



# Modeling 101: Econometrics of Regression and Discrete Choice Models

Constantinos Antoniou

National Technical University of Athens

July 4, 2011

# Outline

- Introduction / Background
- Suitable techniques
  - Model structure
  - Dealing with data properties
  - Diagnostics and presentation of results
- Models to be estimated
- Lessons from on-going case studies
- Standardized views

# Sources and references

- <http://www.sustaincity.org/publications>
  - Picard, N., C. Antoniou and A. de Palma (2010) Econometric Models, SustainCity Deliverable, 2.4, THEMA, Université de Cergy-Pontoise.
  - Picard, N. and Antoniou, C. (2011) Econometric guidance, SustainCity Deliverable, 5.1, THEMA.
- Picard, N. and C. Antoniou (2011). Econometric guidance for developing UrbanSim models. First lessons from the SustainCity project. Proceedings of the 51st European Congress of the Regional Science Association International, 30th August - 3rd September 2011, Barcelona, Spain

<b>Model</b>	<b>Paris</b>	<b>Brussels</b>	<b>Zurich</b>
Household location	Nested: relocation/ dwelling type/ tenure status/ location	Multinomial Logit structure. Besides this, nested structures of choice will be tested in order to account for correlation of attributes across alternatives.	MNL with explaining variables of domains: life style, dwelling type, location (Household relocation: Probabilities for relocation of HH according to income and age)
Job location	Matching workplace/ business	Nested logit; sampling of alternatives	Hierarchical NL of firm location choice (Bodenmann & Axhausen, 2010)
Real estate price	Simultaneous equation (5 types), spatial correlation, Dwelling level	Hedonic model; estimated using “interval regression”. A spatial autoregressive model will be considered	Spatial error model (Löchl and Axhausen, 2010)
Land developpt	Matching project location/land use transition	2-step model: Supply by building type per zone: linear regression/Choice of zone: Multinomial logit	NL with explaining variables of domains: project, developer and development constraints

# Notations and assumptions

- Majority of models fall under two general categories:
  - Linear regression models

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Hedonic regression is a commonly used type, in this context
  - Discrete choice models

$$U_{jn} = X_{jn} \beta + v_{jn}$$

# Notation - Linear regression models

The linear regression model is given by:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where the error terms  $\varepsilon_i$  are assumed to be white noise (normally distributed with zero mean and variance  $\sigma^2$ ). The nonrandom part of the equation describes the dependent variable  $Y_i$  with a straight line. The slope of the line (the regression coefficient)  $\beta$  denotes the increase to the dependent variable per unit change in the corresponding explanatory variable (or regressor). The line intersects the y-axis at the intercept  $\alpha$ .

# Linear regression

- The basic Gauss-Markov assumptions require:
  - Linearity (in the parameters; nonlinearity in the variables is acceptable);
  - Homoscedasticity;
  - Exogenous independent variables;
  - Uncorrelated disturbances; and
  - Normally distributed disturbances
- If the above assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. Best Linear Unbiased Estimator.
  - In other words, it is unbiased and efficient (has the lowest total variance among all unbiased linear estimators).

# Interval regression

- It is often the case, especially concerning income or price data, that information is missing on the exact value of the explained variable.
  - Instead, the only available information is that the explained variable lies in some interval.
- In that case, a maximum likelihood estimator can be used.
- The likelihood then corresponds to the probability that the explained variable lies in the observed interval. The statistical structure of the model and the assumptions are similar to simple linear regression.

# Notation – Discrete choice models

The specification of a random utility model uses the following utility specification (for a decision maker  $n$  choosing alternative  $j$  from a choice set of  $J$  alternatives):

$$U_{jn} = X_{jn}\beta + v_{jn}$$

where  $X_{jn}$  are observable variables that relate to the alternative  $j$  and decision maker  $n$ ,  $\beta$  is a vector of coefficients of these variables, and  $v_{jn}$  is a zero-mean, random term that is iid extreme value. Several assumptions can be made about the distribution and the variance/covariance structure of the error term. The most common assumptions lead to the logit model (i.i.d. Gumbel error terms) and probit model (Normal error terms).

(Following: Ben-Akiva and Lerman, 1985)

# Discrete choice analysis (DCA)

- Decision-Maker
  - Individual
  - Socio-economic characteristics, e.g., age, gender, income
  - In the UrbanSim context: “individual” may be firm, etc.
- Alternatives
  - Choice set identification
  - Discrete vs. continuous
  - Universal vs. individual choice set (availability)
  - Deterministic vs. stochastic (awareness)

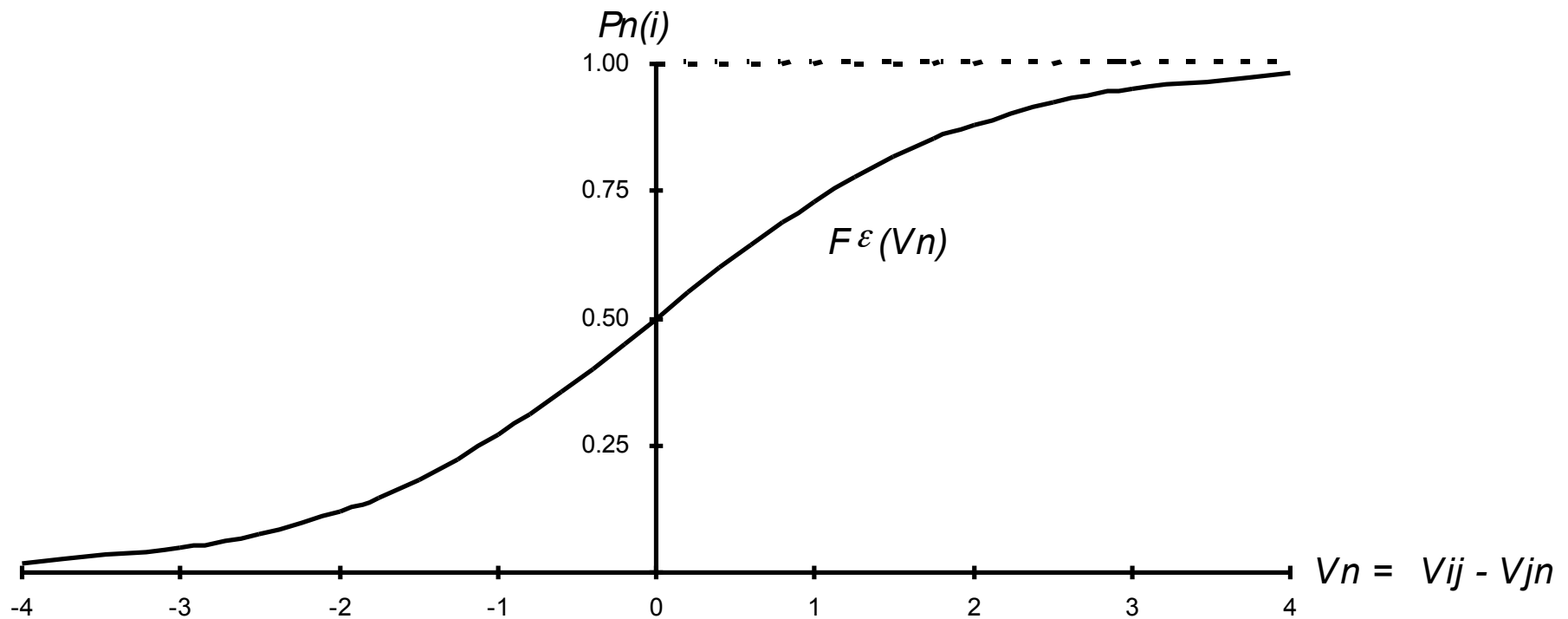
# Choice set

- A decision-maker  $n$  selects one and only one alternative from a choice set

$$C_n = \{ 1, 2, \dots, i, \dots, J_n \}$$

- with  $J_n$  alternatives.

# Binary choice



# Random utility model

- Utility:  $U_{in} = V_{in} + \varepsilon_{in}$ 
  - $V_{in}$  = Systematic utility expressed as a function of observable variables, e.g.,  $\sum \beta_k x_{ink}$
  - $\varepsilon_{in}$  = Random utility component
- Probability:

$$P(i | C_n) = P(U_{in} \geq U_{jn} \quad \forall j \in C_n)$$

$$P(U_{in} = \max_j U_{jn} \quad \forall j \in C_n)$$

*(there are of course other intricacies, like location and scale, normalization and so on)*

# Random term specification

- Assumptions about
  - Distribution
  - Variance/covariance structure
- Typical models:
  - Logit model (i.i.d. Gumbel error terms)
  - Probit model (Normal error terms)
  
  - What are the pros and cons of each?

# Multiple choice

- Choice set  $C_n$ :  $J_n$  alternatives,  $J_n \geq 2$

$$\begin{aligned} P(i | C_n) &= P[V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in C_n] \\ &= P\left[(V_{in} + \varepsilon_{in}) = \max_{j \in C_n} (V_{jn} + \varepsilon_{jn})\right] \\ &= P\left[\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}, \forall j \in C_n\right] \end{aligned}$$

# Multinomial Logit (MNL)

- $\varepsilon_{jn}$  independent and identically distributed (i.i.d.)
- $\varepsilon_{jn} \sim \text{Gumbel}(0, \mu) \quad \forall j$

# Specification of systematic components

- Types of Variables
  - Attributes of alternatives:  $Z_{in}$ , e.g., (for a household location model:) area, age, transit access, parking availability, ...
  - Characteristics of decision-makers:  $S_n$ , e.g., age, gender, income, occupation
  - Therefore:  $X_{in} = h(Z_{in}, S_n)$
- Examples:
  - $X_{in1} = Z_{in1} = \text{area}$
  - $X_{in2} = \log(Z_{in2}) = \log(\text{rent/buy cost})$
  - $X_{in3} = Z_{in2}/S_{n1} = (\text{rent/buy cost}) / \text{income}$
- Functional Form: Linear in the Parameters
- $V_{in} = b_1 X_{in1} + b_2 X_{in2} + \dots + b_k X_{inK}$
- $V_{jn} = b_1 X_{jn1} + b_2 X_{jn2} + \dots + b_k X_{jnK}$

# Independence from Irrelevant Alternatives

- Property of the Multinomial Logit Model
- Example: blue/red apartment\*



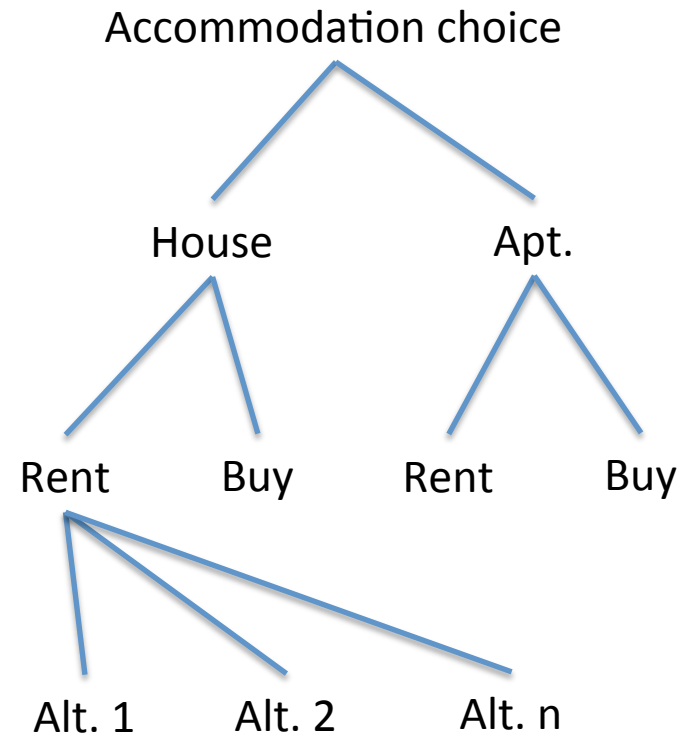
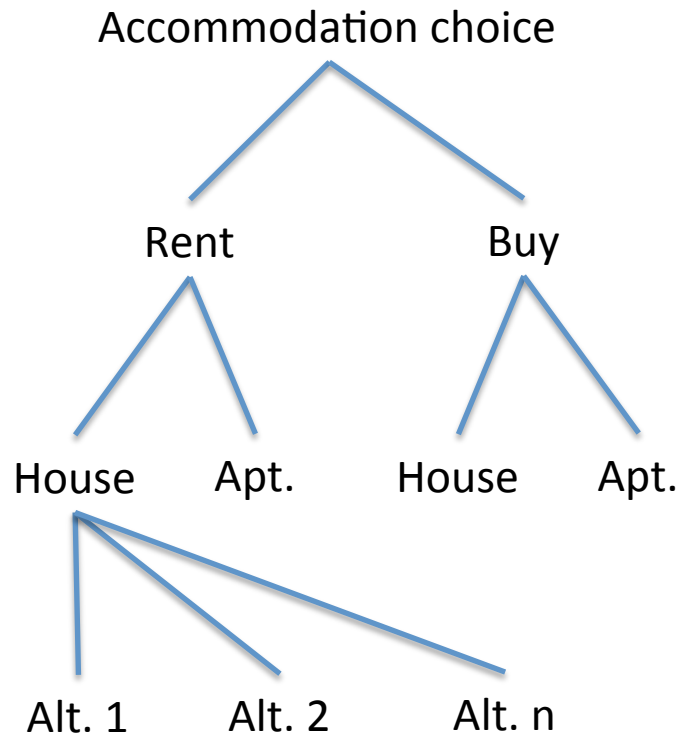
\* reference to the red bus/blue bus paradox

# Nested logit (NL)

- Overcome the IIA Problem of Multinomial Logit when
  - Alternatives are correlated (e.g., red house and blue house)
  - Multidimensional choices are considered (e.g., type of house and rent/buy)
    - Generalization: Partition nests into sub-nests

# Tree representation of nested logit

- Two different tree structures:
  - First house location and then job location? Or vice versa?
  - First type of house and then tenure type (rent/buy)?



# Nested logit model specification

- Deterministic utility term for nest  $C_{mn}$ :

$$V_{C_{mn}} = \tilde{V}_{C_{mn}} + \frac{1}{\mu_m} \ln \sum_{j \in C_{mn}} e^{\mu_m \tilde{V}_{jn}}$$

- Model:  $P(i | C_n) = P(C_{mn} | C_n) P(i | C_{mn})$

– where

$$P(C_{mn} | C_n) = \frac{e^{\mu V_{C_{mn}}}}{\sum_l e^{\mu V_{C_{ln}}}}$$

– and

$$P(i | C_{mn}) = \frac{e^{\mu_m \tilde{V}_{in}}}{\sum_{j \in C_{mn}} e^{\mu_m \tilde{V}_{jn}}}$$

# Mixed MNL (MMNL)

A detailed description of mixed logit is available in Train (2003) and Walker (2001). The specification of a random coefficient mixed logit model uses the following utility specification (for a decision maker  $n$  choosing alternative  $j$  from a choice set of  $J$  alternatives):

$$U_{jn} = X_{jn}\beta + \sigma_j \varepsilon_{jn} + \nu_{jn}$$

where:

$X_{jn}$  are observed variables that relate to the alternative  $j$  and decision maker  $n$ ,

$\beta$  is a vector of coefficients of these variables,

$\varepsilon_{jn}$  is a Gaussian, zero-mean error term, with a standard deviation  $\sigma_j$ , and

$\nu_{jn}$  is a zero-mean, random term that is iid extreme value.

# Latent variables

- The nested Logit model is relevant when the upper level category is observable. This is the case, for example, for dwelling type or tenure type. In some cases, the upper level category is implicit and cannot be observed.
- This is the case, for example, for budget constraints, which prevent the constrained households to borrow in order to buy their dwelling, and so that they are bounded in the tenant category even though their expected utility is lower in this category than in the owner category.
- The modeller cannot know a priori which households rent because they chose so, and which households rent because they are budget constrained.

# DEALING WITH DATA PROPERTIES

# Importance sampling

- In a MNL model, under the IIA assumption, random sampling can be performed when the number of alternatives is too large.
  - Extending random sampling to NL is not straightforward.
- Importance sampling of a zone is equivalent to uniform sampling of dwellings located in the zone.
- Importance sampling should not prevent the same zone to appear twice or more in the choice set
  - Some econometric software restricts this.
  - In case the same zone cannot appear twice in a choice set, this leads to an under-representation of largest alternatives, which becomes more and more severe as the number of alternatives increases.
  - This leads to a bias in the coefficients of all variables correlated with zone size. This bias should be corrected.

# Importance sampling (2)

- Under-representations of large alternatives, and the resulting bias, become more and more severe when the number of alternatives in the individual choice sets is increased
  - The number of alternatives in individual choice sets should not be increased too much
  - E.g. 10 alternatives randomly chosen for each household choice set was a reasonable figure for household location choice in Paris case study

# (Pseudo-)Panel data

- The data that are used in the UrbanSim models come from several time periods
- (Unobserved) heterogeneity across individuals
  - Pooling data across individuals while ignoring heterogeneity (when it is present) will lead to biased and inconsistent estimates of the effects of pertinent variables
- Fixed effects / random effects
  - Fixed effects: estimating a constant term for each individual and choice
    - Can quickly become intractable
  - Random effects: assume fixed term varies according to some probability distribution

# Spatial econometrics

- Spatial effects are one of the main methodological challenges in this field (e.g. hedonic regression)
  - Spatial dependence
    - By-product of measurement errors for observations in contiguous spatial units (aggregate measurement errors “spill-over” unit boundaries)
    - Interdependencies across space
  - Spatial heterogeneity
    - Functional forms and parameters vary with location
    - Are not homogeneous across the data set

# Spatial econometrics (2)

- Neglect of spatial considerations in econometric models not only affects the magnitudes of the estimates and their significance, but may also lead to serious errors in the **interpretation** of standard regression **diagnostics** such as tests for heteroscedasticity
- Proposed frameworks to overcome this:
  - Spatial econometrics models
  - Geographically Weighted Regression (GWR)

# Spatial econometrics (3)

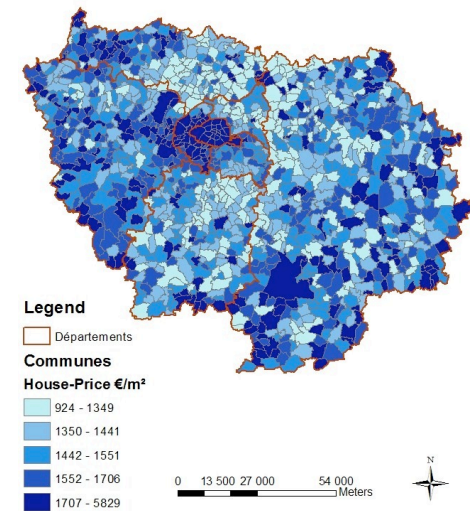
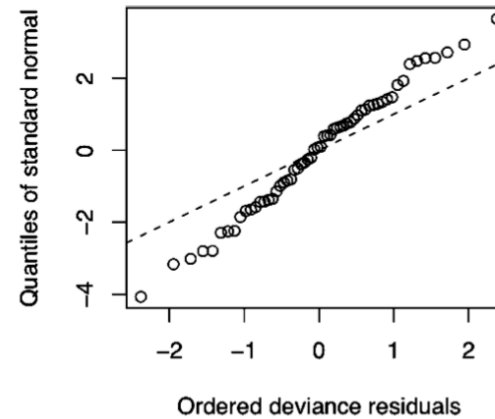
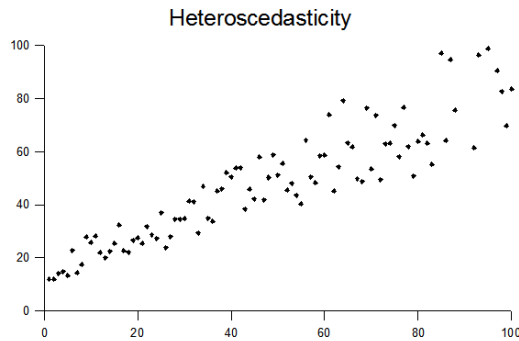
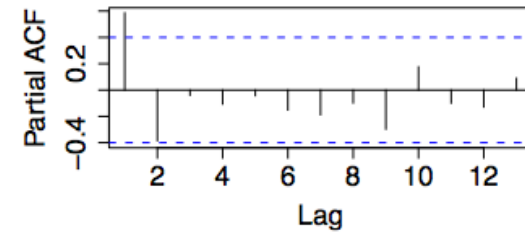
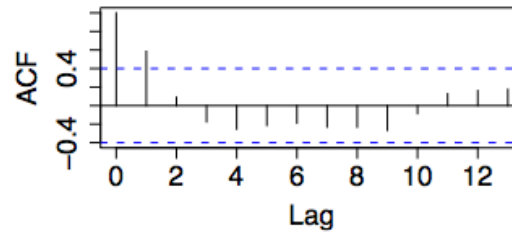
- Spatial econometrics models
  - Spatial Autoregressive Model (SAR)
    - Both the direct and indirect effects of a neighbourhood's housing characteristics are captured through a spatial multiplier
  - Spatial Error Model (SEM)
    - Spatial autocorrelation is assumed to arise from omitted variables that follow a spatial pattern
  - Spatial Mixed Model (mix of SAR and SEM)
  - Spatial Durbin Model
    - Extension of SAR
    - Includes a spatial lag of the dependent variable as well as spatial lags of the explanatory variables
- Geographically Weighted Regression (GWR)
  - Local version of spatial regression that generates parameters disaggregated by the spatial units of analysis
  - This allows assessment of the spatial heterogeneity

# Endogeneity of variables and selection bias

- Endogeneity is a serious problem commonly faced in LUTI models interested in interactions between modules
  - A typical example is given by the prices in the household location choice model, which is correlated with the error term.
  - This problem is caused either by the simultaneous determination of the supply and the demand for dwelling units, or by omitted attributes that are correlated with price.
- When endogeneity results from omitted attributes, the best solution is to include enough explanatory variables in the model of interest
  - Instrumental variables technique can be used to correct for endogeneity, provided that at least one instrument is available for each endogenous variable
  - It often proves to be difficult to find such instruments
  - E.g. in the case of household location, if it can be reasonably assumed that dwellings and offices compete for land, then variables measuring local business tax can be used to instrument dwelling prices
  - Note that a rich enough model can be estimated precisely enough only when sample size is large enough, which typically means at least 50,000 households

# Model diagnostics

- Statistical
- Graphical



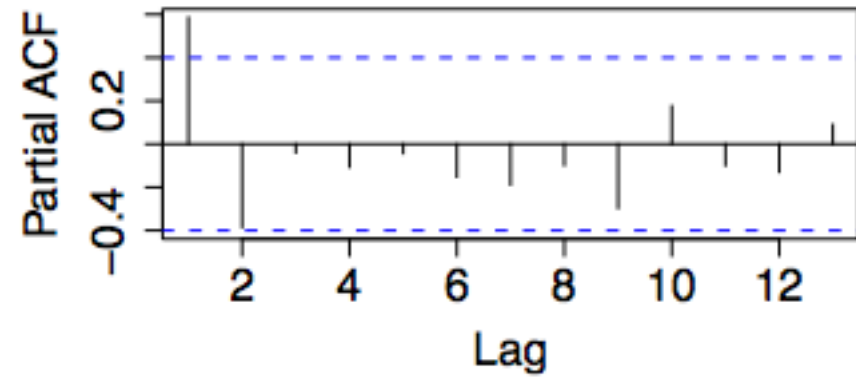
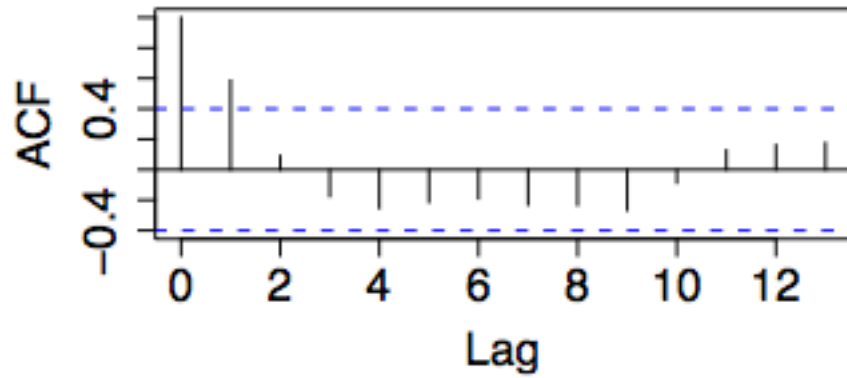
# Model output

- Output information
  - Estimated coefficient
  - Standard error
  - T-statistic
  - p-value
- Summary statistics
  - Regression: corrected  $R^2$
  - DCM estimated using ML:
    - Null and final log-likelihood
    - AIC
    - Degrees of freedom
    - (Corrected) Likelihood ratio tests (whether restrictions should be retained)

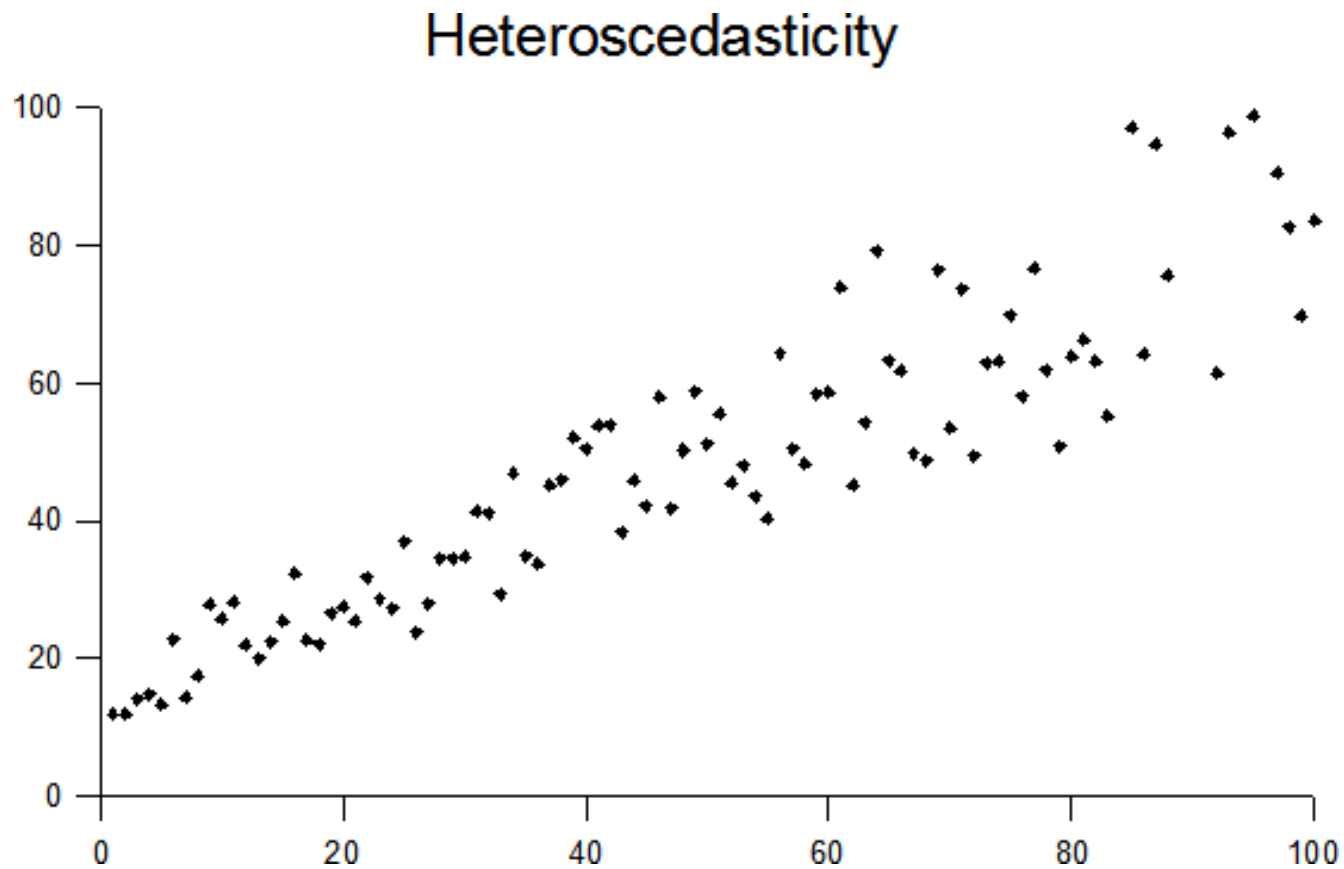
# Violations of model assumptions

- The models have explicit underlying assumptions that need to be satisfied by the data, in order to be valid.
  - A number of violations may often occur, however, resulting in residuals that are not independently and identically distributed.
- (Partial) autocorrelation functions (or “corellograms”) and residual plots can be used to identify and visualize violations such as autocorrelation and heteroscedasticity

# ACF/PACF



# Heteroscedasticity



# Violations of model assumptions (2)

- Formal statistical tests:
  - Shapiro-Wilk test for normality assumption
    - Skewness and kurtosis can provide additional information
  - Box-Ljung test for autocorrelation for various lags
    - “Portmanteu” tests: consider the first few autocorrelations as a whole
  - Breusch-Pagan test for heteroscedasticity
  - Hausman test for endogeneity