

RUNNING A PROPER REGRESSION ANALYSIS

V G R CHANDRAN GOVINDARAJU

UITM

Email: vgrchan@gmail.com

Website: www.vgrchandran.com/default.html



Topics

Running a proper regression analysis

First half of the day:

- 1. What is regression?*
- 2. How to estimate? (Simple and Multiple Regression)*
- 3. Checking the assumptions of regression*

Second half of the day

- 1. Regression with dummy variables*
- 2. Recap: Time Series Econometrics*



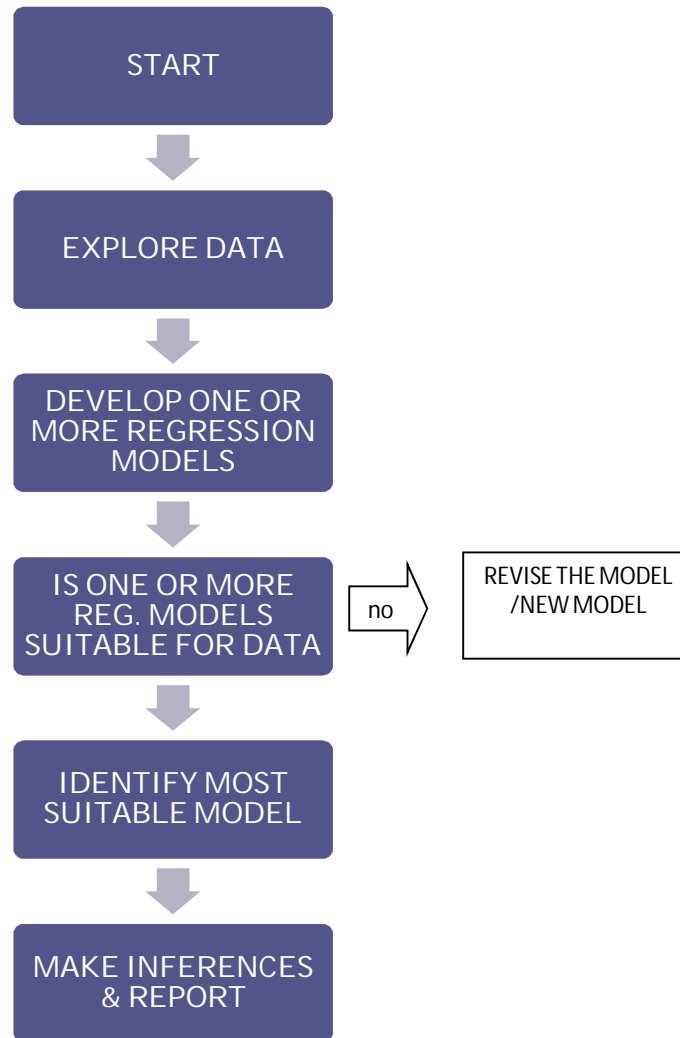
Types of data

- Cross sectional
- Time series
- Panel data

- Where to get the data, DOS and BNM
- Lets download some data

- Data transformation – level data, growth rate, index numbers, nominal to real values, exponential to linear models, etc

What to do after obtaining your data?





Explore the data

- Data cleaning
- Feel your data –
 - Descriptive Statistics
 - Correlations and Plots

What is regression?

- Relationship between two variables (simple) or more than two variables (multiple)
- Models:

$$Y = \alpha + \beta X + \varepsilon$$

- α is the intercept
- β is the coefficient
- ε is the error term

Regression (Simple Example)

n	y	x
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5
7	96	6
8	97	6
9	109	7
10	119	8
11	149	9
12	145	9
13	154	10
14	166	10



What is regression? (continue)

SIMPLE EXAMPLE

- Lets plot the data – scatter plots
- Fitting a regression line.
- Finding the error term (residuals)
- Residuals are the most important part of regression (will let you know why later)



Estimating the alpha and beta.

- Using Excel, SPSS, Eviews, Microfit, STATA, SPLUS
- Will teach how to use Excel, EVIEWS and SPSS (just an overview)
- Interpreting the outputs



Things to evaluate (output)

- Economic Criteria – signs and size of the effects (coefficient) – follows economic theory – demand for food (price variable)
- Coefficient of determinants
- Significance test on parameters (also joint test)
- Model selection criteria
- Functional form
- Econometric criteria - assumptions (do not violate)



Assumptions of linear regression

- Linearity
- Normality
- Autocorrelation/Serial Correlation
- Heterogeneity
- Multicollinearity

Linearity

- Straight enough condition (scatter plots)
- SPSS: Graphs: Scatter: Matrix: enter the dependent (outcome) variable first and then each of the independent variables (categorical/nominal variables don't need to be entered, but do it anyway to see what it looks like).
- SPSS: Analyze: Regression: Linear:
- Ramsey RESET test.

Normality

- We do not need to test each series
- Just test the residuals
- Jarque-Berra statistics or the QQ and PP plots
- We use JB stat.
- Null Hypo: Normal

- What to do if data is not normal?
 - Increase sample size
 - Transform data – e.g. log values – Data may not be normal b'cause of specification problems or functional form. Remember linearity



Serial Correlation/AutoCorrelation

- Likely a problem in time series data especially data with short frequency
- What cause autocorrelation?
 - Omitted variables
 - Misspecification
- Consequences of autocorrelation
 - OLS estimators will be inefficient
 - Variance of the coefficient will be biased and inconsistent



How to detect autocorrelation

- Graphical methods: plot the residual and also draw a scatter plot of residual against residual (-1)
- Durbin-Watson test – Eviews (Null Hypo: no autocorrelation)
- Application –when model includes constant, only first order and no lagged dependent variables
- We have a table to compare (DW stat) to the critical values (but rule of thumb – if the value nears 2 than it is ok)



How to detect autocorrelation

- Breusch-Godfrey test for serial correlation
- It can test higher orders
- Eviews – View/Residual tests/serial correlation
LM test



How to solve autocorrelation?

- Cochrane-Orcutt iterative procedure (beyond our scope – remember I said regression in plain English)
- AR (1)



Test for specification

- Ramsey's RESET test
- We include the predicted dependent variable as one of the regressors
- Lets do it in eviews

Heteroskedasticity

- The opposite of homoskedasticity
- Hetero means unequal; Homo means equal
- Second part of the word “skedasticity” means spread (variance)
- Example: Consumption – rich and poor – rich have better spread (save and consumption) poor have lower spread
- There are many ways to test hetero :
- Graphically – plot residual squared against dependent or independent variable – there must not be a systematic pattern
- However graphical methods can be used for multiple regression

Heteroskedasticity


- The following test can be used:
 - Breusch-Pagan LM test, Glesjer LM Test, Harvey-Godfrey LM test, Park test, Goldfeld-Quabdt test, and White test
 - Lets use the White test
 - Null Hypo: No hetero or homo

Consequence of hetero and ways to correct it

- No change in estimated parameter but standard error is effected (so does the significant)
- Generalized (or weighted) least squares (beyond or discussion)
- Run a heterogeneity corrected regression (lets do a simple (White corrected standard error estimates)
- Alternatively, we can also use dummy variables to account for hetero

Multicollinearity

- Whether there is any relationship between the regressors
- Consequences – parameter is indeterminate if perfect multicollinearity (However, real data do not have perfect multicollinearity)
- Imperfect multicollinearity – when regressors are correlated but less than perfect
- How to detect?
 - Correlation matrices
 - Check the significance of individual coefficient (t-test) and the joint significance (F-test)
 - Run the regression by separating the regressors
 - VIF – Eviews or in SPSS (VIF value of less than 10 is ok)



Structural break and parameter stability test

- Aim is to see whether parameters of the models have been constant over the periods
- Chow test – we have to know the point of the break
- CUSUM and CUSUM Q Test – parameter stability

Regression Analysis with Dummy Variables

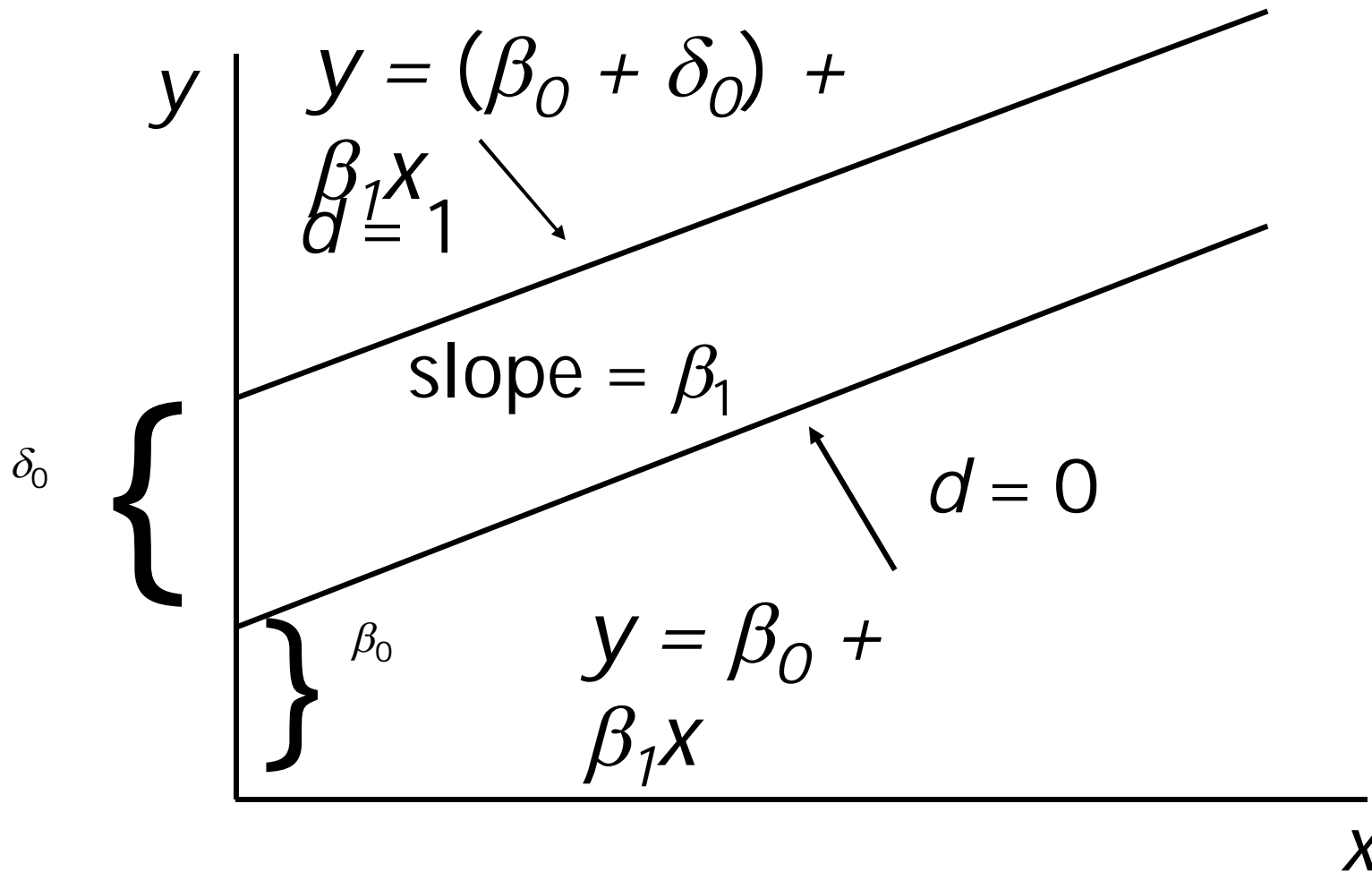
$$\blacklozenge y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \dots + \beta_k x_k + u$$

Dummy Variables

- A dummy variable is a variable that takes on the value 1 or 0
- Examples: male (= 1 if are male, 0 otherwise), south (= 1 if in the south, 0 otherwise), etc.
- Dummy variables are also called binary variables, for obvious reasons

A Dummy Independent Variable

- Consider a simple model with one continuous variable (x) and one dummy (d)
- $y = \beta_0 + \delta_0 d + \beta_1 x + u$
- This can be interpreted as an intercept shift
- If $d = 0$, then $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, then $y = (\beta_0 + \delta_0) + \beta_1 x + u$
- The case of $d = 0$ is the base group

Example of $\delta_0 > 0$ 

Dummies for Multiple Categories

- We can use dummy variables to control for something with multiple categories
- Suppose everyone in your data is either a HS dropout, HS grad only, or college grad
- To compare HS and college grads to HS dropouts, include 2 dummy variables
- $hsgrad = 1$ if HS grad only, 0 otherwise; and $colgrad = 1$ if college grad, 0 otherwise

Multiple Categories (cont)

- Any categorical variable can be turned into a set of dummy variables
- Because the base group is represented by the intercept, if there are n categories there should be $n - 1$ dummy variables
- If there are a lot of categories, it may make sense to group some together
- Example: top 10 ranking, 11 – 25, etc.

Interactions Among Dummies

- Interacting dummy variables is like subdividing the group
- Example: have dummies for male, as well as hsgrad and colgrad
- Add $\text{male}^* \text{hsgrad}$ and $\text{male}^* \text{colgrad}$, for a total of 5 dummy variables \rightarrow 6 categories
- Base group is female HS dropouts
- hsgrad is for female HS grads, colgrad is for female college grads
- The interactions reflect male HS grads and male college grads

More on Dummy Interactions

- Formally, the model is $y = \beta_0 + \delta_1 male + \delta_2 hsgrad + \delta_3 colgrad + \delta_4 male * hsgrad + \delta_5 male * colgrad + \beta_1 x + u$, then, for example:
- If $male = 0$ and $hsgrad = 0$ and $colgrad = 0$
- $y = \beta_0 + \beta_1 x + u$
- If $male = 0$ and $hsgrad = 1$ and $colgrad = 0$
- $y = \beta_0 + \delta_2 hsgrad + \beta_1 x + u$
- If $male = 1$ and $hsgrad = 0$ and $colgrad = 1$
- $y = \beta_0 + \delta_1 male + \delta_3 colgrad + \delta_5 male * colgrad + \beta_1 x + u$

Other Interactions with Dummies

- Can also consider interacting a dummy variable, d , with a continuous variable, x
- $y = \beta_0 + \delta_1 d + \beta_1 x + \delta_2 d * x + u$
- If $d = 0$, then $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, then $y = (\beta_0 + \delta_1) + (\beta_1 + \delta_2) x + u$
- This is interpreted as a change in the slope

Other use of dummy variables

- Seasonal dummy
- Structural breaks
- Shocks
- etc



Lets Recap Our Time Series Analysis

- Unit Root Test
- Cointegration
- Vector Error Correction Model
- Granger Causality

Must Have Books (for new researchers)

Practical Data Analysis

- Gary Koop (2004) Analysis of Economic Data, John Wiley.

Basic Econometrics

- Gary Koop (2008) Introduction to Econometrics, John Wiley
- Samprit Chatterjee, Ali S. Hadi, Bertam Price (2000) Regression Analysis by Example, John Wiley.
- Dimitrios Asteriou and Stephen G. Hall (2007) Applied Econometrics: A Modern Approach Using Eviews and Microfit, Palgrave

Basic Statistics and Regression Models

- De Veaux, Paul Velleman and David Bock, Stats: Data and Models, Pearson. (for basic statistics)



Thank you

QUESTIONS PLEASE

More materials will soon be available (by end of the month) through my website:

www.vgrchandran.com/default.html