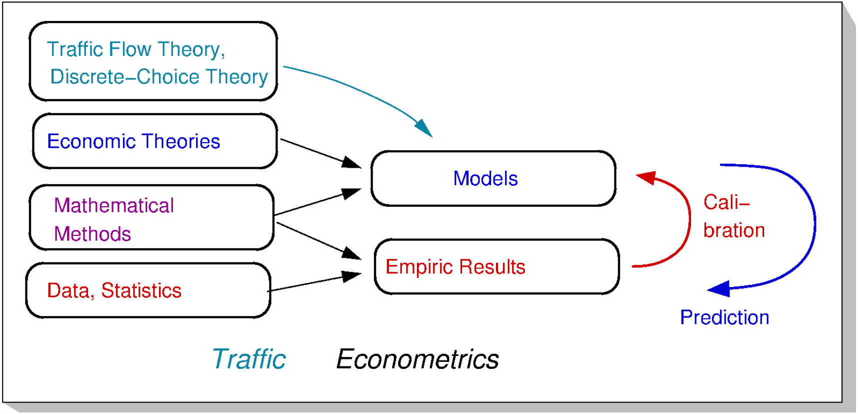
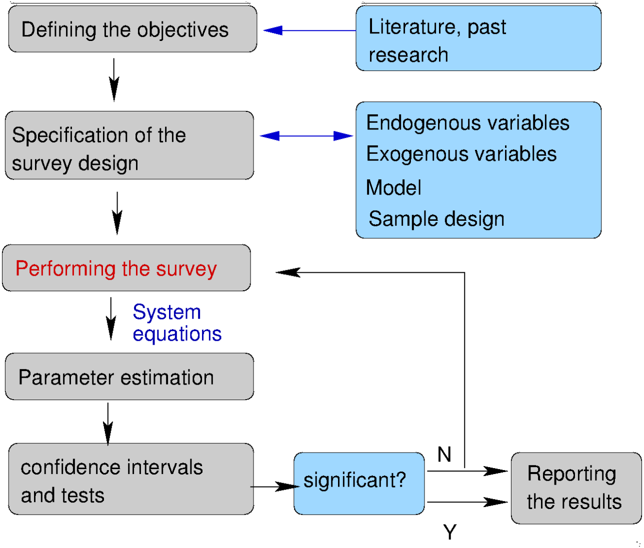
Scope of econometrics – from a mathematical point of view



The field of **Traffic Econometrics** includes all mathematical models and statistical procedures to quantitatively analyze empirical (transportation) data with respect to economic effects.

General procedure of an econometric analysis

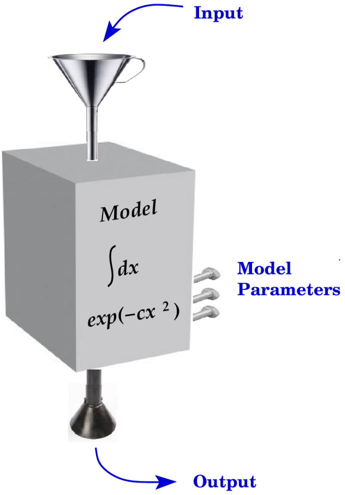


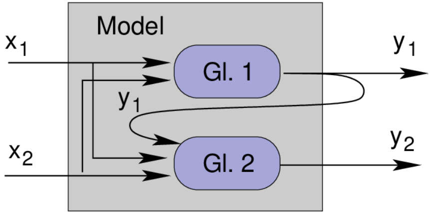
Examples of models and compatible data

- The model must be compatible to the research objective:
  - Model for fuel/energy consumption → real-valued output → e.g., **regression models**
  - Model for trip or mode choice → discrete output → e.g., **discrete-choice models**
  - Classification of different days with respect to the traffic demand profile (mid-workdays, sundays, holidays, ...) → econometric **discriminant analysis**

- If the model is a **discrete-choice model**, the questions of the survey must be sets of alternatives that are ..
  - exclusive: at most one alternative can be ticked
  - complete: at least one alternative may be ticked ⇒ exactly one
  - sufficiently different
  - ? How to formulate a question regarding the kind of schools visited?
  - ? For route choice, we have two routes that only differ in that one route contains a small detour of to go to a bakery. Which criterion is violated?

Information flow of an econometric model





Econometric models

Since econometrics describes things quantitatively, its basic language is mathematics and its basic concept a **(mathematical) model**

$$Y_k = f_k(\tilde{x}_1, ..., \tilde{x}_m, .., \tilde{x}_M, \beta_0, ..., \beta_j, ..., \beta_J) + \epsilon_k = f_k(\tilde{\boldsymbol{x}}, \boldsymbol{\beta}) + \epsilon_k$$

- A model consists of one or more equations $Y_k = f_k(\tilde{\boldsymbol{x}}, \boldsymbol{\beta}) + \epsilon_k$ explaining the output quantity Y_k as a function of the input $\tilde{\boldsymbol{x}}$
- to *tune* the model, there are **model parameters** $\boldsymbol{\beta}$
- a model can be stochastic $(\epsilon_k \neq 0)$ or deterministic $(\epsilon_k = 0)$
- the above formulation is the most general one and includes *all* conceivable econometric models.

Endogenous Variables

- The name **endogenous variable** comes from the Greek *endo* ($\epsilon\nu\delta$) and the suffix *gen* ($\gamma\epsilon\nu o\varsigma$) meaning generated from the inside [the model]
- in systems theory, the endogenous variables are the **output**
- mathematically, the endogenous variable are the **dependent variables**
- logically, they are the **explained variables**

Example mode choice: Y_k: #decisions for mode k (e.g., walking, cycling, public transport, car, combined/others)

Exogeneous variables

- The name **exogenous variable** comes from the Greek *exo* ($\epsilon\xi\omega$) and the suffix *gen* ($\gamma\epsilon\nu o\varsigma$) meaning "coming from the outside"
- in systems theory, the endogenous variables are the **input**
- mathematically, the endogenous variable are the **independent variables**
- logically, they are the **explanatory variables**
- an additive fixed function of explanatory variables is often called a **factor**

Since the exogenous variable describe explanatory factors, they are generally considered to be deterministic with all stochasticity deferred to additional random terms ϵ_k

Example mode choice: $\tilde{x}_m$ can be travel times or costs for the different modes, and also socioeconomic factors such as age, gender, or income

Random terms

The **random terms**, also called **residual terms** ϵ_k (from *residuum*: the rest) summarize all what is not known and cannot be explained by the exogenous/explanatory variables:

Scio nescio (I know that I do not know)

possible reasons for ϵ_k:

- cannot include all factors **(watch out for bias!)**
- the model is not appropriate (e.g., explaining the fuel consumption by a linear function of speed)
- the data used for model calibration contain errors
- in case of human decisions:

man $\neq$ machine; homo $\neq$ *homo oeconomicus*

Example mode choice: we ignored the weather or the additional utility to stop by at a bakery (not provided by public transport), or need of a certain mode (car) for subsequent trips of that day

Model parameters

The **model parameters** $\beta_j,\ j=0,\ldots,J$ *tune* the model to fit the data

- The parameters are determined by fitting the model to *learning data sets*, a process called **calibration**
- To test the *explanatory/prediction power* of a model, the calibrated model is applied to *test data sets* with known output, a process called **validation**
- In contrast to the exogenous variables changing from application to application, the parameters are *fixed* after calibration.

The existence of well validated models is the *raison d' être* for econometrics as such

Example mode choice: Parameters characterize, e.g., the monetary **value of time** (VoT) in €/h

Model functions

The **model function** is a mathematical representation of the economic process under investigation. The model's mathematical structure must reflect reality as well as possible:

- linear *vs.* nonlinear
- deterministic *vs.* stochastic
- single or multiple equations that may be linked, chained, or with feedback
- which exogenous variables are relevant?

The model structure defines the qualitative aspects and the model parameters the quantitative aspects of whatever is to be investigated

Make it as simple as possible but not simpler
(Einstein)

Mathematical structure I: linear *vs.* nonlinear

Four steps from linearity to nonlinearity:

1. **Truly linear models:**

$$Y = \hat{y}(\tilde{\boldsymbol{x}}, \boldsymbol{\beta}) + \epsilon = \sum_{j=0}^{J} \beta_j \tilde{x}_j + \epsilon = \boldsymbol{\beta}' \tilde{\boldsymbol{x}} + \epsilon$$

   Because of linearity, each endogenous variables has its own uncoupled single equation, so it is enough to consider a single component.

2. **Parameter-linear (quasi-linear) models**

$$Y = \hat{y}(\tilde{\boldsymbol{x}}, \boldsymbol{\beta}) + \epsilon = \sum_{j=0}^{J} \beta_j g_j(\tilde{\boldsymbol{x}}) + \epsilon = \sum_{j=0}^{J} \beta_j x_j + \epsilon = \boldsymbol{\beta}' \boldsymbol{x} + \epsilon$$

   If it is reasonable to assume fixed (generally nonlinear) functions $x_j = g_j(\tilde{\boldsymbol{x}})$, we have a linear model with the **factors** x_j becoming the "new" exogenous variables

Example: average distance travelled in vehicles per year

- y: distance [km/year] (deterministic because of "average")
- $\tilde{x}_1$: income [€/year]
- $\tilde{x}_2$: fuel cost [€/liter]
- Two exogenous variables → 4 factors:
  $g_0(\tilde{\boldsymbol{x}}) = 1$, $g_1(\tilde{\boldsymbol{x}}) = \tilde{x}_1$, $g_2(\tilde{\boldsymbol{x}}) = \tilde{x}_2$, $g_3(\tilde{\boldsymbol{x}}) = \tilde{x}_1\tilde{x}_2$,

Model: $$y = \beta_0 + \beta_1\tilde{x}_1 + \beta_2\tilde{x}_2 + \beta_3\tilde{x}_1\tilde{x}_2 = \sum_{j=0}^{3}\beta_j x_j$$

? Discuss the elasticity $\epsilon_2 = \frac{\bar{x}_2}{\bar{y}}\frac{\mathrm{d}y}{\mathrm{d}x_2} = \frac{\beta_2\bar{x}_2}{y} = -0.15$

! 0.15 % decrease in kilometrage per increase of the fuel costs by 1 %

? Discuss the meaning of the factors, particularly the product x_3

! $x_0 = 1$: constant; $x_1 = \tilde{x}_1$: increase with income ($\beta_1 > 0$); $x_2 = \tilde{x}_2$: price sensitivity ($\beta_2 < 0$); $x_3 = \tilde{x}_1\tilde{x}_2$: increase of price sensitivity (becoming less negative) with increasing income ($\beta_3 > 0$)

Mathematical structure I: linear *vs.* nonlinear

3. **Nonlinear models that can be linearized**
   Classical example: unlimited growth

$$G(t) = G_0 e^{t/\tau + \epsilon}$$

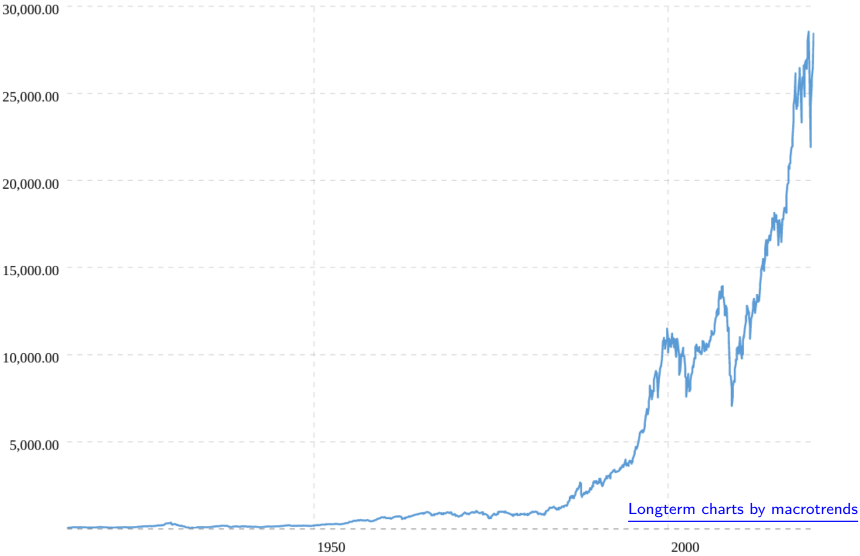
- endogenous variable G: growth measure, e.g., company size, $\#$items sold of a newly introduced product
- exogenous variable t: time
- parameter G_0: initial growth measure
- parameter τ: time for growing by a factor of $e = 2.71...$
- random multiplicative factor e^{ϵ}

Linearisation: $Y(t) = \ln G(t) = \ln G_0 + \frac{t}{\tau} + \epsilon$

Reformulation by setting $x_0 = 1$, $x_1 = t$, $\beta_0 = \ln G_0$, $\beta_1 = 1/\tau$:

Standard form: $Y(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$

Example: Dow Jones Industrial Average linear



Example: Dow Jones Industrial Average log scale

Longterm charts by macrotrends

Mathematical structure I: linear *vs.* nonlinear

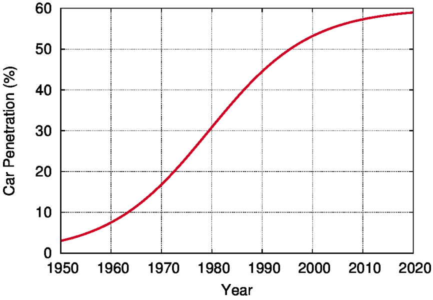
4. **Irreducibly nonlinear models**
Classical example: limited growth

$$y(t) = \frac{y_s}{1 + (y_s/y_0 - 1)e^{-t/\tau}}$$

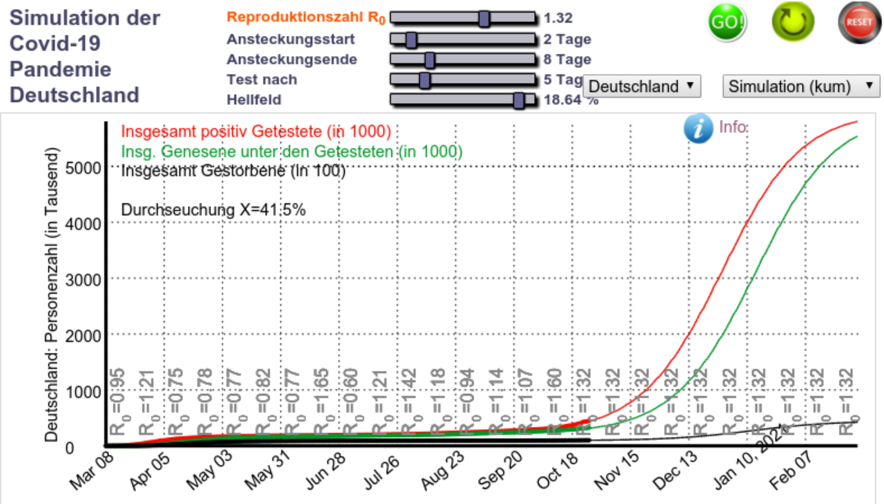
- Solution of the ODE $\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{y(t)}{\tau}\left(1 - \frac{y(t)}{y_s}\right)$ for the initial value $y(t_0) = y_0$
- Plot for $t_0 = 1950$, $y_0 = 3\,\%$, $y_s = 60\,\%$, and $\tau = 10\,\text{years}$

? What might this represent?

! For example, the penetration rate for passenger cars per person.



Actual example: Corona simulation



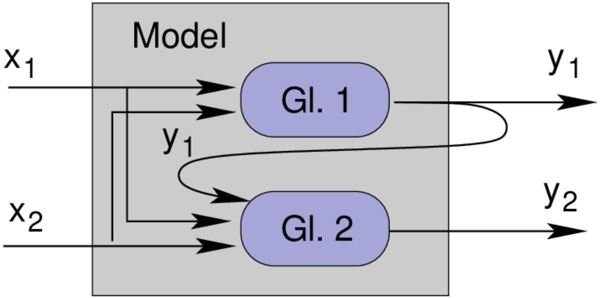
Simplest macroscopic SI (susceptible-infected) model:

$y(t+\tau) = y(t) + R_0 y(t)(1 - y(t))$ → same model as above

Mathematical structure II: Other criteria

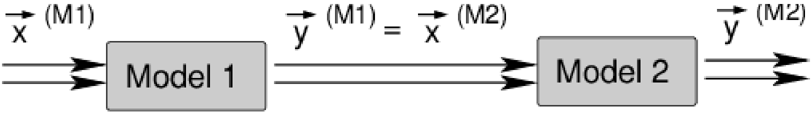
- **deterministic** *vs.* **stochastic** models
- #exogenous variables: 1: **univariate**; ≥ 2 **multivariate** models
- #endogenous variables: single- *vs.* multi-equation models
- Scaling of the endogenous variables: real-valued $\rightarrow$ **regression** models; discrete $\rightarrow$ **discrete choice** models
- Linking, chaining, and feedback

Linking

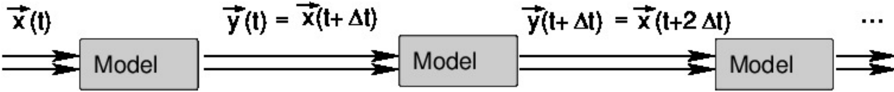


One or more endogenous variables of a model equation serve as exogenous variables of other model equations

Chaining

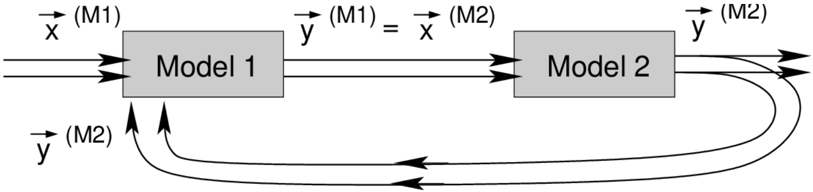


The endogenous variables of Model 1 serve as exogenous variables of Model 2



- Special case of chaining: **time evolution**: the endogenous variables at time t are the exogenous variables at the next time step $t + \Delta t$
- The model itself is generally the same in all steps (autonomous model)

Feedback



Combination of chaining and linking

Complex example: Four-Step Model of transportation planning