Forecasting Methods / Métodos de Previsão

Week 6

ISCTE - IUL, Gestão, Econ, Fin, Contab.

Diana Aldea Mendes

diana.mendes@iscte.pt

March 17, 2011

Multiple regression

- **S**uppose we have: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Problems:
  - Even if straight line relationship were true, we would never get all points on an XY-plot lying precisely on it due to measurement error
  - True relationship probably more complicated, straight line may just be an approximation
  - Important variables which affect Y may be omitted.
- Solutions:
  - Multiple regression (same as simple regression except many independent (explanatory) variables)
  - Nonlinear models (quadratic, log-log, lin-log,...)

Example

Sales-advertising equation can be extended to include variables such as consumers income, price and the price and advertising of competitors' products

Multiple regression

- General case: express a k variable regression model as a series of equations (k equations condensed into a matrix form)

$$y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_k x_{k(i)} + \varepsilon_i, \quad i = 1, ..., n$$

  - β_0 is still the intercept
  - β_1 to β_k all called slope parameters (partial regression slope coefficients)
  - ε_i is the error term (or disturbance), with zero mean and constant variance
  - The levels of these variables for the ith case are labeled $x_{1(i)}, x_{2(i)}, ..., x_{k(i)}$

- Total variability in dependent variable $Y =$ Variability explained by the explanatory variables (X_i) in the regression + Variability that cannot be explained and is left as an error (ε).

Multiple regression

- The dependent variable y has n random observations and can be written as a system of linear equations:

$$\left\{ \begin{array}{c} y_1 = \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + ... + \beta_k x_{k(1)} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + ... + \beta_k x_{k(2)} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + ... + \beta_k x_{k(n)} + \varepsilon_n \end{array} \right.$$

Multiple regression

- Or, in matrix notation

$$Y_{(n\times 1)} = X_{(n\times(k+1))}\beta_{((k+1)\times 1)} + \varepsilon_{(n\times 1)}$$

$$\begin{aligned} Y_{(n\times 1)} &= \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},\ X_{(n\times(k+1))} = \begin{bmatrix} 1 & x_{1(1)} & \dots & x_{k(1)} \\ 1 & x_{1(2)} & \dots & x_{k(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & \dots & x_{k(n)} \end{bmatrix} \\ \beta_{((k+1)\times 1)} &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},\ \varepsilon_{(n\times 1)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \end{aligned}$$

Multiple regression

where

- $Y-$ is the column vector (type $(n \times 1)$) of observations for dependent (response) variable
- $X-$ matrix that portrays the n observations on k independent variables $x_1, ..., x_k$, and the first column of 1's represents the intercept term (type $n \times (k+1)$) (each column represents an independent variable)
- $\beta-$ column vector of unknown parameters (type $k+1$)
- ε column vector of error terms (type $(n \times 1)$
- The sum of the squared residuals is given by

$$SSE = U^T U, \text{ where } \; U = Y - \widehat{Y}$$

Multiple regression

- Multiple Regression analysis uses data $(x_1, ..., x_k$ and y) to make a guess or estimate of what $\beta_0, ..., \beta_k$ are.
- $\hat{\beta}_i$ is the **marginal effect** of x_i on y. It is a measure of how much the explanatory variable x_i influences the dependent variable (measure of how much y tends to change when x_i is changed by one unit)
- In order to obtain the estimates $\hat{\beta}_i$ we need to differentiate

$$SSE = U^T U = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1(i)} - \hat{\beta}_2 x_{2(i)} - ... - \hat{\beta}_k x_{k(i)} \right)^2$$

with respect to the unknowns β_i. Yields $k+1$ simultaneous equations in $k+1$ unknowns (Normal Equations) (OLS), that is

$$\frac{\partial (SSE)}{\partial \beta_j} = 0, \ j = 0, ..., k \quad \text{or} \quad X^T X \beta = X^T Y$$

Multiple regression

- If matrix X has rank $(k+1)$, then $X^T X$ is invertible, and the system has a unique solution. This solution corresponds to the estimates of $\hat{\beta}_i$ and it is given by
$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y$$
- Computer packages will calculate OLS estimates.

Multiple regression

- The standard errors (SE) of the estimated coefficients are:

$$\sigma^2 \longrightarrow s^2 = \frac{SSE}{GL} = \frac{E^T E}{n - (k+1)}$$

and the variance of $\widehat{\beta}_i$ are given by the diagonal elements of the variance-covariance matrix, that is,

$$s^2 \left(X^T X \right)^{-1}$$

Multiple regression

Example

For $k = 2$ and 15 observation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

we obtain the following normal system

$$\begin{aligned} \left(X^T X\right)^{-1} &= \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, \left(X^T Y\right) = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}, \\ SSE &= E^T E = 10.96 \end{aligned}$$

Multiple regression

the β coefficients are then given by:

$$\left(X^TX\right)^{-1}\left(X^TY\right)=\begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}\begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}=\begin{bmatrix} 1.1 \\ -4.4 \\ 19.88 \end{bmatrix}$$

and the standard errors are

$$s^2=\frac{SSE}{n-(p+1)}=\frac{10.96}{15-3}=0.91$$

Variance-covariance matrix

$$s^2\left(X^TX\right)^{-1}=0.91\left(X^TX\right)^{-1}=\begin{bmatrix} 1.83 & 3.20 & -0.91 \\ 3.20 & 0.91 & 5.94 \\ -0.91 & 5.94 & 3.93 \end{bmatrix},$$

Multiple regression

$$
\begin{aligned}
Var\left(\widehat{\beta}_0\right) &= 1.83 \rightarrow SE\left(\widehat{\beta}_0\right) = 1.35 \\
Var\left(\widehat{\beta}_1\right) &= 0.91 \rightarrow SE\left(\widehat{\beta}_1\right) = 0.96 \\
Var\left(\widehat{\beta}_2\right) &= 3.93 \rightarrow SE\left(\widehat{\beta}_2\right) = 1.98
\end{aligned}
$$

Finally we have the model

$$y = \underset{(1.35)}{1.10} - \underset{(0.96)}{4.40} x_{1(i)} + \underset{(1.98)}{19.88} x_{2(i)}$$

Multiple regression

Some particular cases:

- k independent variables (regression model of order k)

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k}_{k \text{ indep. var}} + \varepsilon$$

- 1 independent variable x with several powers

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$$

- Interaction models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

Multiple regression

- **Example: first order model**

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\
E(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2
\end{aligned}
$$

where

  - β_0 is the intersection with the $yy - axis$ (the value of $E(y)$ when $x_1 = x_2 = 0$)
  - β_1 : change in $E(y)$ when x_1 increase with 1 unit and x_2 is fixed
  - β_2 : change in $E(y)$ when x_2 increase with 1 unit and x_1 is fixed

Multiple regression

Example

Explaining House Prices: Data on N = 546 houses sold in Windsor, Canada. Dependent variable, y , is the sales price of the house in Canadian dollars. Four explanatory variables:

- x_1= *the lot size of the property (in square feet)*
- x_2 = *the number of bedrooms*
- x_3 = *the number of bathrooms*
- x_4 = *the number of storeys (excluding the basement).*

Multiple regression

- Statistical Aspects of Multiple Regression
  - Largely the same as for simple regression (the same residual assumptions, the same output interpretation).
- New hypothesis test (from R^2) (F-test for the Overall Model)

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{versus} \\ H_1 &: \text{ at least one } \beta \text{ is not zero} \end{aligned}$$

- It is a global test in order to conclude about the model utility (used to test whether any of the independent variables are linearly associated with y)
- The test statistics (F-statistics) is given by

$$\text{Test statistic: } F = \frac{\frac{SS_{yy} - SSE}{k}}{\frac{SSE}{n-(k+1)}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}} = \frac{\text{mean square (model)}}{\text{mean square (error)}}$$

Multiple regression

- If the $p-$value is less that the significance level, reject the null
- If we reject the null, that means that the chosen model it is adequate and can be applied to our purpose. Does not imply that the model is the best model.
- Conclusion: the Test F permits to conclude that some of the independent variables are important for the regression model (but we don't know exactly which ones)
- The Test t permits to select the significant independent variables.

Multiple regression

Plots

- plot of ε_i (u_i) vs $\hat{y}_i$ Can be used to check for linear relation, constant variance
  - If relation is nonlinear, U-shaped pattern appears
  - If error variance is non constant, funnel shaped pattern appears
  - If assumptions are met, random cloud of points appears
- Plot of ε_i (u_i) vs $x_{j(i)}$ for each j. Can be used to check for linear relation with respect to x_j
  - If relation is nonlinear, U-shaped pattern appears
  - If assumptions are met, random cloud of points appears

Multiple regression

- Plot of ε_i (u_i) vs i. Can be used to check for independence when collected over time
  - If errors are dependent, smooth pattern will appear
  - If errors are independent, random cloud of points appears
- Histogram of ε_i (u_i)
  - If distribution is normal, histogram of residuals will be mound-shaped, around 0

Multiple regression

- **Omitted Variable Bias**: "Omitted variable bias" is a statistical term for the following issues.
- **IF** We exclude explanatory variables that should be present in the regression,
- **AND** these omitted variables are correlated with the included explanatory variables,
- **THEN** the OLS estimates of the coefficients on the included explanatory variables will be biased.

Multiple regression

- Practical Advice for Selecting Independent Variables
  - Include (insofar as possible) all independent variables which you think might possibly explain your dependent variable. This will reduce the risk of omitted variable bias.
  - However, including irrelevant explanatory variables reduces accuracy of estimation and increases confidence intervals.
  - Do *t*-tests (or other hypothesis tests) to decide whether variables are significant. Run a new regression omitting the explanatory variables which are not significant.

Multiple regression

- **Multicollinearity**
- *Intuition:* if two variables are highly correlated they contain roughly the same information. The OLS estimator has trouble estimating two separate marginal effects for two such highly correlated variables.
- *Symptom*: Individual coefficients may look insignificant, but regression as a whole may look significant (e.g. R^2 big, F-stat big, but t-stats on individual coefficients small).
- *Common way to investigate if multicollinearity is a problem*: Looking at a correlation matrix for indep variables can be helpful in revealing extent and source of multicollinearity problem.
- **Note:** high correlation means that correlations between your indep. variables > 0.9 (then you probably have a multicollinearity problem).
- Solutions to multicollinearity problem
    - Get more data (often not possible).
    - Drop out one of the highly correlated variables.

Multiple regression

Example

A regression relating to the effect of studying on student performance.

- y = *student grade on test*
- x_1 = *family income*
- x_2 = *hours studies per day*
- x_3 = *hours studied per week.*
- *But* $x_3 = 7x_2$ *– an exact linear relationship between two explanatory variables (they are perfectly correlated). This is a care of* **perfect multicollinearity**.

Multiple regression

- In practice you will never get perfect multicollinearity, unless you do something that does not make sense (like put in two explanatory variables which measure the exact same thing).

Example

Macroeconomic regression involving the interest rate.

- $x_1 =$ interest rate set by Bank of England
- $x_2 =$ interest rate charged by banks on mortgages.
- x_1 and x_2 will not be exactly the same, but will be very highly correlated (e.g. $R = 0.99$).
- If you include both x_1 and x_2 you will run into a multicollinearity problem. So include one or the other (not both).

Multiple regression - Dummy Variables

- Dummy variables are variables that are created to allow for qualitative effects in a regression model.
- A dummy variable will take the value 1 or 0 according to whether or not the condition is present or absent for a particular observation.
- If the variable has m levels, we include $m - 1$ dummy variables
- Model with no interaction (same slope)
- Model with interaction (allows the slope of y with respect to x to be different for the two groups with respect to the categorical variable)

Multiple regression

- Example: (no interaction) suppose we are investigating the relationship between the wage (Y) and the number of years of experience (X) of workers in a particular industry. Our initial model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Define $d = 1$ for male workers and $d = 0$ for female workers. The overall equation becomes

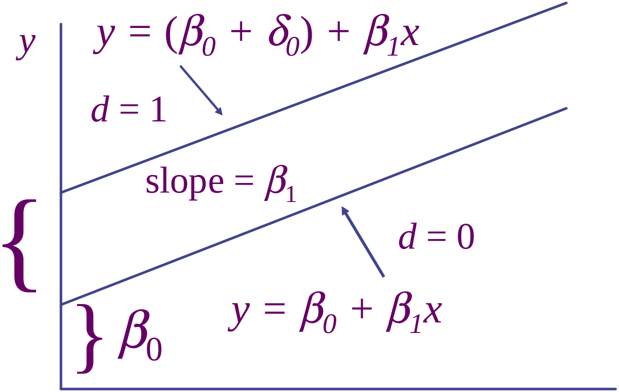
$$y = \beta_0 + \beta_1 x + \delta_0 d + \varepsilon$$

where δ_0 will measure the differential between male and female workers, having taken account of differences in experience.

Multiple regression

- We can run a normal multiple regression with x and d as indep variables. Assuming that δ_0 is positive it means that the regression line for male workers lies above that for female workers
- δ_0 measures the extent of the upward shift.
  - If $d = 0$, then $y = \beta_0 + \beta_1 x + \varepsilon$
  - If $d = 1$, then $y = \left(\beta_0 + \delta_0\right) + \beta_1 x + \varepsilon$

Multiple regression



y
$y = (\beta_0 + \delta_0) + \beta_1 x$
$d = 1$
slope = β_1
$d = 0$
$y = \beta_0 + \beta_1 x$
$\}\ \beta_0$

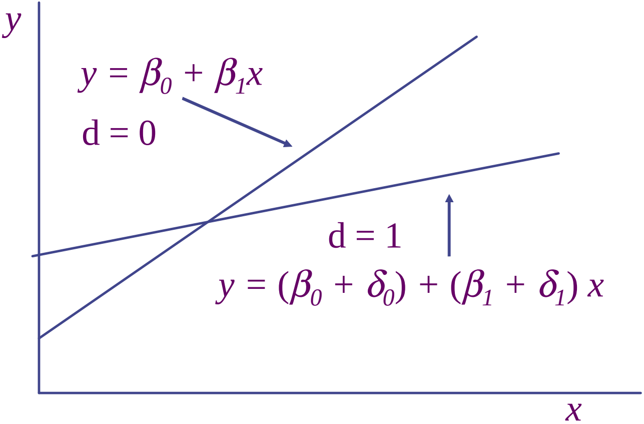
Regression multiple

- Example: interaction between the dummy variable and the indep. variable x

$$y = \beta_0 + \delta_0 d + \beta_1 x + \delta_1 dx + \varepsilon$$

- If $d = 0$, then $y = \beta_0 + \beta_1 x + \varepsilon$
- If $d = 1$, then $y = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)\, x + \varepsilon$

Multiple regression


y
$y = \beta_0 + \beta_1 x$
d = 0
d = 1
$y = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)\, x$
x

Multiple regression - Nonlinear models

- OLS can be used for relationships that are not strictly linear in x and y by using nonlinear functions of x and y (still linear in the parameters)
- For example:
  - Can take the natural log of x, y or both
  - Can use quadratic, cubic, or inverse forms of x
  - Can use interactions of x variables
- How to decide which **nonlinear form**?
- It can be hard to decide which nonlinear form is appropriate. Here are a few pieces of advice.
  - Sometimes economic theory suggests a particular functional form.
  - Experiment with different functional forms and use hypothesis testing procedures or R^2 to decide between them.

Multiple regression

Example

Is there a quadratic pattern? Run OLS regressions on two models (linear):

$$y_i = \beta_0 + \beta_1 X_{1(i)} + \varepsilon_i$$

and (non-linear)

$$y_i = \beta_0 + \beta_1 X_{1(i)} + \beta_2 x_{1(i)}^2 + \varepsilon_i$$

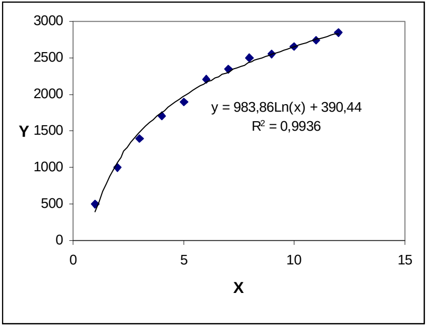
and choose the **quadratic model** if its R^2 is **higher** than the linear model (In order to use it for choosing a model, all models must have the same Y).

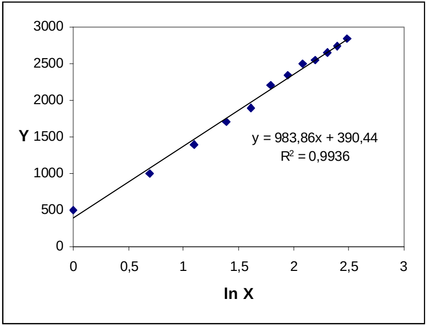
- **Warning**: you can **only** use R^2 to compare models involving nonlinear transformations of the **independent** variables.

Multiple regression

- Three cases, differing in whether y and/or x is transformed by taking logarithms
  - (1) **linear-log** $y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$
  - (2) **log-linear** $\log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - (3) **log-log** $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$
- After creating the new variable(s) $\log(y)$ and/or $\log(x)$, the regression is linear in the new variables and the coefficients can be estimated by OLS.
- In model (1), β_1 is approximately the change in y for a 100 percent change in x
- In model (2), β_1 is approximately the percentage change in y given a 1 unit change in x
- In model (3), β_1 is the elasticity of y with respect to x

Multiple regression





Multiple regression

- Log models are invariant to the scale of the variables since measuring percent changes
- They give a direct estimate of elasticity
- For models with $y > 0$, the conditional distribution is often heteroskedastic or skewed, while $\log(y)$ is much less so
- The distribution of $\log(y)$ is more narrow, limiting the effect of outliers
- What types of variables are often used in log form?
  - Dollar amounts that must be positive
  - Very large variables, such as population
- What types of variables are often used in level form?
  - Variables measured in years
  - Variables that are a proportion or percent

Multiple regression - Weighted Least Squares

- While it is intuitive to see why performing OLS on a transformed equation is appropriate, it can be tedious to do the transformation
- **Weighted least squares** is a way of getting the same thing, without the transformation
- Sometimes multiply/dividing all your indep. variables by some variable (x_i) is enough to fix the problem (heteroskedasticity).
- Idea is to minimize the weighted sum of squares (weighted by $1/x_i$) in order to obtain the parameter estimates

Multiple regression - Modelling strategies

- *"The three golden rules of econometrics are test, test and test." David F. Hendry (1980)*
- Begin with a general model which nests the restricted model and so allows any restrictions to be tested
- These restrictions may be suggested either by theory – or by empirical results
- **TEST 1:** First ensure that the general model does not suffer from any diagnostic problems. Examine the residuals in the general model to ensure that they possess acceptable properties (test for problems of autocorrelation, heteroskedasticity, non-normality, incorrect functional form etc.)
- **TEST 2:** Now test the restrictions implied by the specific model against the general model – either by exclusion tests or other tests of linear restrictions.
- **TEST 3:** If the restricted model is accepted, test its residuals to ensure that this more specific model is still acceptable on diagnostic

Multiple regression

- Frequently (and recently) asked questions!
    - “Should I include all the variables in the database in my model?”
    - “How many explanatory variables do I need in my model?”
    - “How many models do I need to estimate?”
    - “What functional form should I be using?”
    - “Do I need to include lagged variables?”
    - “What are interactive dummies – do I need them?”
    - “Which regression model will work best and how do I arrive at it?”

Multiple regression

- Typical cross-section model
  - Maybe several hundred observations
  - Maybe 10-12 potential explanatory variables, some of which will be dummy variables.
  - So plenty of degrees of freedom but still lots of potential models to try, especially if you consider alternative functional forms, interactive dummies
  - Maybe problems of multicollinearity, heteroskedasticity and non-normality
  - Model selection is not just a matter of maximizing $\bar{R}^2$ over all possible models (or some other criterion)
  - Use economic theory and past studies to identify "core" variables
  - Test exclusion restrictions from a general model but balanced against misspecification tests. "Informed" searches.

Multiple regression

- Typical time series model
  - Maybe only around a hundred observations
  - Maybe four or five potential explanatory variables, some of which may be dummy variables.
  - Relatively few degrees of freedom but still lots of potential models to try, especially if you consider alternative functional forms, lagged variables and interactive dummies
  - As well as problems of multicollinearity, heteroskedasticity and non-normality there may be issues of autocorrelation and non-stationarity
  - Model selection is not just a matter of maximizing $\bar{R}^2$ over all possible models
  - Use economic theory and past studies to identify "core" variables and if possible functional form
  - Test exclusion and other restrictions from a general model but balanced against misspecification tests. "Informed" searches.