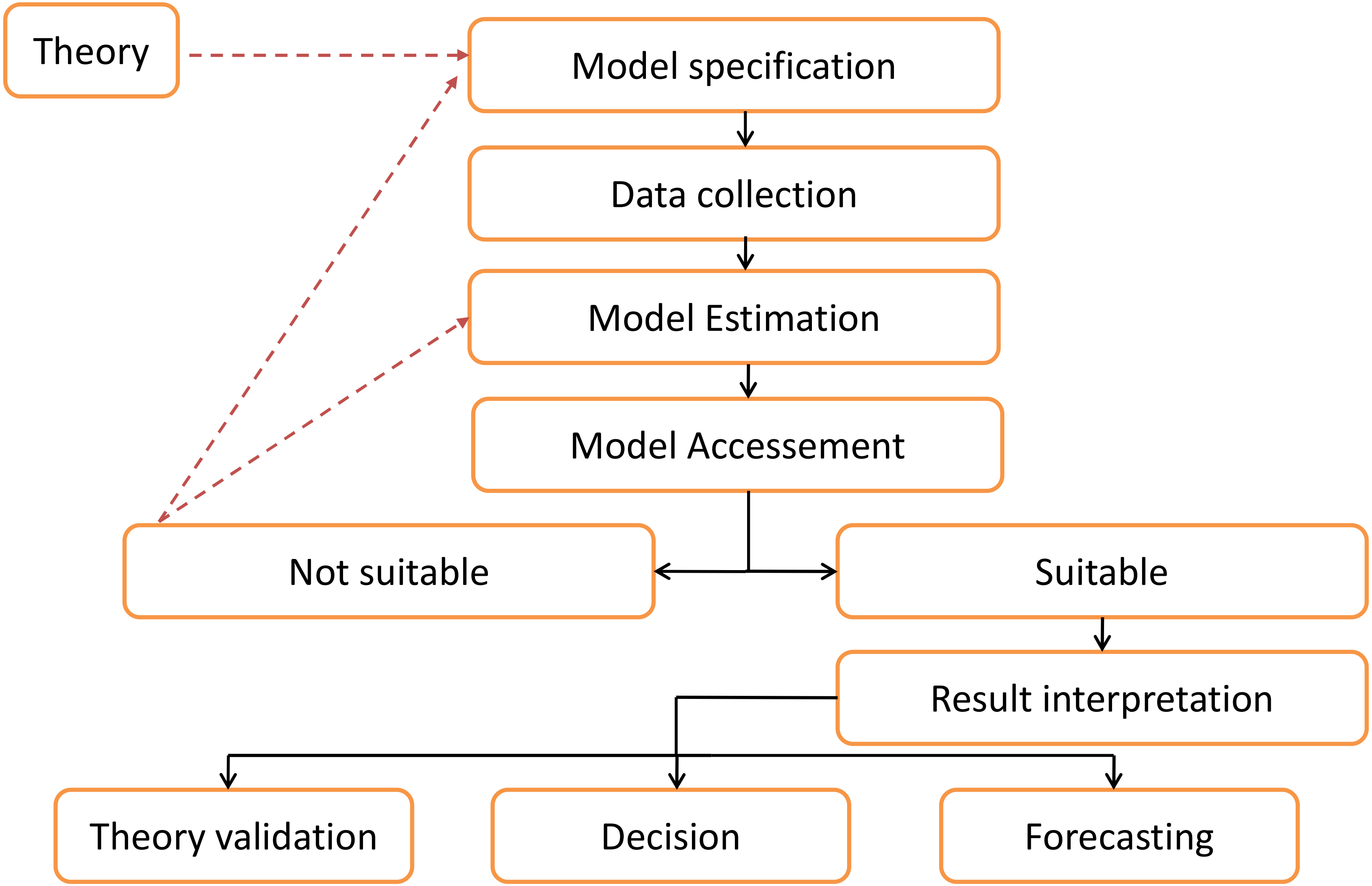
Econometrics

- Concept:
  - Application of statistical techniques to analyse data in areas such as economics, finance, management, etc. with the aim of estimating the relation between a dependent variable / variable of interest and several explanatory variables / determinants
  - Examples:
    - Consumption= f(income, age, …)
    - Wage=f(education, prof. experience, age, …)
    - Debt=f(firm age, total assets, …)

- Aims:
  - Testing the validity of theories
  - Forecasting
  - Evaluate policies

Methodology



Causality & ceteris paribus analysis

A major aim is analysing the **determinants** of the variable of interest (example: firm age, total assets, ….cause the firms debt?):

- Check whether the explanatory variables are **significant (statistically)** to explain the variable of interest, that is, check the existence of **causality**

- The **marginal / partial effect** of each explanatory variable over the variable of interest is measured, ceteris paribus, that is, assuming all the remaining determinants constant.

Multiple regression linear model: aim

Aim: explainning $E(Y|X)$

Y: dependente variable / variable of interest

X: explanatory variables / determinants / regressors

$E(Y|X)$: expected value / conditional mean of Y given X

$E(Y|X)$ is a function of parameters β that are estimated

Specification of the MRLM

Model specification:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i \qquad (i = 1, \cdots, N)$$

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

u: error term – all the determinants of Y, that have not been included in X

β: coefficients to be estimated

k: # of explanatory variables

$k + 1$: # of parameters (in models including a constant term)

N: # of observations

Specification of the MRLM (cont.)

Example: factors in u_i

$$hourly\_wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i$$

u_i contains unobservable factors such as gender, location, motivation, activity sector, … that explain wage, but were not measured

OLS estimation, predicted values and residuals

Predicted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}$

Residuals: $\hat{u}_i = Y_i - \hat{Y}_i$

$\hat{u}$: residual

$\hat{Y}$: estimator of $E(Y|X)$

$\hat{\beta}$: estimator of β

Model estimation:

- Ordinary least squares (OLS): $min \sum_{i=1}^{N} \hat{u}_i^2$
  - The sum of the square of the residuals is minimized with respect to β

Interpretation

In general $\beta_j, j = 1, \dots k$, measures the impact on the conditional mean of Y given X, $E(Y|X)$, due to a variation of the determinant X_j associated to β_j, all the rest equal

Linear model (no variable transformation),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i$$

Partial effect:

$$\Delta X_j = 1 \rightarrow \Delta E(Y|X) = \beta_j, \textit{ceteris paribus}$$

Interpretation

Linear model in the parameters:

$$Y_i^* = \beta_0 + \beta_1 X_{i1}^* + \cdots + \beta_k X_{ik}^* + u_i$$

Y^*	X^*	Interpretation
Y	X_j	$\Delta X_j = 1 \rightarrow \Delta E(Y\|X) = \beta_j$
ln(Y)	X_j	$\Delta X_j = 1 \rightarrow \Delta E(Y\|X) = 100\beta_j\%$
Y	ln(X_j)	$\Delta X_j = 1\% \rightarrow \Delta E(Y\|X) = \frac{\beta_j}{100}$
ln(Y)	ln(X_j)	$\Delta X_j = 1\% \rightarrow \Delta E(Y\|X) = \beta_j\%$
Y	X, X^2	$\Delta X_j = 1 \rightarrow \Delta E(Y\|X) = \beta_x + 2\beta_{x^2}X$

Interpretation

Example: consider the model

$$price_i = \beta_0 + \beta_1 area_i + \beta_2 rooms_i + u_i$$

where the house *price*, in hundred of dollars, depends on *area* (m2) and number of *rooms*. Estimated model

$$\widehat{price}_i = -19.286 + 1.384 area_i + 15.121 rooms_i$$

.

Assuming everything else constant:

- for an unitary variation of area, that is for each additional m^2, in average the house price increases 1.384 hunded of dolars
- for an unitary variation of rooms, that is for each additional room, in average the house price increases 15.121 hunded of dolars

Output in Stata

```
. regress price rooms area

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.95
       Model |  579971.198     2  289985.599           Prob > F      =  0.0000
    Residual |  337883.308    85  3975.09774           R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6232
       Total |  917854.506    87  10550.0518           Root MSE      =  63.048

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rooms |   15.12134   9.488598     1.59   0.115    -3.744538    33.98721
       area |   1.383606   .1489435     9.29   0.000     1.087467    1.679746
       _cons |  -19.2855   31.04753    -0.62   0.536    -81.0163     42.4453
------------------------------------------------------------------------------
```

Interpretation

Example: consider the alternative model

$$ln(\widehat{preço_i}) = 1.289 + 0.810ln(area_i) + 0.038quartos_i$$

Everything else constant:

- An increase of 1% in area, is estimated to generate an increase in the average price of 0.810%
- For each additional room, the average price of houses is expected to increase 3.8%

Note that ln(.) only is defined for positive variables

Interpretation

Output

```
. generate lprice=ln(price)
. generate larea=ln(area)
. regress lprice rooms larea

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   54.47
       Model |  4.50364223     2  2.25182112           Prob > F      =  0.0000
    Residual |  3.51396129    85  .041340721           R-squared     =  0.5617
-------------+------------------------------           Adj R-squared =  0.5514
       Total |  8.01760352    87  .092156362           Root MSE      =  .20332

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rooms |   .0376464   .0303446     1.24   0.218    -.0226868    .0979795
       larea |   .8100637   .0987611     8.20   0.000     .6137002    1.006427
       _cons |    1.28929   .4666125     2.76   0.007     .3615395    2.217041
------------------------------------------------------------------------------
```

Sampling distribution of the OLS estimator for $\hat{\beta}_j$, t statistics and CI

Consider

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\beta_j}\right) \text{ and } \frac{\hat{\beta}_j - \beta_j}{\sigma_{\beta_j}} \sim N(0,1).$$

As σ_{β_j} is unknown, $\hat{\sigma}_{\beta_j}$ is used:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \sim t(N-k-1)$$

The CI for β_j is

$$\left(\hat{\beta}_j - t^{\alpha/2}_{N-k-1}\hat{\sigma}_{\beta_j}; \hat{\beta}_j + t^{\alpha/2}_{N-k-1}\hat{\sigma}_{\beta_j}\right),$$

which means that with $(1-\alpha)100\%$ of confidence, β_j is included in this interval

T test: main versions

Testing the individual significance of $X_j, j = 1, \dots k$:

Hipothesis	t_j	**Note**
$H_0: \beta_j = 0$ $H_1: \beta_j \neq 0$	$\dfrac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$	Included in regression outputs

.

T tests

Example (cont.): Re-consider the Stata output

```
. regress price area rooms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.95
       Model |  579971.198     2  289985.599           Prob > F      =  0.0000
    Residual |  337883.308    85  3975.09774           R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6232
       Total |  917854.506    87  10550.0518           Root MSE      =  63.048

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   1.383606   .1489435     9.29   0.000     1.087467    1.679746
       rooms |   15.12134   9.488598     1.59   0.115    -3.744538    33.98721
       _cons |   -19.2855   31.04753    -0.62   0.536     -81.0163    42.4453
------------------------------------------------------------------------------
```

Testing several linear combinations of coefficients

Hypothesis:

Testing the **joint significance of some regressors**:

$H_0: \beta_{p+1} = \cdots = \beta_k = 0$ $\quad\left(Y_i = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip} + u_i\right)$

$H_1: \exists \beta_j \neq 0, j = p+1, \ldots k$ $\quad\left(Y_i = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip} + \beta_{p+1} X_{ip+1} + \cdots \beta_k X_{ik} + e_i\right)$

2. Testing the **global significance**:

$H_0: \beta_1 = \cdots = \beta_k = 0$ $\quad(Y_i = \beta_0)$

$H_1: \exists \beta_j \neq 0, j = 1, \ldots k$ $\quad(Y_i = \beta_0 + \beta_1 X_{i1} + \cdots \beta_k X_{ik} + e_i)$

Note:

- Equivalent to $H_0: R^2 = 0$ (absence of fit)

Testing several linear combinations of coefficients

Test statistics:

1. Testing the **joint significance of some regressors**

$$F = \frac{(R^2 - R_*^2)/m}{(1 - R^2)/(N - K - 1)} \sim F(m, N - k - 1)$$

where m is the # restrictions (# β's in H_0), R^2, R_*^2 are features of the unrestricted and restricted model, respectively

2. Testing the **global significance** $(R_*^2 = 0)$:

$$F = \frac{R^2/m}{(1 - R^2)/(N - K - 1)} \sim F(m = k, N - k - 1)$$

Rejection region: right side of F distribution

F tests

Example: testing the global significance of the regressors in

$$price_i = \beta_0 + \beta_1 area_i + \beta_2 rooms_i + +u_i$$

$H_0: \beta_1 = \beta_2 = 0$ (absence of global significance)
$H_1: \exists \beta_j \neq 0, j = 1,2$

This test statistics, as well as the respective p-value are given in the regression output (see the results in green). H_0 is rejected. The regressors are jointly significant

Alternatively
$\rightarrow R^2 = 0.6319$

$$F = \frac{0.6319/2}{(1-0.6319)/(88-2-1)} = 72.95$$

At the 5% significance level, the critical value is $F(2{,}85)$ ≃3,15, which leads to the rejection of the null hypothesis

Variation decomposition & determination coefficient

- Analysis of variance: decomposition of the total variation of Y

	SS	DF	MS
Explained	$\sum(\hat{y}_i - \bar{y})^2$	k	MSE
Residual	$\sum \hat{u}_i{}^2$	N-k-1	$MSR = \hat{\sigma}^2$
Total	$\sum(y_i - \bar{y})^2$	N-1	s_y^2

- Determination coefficient: $R^2 = \frac{SSE}{SST}$ $\quad 0 \leq R^2 \leq 1$

- Measures the proportion of the variation of Y explainned by the model

- Adjusted version for degrees of freedom: $\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1}$

- R^2 allows the comparison of models with constant and the same dependent variable, while $\bar{R}^2$ requires the same conditions, but may compare models with different k. Increasing k, necessarily increases R^2: large models are selected

F tests

Example: select one of the two following models

$$price_i = \beta_0 + \beta_1 area_i + \beta_2 rooms_i + \beta_3 garden_i + u_i$$
$$preço_i = \beta_0 + \beta_2 quartos_i + e_i$$

$H_0: \beta_1 = \beta_3 = 0$ $(under\ H_0\ the\ restricted\ model\ is\ selected)$
$H_1: \beta_1 \neq 0\ or\ \beta_3 \neq 0$
$\rightarrow R_*^2 = 0.258$
$\rightarrow R^2 = 0.672$

$$F = \frac{(0.672 - 0.258)/2}{(1 - 0.672)/(88 - 3 - 1)} = 53.06$$

At the 5% significance level, the critical value is $F(2{,}84) \simeq 3.15$. H_0 is rejected. Therefore, the unrestricted model is selected

Example (cont.):

Stata output

```
. regress price area lote rooms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   57.43
       Model |  617018.847     3  205672.949           Prob > F      =  0.0000
    Residual |  300835.658    84  3581.37688           R-squared     =  0.6722
-------------+------------------------------           Adj R-squared =  0.6605
       Total |  917854.506    87  10550.0518           Root MSE      =  59.845

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   1.322524   .1426449     9.27   0.000     1.038859    1.606189
        lote |   .0222365   .0069137     3.22   0.002     .0084878    .0359852
       rooms |    13.7864   9.015998     1.53   0.130    -4.142903     31.7157
       _cons |  -21.72645   29.47964    -0.74   0.463    -80.34994    36.89704
------------------------------------------------------------------------------

. test area lote
(…)
```

Again the t test: testing the equality of two regression coefficients

Example:

$$H_0: \beta_1 = \beta_2 \text{ is based on } t_{\hat{\delta}} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_1 - \hat{\beta}_2}} \sim t(N - k - 1)$$

In the framework of model,

$$\widehat{price}_i = \underset{}{-19.286} + \underset{(0.149)}{1.384}\ area_i + \underset{(9.489)}{15.121}\ rooms_i$$

Test whether the partial effect of area and rooms is equal

```
. quietly regress price rooms area
. test area==rooms
 ( 1)   area - rooms = 0
        F(  1,    85) =    2.06
             Prob > F =    0.1548
```

The hypothesis of the equality of both effects is not rejected

Assumptions of the MRLM and properties of the OLS estimators

Assumptions:

1. Linear model in the parameters: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i$
2. Random sample
3. $E(U|X) = 0$
4. Absence of perfect collinearity
5. Homoskedasticity: $Var(U|X) = \sigma^2$
6. Normality of u: U$\sim Normal(0, \dot{\sigma}^2)$

Properties

Small samples	Asymptotic
1-4: unbiased estimators	1-4: consistent estimators
1-5: unbiased and efficient estimators	1-5: consistent, efficient and normally distributed estimators
1-6: unbiased, efficient and normally distributed estimators	

Explanatory variables: multicolinearity

Cause: two or more regressors are excessively correlated

Problem: the estimate of $\sigma^2_{\beta_j}, j = 1, \ldots k$ is inflated because

$$\sigma^2_{\beta_j} = \frac{\sigma^2}{\left(1 - R_j^2\right)\sum_{i=1}^{N}\left(X_{ij} - \bar{X}_j\right)^2}$$

where R_j^2 is the determination coefficient of the regression of X_j on the remaining regressors. $\sigma^2_{\beta_j}$ is increased when R_j^2 increases.

Note that the OLS estimator for β is consistent

Explanatory variables: ommited or irrelevant

When selecting the set of regressors, take into account:

- **Irrelevant regressors** increase the variance (reduce efficiency) of the estimators

- **Regressor ommission** (note that ommited regresssors are in u) yields:
  - Inconsistency if $E(U|X) \neq 0$ $\rightarrow$ endogeneilty
  - Consistency if $E(U|X) = 0$ $\rightarrow$ exogeneity

Qualitative determinants

There are determinants of the variable of interest with a qualitative nature:

- House prices = f(area, rooms, location quality, existence of garden, …)
- wage = f(age, experience, gender, region, activity sector, …)

This qualitative information is coded by dummy variables, which are binary variables defined as

$$d = \begin{cases} 1 \; if \; the \; attribute \; occurs \\ 0 \; if \; the \; attrubute \; does \; not \; occur \end{cases}$$

Qualitative determinants

2 cathegories: 1 dummy

$$d = \begin{cases} 1\ if\ the\ attribute\ occurs\ (ex\colon male) \\ 0\ if\ the\ attrubute\ does\ not\ occur\ (ex\colon female) \end{cases}$$

M cathegories: M-1 dummies

$$d_1 = \begin{cases} 1\ if\ the\ attribute\ 1\ occurs\ (ex\colon south\ region) \\ 0\ if\ the\ attribute\ does\ not\ occur\ (ex\colon other\ regions) \end{cases}$$

$$d_2 = \begin{cases} 1\ if\ the\ attribute\ 2\ occurs\ (ex\colon central\ region) \\ 0\ if\ the\ attribute\ does\ not\ occur\ (ex\colon other\ regions) \end{cases}$$

In both cases the interpretation of the partial effect is made relative to the reference, which is the ommited cathegory: male, north.

Qualitative determinants

Example:

$$wage_i = \beta_0 + \beta_1 male_i + \beta_2 south_i + \beta_3 centre_i + u_i$$

Ceteris paribus

β_1: difference in wage in a man relative to a woman;

β_2: difference in wage for an individual that lives in the south relative to someone that lives in the north;

β_3: difference in wage for an individual that lives in the center relative to someone that lives in the north

Exact interpretation for log-lin models (ex: use log(wage))

$$\left(e^{\beta_j} - 1\right)100\%$$

Important for large estimates of β_j. For example

- $(e^{0.457} - 1)100\% = 57.93\%$ instead of simply 45.7%
- $(e^{-0.063} - 1)100\% = -6.1\ \%$ instead of simply -6.3%

Interaction variables

Interacton variables: result from the multiplication of a dummy and other variable

Example:

$$wage_i = \beta_0 + \beta_1 male_i + \beta_2 educ_i + \beta_3 male_i * educ_i + u_i$$

Escrevendo o modelo para cada grupo:

- Men (male=1): $wage_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) educ_i + u_i$
- Women (male=0): $wage_i = \beta_0 + \beta_2 educ_i + u_i$

Hence:

- β_0: constant term for women
- β_1: difference in the constant term of men and women
- β_2: wage variation for women for each additional education year
- β_3: difference in the previous effect for men relative to women

Chow test for a structural breack

Chow test for structural break:

- Background:
  - Two groups of individuals: G_A, G_B
  - It is susppected that the effect of the determinants on the variable of interest is different for each group

- Implementation:
  - Consider the dummy $D = \begin{cases} 1\ for\ an\ individual\ of\ G_A \\ 0 \text{ for an individual of group } G_B \end{cases}$
  - Estimate the model:
    $Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_k X_k + \gamma_0 D + \gamma_1 D X_1 + \cdots + \gamma_k D X_k + v$
  - Apply an F test for joint significance:

    H_0: $\gamma_0 = \gamma_1 = \cdots = \gamma_k = 0$ (no structural break)

    H_1: No H_0 (structural break)

Testing the functional form- RESET

Test the validity of the model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i$

F test for the joint significance of a set of artificial variables, where the unrestricted model is the original added by powers of $\hat{Y}$ (or $X\hat{\beta}$, since $\hat{Y} = X\hat{\beta}$): $\hat{Y}^2, \hat{Y}^3 \ldots$

$H_0: \gamma_1 = \cdots = \gamma_k = 0 \ (Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i) \rightarrow$ correct FF $\rightarrow R_*^2$

$H_1: n\ H_0 \ (Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma_1 \hat{Y}_i^2 + \gamma_2 \hat{Y}_i^3 \ldots + v)$ $\rightarrow$ incorrect FF$\rightarrow R^2$

Note: the rejection of H_0 requires that a new FF is considered – the model of H_1 is an artificial regression, it is not a candidate…

<u>Stata</u>

Command for 3 powers after regress

. ovtest

Testing the functional form- RESET

Example: test the functional form of

$$price_i = \beta_0 + \beta_1 area_i + \beta_2 rooms_i + u_i$$

```
regress price rooms area

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.95
       Model |  579971.198     2  289985.599           Prob > F      =  0.0000
    Residual |  337883.308    85  3975.09774           R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6232
       Total |  917854.506    87  10550.0518           Root MSE      =  63.048

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rooms |   15.12134   9.488598     1.59   0.115    -3.744538    33.98721
        area |   1.383606   .1489435     9.29   0.000     1.087467    1.679746
       _cons |   -19.2855   31.04753    -0.62   0.536     -81.0163     42.4453
------------------------------------------------------------------------------

. predict pricehat
(option xb assumed; fitted values)

. generate pricehat2=pricehat^2

. generate pricehat3=pricehat^3
```

Testing the functional form - RESET

Example (cont.):

```
. regress price rooms area pricehat2 pricehat3

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =   41.17
       Model |  610249.039     4  152562.26            Prob > F      =  0.0000
    Residual |  307605.467    83  3706.08996           R-squared     =  0.6649
-------------+------------------------------           Adj R-squared =  0.6487
       Total |  917854.506    87  10550.0518           Root MSE      =  60.878

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rooms |  -58.37904   38.71904    -1.51   0.135    -135.3897    18.63158
        area |  -5.680895   3.613211    -1.57   0.120    -12.86743    1.505637
   pricehat2 |   .0133394   .0076821     1.74   0.086    -.0019399    .0286187
   pricehat3 |  -.0000109   7.20e-06    -1.52   0.133    -.0000252    3.40e-06
       _cons |   675.0476   328.2222     2.06   0.043     22.22683    1327.868
------------------------------------------------------------------------------
```

$H_0: \gamma_1 = \gamma_2 = 0 \rightarrow$ correct FF

$$F = \frac{(0.6649 - 0.631*)/2}{(1 - 0.6649)/(88 - 5)} = 4.08$$

At the 5% level the critical value is $F(2{,}83) = 3.15$

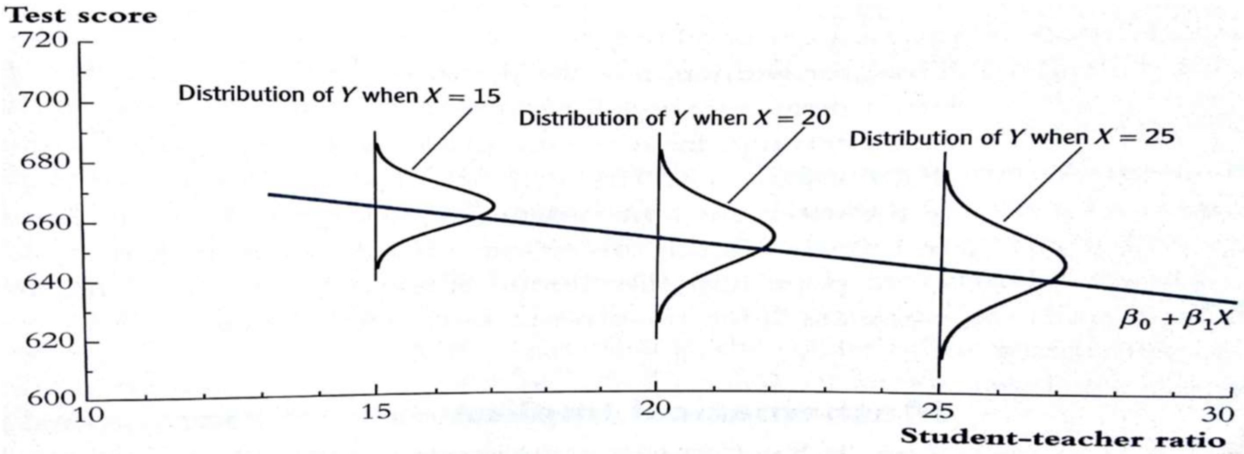
H_0 is rejected: the model functional form is rejected

Heteroskedasticity: definition

Consider the error conditional variance (skedastic function): $Var(U|X)$

Assumption (5): $Var(U|X) = \sigma^2$ (homoskedasticity)
Assumption (5) failure: $Var(U|X) = \sigma^2 h(X)$ (heteroskedasticity)

Illustration:



Assumption (5) issues efficiency and asymptotic normality to the OLS estimators for β. Its failure does not cause inconsistency or unbiadness. Thus, with heteroskedasticity, OLS estimators are unbiased and consistent, but are not efficient or asymptotically normal. Their standard variance formula is no longer valid.

Heteroskedasticity: robust estimation of the covariance matrix

Variance correction

Standard variance (assumes $Var(U|X) = \sigma^2 \mathrm{I}$)

$$Var(\hat{\beta}) = (X'X)^{-1}X'Var(U|X)X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

Robust variance (assumes $Var(U|X) = \Sigma$)

$$Var(\hat{\beta}) = (X'X)^{-1}X'Var(U|X)X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

where $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2)$

Simple implementation in software. Asymptotically valid.

<u>Stata</u>

Add the robust option after regress

```
.regress…, robust
```

Heteroskedasticity: robust estimation of the covariance matrix

Example (cont.): robust variances (estimator of coefficients is the same)

```
. regress price rooms area, robust
Linear regression                                   Number of obs =      88
                                                    F(  2,    85) =   27.22
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.6319
                                                    Root MSE      =  63.048
------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t.    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rooms |   15.12134    8.96599     1.69   0.095    -2.705452    32.94813
        area |   1.383606   .2111629     6.55   0.000     .9637578    1.803455
       _cons |   -19.2855   41.54017    -0.46   0.644    -101.8785    63.30749
------------------------------------------------------------------------------
```

$$\widehat{price}_i = -19.286 + \underset{\substack{(0.149)\\ [0.211]}}{1.384}\, area_i + \underset{\substack{(9.489)\\ [8.966]}}{15.121}\, rooms_i$$

Heteroskedasticity: tests

There are several tests, most of them based on artificial regressions where the dependente variable is $\hat{u}^2$ and implemented as global significance F test.

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i$$

H_0:Homoskedasticity→ $Var(U|X) = \sigma^2$ (use OLS)
H_1:Heteroskedasticity→ $Var(U|X) = \sigma^2 h(X)$ (use robust variances)

Heteroskedasticity: Breusch Pagan test

1. Estimate the model of interest and obtain $\hat{u}^2$
2. Estimate the auxiliary regression:

- $\hat{u}^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_k X_k + e$ and obtain $R^2_{\hat{u}^2}$

3. Test statistics:

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(N - k - 1)} \sim F(k, N - k - 1)$$

Stata

Command after regress

```
.estat hettest, rhs fstat
```

Heteroskedasticity: BP test

Example: test heteroskedastitity in model

$$ln(price_i) = \beta_0 + \beta_1 ln(area_i) + \beta_2 rooms_i + u_i$$

```
gen larea=log(area)
gen lpreço=log(price)
regress  lpreço larea rooms
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,     85) =   54.47
       Model |  4.50364223     2  2.25182112           Prob > F      =  0.0000
    Residual |  3.51396129    85  .041340721           R-squared     =  0.5617
-------------+------------------------------.          Adj R-squared =  0.5514
       Total |  8.01760352    87  .092156362           Root MSE      =  .20332
------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .8100637   .0987611     8.20   0.000     .6137002    1.006427
       rooms |   .0376464   .0303446     1.24   0.218    -.0226868    .0979795
       _cons |    1.28929   .4666125     2.76   0.007     .3615395    2.217041
------------------------------------------------------------------------------
predict uhat, resid
gen uhat2=uhat^2
```

Heteroskedasticity: BP test

Example (cont.):

```
regress uhat2 larea rooms
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =    1.93
       Model |  .026033827     2  .013016913           Prob > F      =  0.1516
    Residual |  .573679783    85  .006749174           R-squared     =  0.0434
-------------+------------------------------           Adj R-squared =  0.0209
       Total |   .59971361    87   .00689326           Root MSE      =  .08215
------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |  -.0607216   .0399045    -1.52   0.132    -.1400625    .0186193
       rooms |   .0227115   .0122608     1.85   0.067    -.0016662    .0470892
       _cons |   .2744371   .1885353     1.46   0.149    -.1004216    .6492957
------------------------------------------------------------------------------
```

H_0 is not rejected: there is evidence on the existence of homoskedasticity

Data

- N firms / individuals: $i = 1, \ldots, N$
- T observations for each individual: $t = 1, \ldots, T$

var7[19]

	cidade	ano	pop	est	renda	rendime
1	1	80	75211	15303	197	11537
2	1	90	77759	18017	342	19568
3	2	80	106743	22462	323	19841
4	2	90	141865	29769	496	31885
5	3	80	36608	11847	216	11455
6	3	90	42099	10265	351	21202
7	4	80	36640	13825	267	14682
8	4	90	46209	18173	588	29044
9	5	80	62134	8175	475	31300
10	5	90	110330	18205	925	56307
11	6	80	118550	12223	276	16291

Advantages:

- Time effects are analysed
- Efficiency gains, as the sample size increases

Cross-sectional / panel data:

- Cross-sectional: independent individuals $\Rightarrow$ independent observations
- Panel: same individuals followed trought time $\Rightarrow$ independent individuals, but for each individual observations are time dependent

Short panel:

- Sample with numerous individuals ($N \longrightarrow \infty$) but with a short time horizon (small T)
- Time dependence for the observation of each individual is allowed. Individuals are independent.

Balanced panel:

- All individuals report in all t ($T_i = T, \forall i$)

Umbalanced panel:

- Information at some moments is missing for some individuals ($T_i \neq T$)
- Major cause: some individuals decide not to provide information after some periods $\longrightarrow$ 'attrition'
- Most estimators can be used

Decomposition of the variation

- The variability of Y_{it} is decomposed into:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\bar{Y})^2=\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\bar{Y}_i+\bar{Y}_i-\bar{Y})^2$$

$$=\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\bar{Y}_i)^2+\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y})^2$$

Variability for individual i
"within variation"

Variability across
individuals
"between variation"

Models for panel data

Model:

$$Y_{it} = \alpha_i + x_{it}'\beta + u_{it} \qquad (i = 1, \dots, N; t = 1, \dots, T)$$

- α_i: individual effects, time invariant and not observed
- x_{it} - explanatory variables that may include:
  - x_{it}: characteristics that are differet across individuals and for each individual change trought time
  - x_i: observed characteristics that do not change in time
  - d_t: time *dummy* at t
  - $d_t.x_{it}$: interaction variables
- u_{it}: idiossycratic error – differs across i and t

Panel data models

Time-dummies

Aim: analyse time effects

For T years, the first year is taken as reference and $(T-1)$ dummies, one for each of the remaining years, are considered

- Ex: panel data for 2016, 2017, 2018. Two dummies are considered, D2017 and D2018 which assume the value 1 at the respective year and 0 otherwise.
  - The coefficient of D2017 informs on the variation on the mean of Y in 2017 relative to 2016, caused by other factors than the regressors.
  - The coefficient of D2018 informs on the variation on the mean of Y in 2018 relative to 2016, caused by other factors than the regressors.

Panel data models

Model $Y_{it} = \alpha_i + x'_{it}\beta + u_{it}$ may be written as

$$Y_{it} = x'_{it}\beta + (\alpha_i + u_{it})$$

where the error term has two components, α_i, u_{it}, with α_i correlated or not with the explanatory variables:

Random effects:

- Assumption: α_i and x_{it} are not correlated
- Estimators addressed here: Pooled and Random effects

Fixed effects:

- Assumption: α_i and x_{it} may be correlated
- Estimators addressed here: Fixed effects or “Within” and First diffrences

- Model:

$$Y_{it} = \alpha + x_{it}'\beta + \underbrace{(\alpha_i - \alpha + u_{it})}_{v_{it}}$$

- Estimation:
  - OLS with cluster or similar option for the variace

Stata

regress $Y X_1 \dots X_k$, vce(cluster *clustvar*)

Panel data models
Random effects estimador

Model:

$$Y_{it} = \alpha + x_{it}'\beta + (\alpha_i - \alpha + u_{it})$$

with $Var(\alpha_i) = \sigma_\alpha^2$ and $Var(u_{it}) = \sigma_u^2$

This estimator is efficient. $cor(u_{it}, u_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_u^2)$ is exploited. In contrast, the pooled OLS estimator does not exploit the panel nature of the data, apart from the variance calculation in cluster robust form.

Estimation: generalized LS

$$Y_{it} - \hat{\theta}_i\bar{Y}_i = (1 - \hat{\theta}_i)\alpha + (x_{it} - \hat{\theta}_i\bar{x}_i)'\beta + v_{it}$$

where $\hat{\theta}_i = 1 - \sqrt{\hat{\sigma}_u^2/(T_i\hat{\sigma}_\alpha^2 + \hat{\sigma}_\alpha^2)}$ and $v_{it} = (1 - \hat{\theta}_i)\alpha_i + (u_{it} - \hat{\theta}_i\bar{u}_i)$

Stata

xtreg $Y X_1 \ldots X_k$, vce(cluster *clustvar*)

Model:

$$Y_{it} - \bar{Y}_i = (x_{it} - \bar{x}_i)'\beta + (u_{it} - \bar{u}_i)$$

Estimation: OLS applied to the transformed variables with a cluster version for the variance

Stata

xtreg $Y X_1$ ... X_k, fe vce(cluster *clustvar*)

Despite the robustness, given that random effects are not required, this estimator has the disadvantage of eliminating from the model:

- All time invariant explanatory variables
- All time variant explanatory valibles that change in time by a constant: age, experience in models including a constant term.

First difference estimator

Model:

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (u_{it} - u_{i,t-1}) \Leftrightarrow$$
$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it}$$

Stata

regress *D.Y D.X*$_1$ ... *D.X*$_k$, vce(cluster *clustvar*)

Estimation: OLS applied to transformed variables, with cluster option for variance estimation

Displays the same disadvantages of the FE estimator and in fact is numerically equal to the FE estimator for T=2.

Testing whether effects are fixed or random

Hausman test:

$H_0: E(\alpha_i | x_{it}) = 0$ (RE & FE consistent, but only RE efficient)

$H_1: E(\alpha_i | x_{it}) \neq 0$ (FE consistent, RE inconsistent)

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RA})'\left[V\left(\hat{\beta}_{FE}\right) - V\left(\hat{\beta}_{RE}\right)\right]^{-1}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \sim \chi_k^2$$

Stata

(applies to models estimated without robust or cluster options)

xtreg $Y X_1 \ldots X_k$, fe
estimates store *ModelFE*
xtreg $Y X_1 \ldots X_k$
estimates store *ModelRE*
hausman *ModelFE ModelRE*

Policy analysis based on panel data, T=2

Consider a sample where individuals are observed twice (observed before and after the programme implementation) and we have individuals of two types: affected (cases / treated) and not affected (controls)

Model:

$$y_{it} = \alpha + \delta d2 + \beta prog_{it} + \alpha_i + u_{it}$$

where $prog = 1$ if treated/affected and d2=1 after the programme implementation

Model based on differences:

$$\Delta y_{it} = \delta + \beta prog_{it} + \Delta u_{it}$$

Effect of the programme: β

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), *scrap*, changes as a consequence of the participation in a training programme, (*Grant=1* if training was received), *in* 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations bellow coefficients)

$$\Delta ln(\widehat{scrap}) = -\underset{(0.097)}{0.057} - \underset{(0.164)}{0.317}\ grant, n = 54, R^2 = 0.067$$

- Trainning reduced the scrap rate in $(e^{0.317} - 1)100\% = 27.2\%$
- The scrap rate reduced in $(e^{0.057} - 1)100\% = 5.9\%$ due to factors which are not the training programme participation

•

Dynamic models for panel data: introduction

Models where the lagged dependente variable $Y_{i,t-1}$, $Y_{i,t-2}$, appears as explanatory variable

- Example – Autoregressive of order 1, $AR(1)$, model:
$$Y_{it} = \gamma_1 Y_{i,t-1} + \alpha_i + u_{it}$$

- All the estimators based on static models presented previously are inconsistent

Dynamic models for panel data: introduction

Model:

$$\Delta Y_{it} = \gamma_1 \Delta Y_{i,t-1} + \Delta x_{it}'\beta + \Delta u_{it}, t = 3, \dots, T$$

Assumption: absence of autocorrelation in $u_{it} \Longrightarrow \Delta u_{it}$ has autocorrelation of order 1:

$Cov(u_{it}, u_{i,t-1}) = 0$ so that

$$Cov(\Delta u_{it}, \Delta u_{i,t-1}) = Cov\left((u_{it} - u_{i,t-1})(u_{it-1} - u_{i,t-2})\right) = -Cov(u_{i,t-1}, u_{i,t-1}) \neq 0$$

Most well known estimadores (instrumental variable estimators):

- Anderson-Hsiao (1981)
- Arellano-Bond (1991) – 'Difference GMM'
- Blundell-Bond (1998) – 'System GMM'

Dynamic models for panel data: introduction

Anderson-Hsiao (1981):

2 types of IV:

- $Y_{i,t-2}$

Stata
ivregress gmm $D.Y(DL.Y = L2.Y)\ D.X_1\ ...\ D.X_k$

- $\Delta Y_{i,t-2}$ (1 observation is lost, but yields is general more efficient estimators)

Stata
xtivreg $\mathrm{D}.Y(\mathrm{DL}.Y = \mathrm{DL2}.Y)\ \mathrm{D}.X_1\ ...\ \mathrm{D}.X_k$
or
xtivreg $Y(\mathrm{L}.Y = \mathrm{L2}.Y)\ X_1\ ...\ X_k$, fd

Dynamic models for panel data: introduction

Arellano-Bond (1991):

- Suggests the use of all lags of $Y_{i,t}$ as IV:
  - $t = 3: Y_{i,1}$
  - $t = 4: Y_{i,2}, Y_{i,1}$
  - …
  - $t = T: Y_{i,T-2}, \dots, Y_{i,2}, Y_{i,1}$
- # VI= $(T-1)(T-2)/2$
- One may decide to use only part of the available IV
- More efficient than Anderson-Hsiao's (1981) estimators

Stata
xtabond $Y X_1 \dots X_k$, maxldep(#) twostep vce(robust)

Blundell-Bond (1998):

- IV suggested: $\Delta Y_{i,2},\ldots,\Delta Y_{i,T-1}$
- # VI= $\frac{(T-1)(T-2)}{2} + (T-2)$
- One may decide to use only part of the available IV
- More efficient than Arellano-Bond's (1991) estimators but requires additional assumptions

Stata
xtdpdsys $Y X_1 \ldots X_k$, maxldep(#) twostep vce(robust)

Dynamic models for panel data: introduction

Relevant tests:

- Hansen's J test of overidentification

> Stata
> (after xtabond or xtdpdsys, with variance estimated in a standard way)
> estat sargan

- Autocorrelation test

> Stata
> (after xtabond or xtdpdsys)
> estat abond, artests(3)