Phd Program in Transportation

Transport Demand Modeling

Filipe Moura

MODULE 2
Multiple Linear Regression

Outline

1. Learning objectives
2. What is MR analysis? What does it look like? Interpreting the LR?
3. Example
4. Linear regression assumptions
5. Curve fitting
6. Mathematical properties
7. Statistical properties
8. Estimation
9. Goodness of fit
10. Statistical tests
11. Begin doing your Home assignment for this module

Introduction

- It is one of the most widely applied econometric techniques:
  - Suitable for modeling a wide variety of relationships between variables.
  - In many practical applications the assumptions of linear regression are often suitably satisfied.
  - Its outputs are relatively easy to interpret and communicate.
  - The estimation of regression models is relatively easy, the routines for its estimation are available in a vast amount of software packages.

- The main problem is that linear regression can also be overused or misused, when its assumptions are not strictly met, and the correct alternatives are not known, understood, or applied.

What is Multiple Regression Analysis?

- Statistical technique used to analyze the relationship between a **single dependent** variable (aka *criterion/regressand*) and **several independent** variables (aka *predictors/regressors*)
- Objective of MRA:
  - Known values of the IV are used to predict the values of the DV selected by the analyst
- Each IV is weighted by the regression procedure to ensure maximal prediction from the overall set of IV
  - Weights denote the relative contribution of each IV to the overall prediction and helps in the interpretation as to the influence of each IV in making the prediction
  - Still, correlation between IV's can complicate the interpretative process

How does the LR model look like?

Formal representation

- Linear

$$y_n = \alpha + \beta x_n + \varepsilon_n, \quad n = 1,\ldots,N \quad observations$$

$$y = \alpha + \beta \ln x + \varepsilon$$

$$y = \alpha + \beta x^2 + \varepsilon$$

- Multiple linear

$$y = \alpha + \sum_{n=1}^{N} \beta_n x_n + \varepsilon, \quad n = 1,\ldots,N \, variables$$

- Nonlinear

$$y = \alpha + \frac{1}{x + \beta} + \varepsilon$$

How does the LR model look like?
Formal representation **(cont'd)?**

- What about… $y = e^{\alpha} x^{\beta} e^{\varepsilon}$

$$y = \frac{1}{1 + e^{\alpha + \beta x + \varepsilon}}$$

- Linear transformation of the nonlinear expressions is one common solution to use the linear regression procedures

$$\ln y = \alpha + \beta \ln x + \varepsilon$$

$$\frac{1}{y} - 1 = e^{\alpha + \beta x + \varepsilon} \Leftrightarrow \ln\left(\frac{1}{y} - 1\right) = \alpha + \beta x + \varepsilon$$

Interpreting the regression model

- Regression coefficients (B)
  - Estimated change in the dependent variable for a unit change in the independent variable.
  - Its value indicates the extent to which the IV is associated with the DV.
- Intercept
  - The intercept has explanatory value only within the range of values for the IV.
  - The intercept has interpretative value **only if zero is conceptually valid value for the IV**.
  - If the IV cannot have a true value of zero, the intercept only **aids in the improvement of the prediction process** and has **no explanatory value**.

FEUP

Example

- Objective: Make you familiar with linear regression models
- Your task: Estimate a linear regression model that predicts trips per occupied dwelling unit.
- Data [file in TDM's website: TDM_LR_Chicago_Example.xls]:
  - Trip production of 57 Traffic Assignment Zones of Chicago in 1960's
  - TODU: Motorized Trips (private car or Public Transportation) per occupied dwelling unit
  - ACO: Average car ownership (cars per dwelling)
  - AHS: Average household size
  - Three zonal social indices: SRI: Social Rank Index; UI: Urbanization; Index; SI: Segregation Index

Example - Social Rank Index

- This index contains 2 elements:
  1. the proportion of blue-collar workers, defined as the ratio of craftsmen, operatives, and laborers to all employees; and
  2. educational level as measured by the proportion of persons 25 years and older completing eight or fewer years of schooling.
- The social rank index is inversely related to both ratios
  - it attains a maximum value where no residents fall into the blue-collar jobs category, and
  - all adult residents have more than eight years of education

FEUP

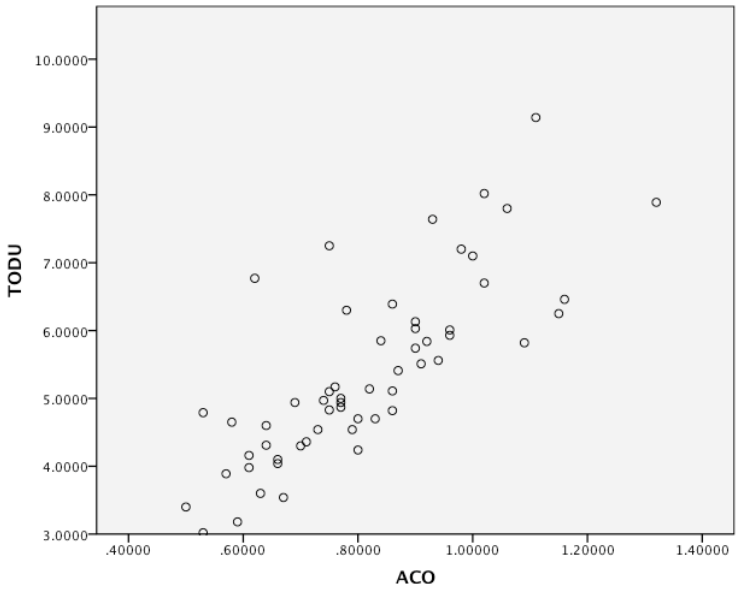
Example (cont'd) - Urbanization Index

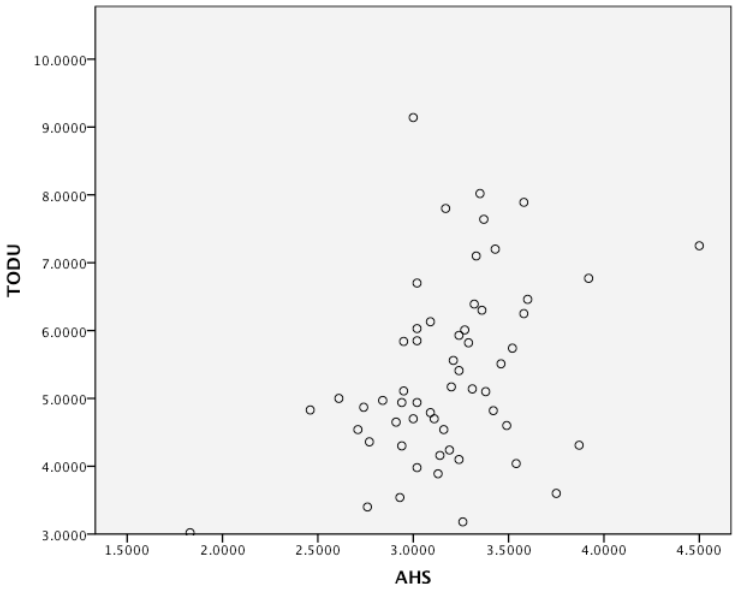
- This index contains 3 elements:
  1. fertility rate, defined as the ratio of children under 5 years of age to the female population of childbearing age
  2. female labor force participation rate, meaning the % of women who are in the labor force
  3. % of single family units to total dwelling units
- The degree of urbanization index would be increased by
  a) lower fertility rate,
  b) higher female labor force participation rate, and
  c) higher proportion of single dwelling units.
- This index measures in a rather negative way the degree of attachment to the home.
  - High values for this index imply less attachment to the home because of fewer children, higher likelihood of women being employed, and less permanency of dwelling unit type in terms of average tenure.

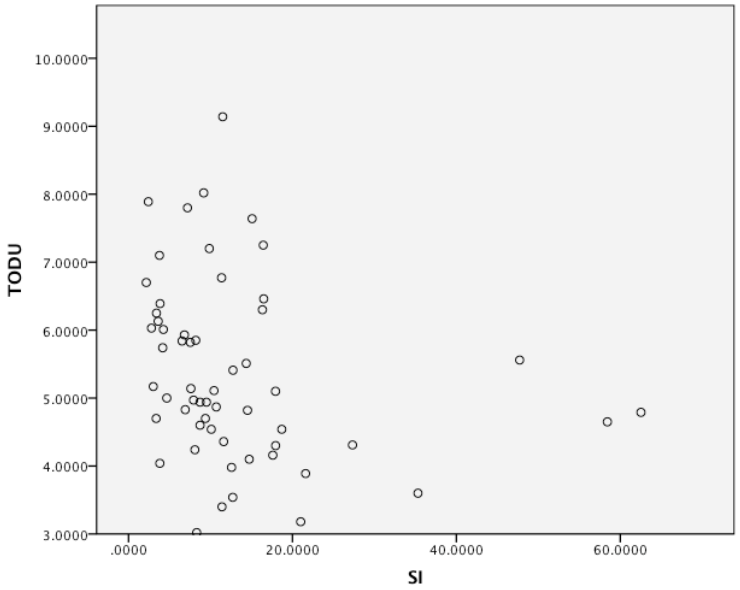
Example (cont'd) - Segregation Index

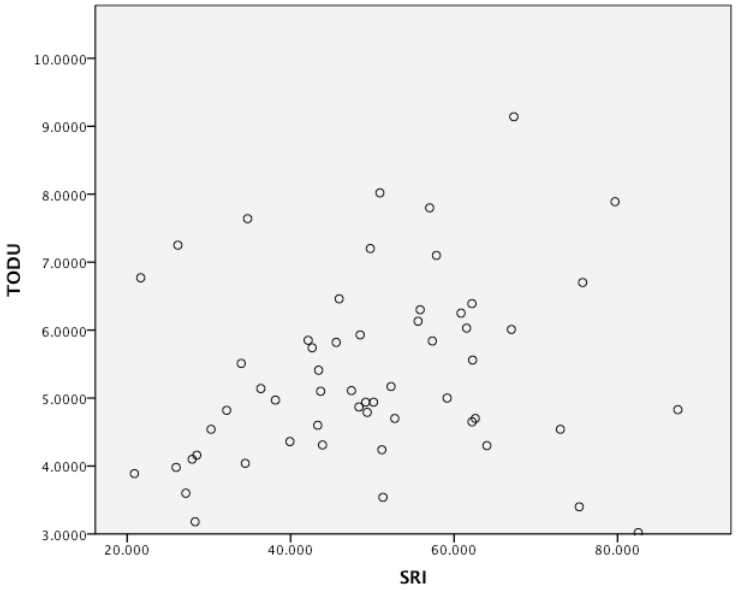
- This index is defined as the proportion of an area of residents who belong to certain minority groups, such as non-whites, foreign-born Eastern Europeans, etc.
- It measures the extent to which these minority groups live in relative isolation.
  - High values for this index imply that those communities are less prone to leaving their living areas and as such to having lower levels of mobility

Example – Preliminary analysis



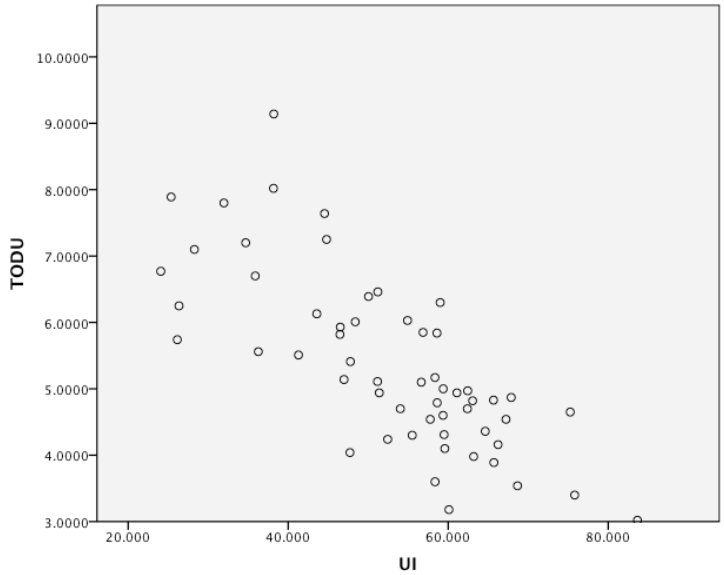






FEUP

Example (cont'd) – Preliminary analysis



- The pair wise comparison of the IV with the DV (TODU) helps in determining causal relationships
- In this case, all IV have logical and quite clear relationships with the TODU, except for the SRI that is more "fuzzy"

Example - Results

- Regression model including full set of IV

$$TODU = \underset{(1.276)}{2.817} + \underset{(3.813)}{3.647} \times ACO + \underset{(0.785)}{0.324} \times AHS + \underset{(0.574)}{0.005} \times SI + \underset{(0.924)}{0.008} \times SRI - \underset{(-2.720)}{0.036} \times UI + \varepsilon$$

Example – Output from SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.839[a]	.704	.675	.7553961	.704	24.283	5	51	.000

a. Predictors: (Constant), UI, SRI, SI, AHS, ACO

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	69.281	5	13.856	24.283	.000[a]
	Residual	29.102	51	.571		
	Total	98.383	56			

a. Predictors: (Constant), UI, SRI, SI, AHS, ACO

b. Dependent Variable: TODU

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.817	2.208		1.276	.208
	ACO	3.647	.957	.489	3.813	.000
	AHS	.324	.412	.095	.785	.436
	SI	.005	.009	.049	.574	.569
	SRI	.008	.009	.097	.924	.360
	UI	−.036	.013	−.368	−2.720	.009

a. Dependent Variable: TODU

Example – Output from SPSS

Model S

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		Sig. F Change
1	.839[a]	.704	.675	.7553961	51	.000

a. Predictors: (Constant), UI, SRI, SI, AHS, ACO

Proportion of variability in a data set that is accounted for by the statistical model.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	69.281	5	13.856	24.283	
	Residual	29.102	51	.571		
	Total	98.383	56			

a. Predictors: (Constant), UI, SRI, SI, AHS, ACO

b. Dependent Variable: TODU

Indicates if there is any (F>2.39 @ α=5%) statistical relationship (based on the variation of the means) between IV and DV

Are signals logical?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.817	2.208		1.276	.208
	ACO	3.647	.957	.489	3.813	.000
	AHS	.324	.412	.095	.785	.436
	SI	.005	.009	.049	.574	.569
	SRI	.008	.009	.097	.924	.360
	UI	-.036	.013	-.368	-2.720	.009

a. Dependent Variable: TODU

Regression of std variables:
- stdIV=(IV-mean)/STDEV
- stdDV=(DV-mean)/STDEV

Statistical significance of DV

Ordinary Least Squares (OLS) regression models: Assumptions

1. The dependent variable is continuous (measured in an interval or ratio scale).
   - Variables measured in ordinal or nominal scales should not be modeled using linear regression.
     - Ordinal Scales – Ordinal Logit or Ordinal Probit models;
     - Count variables (nonnegative integers) – Poisson or negative binomial regression
     - Nominal Scales - Multinomial Logit
2. Linear-in-parameters relationship between DV and IVs
   - Simple linear regression: $y_n = \alpha + \beta x_n + \varepsilon_n$
   - In most applications, the DV is a function of many IVs. Matrix notation and calculation is more appropriate

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \dots & \dots & \dots \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix} \times \begin{bmatrix} \beta_1 & \dots & \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

Type I and type II errors, in statistics (II)

- Type I error
  - **You should accept** the null hypothesis when it erroneously appears to be **False** ⇔ **False negative**
  - A type I error leads one to conclude that something or a relationship does not exists when in reality it does => **UNDER-ESTIMATION of the impact of the IV on the DV.**
  - Example:
    - The capacity of a link is not sufficient to cope with traffic loads when in reality it does
    - The impact of the IV on the DV is not significant when in reality it is

Type I and type II errors, in statistics (I)

- Type II error
  - **You should reject** the null hypothesis when it erroneously appears to be **true** ⇔ **False positive**
  - A type II error leads one to conclude that something or a relationship exists when in reality it doesn't => **OVER-ESTIMATION of the impact of the IV on the DV.**
  - Example:
    - The capacity of a link is sufficient to cope with traffic loads when in reality it doesn't
    - The impact of the IV on the DV is significant when in reality it isn't

FEUP

OLS regression models: Assumptions (cont'd)

3. Linear-in-parameters relationship between DV and IVs
   - If the relationship between IV and DV is not linear, the results of the linear regression analysis will **under-estimate** the true relationship between the IV and DV.
   - This under-estimation carries two risks:
     1. Increased chance of a Type I error for that IV, i.e. increased chance of the true impact of IV on the DV being **under-estimated**
     2. Increased risk of Type II errors (**over-estimation**) for other IVs that share variance with that IV, i.e. increased chance of the true impact of IV on the DV being over-estimated
4. DV variable should be normally distributed (otherwise other regression methods are more appropriate):
   - For datasets smaller than 2000 elements, we use the Shapiro-Wilk test, otherwise, the Kolmogorov-Smirnov test is used.
   - Both tests can be obtained from the Analyze/Descriptive Statistics/Explore command in SPSS

OLS regression models: Assumptions (cont'd)

5. Observations Independently and Randomly Sampled
   - It can be relaxed if remedial actions are taken.
   - It is an assumption necessary to make inferences about the population of interest (the **data should be randomly sampled from the population**).
   - The probability that an observation is selected is unaffected by other observations selected into the sample (**Independence**).
6. Uncertain Relationship between Variables
   - The difference between the equation of a straight-line and a linear regression model is the addition of a **stochastic, disturbance, or disturbance term, ε**.
   - ε consists of several elements:
     - variables omitted from the model (assumed to be of small importance)
     - measurement errors in the DV (the independent variables are assumed to be measured without error) (there are tests to verify this – not covered here)
     - random variation in the underlying data-generating process

OLS regression models: Assumptions (cont'd)

7. ε is independent of X and Expected Value Zero

$$E[\varepsilon_n] = 0 \qquad VAR[\varepsilon_n] = \sigma^2$$

- ε is independent across observations and the error variance is constant across IV.
  - This is the homoscedasticity assumption (*homogeneity of variance*)
  - The net effect of model uncertainty is not systematic across observations
- When disturbances are heteroscedastic then alternative modeling approaches should be used, or in same cases the transformation of variables.
- Heteroscedasticity is worth correcting only when the problem is severe
  - No reason for throwing away an otherwise "good" model
  - The risk of high heteroscedasticity is increasing the possibility of type II error for the IV (over-estimation of the impact of IV on the DV)

8. Disturbance Terms Not Autocorrelated (ε independent across observations)

$$COV[\varepsilon_{i,}, \varepsilon_j] = 0 \quad if\ i \neq j$$

- Common violations of this assumption occur when observations are repeated on individuals. Observations across time often possess autocorrelated disturbances as well.

OLS regression models: Assumptions (cont'd)

9. Regressors and Disturbances Uncorrelated

$$COV\left[X_{i,},\varepsilon_j\right]=0 \quad for \quad all \quad i \quad and \quad j$$

- Exogeneity of the regressors (they are not correlated with the error)
  - The values of the regressors are determined by influences "outside of the model."
  - *y* does not directly influence the value of an exogenous regressor.

10. Disturbances Approximately Normally Distributed

$$\varepsilon_j \approx N\left(0,\sigma^2\right)$$

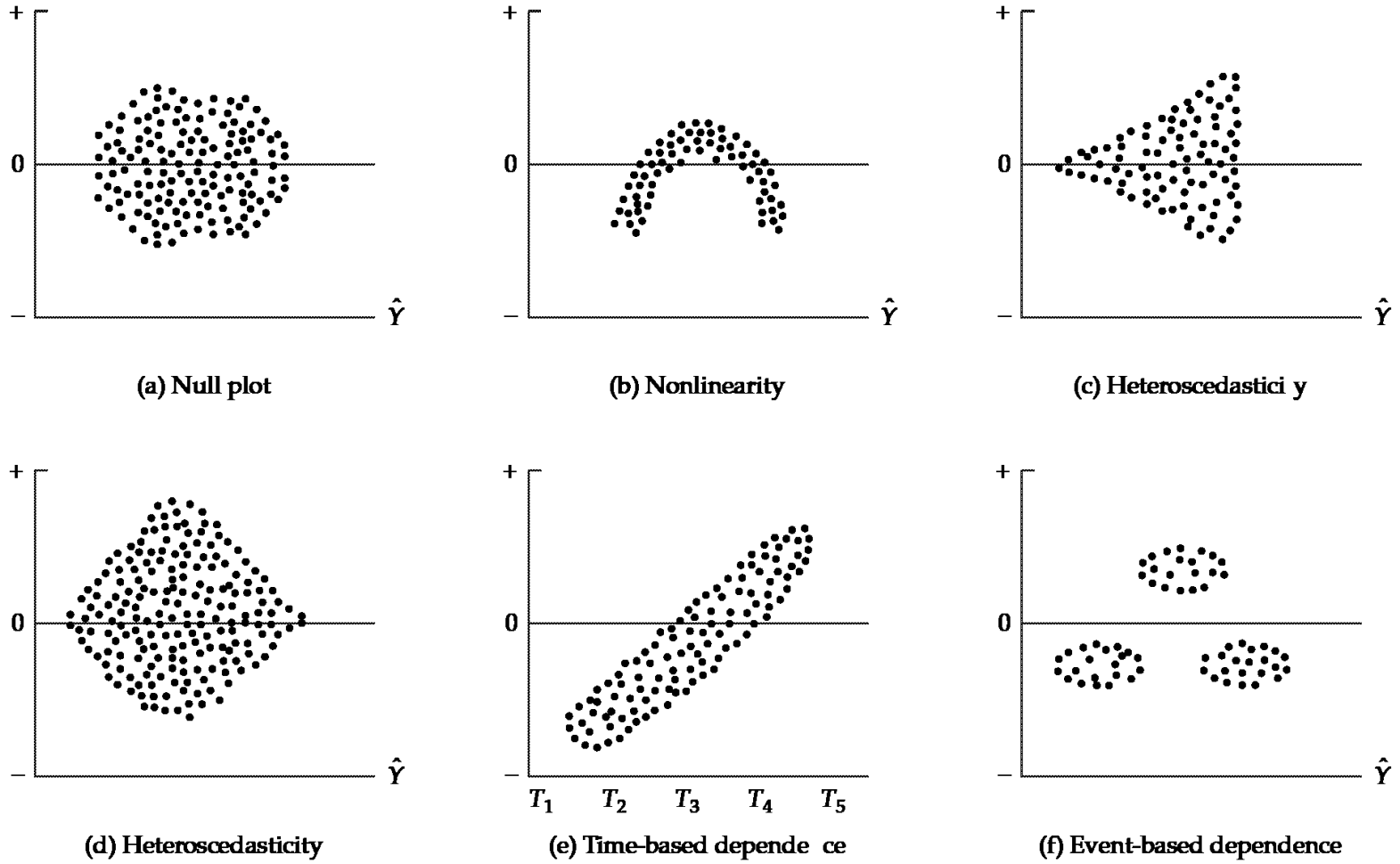
- The disturbance terms are required to be approximately normally distributed in order to make inferences about the parameters from the model
- The error terms are independently and identically distributed as normal (i.i.d. normal).

OLS regression models: Assumptions (cont'd)

Statistical Assumption	Figure next slide	Mathematical Expression
Linear-in-parameters relationship between DV and IVs	(b)	$y_n = \alpha + \beta x_n + \varepsilon_n$
Zero mean of ε	n.a.	$E[\varepsilon_n] = 0$
Homoscedasticity of ε	(c) ; (h)	$VAR[\varepsilon_n] = \sigma^2$
Nonautocorrelation of ε	(e) ; (f)	$COV[X_{i,}, \varepsilon_j] = 0 \quad for \quad all \quad i \quad and \quad j$
Uncorrelatedness of regressor and disturbances	n.a.	$COV[\varepsilon_{i,}, \varepsilon_j] = 0 \quad if \; i \neq j$
Normality of disturbances	(g)	$\varepsilon_j \approx N(0, \sigma^2)$

OLS regression models: Assumptions (cont'd)

Graphical analysis of residuals



Source: Hair et al (1995)

OLS regression models: Curve fitting

- Formal representation of the simple LR model

$$y_n = \alpha + \beta x_n + \varepsilon_n, \quad n = 1,\ldots,N \quad observations$$

where

y_n is the observed value of the dependent variable for the n^{th} observation,
x_n is the value of the independent variable for the n^{th} observation,
ε_n *is the residual of* the n^{th} observation, *and*
α and β are unknown parameters.

- We do not know (**and will probably never know**) the true values of α and β.
  - What the regression model will produce is estimates for these parameters
  - They are only estimates because the regression model uses a sample (N observations) of the entire population

$$E[y_n \mid x_n] = E[\alpha + \beta x_n + \varepsilon_n] \Leftrightarrow \hat{y} = \hat{\alpha} + \hat{\beta} x$$

OLS regression models: Curve fitting (cont'd)

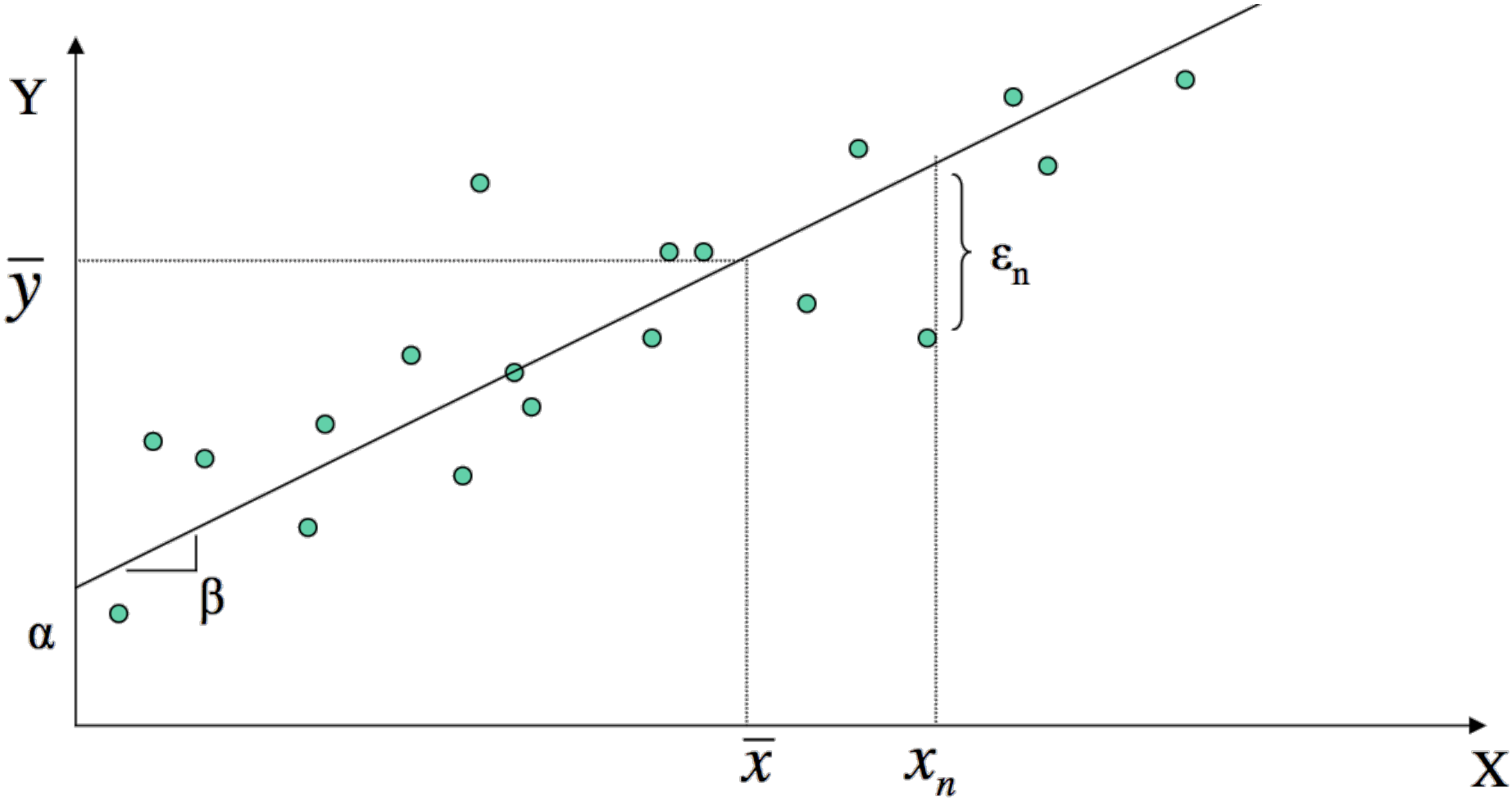
- Ordinary Least Squares (OLS)
  - The technique used in models that attempt to **minimize the sum of squared residuals** and that produces the estimators of the LR model
  - $\hat{\alpha}$ is the estimator of the intercept, and $\hat{\beta}$ is the estimator of the slope in a simple linear regression model
- Algebraic expression of the OLS regression model

$$y_n = \alpha + \beta x_n + \varepsilon \Leftrightarrow \varepsilon = y_n - \alpha - \beta x_n$$

$$\Rightarrow Min_{\alpha,\beta} \sum_{n=1}^{N} \varepsilon_n^2 = Min_{\alpha,\beta} \sum_{n=1}^{N} \left(y_n - \alpha - \beta x_n\right)^2$$

  - Attention:
    - Large ε are highly weighted
    - Outliers should be treated carefully before exclusion (there are tests for this)

OLS regression models: Curve fitting (cont'd)



FEUP

OLS regression models: Mathematical properties

- 1st order conditions: Estimator of α

$$\frac{\delta}{\delta\alpha}\left[\sum_{n=1}^{N}\left(y_n-\alpha-\beta x_n\right)^2\right]=-2\sum_{n=1}^{N}\left(y_n-\hat{\alpha}-\hat{\beta}\,x_n\right)=0$$

$$\Rightarrow\sum_{n=1}^{N}\left(y_n-\alpha-\beta x_n\right)=0\Rightarrow\sum_{n=1}^{N}y_n-N\hat{\alpha}-\hat{\beta}\sum_{n=1}^{N}x_n=0$$

$$\Rightarrow\hat{\alpha}=\frac{\sum_{n=1}^{N}y_n}{N}-\frac{\hat{\beta}\sum_{n=1}^{N}x_n}{N}=\overline{Y}-\hat{\beta}\,\overline{X}\qquad(1)$$

OLS regression models: Mathematical properties (cont'd)

- 1st order conditions: Estimator of β

$$\frac{\delta}{\delta\beta}\left[\sum_{n=1}^{N}\left(y_n-\alpha-\beta x_n\right)^2\right]=-2\sum_{n=1}^{N}x_n\left(y_n-\hat{\alpha}-\hat{\beta}\,x_n\right)=0\Rightarrow\sum_{n=1}^{N}x_n y_n-\hat{\alpha}\sum_{n=1}^{N}x_n-\hat{\beta}\sum_{n=1}^{N}x_n^2=0$$

Substituting the expression for $\hat{\alpha}$, *we obtain*:

$$\Rightarrow\sum_{n=1}^{N}x_n y_n-\left[\frac{\sum_{n=1}^{N}y_n}{N}-\hat{\beta}\frac{\sum_{n=1}^{N}x_n}{N}\right]\sum_{n=1}^{N}x_n-\hat{\beta}\sum_{n=1}^{N}x_n^2=0\Rightarrow\sum_{n=1}^{N}x_n y_n-\frac{\sum_{n=1}^{N}y_n\sum_{n=1}^{N}x_n}{N}+\hat{\beta}\frac{\left(\sum_{n=1}^{N}x_n\right)^2}{N}-\hat{\beta}\sum_{n=1}^{N}x_n^2=0$$

$$\Rightarrow N\sum_{n=1}^{N}x_n y_n-\sum_{n=1}^{N}y_n\sum_{n=1}^{N}x_n+\hat{\beta}\left(\sum_{n=1}^{N}x_n\right)^2-\hat{\beta}N\sum_{n=1}^{N}x_n^2=0\Rightarrow\hat{\beta}=\frac{N\sum_{n=1}^{N}x_n y_n-\sum_{n=1}^{N}y_n\sum_{n=1}^{N}x_n}{N\sum_{n=1}^{N}x_n^2-\left(\sum_{n=1}^{N}x_n\right)^2}$$

$$\Rightarrow\hat{\beta}=\frac{\sum_{n=1}^{N}\left(x_n-\bar{x}\right)\left(y_n-\bar{y}\right)}{\sum_{n=1}^{N}\left(x_n-\bar{x}\right)^2}=\frac{COV(x,y)}{VAR(x)}\qquad(2)$$

Required condition!

$$\sum_{n=1}^{N}\left(x_n-\bar{x}\right)^2\neq 0$$

OLS regression models: Mathematical properties (cont'd)

- We can also check the 2nd order conditions to ensure that we indeed have a minimum:

$$\frac{\delta^2}{\delta\alpha^2}\left[\sum_{n=1}^{N}\left(y_n-\alpha-\beta x_n\right)^2\right]=-2\sum_{n=1}^{N}\left(-1\right)=2N>0$$

$$\frac{\delta^2}{\delta\beta^2}\left[\sum_{n=1}^{N}\left(y_n-\alpha-\beta x_n\right)^2\right]=-2\sum_{n=1}^{N}\left(-x_n^2\right)=2\sum_{n=1}^{N}x_n^2>0$$

Estimation (I)

- Inferences are made with the β parameters in classical linear regression
- Considering as an example B_1 (*estimator of true beta*)
  - The sampling distribution of B_1 is the distribution of values that would result from repeated samples drawn from the population with levels of the independent variables held constant.
  - It can be deducted that the sampling distribution of B_1 is approximately normal (*X* being the associated IV).

$$B_1 \approx N\left(\beta_1, \frac{\sigma^2}{\sum (X_i - X)^2}\right)$$

Estimation (II)

- Since the population variance σ^2 is typically unknown, an estimate called **mean squared error (MSE)** is calculated. MSE is an estimate of the variance in the regression model

$$MSE = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n - p}$$

  - n = sample size and p = number of estimated model parameters.

- It can be shown that MSE is an unbiased estimator of σ^2

$$E[MSE] = \sigma^2$$

Estimation (III)

- Since B_k is normally distributed, and β_k is a constant and $\beta_k = E(B_k)$

$$Z^* = \frac{B_k - \beta_k}{\sigma\{B_k\}}$$

- In practice the true StDev in the denominator is not known and is estimated using MSE

$$t^* = \frac{B_k - \beta_k}{\sqrt{\frac{MSE}{\sum (X_i - \overline{X})^2}}} = \frac{B_k - \beta_k}{s\{B_k\}} \approx t(\alpha; n-p)$$

  - α = level of significance, $(n-p)$ = associated degrees of freedom.
- Enables a statistical test of the probabilistic evidence in favor of specific values of β_k.

Estimation (IV)

$$B_k \pm t\left(1-\frac{\alpha}{2}; n-p\right) s\{B_k\}$$

- The confidence interval provides the long-run probability that the true value of β lies in the computed confidence interval, conditioned on the same levels of X being sampled.

$$H_0: \beta_k = 0$$

$$H_a: \beta_k \neq 0.$$

$$\text{If } |t^*| \leq t_{crit}\left(1-\frac{\alpha}{2}; n-p\right), \text{ conclude } H_0$$

$$\text{If } |t^*| > t_{crit}\left(1-\frac{\alpha}{2}; n-p\right), \text{ conclude } H_a$$

Estimation: Standardized Regression Models

- Due to differences in scale u*sing the original measurement units of X* will not provide an indication of which ones have largest relative impact on *Y*.

$$X_1' = \frac{X_1 - \overline{X}}{s\{X_1\}}$$

- Standardization solves this problem. The estimated regression parameters in a standardized regression model are interpreted as a change in the response variable per unit change of one standard deviation of the independent variable. Standardized variables are created with expected values equal to 0 and variances equal to 1
- Standardization strictly works on continuous variables, those measured on interval or ratio scales.

Estimation: Validation of Regression Assumptions

- **Linearity**
  - Checked informally using several plots
    - independent variables on the X vs. residuals on the Y,
    - model predicted (fitted) values on the X vs. residuals on the Y.
  - Curvilinear trends in the disturbances are an evidence of non-linear relations.

Estimation: Validation of regression assumptions (I)

FEUP

❒ Homoscedastic Disturbances

- The consequence of a heteroscedastic regression is reduced precision of beta parameter estimates. Regression parameters will be less efficient under this circunstances.
- MSE will be larger for a heteroscedastic regression (smaller t*).

- Scatter plots are used to assess homoscedasticity. A plot of model fitted values vs. errors is typically inspected first. If heteroscedasticity is detected, then plots of the disturbances vs. independent variables or partial variate plots should be conducted to identify were the problem occurs.

Estimation: Validation of regression assumptions (II)

- **Uncorrelated Disturbances**
  - Correlation of disturbances across time is called serial correlation. Plot of disturbances vs. time, or a plot of disturbances vs. ordered observations (over space).
    - Serially correlated disturbances will reveal a trend over time, with peaks and valleys in the disturbances that typically repeat themselves over fixed intervals.
  - Durbin–Watson statistic. This statistic is calculated from the disturbances of an OLS regression.
  - Durbin–Watson ≈ 2,0 (+-0,2) No autocorrelation
  - Durbin–Watson ≠2,0 (+-0,2) Presence of autocorrelation
    - See Danomar N. Gujarati "Basic Econometrics (3rd ed." (1995) – pg. 420, for more in depth information

Estimation: Validation of regression assumptions (III)

❒ Exogenous Independent Variables

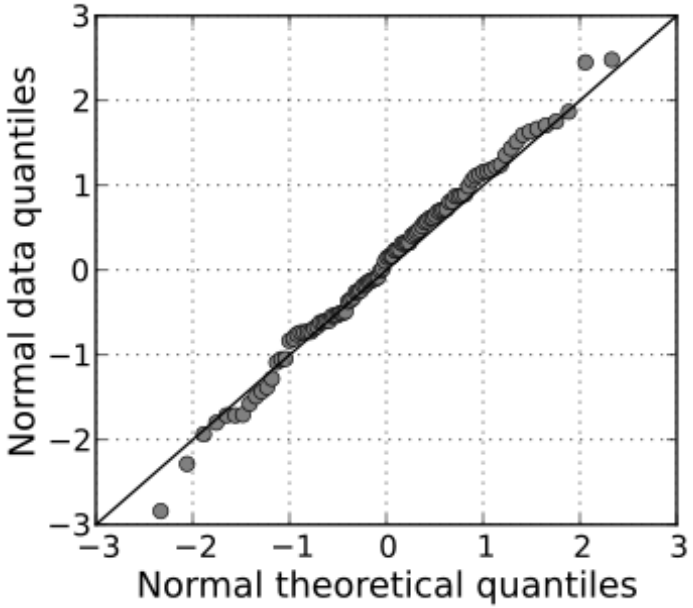
- The value of an exogenous variable is determined by factors outside the model (i.e., somehow "*hidden*" in the error term).
- When endogeneity is present, however, the covariance between *X and* ε *is nonzero* and the least squares estimate is biased.
- The direction of bias depends on the covariance between *X and* ε. *A* negative covariance will result in a negative bias (Type I error), and a positive in a positive bias (Type II error).

Estimation: Validation of regression assumptions (IV)

- **Normally Distributed errors**
  - If making inferences is not important, then the assumption of normality can be dropped without consequence.
  - Normality can be assessed through nongraphical, graphical, and nonparametric methods
  - Summary statistics of the disturbances, including minimum, first and third quartiles, median, and maximum values of the disturbances (normal distribution is symmetric and mean≈median ≈ 0).
  - Histograms of the disturbances – should reveal the familiar bell-shaped curve. The number of observations above and bellow zero should be approximately equivalent and mirror each other.

Estimation: Validation of regression assumptions (V)

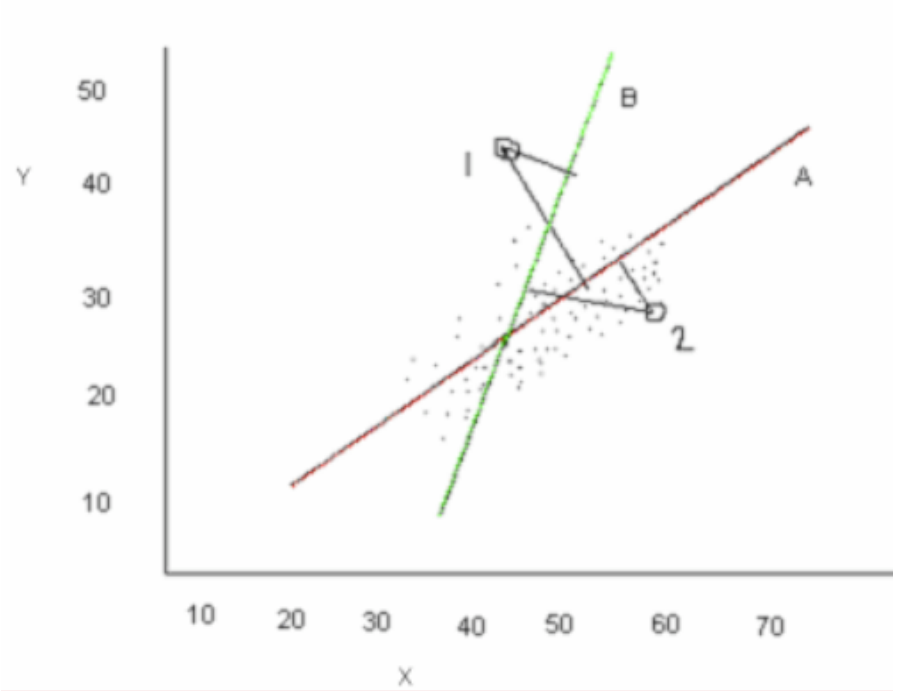
- **Normality of disturbances**
  - Normal probability quantile-quantile (Q-Q) plots of the disturbances. Normal Q-Q plots are constructed such that normally distributed disturbances will plot on a perfectly straight line.



Source: http://en.wikipedia.org/wiki/Q-Q_plot

Detection of influential observations (*outliers*)

- A least squares model can be distorted by a single observation. The fitted line or surface might be tipped so that it no longer passes through the bulk of the data.
- In order to reduce the effect of a very large error it will introduce many small or moderate errors.
- For example, the point 1 is actually an outlier and in presence of this point the regression line is dragged out to that point resulting the point 2 as an outlier, though it is a clean point, because now the distance of this point from the line is longest.



Estimation: Regression Outliers (II)

- Removing observations from the regression

  Criticisms - “the data were fit to the model.”

- Leaving the observations in the model

  Criticism - “lack of fit.”

- We should fully document and completely justify the removing of any and all data from an analysis is good practice.

Detection of influential observations (*outliers*)

- There are two types of outliers depending on the variable in which it occurs:
  - Outliers in the response variable represent model failure.
  - Outliers with respect to the predictors are called leverage points
    - these can affect the regression model
- There are many methods for detection of outliers available in the literature. Some statistics that are obtained through row deletion method of regression matrix.
  - It is examined in turn how the deletion of each row affects the estimated coefficients, the predicted values (fitted values), the residuals, and the estimated covariance structure of the coefficients.
  - These methods estimate the influence of an observation on the regression outcomes and uses cut-off values called ***leverage*** values, that help in identifying those observations that are far away from corresponding average predictor values
- Boxplots can help visually to detect outliers
  - values 3x bigger than P75 (thumb rule)

Detection of influential observations (*outliers*)

- Cook's Distance
  - Cook (1977) proposed a statistic for detection of outlier as follows:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ \mathrm{MSE}}.$$

    - $\hat{Y}_j$ is the prediction from the full regression model for observation j;
    - $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted;
    - *MSE* is the mean square error of the regression model;
    - *p* is the number of fitted parameters in the model
  - An interpretation is that *Di* measures the sum of squared changes in the predictions when observation *i* is not used in estimating β. *Di* approximately follows F (p, n-p) distribution.
  - The cut off value of Cook-Statistic is 4/n.

Detection of influential observations (*outliers*)

- DDFITS
  - It is the difference between the predicted responses from the model constructed using complete data and the predicted responses from the model constructed by setting the i^{th} observation aside.
  - Unlike Cook's distance, it does not look at all of the predicted values with the i^{th} observation set aside. It looks only at the predicted values for the i^{th} observation.

$$DFFITS = \frac{\hat{y}_i - \widehat{y_{i(i)}}}{s_{(i)}\sqrt{h_{ii}}}$$

  - $\hat{Y}_j$ is the prediction from the full regression model for observation j;
  - $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted;
  - *s(i)* is the standard error estimated without the point in question, and
  - h_{ii} is the leverage for the point (p/n), the number of parameters divided by the number of points.
  - The cut off value of DFFIT is $2\sqrt{\frac{p}{n}}$,

Goodness-of-Fit Indicators (I)

- Goodness-of-fit (GOF) statistics are useful for comparing
  - results across multiple studies,
  - competing models within a single study,
  - and for providing feedback on the extent of knowledge about the uncertainty involved with the phenomenon of interest

Goodness-of-Fit Indicators (II)

- Sum of square errors (variation of the fitted regression line around the observations)

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

- Regression sum of squares (variation of the fitted regression line around $\bar{Y}$)

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$

- Total sum of squares (the variation of each observation around $\bar{Y}$)

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$$

$$\boxed{SST = SSR + SSE}$$

Goodness-of-Fit Indicators (III)

❒ Coefficient of determinantion R^2

- It varies between [0;1] is the proportion of total variance explained

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

- Since R^2 can only increase when variables are added to the regression model an adjusted measure ($R^2_{Adjusted}$) is proposed to account for the degrees of freedom changes as a result of different numbers of model parameters.

$$R^2 = 1 - \frac{SSE\,/(n-p-1)}{SST\,/(n-1)}$$

- The R^2 and $R^2_{Adjusted}$ provide only relevant comparisons with previous models that have been estimated on the phenomenon under investigation.
- The absolute values of R^2 and $R^2_{Adjusted}$ measures are not sufficient measures to judge the quality of a model.

Goodness-of-Fit Indicators (IV)

❒ F test

- It is general and flexible approach to test the statistical difference between competing models.
- First, a full or unrestricted model is estimated. The sum of square errors for the full model is

$$SSE_F = \sum_{i=1}^{n} (Y_i - \hat{Y}_{Fi})^2$$

- A reduced model is then estimated (viable competitor to the full model but with fewer variables or only a constant)

$$SSE_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_{Ri})^2$$

- The logic of the F test is to compare the values of SSE_R and SSE_F

Goodness-of-Fit Indicators (II)

F test

- If $SSE_R = SSE_F$, the full model has done nothing to improve the fit of the model. The reduced model is superior
- In statistical terms, the null hypothesis (H_0) is that all of the additional parameters in the full model are not significant (or $\beta_k=0$)

$$H_0: \text{all } \beta_k = 0$$

$$H_a: \text{all } \beta_k \neq 0.$$

$$F^* = \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}} \approx F(1-\alpha; df_R - df_F, df_F)$$

$$\text{If } F^* \leq F(1-\alpha; df_R - df_F, df_F), \text{ then conclude } H_0$$

$$\text{If } F^* \geq F(1-\alpha; df_R - df_F, df_F), \text{ then conclude } H_a$$

Multicollinearity in the Regression

- **Multicollinearity** exists when **IV are correlated with each other** or when **IV are correlated with omitted variables** (somehow included in the error term) that are related to the dependent variable (eg resulting in inefficient parameters).
- How to evaluate if there is any multicollinearity?
  - Pairwise correlation between variables could be used to detect multicollinearity (when r>0,75, there is a strong sign of problems)
  - Variance Inflation Factor (VIF) - Values >5 or >10 indicate collinearity problems
  - Condition Index or Condition Number - values >15 indicate problems and >30 serious problems

Multicollinearity in the Regression: Variance Inflation Index (VIF)

- It provides an index that measures **how much the variance of an estimated regression coefficient** (the square of the estimate's standard deviation) **is increased because of collinearity**.
  - Consider the following linear model with k independent variables:

$$Y = \beta 0 + \beta 1\, X1 + \beta 2\, X\, 2 + \ldots + \beta k\, Xk + \varepsilon.$$

  - It can be shown that the variance of the j^{th} β is given by:

$$\operatorname{var}(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)\widehat{\operatorname{var}}(X_j)} \cdot \frac{1}{1-R_j^2},$$

    - where R_j^2 is obtained for the regression of X_j on the other covariates/regressors (a regression that does not involve the response variable Y)
    - The R_j^2 indicates how predictable the j^{th} IV is from the set of other IVs.
  - *VIF=1/(1-R_j^2)*
  - *Tolerance = 1-R_j^2*

Multicollinearity in the Regression: Condition Index (I)

- Most multivariate statistical approaches involve decomposing a **correlation matrix** into **linear combinations of variables** (compare the variances of the IV in the case of the MRA).
- The SPSS will produce a table like this (for the Chicago case study):

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	ACO	AHS	SI	SRI	UI
1	1	5.387	1.000	.00	.00	.00	.01	.00	.00
	2	.444	3.481	.00	.00	.00	.69	.01	.00
	3	.084	7.989	.00	.05	.01	.05	.06	.11
	4	.074	8.544	.00	.00	.01	.15	.38	.05
	5	.009	24.037	.01	.79	.21	.09	.37	.19
	6	.002	59.882	.99	.15	.77	.00	.18	.65

a. Dependent Variable: TODU

Multicollinearity in the Regression: Condition Index (I)

- Eigenvalue (λ)
  - Factor analysis yields a set of factors that are linear combinations with different combinations of load factors for each Xi
  - Each of these factors identified explains a % variance of the overall variance of the observed values of the Xi independent variables.

$$\%VAR_{f_i} = \frac{Var(f_i)}{Var(X_{1,\ldots,n})}$$

$$\lambda_i = \frac{\%VAR_{f_i}}{\overline{\%VAR}}$$, where $\overline{\%VAR}$ is the expected average % explanation of the overall variance of the IVs.

- Condition index or con Index (CI_i)

$$CI_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}}$$

Recommended readings

- Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) "Statistical and econometric Methods for Transportation Data Analysis", CRC – Chapter 3 and Annex A
- **Hair, Joseph P. et al (1995) "Multivariate Data Analysis with Readings", Fourth Edition, Prentice Hall - Chapter 3**
- João Maroco, Regina Bispo (2003) "Estatística Aplicada às Ciências Sociais e Humanas", Ed. Manuais Escolares – Capítulo 13

FEUP