Quantitative Data Analysis

Master in Management - Academic Year 2023/2024

Ricardo Moura

ISEG

March 11, 2024

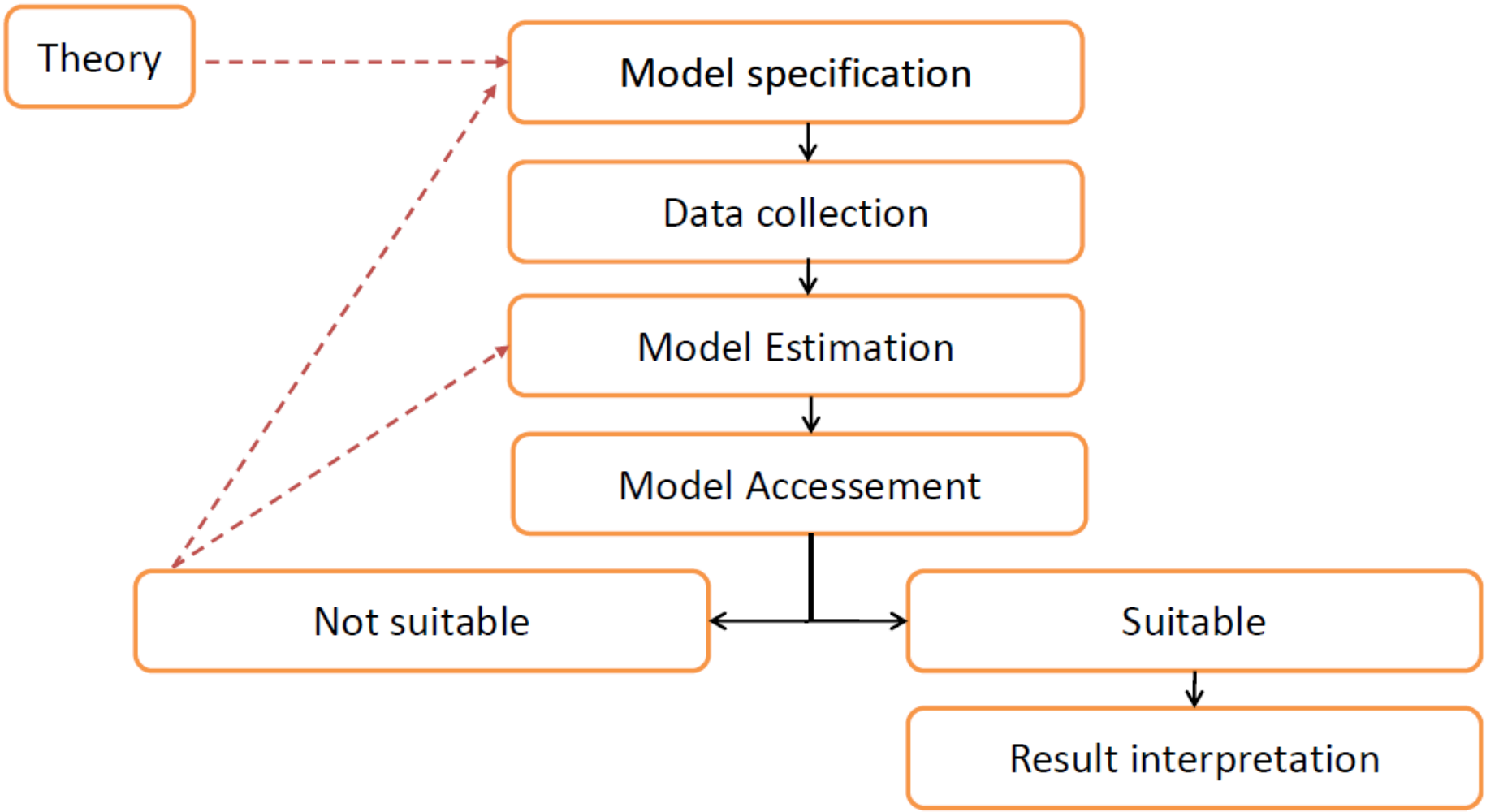
Econometrics

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



Methodology

1 - Problem definition

2 - Specicify Econometric Model, select variables

3 - Gather data and cross data, and preprocess data

4 - Estimate model (Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), Generalized Method of Moments (GMM)). Estimate parameters.

5 - Model Validation - hypothesis testing, check model assumptions. Check multicolinearity, heteroscedasticity, autocorrelation and model misspecification.

6 - Model refinement - go to 4 or 2.

7 - Result Interpretation and if needed forecast and prediction

Causality & *ceteris paribus* analysis

One of the **major aims** is analyse 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 linear regression model

Aim: Explain $E(Y|\mathbf{X})$

Y - dependent variable/response variable/variable of interest

$\mathbf{X}$ - explanatory variables/determinants/regressors/covariates

$E(Y|\mathbf{X})$ - expected value/ conditional mean of Y given $\mathbf{X}$

$E(Y|\mathbf{X})$ is a function of parameters $\boldsymbol{\beta}$ that need to be estimated

Specification Multi-LR model

Sometimes refered to as MLR model (do not confuse with Multivariate Linear Regression Model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i, i = 1, \ldots, n$$

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Where:

- u: error term associated to all the determinants of Y that were not included in $\mathbf{X}$
- β: $p + 1$ coefficients that need to be estimated
- p: number of explanatory variables

OLS estimation

The real value of $E(y|\mathbf{x})$ can be estimated by the **estimator** $\hat{y}$ and the **predicted values** can be calculated by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$.
Thus, the **residuals** are given by $\hat{\mathbf{u}}$ with values given by

$$\hat{u}_i = y_i - \hat{y}_i.$$

Model Estimation By Ordinary Least Squares (**OLS**)

Estimate β by minimizing the sum of the square of the residuals, *i.e.*,

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \hat{u}_i^2 \right\}$$

Interpretation

Generalizing, β_j, measures the impact on $E(Y|\mathbf{X})$, due to a variation of the determinant X_j associated, when all the rest determinants are left equal.

Considering no variable transformation, we have a linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i$$

and the partial effects are

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

that is, for every unit of variation in X_j there is an effect of β_j units in the conditional mean of Y given $\mathbf{X}$, assuming all the rest of the determinants stay constant.

Interpretation

Linear model in the parameters:

$$y_i^* = \beta_0 + \beta_1 x_{i1}^* + \cdots + \beta_p x_{ip}^* + u_i$$

Example: Suppose we have a regression model where the dependent variable is the natural logarithm of income (ln(income)), and we're interested in how years of education (x) affect income. The regression equation might look something like this:

$$\ln(\text{income}) = 1 + 0.1(\text{years of education}) + u$$

For years of education $= 10$ and for 11, we have $\ln(\text{income}_1) = 2$ and $\ln(\text{income}_2) = 2.1$, thus
$\ln(\text{income}_1) - \ln(\text{income}_2) = 0.1 \Leftrightarrow \frac{\ln(\text{income}_1)}{\ln(\text{income}_2)} = 0.1 \Leftrightarrow \frac{\text{income}_1}{\text{income}_2} = e^{0.1} \approx 1.1$
That means that the variation is in fact $0.1 \times 100 = 10\%$

Interpretation

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

$$\text{price}_i = \beta_0 + \beta_1 \text{km}_i + \beta_2 \text{cc}_i + u_i$$

where the price of a used car, in euros, depends on mileage, measured in (km), engine capacity, measured in cubic centimeters (cc).

Estimated model:

$$\widehat{\text{price}}_i = 10036.77 - 0.065\text{km}_i + 6.148\text{cc}_i$$

Assuming everything else constant:

- for an **unitary** variation (increase) of km, that is for each additional km, in average the car price **decreases** 0.065 euros
- for an **unitary** variation (increase) of cc, that is for each additional cubic centimetre, in average the car price **increases** 6.148 euros

Interpretation

Output in Stata

```
regress price km cc
```

Source	SS	df	MS	Number of obs	=	82
				F(2, 79)	=	27.64
Model	2.2335e+09	2	1.1167e+09	Prob > F	=	0.0000
Residual	3.1915e+09	79	40398780.1	R-squared	=	0.4117
				Adj R-squared	=	0.3968
Total	5.4250e+09	81	66975238.5	Root MSE	=	6356

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
km	-.0654935	.0098596	-6.64	0.000	-.0851184	-.0458685
cc	6.148159	1.133155	5.43	0.000	3.892671	8.403648
_cons	10036.77	2038.148	4.92	0.000	5979.936	14093.6

Interpretation

Example: consider the alternative model

$$\widehat{ln(\text{price}_i)} = 12.002 - 0.308 ln(\text{km})_i + 0.0004\text{cc}_i$$

Assuming everything else constant:

- An increase of **1%** in km, is estimated to generate a **decrease** in the average price of **0.308%**
- For each **additional** cc, the average price is expected to increase **0.0004*100=0.04%**

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

Interpretation

Output

```
. gen lprice=ln(price)
. gen lkm=ln(km)

. regress lprice lkm cc
```

Source	SS	df	MS	Number of obs	=	82
				F(2, 79)	=	18.49
Model	12.4785888	2	6.2392944	Prob > F	=	0.0000
Residual	26.6615738	79	.337488276	R-squared	=	0.3188
				Adj R-squared	=	0.3016
Total	39.1401626	81	.483211884	Root MSE	=	.58094

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lkm	-.3084663	.0563883	-5.47	0.000	-.4207044	-.1962283
cc	.0004137	.0001013	4.08	0.000	.0002121	.0006153
_cons	12.00184	.6192318	19.38	0.000	10.76929	13.23439

OLS estimator for β_j

In this model it is considered that

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\sigma_{\beta_j}} \sim N(0,1)$$

Since, σ_{β_j} is unknown, we may use its estimator $\hat{\sigma}_{\beta_j}$ and thus:

$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \sim t_{n-p-1}$$

Therefore the CI for β_j is

$$\left(\hat{\beta}_j - t_{n-p-1;\alpha/2}\hat{\sigma}_{\beta_j}; \hat{\beta}_j + t_{n-p-1;\alpha/2}\hat{\sigma}_{\beta_j}\right)$$

for $(1-\alpha) \times 100\%$ confidence level.

T test:Main Version

We usually test the individual significance of x_j, by using

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$$

resuming the test to

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$$

which is included in the Stata output

```
regress price km cc

...
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0654935   .0098596    -6.64   0.000    -.0851184   -.0458685
          cc |   6.148159   1.133155     5.43   0.000     3.892671    8.403648
       _cons |   10036.77   2038.148     4.92   0.000     5979.936     14093.6
------------------------------------------------------------------------------
```

Testing linear combinations of coefficients

1. Testing the joint significance of **some regressors** ($k < p$):

$$H_0 : \beta_k + 1 = \cdots = \beta_p = 0 \quad vs. \quad H_1 : \exists \beta_j \neq 0, j = k+1, \ldots, p$$

Here, H_0 is the same as "is the model defined solely by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i"$$

2. Testing the **global significance**:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad vs. \quad H_1 : \exists \beta_j \neq 0, j = 1, \ldots, p$$

Here, H_0 is the same as "is the model defined solely by:

$$y_i = \beta_0"$$

and it is equivalent to $H_0 : R^2 = 0$ which represents the **absence of fit**

Testing linear combinations of coefficients

1. Testing the joint significance of some regressors

$$F = \frac{\frac{R^2 - R_*^2}{m}}{\frac{1-R^2}{n-p-1}} \sim F_{m,n-p-1}$$

where m is the number of restrictions in H_0, that is, the number of $\beta's$ equal to zero in H_0, and R^2 and R_*^2 are features of the unrestricted and restricted model, respectively.

2. Testing the global significance:

$$F = \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}} \sim F_{p,n-p-1}$$

Rejection Region: Right side of F distribution

Testing linear combinations of coefficients

Example: Consider the model

$$\text{price}_i = \beta_0 + \beta_1 \text{km}_i + \beta_2 \text{cc}_i + u_i$$

Let us test $H_0 : \beta_1 = \beta_2 = 0$, that is, the absence of **global significance**. Since, $R^2 = 0.4117$, we have

$$F = \frac{0.4117/2}{(1 - 0.4117)/(82 - 2 - 1)} = 27.64$$

Since, $f_{2,79;0.05} \approx 3.11$, we conclude that we must **reject** H_0. Thus we have **statistical evidences** that the regressors are **jointly significant**.

This result can be seen in the Stata output.

Testing linear combinations of coefficients

Example: Select one of the two models

$$\text{price}_i = \beta_0 + \beta_1 \text{km}_i + \beta_2 \text{cc}_i + \beta_3 \text{power}_i + u_i$$

or

$$\text{price}_i = \beta_0 + \beta_2 \text{cc}_i + e_i \text{ (restricted model)}$$

Test: $H_0 : \beta_1 = \beta_3 = 0$, under H_0 we are **selecting** the **restricted model**
Collect, $R^2_* = 0.0871$ and $R^2 = 0.6276$
Calculate

$$F = \frac{(0.6276 - 0.0871)/2}{(1 - 0.6276)/(82 - 3 - 1)} = 57.01$$

Since, $f_{2,79;0.05} \approx 3.11$, we conclude that we must **reject** H_0. Thus, we have statistical information to select the **unrestricted** model.

This result can be seen in the Stata output.

Example (cont.):

Stata output

```
. regress price km cc power
      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(3, 78)        =     43.81
       Model |  3.4046e+09         3  1.1349e+09   Prob > F        =    0.0000
    Residual |  2.0204e+09        78  25902991.4   R-squared       =    0.6276
-------------+----------------------------------   Adj R-squared   =    0.6132
       Total |  5.4250e+09        81  66975238.5   Root MSE        =    5089.5
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t.    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0632483    .007902    -8.00   0.000    -.0789799   -.0475166
          cc |  -.1601691   1.305195    -0.12   0.903    -2.758612    2.438274
       power |   94.70915    14.0856     6.72   0.000     66.66687    122.7514
       _cons |   9190.395   1636.871     5.61   0.000     5931.636    12449.16
------------------------------------------------------------------------------

. test km power
 ( 1)  km = 0
 ( 2)  power = 0

       F(  2,     78) =   57.01
            Prob > F =    0.0000
```

Variation Decomposition & determination coefficient

Source of Variation	SS	df	MS
Explained (Regression)	$SSR = \sum(\hat{y}_i - \bar{y})^2$	p	$MSR = \frac{SSR}{p}$
Residual (Error)	$SSE = \sum \hat{u}_i^2$	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$
Total	$SST = \sum(y_i - \bar{y})^2$	$n-1$	

Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \Rightarrow 0 \leq R^2 \leq 1$$

Measures the proportion/percentage of the variation of y explained by the model

Adjusted Coefficient of Determination

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

Provides better comparison between muliple regression models with different number of explanatory variables, p.

Variation Decomposition & determination coefficient

The R^2 can be used to compare models with constant and the same explanatory variables.

If p is increased, R^2 also increases, and it does not mean that the model with more variables is better. If a decision is made based on R^2, of course that larger models are going to be selected, even if it is a wrong option.

High R^2 can be produced if there is small SSE (ideally) but also if there is large SST

From the same y_i, larger R^2 models provide better explanation of Y.

"Generally, experienced analysts have found that R^2 is 0.80 or above for models based on time-series data. Cross-section data models (e.g., cities, states, firms) have values in the 0.40 to 0.60 range, and models based on data from individual people often have R^2 values in the 0.10 to 0.20 range." Newbold, 8th edition

Testing the equality of two regression coefficients

Example:

$$H_0 : \beta_1 = \beta_2 \quad \textit{versus} \quad H_0 : \beta_1 \neq \beta_2$$

Based on

$$t_{\hat{\delta}} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_1 - \hat{\beta}_2}} \sim t_{n-p-1}$$

for $\hat{\delta} = \hat{\beta}_1 - \hat{\beta}_2$

Test whether the partial effect of km and cc is equal

```
. quietly regress price km cc
. test km=cc
 ( 1)  km - cc = 0
       F(  1,    79) =   29.89
            Prob > F =    0.0000
```

We reject that both effects are equal

Assumptions of the Multi-LR

Assumptions:

1. Linear model in the parameters
2. The sample was collected randomly
3. Given the values of **x** the expected value of the residuals is zero
4. Absence of colinearity(multicolinearity) ($X_1 = \lambda X_2$ for some λ)
5. Homoskedasticity: Uniform variance of the residuals.
6. Normality of U, i.e., $U \sim N(0, \sigma^2)$
   Properties: 1-4 respect assumptions (small sample: unbiased estimators; Asymptotic: consistent)

   1-5 respect assumptions (small sample: unbiased and efficient; Asymptotic:consistent, efficient abd normally distributed)

   1-6 respect assumptions (small samples:unbiased, efficient and normally distributed estimators)

Multicolinearity

If two or more regressors are excessively correlated, the estimate of variance of β_j will be inflated.
If they are highly correlated, we cannot identify which of the variables is causing the change in the response variable.

When selecting the set of regressors, take into account:

Irrelevant/redundant regressors
Regressor ommission - ommited regressors will be in the residuals
If they are relevant and are ommited, $E(u|x) \neq 0$ leading to endogeneity (inconsistency)
If they are irrelevant and are ommited, $E(u|x) = 0$ leading to exogeneity (consistency)

Qualitative regressors

There are regressors 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(indicator) variables, which are binary variables defined as

$$d = \begin{cases} 1, \text{if the attribute occurs} \\ 0, \text{otherwise} \end{cases}$$

2 categories leads to 1 dummy

M categories leads M-1 dummies

The interpretation of the partial effects is relative to the omitted category (the reference category)

Qualitative regressors

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

Ceteris Paribus
β_1: the difference in wage of a man relative to a woman;
β_2: a difference in wage of someone that lives in south relative to someone that lives in the north
β_3: a difference in wage of someone that lives in center relative to someone that lives in the north

When using log-linear models ($log(wage_i)$) the difference is calculated as $(e^{\hat{\beta}_j} - e^0)100\%$ - mainly for large values of $\hat{\beta}_j$

β_1: the change of $(e^{\hat{\beta}_1} - 1)100\%$ in the wage in a man relative to a woman (not $\hat{\beta}_1 \times 100\%$)

Interaction variables

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

Example: $wage_i = \beta_0 + \beta_1 male_i + \beta_2 educ_i + \beta_3 \mathbf{male_i} * \mathbf{educ_i} + u_i$ We will have different models for each group:

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

Structural Break

When we have to groups G_A and G_B, we might suspect that the effect of the regressors on the response variable is different for each group

Perform a **Chow test**:

Consider dummy $D = \begin{cases} 1, \text{if it is from } G_A \\ 0, \text{if it is from } G_B \end{cases}$

Estimate the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma_0 + \gamma_1 d_i x_{i1} + \cdots + \gamma_p d_i x_{ip} + \omega_i$$

Apply a F test for the joint significance of γ's:

$$H_0 : \gamma_0 = \gamma_1 = \cdots = \gamma_p = 0 \quad \textit{versus} \quad H_1 : \textit{No } H_0$$

that is

$$H_0 : \textit{No Structural break} \quad \textit{versus} \quad H_1 : \textit{Structural break}$$

H_0 not rejected the model: is reduced to $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i$

The Regression Equation Specification Error Test (RESET)

Diagnostic test used to test for model misspecification, particularly concerning the functional form of the model. The objective is to test the joint the significance of a set of artificial variables.

First we consider the model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i$ and then we consider the alternative model, based in powers of $\hat{Y}$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \cdots + v_i$$

Then we test

$$H_0 : \gamma_1 = \cdots = \gamma_p = 0 \quad \textit{versus} \quad H_1 : \textit{No } H_0$$

If we do not reject H_0, the first functional form is note rejected. If we reject, we should consider a new functional form (the one given is artificial and not a candidate)

Stata: `ovtest` - test to consider 3 powers used after `regress`

The Regression Equation Specification Error Test (RESET)

If H_0 is not rejected, it suggests that there's no evidence of misspecification, meaning the model may be appropriately specified. But there is no guarantee that all relevant variables are included. The test did not detect specific types of misspecification.

Otherwise, H_0 is rejected, it indicates potential misspecification of the model. Relevant variables may have been omitted, the functional form of the model may be incorrect (it might have to include quadratic or interaction terms), or there are other reasons. Rejecting the null hypothesis in the RESET test specifically suggests that adding polynomial or interaction terms of the predicted values may help to explain the dependent variable, implying the original model may be missing important predictors or relationships.

RESET example

```
. regress price km cc

      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(2, 79)        =     27.64
       Model |  2.2335e+09         2  1.1167e+09   Prob > F        =    0.0000
    Residual |  3.1915e+09        79  40398780.1   R-squared       =    0.4117
-------------+----------------------------------   Adj R-squared   =    0.3968
       Total |  5.4250e+09        81  66975238.5   Root MSE        =      6356
                                                  .
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0654935   .0098596    -6.64   0.000    -.0851184   -.0458685
          cc |   6.148159   1.133155     5.43   0.000     3.892671    8.403648
       _cons |   10036.77   2038.148     4.92   0.000     5979.936     14093.6
------------------------------------------------------------------------------
. ovtest
Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables
                  F(3, 76) =      0.60
                  Prob > F =      0.6189
```

We do not reject H_0 thus the model's functional form is supported by the data

$$price_i = \beta_0 + \beta_1 km_i + \beta_2 cc_i + u_i$$

RESET example one power

```
. regress price km cc

      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(2, 79)        =     27.64
       Model |  2.2335e+09         2  1.1167e+09   Prob > F        =    0.0000
    Residual |  3.1915e+09        79  40398780.1   R-squared       =    0.4117
-------------+----------------------------------   Adj R-squared   =    0.3968
       Total |  5.4250e+09        81  66975238.5   Root MSE        =      6356
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0654935   .0098596    -6.64   0.000    -.0851184   -.0458685
          cc |   6.148159   1.133155     5.43   0.000     3.892671    8.403648
       _cons |   10036.77   2038.148     4.92   0.000     5979.936    14093.6
------------------------------------------------------------------------------
. predict pricehat
(option xb assumed; fitted values)

. generate pricehat2=pricehat^2
(2 missing values generated)
```

RESET example one power

```
. regress price km cc pricehat2

      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(3, 78)        =     18.26
       Model |  2.2382e+09         3   746073837   Prob > F        =    0.0000
    Residual |  3.1868e+09        78  40856061.7   R-squared       =    0.4126
-------------+----------------------------------   Adj R-squared   =    0.3900
       Total |  5.4250e+09        81  66975238.5   Root MSE        =    6391.9

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0750769   .0298575    -2.51   0.014    -.1345187   -.0156351
          cc |   7.246896   3.424082     2.12   0.037     .4300728    14.06372
   pricehat2 |  -5.49e-06   .0000161    -0.34   0.735    -.0000376    .0000266
       _cons |   10343.66   2239.294     4.62   0.000     5885.567    14801.75
------------------------------------------------------------------------------
```

We do not reject H_0 thus the model's functional form is supported by the data

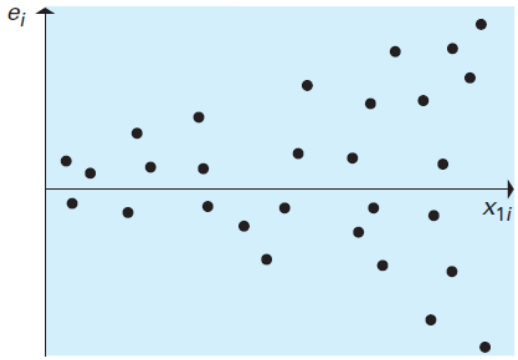
$$price_i = \beta_0 + \beta_1 km_i + \beta_2 cc_i + u_i$$

Heteroskedasticity

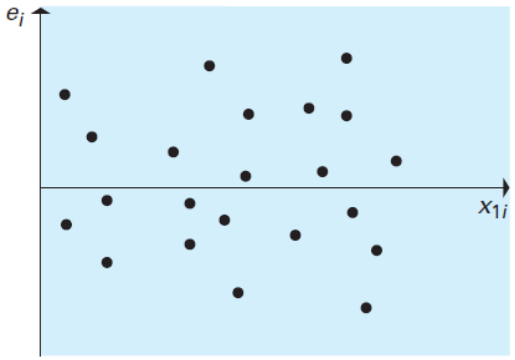
Consider the conditional variance of the error (skedastic function)

$$Var(u|\mathbf{x})$$

If assumption (5) is verified we have $Var(u|\mathbf{x}) = \sigma^2$ (**homoskedasticity**), if not $Var(u|\mathbf{x}) = \sigma^2 h(x)$ (**heteroskedasticity**)



(a) Heterosecedasticity



(b) No Apparent Heterosecedasticity

If homoskedasticity is met, then the OLS estimators of β will be efficient and asymptotically normal. If not, they will not be efficient neither asymptotically normal, but still unbiased and consistent.(Their standard variance formula is no longer valid)

Heteroskedasticity

Homoskedasticity - We may assume $Var(U|\mathbf{X}) = \sigma^2 \mathbf{I}_n$ and the standard variance is

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

Heteroskedasticity - We have to correct the variance, $Var(U|\mathbf{X}) = \mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$, and present the robust variance

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

Thankfully, in stata, we just have to incorporate
`regress ..., robust`

Heteroskedasticity

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

$$y_i = \beta_0 + \beta_1 x_{i1} + dots + \beta_p x_{ip} + u_i$$

$$H_0 : Var(U|\mathbf{X}) = \sigma^2 \quad versus \quad Var(U|\mathbf{X}) = \sigma^2 h(\mathbf{X})$$

Not rejecting H_0 we may use OLS variances.
Rejecting we must use robust variances.

Breusch Pagan test

1. Estimate the model of interest
2. Obtain $\hat{u}^2$ (remember $\hat{u}_i^2 = (y_i - \hat{y}_i)^2$)
3. Estimate auxiliary regression

$$\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_p x_p + e$$

4. Obtain $R^2_{\hat{u}^2}$
5. Use test statistic:

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

Reject for large values of F.
Stata: `estat hettest, rhs fstat`
`hettest` (heteroskedasticity test), `rhs` (right hand side to be performed in the explanatory variables) and `fstat` (F statistic version droping normal assumption)

BP example

```
. regress price km cc
      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(2, 79)        =     27.64
       Model |  2.2335e+09         2  1.1167e+09   Prob > F        =    0.0000
    Residual |  3.1915e+09        79  40398780.1   R-squared       =    0.4117
-------------+----------------------------------   Adj R-squared   =    0.3968
       Total |  5.4250e+09        81  66975238.5   Root MSE        =      6356
                                                .
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0654935   .0098596    -6.64   0.000    -.0851184   -.0458685
          cc |   6.148159   1.133155     5.43   0.000     3.892671    8.403648
       _cons |   10036.77   2038.148     4.92   0.000     5979.936     14093.6
------------------------------------------------------------------------------
. estat hettest, rhs fstat
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: km cc
         F(2 , 79)     =     3.34
         Prob > F      =   0.0407
```

BP example without hettest

```
 regress price km cc
      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(2, 79)        =     27.64
       Model |  2.2335e+09         2  1.1167e+09   Prob > F        =    0.0000
    Residual |  3.1915e+09        79  40398780.1   R-squared       =    0.4117
-------------+----------------------------------   Adj R-squared   =    0.3968
       Total |  5.4250e+09        81  66975238.5   Root MSE        =      6356
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -.0654935   .0098596    -6.64   0.000    -.0851184   -.0458685
          cc |   6.148159   1.133155     5.43   0.000     3.892671    8.403648
       _cons |   10036.77   2038.148     4.92   0.000     5979.936    14093.6
------------------------------------------------------------------------------
. predict uhat, resid
. gen uhat2=uhat^2
```

$\hat{u} = pricce - \widehat{price}$

BP example without `hettest`

```
. regress uhat2 km cc
      Source |       SS           df       MS      Number of obs   =        82
-------------+----------------------------------   F(2, 79)        =      3.34
       Model |  5.7968e+16         2  2.8984e+16   Prob > F        =    0.0407
    Residual |  6.8639e+17        79  8.6885e+15   R-squared       =    0.0779
-------------+----------------------------------   Adj R-squared   =    0.0545
       Total |  7.4436e+17        81  9.1896e+15   Root MSE        =    9.3e+07

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          km |  -326.1195   144.5923    -2.26   0.027    -613.9233   -38.31567
          cc |   32557.12   16617.94     1.96   0.054    -520.0646    65634.31
       _cons |   2.41e+07   2.99e+07     0.81   0.422    -3.54e+07    8.36e+07
------------------------------------------------------------------------------
```

Since p-value is smaller that 0.05, we reject homoskedasticity assumption for a significance level of 5%, thus having evidence of the existence of heteroskedasticity

Normality

The issue of residual normality in Multiple Linear Regression models is often underestimated or not widely discussed in econometrics.

Focus in large samples (CLT)

Robustness of Linear Regression - Even if the assumption is not met, the model can still be considered valid because it is robust.