University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Statistics for Economics and Business

Stefano Bonnini & Valentina Mini

Multiple Linear Regression Analysis: Practical issues

Lecture 8 – 8th of March 2019

Step-by-step analysis

#0. prepare and describe the dataset and the variables of interest

#1. graphical analysis of relationships (please, comment)

#2. specify the model (command and definition of your equation model)

#3. interpreting the coefficients

#4. range of xi and possible prediction

#5. the goodness of fit of your model (R2)

#6. try to compare different models (R2 adj)

#7. the significance of your model (F test)

#8. the significance of each single coefficient (t-test)

#9. graphical analysis of LRM conditions (e)

#10. check for the multicollinearity (VIF)

#11. please comment the obtained results considering the initial Research Question

A case study

Central research question:

Do the sugar, cereals and meat consumptions affect the Alcoholic beverages consumption?

Database: *eating*

The **model** we should use is a MLRM

0. prepare and describe the dataset

- Change directory (looking for the right one)
- View data structure and select those you are interested in
- Describe data and variables you are going to use

1. Graphical analysis

- Rename the variables
- Plot each xi and y
- Observe and comment the path described by the graphs

2. Specify the model

The specific model is:

Alcoholic.Beverage=b0+b1*Cereals+b2*Sugar+b3*Meat

Using the renamed variables you may write:

$$y = b0 + b1x1 + b2x2 + b3x3$$

Perform the model using lm command

3. Interpreting the coefficients

- b_0= when x_1,x_2,x_k are all equal 0, y will be equal b_0
- b_1 =for each increase of 1 unit *x1*, we expect an increase equals *b1* in y, when all other xi are held constant.
- b2 = for each increase of 1 unit x2, we expect an increase equals b2 in y, when all other xi are held constant.
- bk = for each increase of 1 unit xk, we expect an increase equals bk in y, when all other xi are held constant.

4. Range of x_i and possible predictions

- Explore the range of your explanatory variables
- Compute the prediction using your equation model

5. The goodness of fit of our model (Rsqr)

- Check the value of R sqr : it reports the proportion of total variation in Y explained by all Xi variables taken together

- Remember the Rsqr range (0;1)

6. Comparing different models (Radj)

- Try to perform different MLR models (by adding or subtracting $x_{i\text{-}es}$ from your original model)
- Check the R adj to compare those different models
- The higher the R adj – the better is the model

7. The significance of our model (F test)

- Check the F-test: it shows if there is a linear relationship between all of the X variables considered together and Y
- *Take into account alpha (the significance level) and the F-test' p-value*

8. The significance of each single coefficient (t)

- For each coefficient check the t-value and the associated p-value
- Make inference about each single coefficient

9. Graphical analysis of MLRM conditions

Among others conditions (see previous lectures) we should check the normal distribution of our error terms (residuals) : the QQ-plot is useful to visually check the normality of the data.

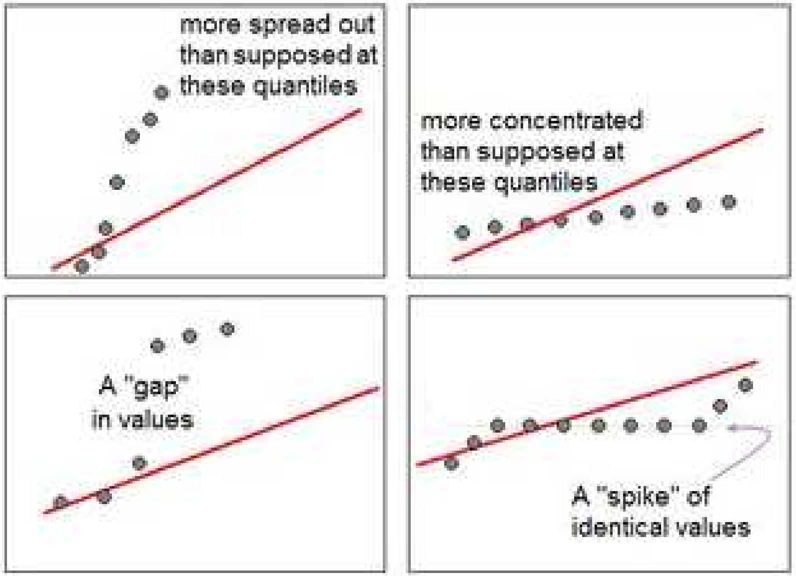
The QQ-plot draws the correlation between a given sample and the normal distribution.

If the values lie along a line the distribution has the same shape (up to location and scale) as the theoretical distribution we have supposed.

As sample sizes become larger, generally speaking the plots 'stabilize' and the features become more clearly interpretable rather than representing noise. [With some very heavy-tailed distributions, the rare large outlier might prevent the picture stabilizing nicely even at quite large sample sizes.]

If we perform qqline(e1) we add a reference line

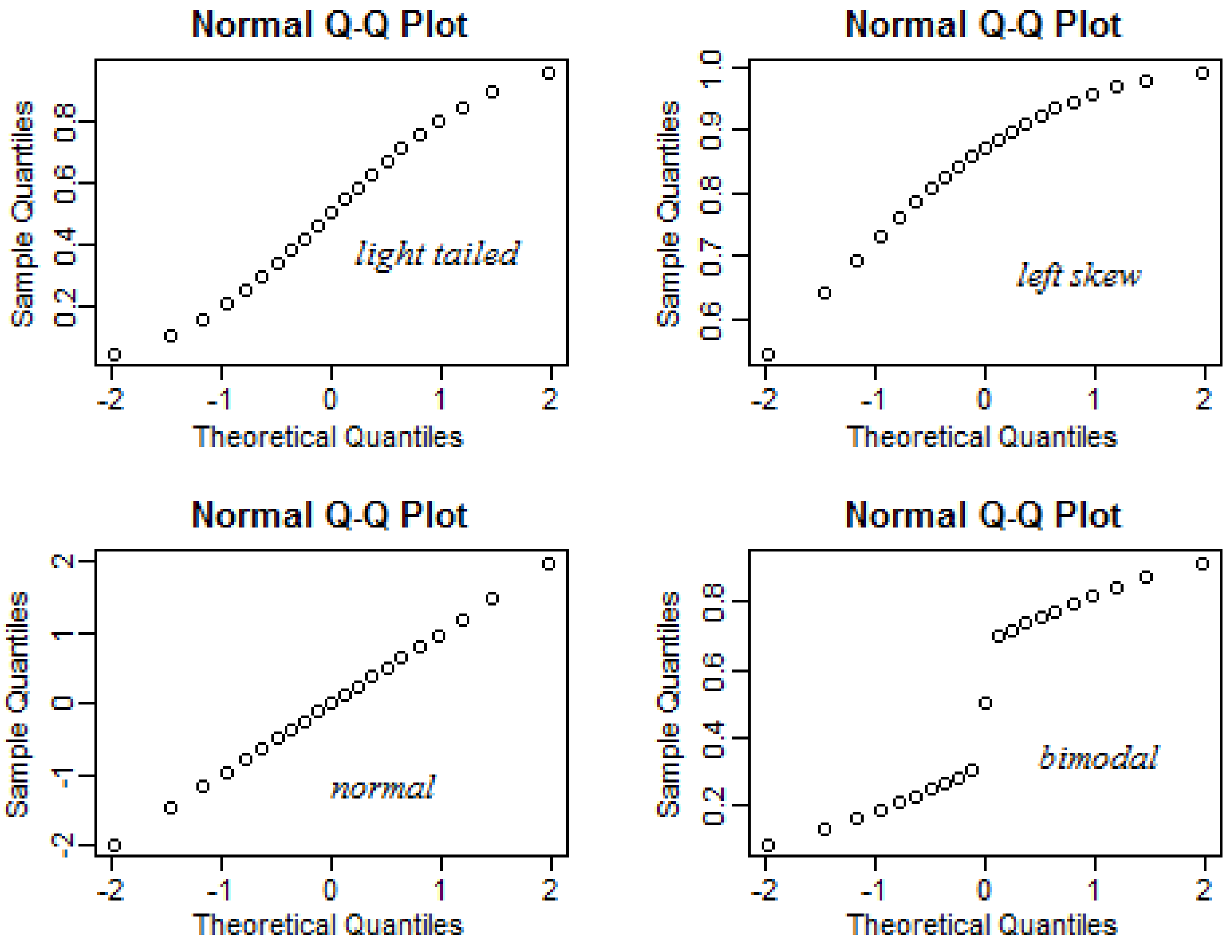
9. Graphical analysis of MLRM conditions (example)



As we see, less concentrated points increase more and more concentrated points than supposed increase less rapidly than an overall linear relation would suggest, and in the extreme cases correspond to a gap in the density of the sample (shows as a near-vertical jump) or a spike of constant values (values aligned horizontally). This allows us to spot a heavy tail or a light tail and hence, skewness greater or smaller than the theoretical distribution, and so on.

9. Graphical analysis of MLRM conditions (example)

Here's what QQ-plots look like (for particular choices of distribution) *on average*:



10. Check for the multicollinearity (VIF)

- In multiple regression (Chapter @ref(linear-regression)), two or more predictor variables might be correlated with each other. This situation is referred as *collinearity*.
- There is an extreme situation, called **multicollinearity**, where collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. This means that there is redundancy between predictor variables.
- In the presence of multicollinearity, the solution of the regression model becomes unstable.
- For a given predictor (x), multicollinearity can assessed by computing a score called the **variance inflation factor** (or **VIF**), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.
- **As a rule of thumb, a VIF value that exceeds 5 indicates a problematic amount of collinearity (James et al. 2014).**

- When faced to multicollinearity, the concerned variables should be **removed,** since the presence of multicollinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables (James et al. 2014,P. Bruce and Bruce (2017)).

11. Comment results considering the RQ

Considering your analysis you must be able to give an answer to the initial research question:

do the sugar, cereals and meat consumptions affect the Alcoholic beverages consumption?

At which level of confidence are you able to make inference?

What are the strength and weak points of your model?

What do you suggest for future researches?

LAB USING R