 1$ so that the restricted ML estimator is $\hat{\theta}_r = (1, \hat{\beta}_r)'$, whereas the unrestricted ML estimator is $\hat{\theta}_u = (\hat{\alpha}_u, \hat{\beta}_u)'$.

Likelihood ratio test procedure:

This test is based on both the unrestricted ML estimator $\hat{\theta}_u$ and the restricted ML estimator $\hat{\theta}_r$. In the following, the value of the log-likelihood function at the restricted ML estimator is denoted by $\log L(\hat{\theta}_r)$ and the value of the log-likelihood function at the unrestricted ML estimator is denoted by $\log L(\hat{\theta}_u)$, whereby $\log L(\hat{\theta}_r) \leq \log L(\hat{\theta}_u)$. The null hypothesis $H_0: c(\theta) = 0$ implies that these values are very similar, whereas the alternative hypothesis implies that the values of the restricted and unrestricted log-likelihood functions are strongly different.

Likelihood ratio test statistic:

$$\text{LRT} = 2 \left[\log L(\hat{\theta}_u) - \log L(\hat{\theta}_r) \right]$$

If $H_0: c(\theta) = 0$ is true, it follows that the likelihood ratio test statistic is asymptotically χ^2 distributed with q degrees of freedom, i.e.:

$$\text{LRT} \xrightarrow{d} \chi_q^2$$

Thus, the null hypothesis is (for a large sample size n) rejected in favor of the alternative hypothesis at the significance level α if:

$$\text{LRT} > \chi_{q;1-\alpha}^2$$

The main advantage of this test is that it is easy to perform. The practical disadvantage is that two models have to be estimated separately.

1.5 Multinomial logit models

Fundamental distribution assumption:

The error terms ε_{ij} are independently and identically standard extreme value distributed over all categories $j = 1, \dots, J$ (and all $i = 1, \dots, n$). With this assumption a single difference of two ε_{ij} has a standard logistic distribution. In the special case of $J = 2$ multinomial logit models fall back to the binary logit model.

Choice probabilities in general multinomial logit models ($i = 1, \dots, n; j = 1, \dots, J$):

$$p_{ij}(X_i, \beta, \gamma) = P(y_{ij}=1|X_i, \beta, \gamma) = \frac{e^{\beta_j'x_i + \gamma'z_{ij}}}{\sum_{m=1}^J e^{\beta_m'x_i + \gamma'z_{im}}}$$

This means that the choice probabilities have a simple closed-form expression.

However, these choice probabilities comprise too many parameters in β and thus are not formally identified since any constant can be added to each of the parameter vectors β_1, \dots, β_J without changing the probabilities, i.e. only the differences between β_1, \dots, β_J are relevant. Therefore, one of these vectors has to be parameterized. Common approaches are to set the parameter vector for alternative 1 or for alternative J to zero, i.e. $\beta_1 = 0$ or $\beta_J = 0$. In the following, the second approach is considered.

On the basis of this normalization $\beta_J = 0$, the category J is the base category (or baseline) and provides the reference point for all other alternatives. This has to be considered for the interpretation of the estimation results (see later). If the numerator and denominator of the choice probabilities are divided by $e^{\beta_J'x_i + \gamma'z_{iJ}} = e^{0 + \gamma'z_{iJ}} = e^{\gamma'z_{iJ}}$, it follows:

$$p_{ij}(X_i, \beta, \gamma) = \frac{e^{\beta_j'x_i + \gamma'(z_{ij} - z_{iJ})}}{1 + \sum_{m=1}^{J-1} e^{\beta_m'x_i + \gamma'(z_{im} - z_{iJ})}} \quad \text{for } j = 1, \dots, J-1$$

$$p_{iJ}(X_i, \beta, \gamma) = \frac{1}{1 + \sum_{m=1}^{J-1} e^{\beta_m'x_i + \gamma'(z_{im} - z_{iJ})}}$$

→ These choice probabilities refer to the most flexible multinomial logit model approach which includes both individual characteristics and alternative specific attributes as explanatory variables. In many empirical studies, however, only one class of explanatory variables is examined. While the term “multinomial logit model” is not consistently used, it often refers to model approaches that include individual characteristics. Approaches with only alternative specific attributes as explanatory variables are often called “conditional logit models”.

Choice probabilities in (pure) multinomial logit models ($i = 1, \dots, n$; $j = 1, \dots, J$):

$$p_{ij}(x_i, \beta) = \frac{e^{\beta_j' x_i}}{\sum_{m=1}^J e^{\beta_m' x_i}}$$

Based on the aforementioned parameterization $\beta_J = 0$, the choice probabilities in such approaches can be alternatively written as follows:

$$p_{ij}(x_i, \beta) = \frac{e^{\beta_j' x_i}}{1 + \sum_{m=1}^{J-1} e^{\beta_m' x_i}} \quad \text{for } j = 1, \dots, J-1$$

$$p_{iJ}(x_i, \beta) = \frac{1}{1 + \sum_{m=1}^{J-1} e^{\beta_m' x_i}}$$

The inclusion of the ML estimator $\hat{\beta}$ into the choice probabilities leads to the corresponding estimator $\hat{p}_{ij}(x_i, \hat{\beta})$ of the choice probabilities for all categories $j = 1, \dots, J$. According to these formulas, the (estimators of) choice probabilities for alternative j imply that they do not only depend on the (estimator of the) parameter vector β_j , but on all other (estimators of) parameter vectors.

The parameter estimators cannot be interpreted like the estimators of the effect of the respective explanatory variable in the case of linear regression models. Instead, it follows for the estimator of the (partial) marginal probability effect of a (continuous) individual characteristic x_{ih} as explanatory variable in (pure) multinomial logit models ($i = 1, \dots, n$):

$$\frac{\partial \hat{p}_{ij}(x_i, \hat{\beta})}{\partial x_{ih}} = \hat{p}_{ij}(x_i, \hat{\beta}) \left[\hat{\beta}_{jh} - \sum_{m=1}^{J-1} \hat{p}_{im}(x_i, \hat{\beta}) \hat{\beta}_{mh} \right] \quad \text{for } j = 1, \dots, J-1$$

$$\frac{\partial \hat{p}_{iJ}(x_i, \hat{\beta})}{\partial x_{ih}} = -\hat{p}_{iJ}(x_i, \hat{\beta}) \sum_{m=1}^{J-1} \hat{p}_{im}(x_i, \hat{\beta}) \hat{\beta}_{mh}$$

Interpretation:

- This formula refers to the estimator of the effect of a small infinitesimal increase of x_{ih} on the change of the probability to choose alternative j
- This estimator of the marginal probability effect does not only depend on the ML estimator $\hat{\beta}_{jh}$ for j , but also on the estimators of the choice probabilities and thus the parameters for all other categories. Furthermore, it varies with different values of all individual characteristics.
- $\hat{\beta}_{jh}$ not even indicates the direction of the estimator of marginal probability effects (in contrast to the case of binary logit models), i.e. a positive (negative) $\hat{\beta}_{jh}$ does not necessarily lead to positive (negative) estimators of the effects

Based on y_1, \dots, y_n and x_1, \dots, x_n , it follows for the estimator of the average (partial) marginal probability effect ($AMPE_{hj}$) of the individual characteristic x_{ih} across all i in (pure) multinomial logit models:

$$\hat{AMPE}_{hj} = \frac{1}{n} \sum_{i=1}^n \hat{p}_{ij}(x_i, \hat{\beta}) \left[\hat{\beta}_{jh} - \sum_{m=1}^{J-1} \hat{p}_{im}(x_i, \hat{\beta}) \hat{\beta}_{mh} \right] \quad \text{for } j = 1, \dots, J-1$$

$$\hat{AMPE}_{hj} = \frac{1}{n} \sum_{i=1}^n \left[-\hat{p}_{ij}(x_i, \hat{\beta}) \sum_{m=1}^{J-1} \hat{p}_{im}(x_i, \hat{\beta}) \hat{\beta}_{mh} \right]$$

The (partial) marginal probability effect at the means of the individual characteristics across $i = 1, \dots, n$ can be correspondingly estimated.

For discrete individual characteristics (including dummy variables) as explanatory variables and for larger changes of continuous characteristics, the estimator of marginal probability effects can lead to very inaccurate results. The estimator of a discrete change of the choice probabilities $p_{ij}(x_i, \beta)$ due to a discrete change Δx_{ih} in (pure) multinomial logit models is as follows (for $j = 1, \dots, J-1$):

$$\begin{aligned} \Delta \hat{p}_{ij}(x_i, \hat{\beta}) &= \Delta P(y_{ij}=1|x_i, \hat{\beta}) = P(y_{ij}=1|x_i + \Delta x_{ih}, \hat{\beta}) - P(y_{ij}=1|x_i, \hat{\beta}) \\ &= \frac{e^{\hat{\beta}'_j x_i + \hat{\beta}_{jh} \Delta x_{ih}}}{1 + \sum_{m=1}^{J-1} e^{\hat{\beta}'_m x_i + \hat{\beta}_{mh} \Delta x_{ih}}} - \frac{e^{\hat{\beta}'_j x_i}}{1 + \sum_{m=1}^{J-1} e^{\hat{\beta}'_m x_i}} \end{aligned}$$

Since the sum over the estimated choice probabilities for all alternatives must be equal to one, the change of one estimator of probabilities is determined by the $J-1$ other changes so that it follows for the estimator of a discrete change of the choice probability $p_{ij}(x_i, \beta)$ due to Δx_{ih} :

$$\Delta \hat{p}_{ij}(x_i, \hat{\beta}) = - \sum_{j=1}^{J-1} \Delta \hat{p}_{ij}(x_i, \hat{\beta})$$

Remarks:

- As in the case of estimated marginal probability effects, the sign of the estimated change $\Delta \hat{p}_{ij}(x_i, \hat{\beta})$ for all $j = 1, \dots, J$ due to a discrete change Δx_{ih} of the individual characteristic x_{ih} need not coincide with the sign of the corresponding ML estimator $\hat{\beta}_{jh}$ for j . If e.g. $\hat{\beta}_{jh}$ is positive, the numerator of the first term in $\Delta \hat{p}_{ij}(x_i, \hat{\beta})$ increases with increasing x_{ih} . However, it is possible that the denominator increases even more due to the values $\hat{\beta}_{mh}$ ($\forall m \neq j$).
- As in the case of estimated marginal probability effects, the estimated changes $\Delta \hat{p}_{ij}(x_i, \hat{\beta})$ vary with different values not only of x_{ih} but also with different values of all other individual characteristics and thus across different observations
- On this basis, average discrete changes of $p_{ij}(x_i, \beta)$ ($j = 1, \dots, J$) across all i and corresponding discrete changes of $p_{ij}(x_i, \beta)$ at the means of the individual characteristics across $i = 1, \dots, n$ can be estimated

While the ML estimator $\hat{\beta}_{jh}$ neither indicates the extent nor the direction of the effect of an individual characteristic x_{ih} on the estimator $\hat{p}_{ij}(x_i, \hat{\beta})$ of the choice probability for alternative j , it nevertheless has an important informative value. This can be recognized by dividing the estimator $\hat{p}_{ij}(x_i, \hat{\beta})$ of the choice probability for j and the corresponding estimator $\hat{p}_{iJ}(x_i, \hat{\beta})$ for the base category J . For the so-called odds it follows for $j = 1, \dots, J-1$:

$$\frac{\hat{p}_{ij}(x_i, \hat{\beta})}{\hat{p}_{iJ}(x_i, \hat{\beta})} = \frac{e^{\hat{\beta}'_j x_i}}{1 + \sum_{m=1}^{J-1} e^{\hat{\beta}'_m x_i}} = e^{\hat{\beta}'_j x_i} = e^{\hat{\beta}'_{j1} x_{i1} + \dots + \hat{\beta}'_{jk_1} x_{ik_1}}$$

Interpretation:

This formula shows that although the ML estimator $\hat{\beta}_{jh}$ does not indicate the effect of x_{ih} on the estimator $\hat{p}_{ij}(x_i, \hat{\beta})$ of the choice probability for j alone, it indicates the direction of the effect on $\hat{p}_{ij}(x_i, \hat{\beta})$ relative to the estimator $\hat{p}_{iJ}(x_i, \hat{\beta})$ of the choice probability for the base category J . If $\hat{\beta}_{jh}$ is positive (negative), an increase of x_{ih} increases (decreases) the odds, i.e. $\hat{p}_{ij}(x_i, \hat{\beta})$ relative to $\hat{p}_{iJ}(x_i, \hat{\beta})$.

The previous analysis of the estimation of the probability effects relative to the base category can be extended to the discussion of the odds for two arbitrary alternatives j and r . It follows ($\forall r \neq j$):

$$\frac{\hat{p}_{ij}(x_i, \hat{\beta})}{\hat{p}_{ir}(x_i, \hat{\beta})} = \frac{\frac{e^{\hat{\beta}'_j x_i}}{1 + \sum_{m=1}^{J-1} e^{\hat{\beta}'_m x_i}}}{\frac{e^{\hat{\beta}'_r x_i}}{1 + \sum_{m=1}^{J-1} e^{\hat{\beta}'_m x_i}}} = \frac{e^{\hat{\beta}'_j x_i}}{e^{\hat{\beta}'_r x_i}} = e^{(\hat{\beta}'_j - \hat{\beta}'_r) x_i} = e^{(\hat{\beta}_{j1} - \hat{\beta}_{r1}) x_{i1} + \dots + (\hat{\beta}_{jk_1} - \hat{\beta}_{rk_1}) x_{ik_1}}$$

Interpretation:

This formula implies that the difference between the two ML estimators $\hat{\beta}_{jh}$ and $\hat{\beta}_{rh}$ indicates the direction of the effect of x_{ih} on the estimator $\hat{p}_{ij}(x_i, \hat{\beta})$ of the choice probability for category j relative to the estimator $\hat{p}_{ir}(x_i, \hat{\beta})$ of the choice probability for category r . If $\hat{\beta}_{jh}$ is greater (less) than $\hat{\beta}_{rh}$, an increase of x_{ih} increases (decreases) $\hat{p}_{ij}(x_i, \hat{\beta})$ relative to $\hat{p}_{ir}(x_i, \hat{\beta})$.

Choice probabilities in conditional logit models ($i = 1, \dots, n; j = 1, \dots, J$):

$$p_{ij}(z_i, \gamma) = \frac{e^{\gamma' z_{ij}}}{\sum_{m=1}^J e^{\gamma' z_{im}}}$$

The inclusion of the ML estimator $\hat{\gamma}$ into these choice probabilities leads to the corresponding estimator $\hat{p}_{ij}(z_i, \hat{\gamma})$ of the choice probabilities for all categories $j = 1, \dots, J$.

Differences to (pure) multinomial logit models:

- The ML estimator $\hat{\gamma}$ is no longer choice-specific so that no normalization is necessary
- The estimators of the choice probabilities for alternative j do not only depend on the attributes z_{ij} , but also on all other alternative specific attributes in $z_i = (z'_{i1}, \dots, z'_{iJ})'$
- Since the alternative specific attributes vary across the categories and the individuals, the ML estimation of conditional logit models with econometric software packages such as Stata requires another specific data organization (“long format” in contrast to “wide format” for pure multinomial logit models)

Example: Data organization in the conditional logit model

In order to examine the effect of the daily travel price (in Euro) and daily travel time (in minutes) on the choice between the use of car alone, carpool, bus, and train for the journey to work, the following table shows an exemplary data organization for the first three persons:

Person i	Transport modes	Choice	Travel price	Travel time
1	Car alone	0	6	50
1	Carpool	0	3	50
1	Bus	0	7	60
1	Train	1	9	30
2	Car alone	1	12	70
2	Carpool	0	4	70
2	Bus	0	7	90
2	Train	0	6	80
3	Car alone	0	3	20
3	Carpool	1	1	20
3	Bus	0	4	30
3	Train	0	5	20
⋮	⋮	⋮	⋮	⋮

Estimator of the (partial) marginal probability effect of a (continuous) alternative specific attribute z_{ijh} of alternative j on the choice of the same alternative j in conditional logit models ($i = 1, \dots, n, j = 1, \dots, J$):

$$\frac{\partial \hat{p}_{ij}(z_i, \hat{\gamma})}{\partial z_{ijh}} = \hat{p}_{ij}(z_i, \hat{\gamma}) [1 - \hat{p}_{ij}(z_i, \hat{\gamma})] \hat{\gamma}_h$$

Estimator of the (partial) marginal probability effect of a (continuous) alternative specific attribute z_{imh} of alternative m on the choice of another alternative j in conditional logit models ($i = 1, \dots, n, j = 1, \dots, J$):

$$\frac{\partial \hat{p}_{ij}(z_i, \hat{\gamma})}{\partial z_{imh}} = -\hat{p}_{ij}(z_i, \hat{\gamma}) \hat{p}_{im}(z_i, \hat{\gamma}) \hat{\gamma}_h \quad \forall m \neq j$$

In contrast to (pure) multinomial logit models, the sign of parameter estimators gives information about the direction of estimated marginal probability effects:

- If $\hat{\gamma}_h$ (e.g. the estimated parameter for price) is positive (negative), an increase of an attribute z_{ijh} for category j (e.g. price for bus) leads to an increase (decrease) of $\hat{p}_{ij}(z_i, \hat{\gamma})$ for the same category j (e.g. the estimated choice probability for bus)
- If $\hat{\gamma}_h$ (e.g. the estimated parameter for price) is positive (negative), an increase of an attribute z_{imh} for category m (e.g. price for train) leads to a decrease (increase) of $\hat{p}_{ij}(z_i, \hat{\gamma})$ for another category j (e.g. the estimated choice probability for bus)

→ As already discussed above, general multinomial logit models can include both individual characteristics and alternative specific attributes as explanatory variables. In this case all previous interpretations from the (pure) multinomial and conditional logit models hold true. Similar to conditional logit models it is important to consider the specific data organization.

Example: Data organization in the general multinomial logit model

The previous example of the analysis of the choice between the use of car alone, carpool, bus, and train for the journey to work now additionally includes the individual characteristic age (in years) as explanatory variable. The following table shows an exemplary data organization for the first two persons:

Person i	Transport modes	Choice	Travel price	Travel time	Age
1	Car alone	0	6	50	32
1	Carpool	0	3	50	32
1	Bus	0	7	60	32
1	Train	1	9	30	32
2	Car alone	1	12	70	51
2	Carpool	0	4	70	51
2	Bus	0	7	90	51
2	Train	0	6	80	51
⋮	⋮	⋮	⋮	⋮	⋮

1.6 Applications

Example 1: Determinants of secondary school choice (I)

By using a (pure) multinomial logit model, the effect of the following individual characteristics on the choice of 675 pupils in Germany between the three secondary school types Hauptschule, Realschule, and Gymnasium is analyzed:

- Years of education of the mother (motheduc)
- Dummy variable for labor force participation of the mother (mothlnlf) that takes the value one if the mother is employed
- Logarithm of household income (loghhincome)
- Logarithm of household size (loghhsize)
- Rank by age among the siblings (birthorder)
- Year dummies for 1995-2002

The three alternatives of the multinomial dependent variable secondary school (schooltype) take the values one for Hauptschule, two for Realschule, and three for Gymnasium, whereby Hauptschule is chosen as base category. As a consequence, two vectors of parameters for the alternatives Realschule and Gymnasium are estimated. The ML estimation of the multinomial logit model with Stata leads to the following results:

Example 1: Determinants of secondary school choice (II)

```
mlogit schooltype motheduc mothinlf loghhincome loghhsizе birthorder year1995 year1996
year1997 year1998 year1999 year2000 year2001 year2002, base(1)
```

```
Multinomial logistic regression          Number of obs   =          675
                                          LR chi2(26)     =          221.20
                                          Prob > chi2     =           0.0000
Log likelihood = -622.24169              Pseudo R2       =           0.1509
```

schooltype	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Hauptschule	(base outcome)					
Realschule						
motheduc	.2987624	.0789985	3.78	0.000	.1439282	.4535967
mothinlf	-.3910109	.2284116	-1.71	0.087	-.8386895	.0566677
loghhincome	.4073928	.2254481	1.81	0.071	-.0344774	.8492631
loghhsizе	-1.145075	.4452722	-2.57	0.010	-2.017792	-.2723571
birthorder	-.1229556	.1255023	-0.98	0.327	-.3689357	.1230245
year1995	.0767306	.4236393	0.18	0.856	-.753587	.9070483
year1996	.2193122	.4457021	0.49	0.623	-.6542478	1.092872
year1997	-.1001198	.4370522	-0.23	0.819	-.9567264	.7564868
year1998	.4394545	.4872894	0.90	0.367	-.5156151	1.394524
year1999	.5753325	.4617021	1.25	0.213	-.3295869	1.480252
year2000	.1609341	.4550522	0.35	0.724	-.7309519	1.05282
year2001	.4590218	.4554232	1.01	0.314	-.4335913	1.351635
year2002	.1428137	.4488494	0.32	0.750	-.7369149	1.022542
_cons	-5.795185	2.312289	-2.51	0.012	-10.32719	-1.263182

Example 1: Determinants of secondary school choice (III)

Gymnasium							
motheduc		.6554335	.0811063	8.08	0.000	.496468	.814399
mothlnlf		-.3775209	.2353811	-1.60	0.109	-.8388595	.0838177
loghhincome		1.710194	.2832594	6.04	0.000	1.155016	2.265372
loghhsz		-1.471701	.4843607	-3.04	0.002	-2.42103	-.5223712
birthorder		-.2736975	.1359262	-2.01	0.044	-.540108	-.007287
year1995		.0761265	.4281856	0.18	0.859	-.7631019	.915355
year1996		.1559678	.4471054	0.35	0.727	-.7203426	1.032278
year1997		-.6671805	.4582662	-1.46	0.145	-1.565366	.2310047
year1998		.1200219	.4989005	0.24	0.810	-.8578052	1.097849
year1999		-.3979308	.4970546	-0.80	0.423	-1.37214	.5762783
year2000		-.0598782	.4663055	-0.13	0.898	-.9738202	.8540638
year2001		.0855021	.4662402	0.18	0.854	-.8283119	.9993161
year2002		-.3087189	.4494414	-0.69	0.492	-1.189608	.5721701
_cons		-23.05832	2.962141	-7.78	0.000	-28.86401	-17.25263

(The presentation of estimation results in empirical studies commonly comprises the parameter estimates, the z-statistics or estimated standard deviations of the estimated parameters, and some information about the significance of the rejection of the null hypothesis that the parameter is zero).

Example 1: Determinants of secondary school choice (IV)

Interpretation:

- The value of 221.20 for the likelihood ratio test statistic implies that the null hypothesis that all 26 parameters are zero (which would imply that no explanatory variable has an effect on the choice of Realschule or Gymnasium relative to Hauptschule) can be rejected at any common significance level
- The parameter estimates for motheduc are positive for both alternatives Realschule and Gymnasium and highly significantly different from zero due to the z-statistics of 3.78 for Realschule and 8.08 for Gymnasium
- These parameter estimates therefore imply that the years of education of the mother have a strong significantly positive effect on the (probability of the) choice of Realschule compared with Hauptschule and additionally on the (probability of the) choice of Gymnasium compared with Hauptschule
- The negative value of the difference $0.299 - 0.655 = -0.356$ of the parameter estimates for motheduc for Realschule and Gymnasium implies that the years of education of the mother have a negative effect on the choice of Realschule relative to Gymnasium or conversely a positive effect on the choice of Gymnasium relative to Realschule. The significance of these effects has to be analyzed by choosing Realschule or Gymnasium as base category.

Example 1: Determinants of secondary school choice (V)

Wald and likelihood ratio tests:

As an example, the null hypothesis that motheduc has no effect on the secondary school choice, i.e. that the two corresponding parameters are zero, is tested. The command for the Wald test in Stata is:

```
test motheduc

( 1)  [Hauptschule]o.motheduc = 0
( 2)  [Realschule]motheduc = 0
( 3)  [Gymnasium]motheduc = 0
      Constraint 1 dropped

      chi2( 2) =    73.70
      Prob > chi2 =    0.0000
```

For the application of the likelihood ratio test, the Stata command “estimates store unrestr” after the unrestricted ML estimation and the command “estimates store restr” after the restricted ML estimation (“mlogit schooltype mothinf logh-hincome loghhszize birthorder year1995 year1996 year1997 year1998 year1999 year2000 year2001 year2002, base(1)”) are necessary (the choice of the names is arbitrary). The command for the likelihood ratio test in Stata is then:

```
lrtest unrestr restr
```

```
Likelihood-ratio test                                LR chi2(2) =    107.99
(Assumption: restr nested in unrestr)                Prob > chi2 =    0.0000
```

Example 1: Determinants of secondary school choice (VI)

The estimation of the average marginal probability effect of motheduc across all 675 pupils on the choice of Gymnasium with Stata leads to the following results:

```
margins, dydx(motheduc) predict(outcome(3))
```

```
Average marginal effects          Number of obs   =          675
```

```
Model VCE      : OIM
```

```
Expression    : Pr(schooltype==3), predict(outcome(3))
```

```
dy/dx w.r.t. : motheduc
```

		Delta-method				[95% Conf. Interval]	
	dy/dx	Std. Err.	z	P> z			
motheduc	.0886491	.0084999	10.43	0.000	.0719897	.1053085	

This value of 0.0886 means that an increase of the years of education of the mother by one (unit) leads to an approximately estimated increase of the choice probability for Gymnasium by 8.86 percentage points. The corresponding values for Hauptschule and Realschule are -0.0797 and -0.0089. These values differ from the estimates of the marginal probability effect at the means of the individual characteristics across all 675 pupils. For the effects on the choice of Gymnasium the estimation with Stata leads to the following results: 38

Example 1: Determinants of secondary school choice (VIII)

The estimation of the average probabilities of the choice of Gymnasium across all 675 pupils for the minimum and maximum values of motheduc = 7 and motheduc = 18 years with Stata leads to the following results:

```
margins, at(motheduc=7) predict(outcome(3))
```

```
Predictive margins                                Number of obs   =           675
Model VCE      : OIM
Expression     : Pr(schooltype==Gymnasium), predict(outcome(3))
at             : motheduc           =           7
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
---+-----						
_cons	.0747206	.0196424	3.80	0.000	.0362222	.1132189

```
margins, at(motheduc=18) predict(outcome(3))
```

```
Predictive margins                                Number of obs   =           675
Model VCE      : OIM
Expression     : Pr(schooltype==Gymnasium), predict(outcome(3))
at             : motheduc           =           18
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
---+-----						
_cons	.9023163	.0326609	27.63	0.000	.838302	.9663305

Example 1: Determinants of secondary school choice (IX)

In contrast, the estimation of e.g. the probability of the choice of Gymnasium for the maximum value of motheduc = 18 years at the means of the other individual characteristics with Stata leads to the following results:

```
margins, at((means)_all motheduc=18) predict(outcome(3))
```

```
Adjusted predictions                               Number of obs   =           675
```

```
Model VCE      : OIM
```

```
Expression     : Pr(schooltype==3), predict(outcome(3))
```

```
at             : motheduc           =           18
                mothlnlf           =    .5525926 (mean)
                loghhincome        =   11.05839 (mean)
                loghhsize          =    1.412881 (mean)
                birthorder         =     1.76 (mean)
                year1995           =    .1377778 (mean)
                year1996           =     .12 (mean)
                year1997           =    .1111111 (mean)
                year1998           =    .0888889 (mean)
                year1999           =    .1007407 (mean)
                year2000           =    .1037037 (mean)
                year2001           =    .1185185 (mean)
                year2002           =    .117037 (mean)
```

		Delta-method				
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
__cons	.918859	.0294339	31.22	0.000	.8611696	.9765485

Example 1: Determinants of secondary school choice (X)

The analysis of discrete changes of the choice probabilities for Hauptschule, Realschule, and Gymnasium due to a discrete change of motheduc requires the estimation of differences between the probabilities (an alternative for a discrete explanatory variable such as mothinf is the ML estimation with Stata by prefixing “i.” as well as the use of the commands as before, like “margins, dydx(mothinf) predict(outcome(3))”). For example, the average probabilities across all 675 pupils for several values of motheduc can be estimated. The following table reports the results:

motheduc (in years)	Hauptschule	Realschule	Gymnasium
7	0.6788	0.2465	0.0747
9	0.4966	0.3171	0.1864
10	0.3953	0.3339	0.2708
11	0.2975	0.3322	0.3703
12	0.2112	0.3121	0.4766
13	0.1418	0.2777	0.5805
14	0.0906	0.2353	0.6741
15	0.0555	0.1914	0.7531
16	0.0329	0.1506	0.8165
18	0.0108	0.0869	0.9023

Example 1: Determinants of secondary school choice (XI)

Interpretation:

- The increase from the minimum value of seven years to the maximum value of 18 years of education of the mother decreases the estimated average choice probabilities for Hauptschule and Realschule by 66.80 and 15.96 percentage points (from 0.6788 to 0.0108 and from 0.2465 to 0.0869), but increases the estimated average choice probability for Gymnasium by 82.76 percentage points (from 0.0747 to 0.9023). In the case of Gymnasium this means an immense increase of more than 1100%.
- The estimated change of the average choice probabilities for an increase of the years of education of the mother from nine to ten (which can be interpreted as the effect of “mittlere Reife”, i.e. the Realschule degree for the mother) is -10.13 percentage points for the case of Hauptschule and 8.44 percentage points for the case of Gymnasium
- The values for an increase of motheduc from ten to 13 years (which can be interpreted as the effect of “Abitur”, i.e. the Gymnasium degree for the mother) are -0.2535 for Hauptschule and 0.3097 for Gymnasium
- The values for an increase of motheduc from 13 to 16 years (which can be interpreted as the effect of an university degree for the mother) are -0.1089 for Hauptschule and 0.2360 for Gymnasium

Example 1: Determinants of secondary school choice (XII)

Alternatively estimated variances of the estimated parameters:

```
mlogit schooltype motheduc mothinflf loghhincome loghhsizе birthorder year1995 year1996 year1997  
year1998 year1999 year2000 year2001 year2002, base(1) robust
```

```
Multinomial logistic regression                Number of obs   =           675  
                                                Wald chi2(26)   =           115.21  
                                                Prob > chi2     =            0.0000  
Log pseudolikelihood = -622.24169             Pseudo R2      =            0.1509
```

schooltype	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
Hauptschule	(base outcome)					
Realschule						
motheduc	.2987624	.0854129	3.50	0.000	.1313561	.4661687
mothinflf	-.3910109	.2310222	-1.69	0.091	-.8438061	.0617843
loghhincome	.4073928	.2341138	1.74	0.082	-.0514617	.8662474
loghhsizе	-1.145075	.438372	-2.61	0.009	-2.004268	-.2858814
birthorder	-.1229556	.1229718	-1.00	0.317	-.363976	.1180648
year1995	.0767306	.4160286	0.18	0.854	-.7386705	.8921318
year1996	.2193122	.4372284	0.50	0.616	-.6376397	1.076264
year1997	-.1001198	.4206516	-0.24	0.812	-.9245817	.7243421
year1998	.4394545	.4871378	0.90	0.367	-.515318	1.394227
year1999	.5753325	.459803	1.25	0.211	-.3258649	1.47653
year2000	.1609341	.4364891	0.37	0.712	-.6945688	1.016437
year2001	.4590218	.4513467	1.02	0.309	-.4256015	1.343645
year2002	.1428137	.4384459	0.33	0.745	-.7165244	1.002152
_cons	-5.795185	2.411338	-2.40	0.016	-10.52132	-1.06905

Example 1: Determinants of secondary school choice (XIII)

Gymnasium							
motheduc		.6554335	.0898482	7.29	0.000	.4793342	.8315328
mothlnlf		-.3775209	.2406994	-1.57	0.117	-.849283	.0942412
loghhincome		1.710194	.3330751	5.13	0.000	1.057379	2.363009
loghhsz		-1.471701	.5114952	-2.88	0.004	-2.474213	-.4691885
birthorder		-.2736975	.1399788	-1.96	0.051	-.5480508	.0006558
year1995		.0761265	.4296538	0.18	0.859	-.7659795	.9182326
year1996		.1559678	.4371795	0.36	0.721	-.7008884	1.012824
year1997		-.6671805	.4614188	-1.45	0.148	-1.571545	.2371837
year1998		.1200219	.5041279	0.24	0.812	-.8680506	1.108094
year1999		-.3979308	.4953305	-0.80	0.422	-1.368761	.5728991
year2000		-.0598782	.4534223	-0.13	0.895	-.9485696	.8288131
year2001		.0855021	.4658155	0.18	0.854	-.8274795	.9984837
year2002		-.3087189	.4486799	-0.69	0.491	-1.188115	.5706776
_cons		-23.05832	3.484456	-6.62	0.000	-29.88773	-16.22891

The Stata command “robust” includes the robust estimation of the variances of the estimated parameters from the quasi maximum likelihood theory, i.e. the (“sandwich”) estimator that includes both the Hessian matrix and the score at the ML estimator. In this case, likelihood ratio tests are not appropriate and not possible with Stata. By default, the estimated variances are only based on the inclusion of the Hessian matrix at the ML estimator.

Example 2: Determinants of travel mode choice (I)

By using a conditional logit model, the effect of the following two alternative specific attributes on the choice between the four travel modes air (i.e. travel by plain), train (i.e. travel by train), bus (i.e. travel by bus), and car (i.e. travel by car) is examined on the basis of data from 210 persons travelling between Sydney and Melbourne:

- Travelcost (i.e. generalized costs of travel in US dollars, which is equal to the sum of in vehicle costs and the product of travel time and the value of travel time savings)
- Termtime (i.e. terminal time in minutes, which is zero for car transportation)

In addition to such attributes, conditional logit models should generally include alternative specific constants in order to capture initial preferences for the different alternatives. Similar to the case of the parameters of individual characteristics in (pure) multinomial logit models, only J-1 alternative specific constants can be included so that category J is again the base category. The ML estimations of the conditional logit models (using air as base category, respectively) lead to the following results (in line with the table on page 30, “travelmode” is a possible name for the identification of the four alternatives, “choice” is a possible name for the dependent variable, and “id” is a possible name for the identification of the persons in the sample):

Example 2: Determinants of travel mode choice (II)

asclogit choice travelcost, case(id) alternatives(travelmode) noconstant

```
Alternative-specific conditional logit      Number of obs      =      840
Case variable: id                        Number of cases     =      210
Alternative variable: travelmode          Alts per case: min =         4
                                           avg =         4.0
                                           max =         4

                                           Wald chi2(1)       =      19.23
Log likelihood = -280.62752                Prob > chi2        =      0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
travelmode						
travelcost	-.0146285	.0033362	-4.38	0.000	-.0211673	-.0080897

Example 2: Determinants of travel mode choice (III)

asclogit choice travelcost, case(id) alternatives(travelmode)

```
Alternative-specific conditional logit      Number of obs      =      840
Case variable: id                        Number of cases     =      210
Alternative variable: travelmode          Alts per case: min =      4
                                           avg =      4.0
                                           max =      4
                                           Wald chi2(1)      =      25.07
Log likelihood = -269.87751              Prob > chi2        =      0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
travelmode						
travelcost	-.0199335	.0039812	-5.01	0.000	-.0277365	-.0121305
air	(base alternative)					
train						
_cons	.6307709	.218141	2.89	0.004	.2032224	1.058319
bus						
_cons	-.3661339	.2374276	-1.54	0.123	-.8314835	.0992156
car						
_cons	-.0827706	.1898705	-0.44	0.663	-.4549098	.2893687

Example 2: Determinants of travel mode choice (IV)

asclogit choice travelcost termtime, case(id) alternatives(travelmode)

```
Alternative-specific conditional logit      Number of obs      =      840
Case variable: id                        Number of cases    =      210
Alternative variable: travelmode          Alts per case: min =         4
                                           avg =         4.0
                                           max =         4
                                           Wald chi2(2)      =      100.76
Log likelihood = -199.97662                Prob > chi2        =      0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
travelmode						
travelcost	-.0157837	.0043828	-3.60	0.000	-.0243739	-.0071936
termtime	-.0970905	.0104351	-9.30	0.000	-.1175429	-.0766381
air	(base alternative)					
train						
_cons	-1.853358	.3700925	-5.01	0.000	-2.578726	-1.12799
bus						
_cons	-2.565624	.3843251	-6.68	0.000	-3.318887	-1.812361
car						
_cons	-5.776359	.6559187	-8.81	0.000	-7.061936	-4.490782

Example 2: Determinants of travel mode choice (V)

Interpretation:

- The costs of a travel mode j significantly decreases the probability of the choice of j (= estimated own cost effect) and increases the probability of the choice of another travel mode $m \neq j$ (= estimated cross cost effect), *ceteris paribus*. Terminal time has also a significantly negative effect on the own alternative.
- The initial preferences for bus and car are not significantly different relative to air if it is only controlled for costs
- In contrast, the initial preferences are significantly lower for train, bus, and especially car relative to air if it is additionally controlled for termtime

Estimation and interpretation of WTP (in the last conditional logit model):

- $W\hat{T}P_{\text{termtime}} = -(-0.097/(-0.016)) = -6.15$
 - Therefore, the result suggests that the persons are on average willing to pay 6.15 US dollars for a one minute less terminal time
-

Example 2: Determinants of travel mode choice (VI)

Remarks to estimated WTP:

- The estimations are only useful if both corresponding parameters are significantly different from zero
- Different base categories in the case of discrete attributes lead to different WTP estimates

Wald and likelihood ratio tests:

As an example, the null hypothesis that neither travelcost nor termtime has any effect on the travel mode choice in model (3), i.e. that the two corresponding parameters are zero, is tested. The command for the Wald test in STATA is (this Wald test statistic is already reported in the underlying ML estimation with Stata since travelcost and termtime are the only explanatory variables so that the tested null hypotheses are identical):

```
test travelcost=termtime=0

( 1)  [travelmode]travelcost - [travelmode]termtime = 0
( 2)  [travelmode]travelcost = 0

      chi2( 2) =   100.76
      Prob > chi2 =    0.0000
```

Example 2: Determinants of travel mode choice (VII)

With respect to the application of the likelihood ratio test, the Stata command “estimates store unrestr” after the unrestricted ML estimation and the command “estimates store restr” after the restricted ML estimation are again necessary. The command for the likelihood ratio test in Stata is then:

```
lrtest unrestr restr
```

```
Likelihood-ratio test                LR chi2(2)  = 167.56
Assumption: restr nested in unrestr)  Prob > chi2 =   0.0000
```

Estimation of marginal probability effects:

- The estimation of average marginal probability effects (and generally a “margins” command) is only possible with Stata 16, but not with Stata 15(!)
- The Stata command “estat mfx” reports the estimated marginal probability effects at the means of the explanatory variables
- While this refers to all explanatory variables, the additional Stata command “varlist()” allows the limitation on a subset of explanatory variables
- The marginal probability effects can also be estimated at specific values of the explanatory variables

The estimation of marginal probability effects for travelcost at the means of the explanatory variables in model (3) with Stata leads to the following results: 52

Example 2: Determinants of travel mode choice (VIII)

estat mfx, varlist(travelcost)

Pr(choice = air|1 selected) = .25665456

variable		dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
travelcost							
air		-.003011	.000881	-3.42	0.001	-.004737 -.001285	102.65
train		.001232	.000379	3.25	0.001	.00049 .001975	130.2
bus		.000429	.000151	2.84	0.005	.000133 .000725	115.26
car		.00135	.000434	3.11	0.002	.000499 .002201	95.414

Pr(choice = train|1 selected) = .30421465

variable		dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
travelcost							
air		.001232	.000379	3.25	0.001	.00049 .001975	102.65
train		-.003341	.000954	-3.50	0.000	-.005211 -.001471	130.2
bus		.000508	.000178	2.85	0.004	.000159 .000857	115.26
car		.0016	.000497	3.22	0.001	.000627 .002574	95.414

Pr(choice = bus|1 selected) = .10584909

variable		dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
travelcost							
air		.000429	.000151	2.84	0.005	.000133 .000725	102.65
train		.000508	.000178	2.85	0.004	.000159 .000857	130.2
bus		-.001494	.00049	-3.05	0.002	-.002453 -.000534	115.26
car		.000557	.000196	2.84	0.004	.000173 .000941	95.414

Example 2: Determinants of travel mode choice (IX)

Pr(choice = car|1 selected) = .3332817

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
travelcost								
air	.00135	.000434	3.11	0.002	.000499	.002201		102.65
train	.0016	.000497	3.22	0.001	.000627	.002574		130.2
bus	.000557	.000196	2.84	0.004	.000173	.000941		115.26
car	-.003507	.00101	-3.47	0.001	-.005486	-.001528		95.414

Interpretation:

- At the means of the explanatory variables the estimated choice probabilities for the four travel modes are $\hat{p}_{i1}(\bar{z}, \hat{\gamma}) = 0.2567$ for air, $\hat{p}_{i2}(\bar{z}, \hat{\gamma}) = 0.3042$ for train, $\hat{p}_{i3}(\bar{z}, \hat{\gamma}) = 0.1058$ for bus, and $\hat{p}_{i4}(\bar{z}, \hat{\gamma}) = 0.3333$ for car
- It follows e.g. for the estimated marginal probability effects of travelcost for car on the choice of car and train (at the means of the explanatory variables):

$$\hat{p}_{i4}(\bar{z}, \hat{\gamma})[1 - \hat{p}_{i4}(\bar{z}, \hat{\gamma})]\hat{\gamma}_1 = 0.3332817 \cdot (1 - 0.3332817) \cdot (-0.0157837) = -0.003507$$

$$-\hat{p}_{i4}(\bar{z}, \hat{\gamma})\hat{p}_{i2}(\bar{z}, \hat{\gamma})\hat{\gamma}_1 = -0.3332817 \cdot 0.30421465 \cdot (-0.0157837) = 0.0016$$

These values imply that an increase of travelcost for car by 1 US dollar leads to an approximately estimated decrease (increase) of the choice probability for car (train) by 0.35 (0.16) percentage points.

Example 2: Determinants of travel mode choice (X)

As before, the effect of travelcost and termtime on the choice between air (base category), train, bus, and car is examined. However, the individual characteristic household income (in 1000 US dollars) is now (besides alternative specific constants) included as additional explanatory variable. In such general multinomial logit models, all Stata commands as in the case of conditional logit models can be used (note: the inclusion of the i. prefix is not possible in the “asclogit” command). On the basis of the ML estimation without and with robustly estimated variances of the estimated parameters, the following tests and estimations are considered besides WTP estimations:

- The Wald test for the null hypothesis that neither travelcost nor termtime has an effect on the travel mode choice (including robustly estimated variances)
- The Wald test for the null hypothesis that neither travelcost nor income has an effect on the travel mode choice (including robustly estimated variances)
- The corresponding likelihood ratio test for the null hypothesis that neither travelcost nor income has any effect on the travel mode choice (only possible on the basis of ML estimations without robustly estimated variances)
- The estimation of marginal probability effects for travelcost and income at the means of the explanatory variables (with robustly estimated variances)

The corresponding Stata commands lead to the following results:

Example 2: Determinants of travel mode choice (XI)

```
asclogit choice travelcost termtime, case(id) alternatives(travelmode) casevars(income)
```

```
Alternative-specific conditional logit      Number of obs      =          840
Case variable: id                          Number of cases     =          210
Alternative variable: travelmode           Alts per case: min =           4
                                           avg =          4.0
                                           max =           4
                                           Wald chi2(5)       =       105.78
                                           Prob > chi2        =         0.0000
Log likelihood = -189.52515
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
travelmode						
travelcost	-.0109274	.0045878	-2.38	0.017	-.0199192	-.0019355
termtime	-.0954606	.0104732	-9.11	0.000	-.1159876	-.0749335
-----+-----						
air	(base alternative)					
-----+-----						
train						
income	-.0511884	.0147352	-3.47	0.001	-.0800689	-.0223079
_cons	-.3249561	.5763335	-0.56	0.573	-1.454549	.8046369
-----+-----						
bus						
income	-.0232107	.0162306	-1.43	0.153	-.055022	.0086006
_cons	-1.744529	.6775004	-2.57	0.010	-3.072406	-.4166531
-----+-----						
car						
income	.0053735	.0115294	0.47	0.641	-.0172237	.0279707
_cons	-5.874813	.8020903	-7.32	0.000	-7.446882	-4.302745
-----+-----						

→ $W\hat{T}P_{\text{termtime}} = -(-0.095/(-0.011)) = -8.74$ (which is higher than before)

Example 2: Determinants of travel mode choice (XII)

```
asclogit choice travelcost termtime, case(id) alternatives(travelmode) casevars(income) robust
```

```
Alternative-specific conditional logit      Number of obs      =      840
Case variable: id                          Number of cases     =      210
Alternative variable: travelmode           Alts per case: min =      4
                                           avg =      4.0
                                           max =      4
                                           Wald chi2(5)       =      74.24
Log pseudolikelihood = -189.52515          Prob > chi2         =      0.0000
```

(Std. Err. adjusted for clustering on id)

choice	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
travelmode						
travelcost	-.0109274	.0049767	-2.20	0.028	-.0206815	-.0011732
termtime	-.0954606	.014622	-6.53	0.000	-.1241191	-.066802
-----+-----						
air	(base alternative)					
-----+-----						
train						
income	-.0511884	.0153147	-3.34	0.001	-.0812046	-.0211722
_cons	-.3249561	.6562138	-0.50	0.620	-1.611111	.9611992
-----+-----						
bus						
income	-.0232107	.013753	-1.69	0.091	-.0501661	.0037447
_cons	-1.744529	.6312441	-2.76	0.006	-2.981745	-.5073138
-----+-----						
car						
income	.0053735	.0099531	0.54	0.589	-.0141343	.0248813
_cons	-5.874813	.9180023	-6.40	0.000	-7.674065	-4.075562
-----+-----						

Example 2: Determinants of travel mode choice (XIII)

```
test travelcost termtime
```

```
( 1)  [travelmode]travelcost = 0
( 2)  [travelmode]termtime = 0
      chi2( 2) =    49.03
      Prob > chi2 =    0.0000
```

```
test travelcost income
```

```
( 1)  [travelmode]travelcost = 0
( 2)  [train]income = 0
( 3)  [bus]income = 0
( 4)  [car]income = 0
      chi2( 4) =    26.10
      Prob > chi2 =    0.0000
```

```
asclogit choice travelcost termtime, case(id) alternatives(travelmode) casevars(income)
```

```
estimates store unrestricted
```

```
asclogit choice termtime, case(id) alternatives(travelmode)
```

```
estimates store restricted
```

```
lrtest unrestricted restricted
```

```
Likelihood-ratio test                    LR chi2(4) =    34.58
(Assumption: restricted nested in unrestricted) Prob > chi2 =    0.0000
```

Example 2: Determinants of travel mode choice (XIV)

```
estat mfx, varlist(travelcost income)
```

```
Pr(choice = air|1 selected) = .27068612
```

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
-----+-----						
travelcost						
air	-.002157	.000957	-2.25	0.024	-.004034 - .000281	102.65
train	.000819	.000386	2.12	0.034	.000063 .001575	130.2
bus	.000347	.000165	2.10	0.036	.000022 .000671	115.26
car	.000992	.000448	2.21	0.027	.000113 .00187	95.414
-----+-----						
casevars						
income	.004085	.001898	2.15	0.031	.000366 .007804	34.548
-----+-----						

```
Pr(choice = train|1 selected) = .27684656
```

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
-----+-----						
travelcost						
air	.000819	.000386	2.12	0.034	.000063 .001575	102.65
train	-.002188	.00105	-2.08	0.037	-.004245 -.000131	130.2
bus	.000355	.00019	1.87	0.062	-.000017 .000727	115.26
car	.001014	.000513	1.98	0.048	7.9e-06 .00202	95.414
-----+-----						
casevars						
income	-.009993	.002634	-3.79	0.000	-.015157 -.00483	34.548
-----+-----						

Example 2: Determinants of travel mode choice (XV)

Pr(choice = bus|1 selected) = .11723223

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
-----+-----						
travelcost						
air	.000347	.000165	2.10	0.036	.000022 .000671	102.65
train	.000355	.00019	1.87	0.062	-.000017 .000727	130.2
bus	-.001131	.000548	-2.06	0.039	-.002206 -.000056	115.26
car	.000429	.000214	2.01	0.044	.000011 .000848	95.414
-----+-----						
casevars						
income	-.000952	.00118	-0.81	0.420	-.003265 .001362	34.548
-----+-----						

Pr(choice = car|1 selected) = .3352351

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
-----+-----						
travelcost						
air	.000992	.000448	2.21	0.027	.000113 .00187	102.65
train	.001014	.000513	1.98	0.048	7.9e-06 .00202	130.2
bus	.000429	.000214	2.01	0.044	.000011 .000848	115.26
car	-.002435	.001125	-2.16	0.030	-.004641 -.00023	95.414
-----+-----						
casevars						
income	.00686	.002154	3.19	0.001	.002639 .011081	34.548
-----+-----						

Example 2: Determinants of travel mode choice (XVI)

Now, only the individual characteristic income is included as explanatory variable. The corresponding Stata command (without robustly estimated variances of the estimated parameters) leads to the following results:

```
asclogit choice, case(id) alternatives(travelmode) casevars(income)
```

```
Alternative-specific conditional logit      Number of obs      =      840
Case variable: id                        Number of cases    =      210
Alternative variable: travelmode          Alts per case: min =        4
                                           avg =       4.0
                                           max =        4
                                           Wald chi2(3)      =      33.82
Log likelihood = -261.74506                Prob > chi2        =      0.0000
```

choice		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]		
air		(base alternative)						
train								
	income	-.059059	.0116784	-5.06	0.000	-.0819482	-.0361698	
	_cons	1.963424	.4195611	4.68	0.000	1.141099	2.785748	
bus								
	income	-.0353522	.0128212	-2.76	0.006	-.0604814	-.010223	
	_cons	.5991678	.4909705	1.22	0.222	-.3631167	1.561452	
car								
	income	.0014204	.0098938	0.14	0.886	-.0179711	.0208118	
	_cons	-.0425218	.454562	-0.09	0.925	-.933447	.8484033	

Example 2: Determinants of travel mode choice (XVII)

These estimation results are identical to the corresponding results in a (pure) multinomial logit model. However, the data organization for the conditional logit or general multinomial logit model does not allow the “mlogit” command. This application is only possible with the data organization as discussed above (“wide” format). The corresponding Stata commands lead to the following results:

```
mlogit travelmode income, base(1)
```

```
Multinomial logistic regression          Number of obs   =          210
                                          LR chi2(3)      =          44.03
                                          Prob > chi2     =          0.0000
Log likelihood = -261.74506              Pseudo R2       =          0.0776
```

travelmode	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
air	(base outcome)					
train						
income	-.059059	.0116784	-5.06	0.000	-.0819482	-.0361698
_cons	1.963424	.4195611	4.68	0.000	1.141099	2.785749
bus						
income	-.0353522	.0128212	-2.76	0.006	-.0604814	-.010223
_cons	.5991669	.4909705	1.22	0.222	-.3631176	1.561452
car						
income	.0014204	.0098938	0.14	0.886	-.0179711	.0208118
_cons	-.0425218	.454562	-0.09	0.925	-.9334469	.8484034

Example 3: Determinants of secondary school choice (I)

Again, a (pure) multinomial logit model for the choice between Hauptschule, Realschule, and Gymnasium is considered, but only with the three explanatory variables motheduc, mothlnlf, and loghhincome. The ML estimation (with robustly estimated variances of the estimated parameters) with the Stata “mlogit” command and the former data organization leads to the following results:

```
mlogit schooltype motheduc mothlnlf loghhincome, base(1) robust
```

```
Multinomial logistic regression           Number of obs   =           675
                                           Wald chi2(6)    =           93.91
                                           Prob > chi2     =           0.0000
Log pseudolikelihood = -638.70703          Pseudo R2       =           0.1285
```

schooltype	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
Hauptschule	(base outcome)					
-----+-----						
Realschule						
motheduc	.3425135	.0849103	4.03	0.000	.1760923	.5089347
mothlnlf	-.1808054	.2113319	-0.86	0.392	-.5950083	.2333975
loghhincome	.1212413	.185545	0.65	0.513	-.2424202	.4849027
_cons	-4.878992	2.090943	-2.33	0.020	-8.977164	-.7808198
-----+-----						
Gymnasium						
motheduc	.6814142	.0916188	7.44	0.000	.5018447	.8609838
mothlnlf	-.094855	.218878	-0.43	0.665	-.5238481	.3341381
loghhincome	1.169913	.2689158	4.35	0.000	.6428477	1.696978
_cons	-20.19249	3.042712	-6.64	0.000	-26.1561	-14.22889

Example 3: Determinants of secondary school choice (II)

These estimation results can be (widely) replicated with the “asclogit” command on the basis of the alternative data organization (“long” format):

```
asclogit choice, case(id) alternatives(alt) casevars(motheduc mothlnlf loghhincome) base(1) robust
```

```
Alternative-specific conditional logit      Number of obs      =      2025
Case variable: id                        Number of cases     =      675
Alternative variable: alt                 Alts per case: min =      3
                                           avg =      3.0
                                           max =      3
```

```
Log pseudolikelihood = -638.70703          Wald chi2(6)       =      93.91
                                           Prob > chi2        =      0.0000
```

(Std. Err. adjusted for clustering on id)

choice	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1	(base alternative)					
2						
motheduc	.3425136	.0849103	4.03	0.000	.1760924	.5089348
mothlnlf	-.1808054	.2113319	-0.86	0.392	-.5950083	.2333974
loghhincome	.1212413	.185545	0.65	0.513	-.2424202	.4849028
_cons	-4.878993	2.090943	-2.33	0.020	-8.977166	-.7808206
3						
motheduc	.6814144	.0916188	7.44	0.000	.5018447	.860984
mothlnlf	-.0948551	.2188781	-0.43	0.665	-.5238482	.334138
loghhincome	1.169913	.2689158	4.35	0.000	.6428477	1.696978
_cons	-20.19249	3.042712	-6.64	0.000	-26.1561	-14.22889
