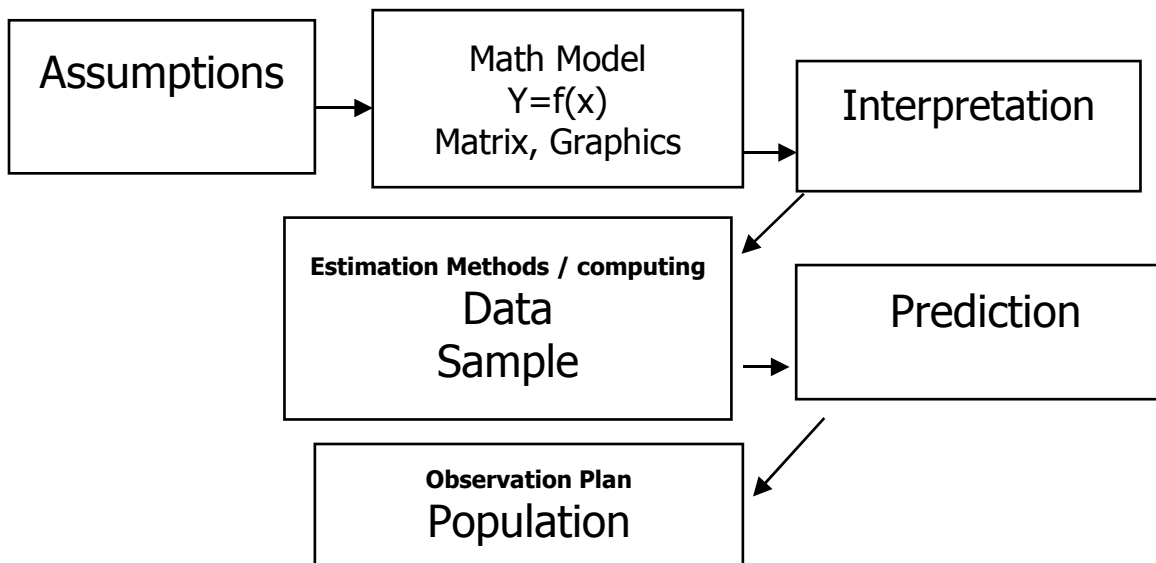


Part 1: Introduction to regression modeling and the RM4Es framework

SUMMARY: This Part introduces the RM4Es framework for understanding model building processes, and starts our discussion with examples on simple regression.

When we discuss model buildings in social science, it is important for us to understand two important issues. The first is the logic of model abstraction, as illustrated by the following diagram 1:1. The second is that model building is a step by step searching process.

Figure 1:1



Here, by a model, we mean an abstract image of reality whose construction has often been guided by theories or empirical observations from past (Tuma & Hannan 1984). A model has some characteristics of the reality, but never has the full characteristics of the reality. In other words, a model is a simplified structure of the reality that can help us to make inferences. For any model to become a reasonable representation of the reality, we often have some assumptions. As soon as we have a model constructed, we want to use our collected data to estimate our model. Then, the estimated model can be used to interpret the reality under study or to make future predictions. The above figure 1.1 illustrates the relationship among all the elements we discussed: assumptions, model, data, estimation, observation plan, population, sampling, interpretation and prediction. Among all these mentioned elements, if we leave the data collection aside and focus our attention on the model building process, four of them are very important and they are (1) function representation or mathematical equations, (2) explanations, (3) assumptions that are often used to conduct model evaluations, and (4) estimation methods.

As an example, let us assume we are interested in building a simple regression model. Simple regression is a model where a response can be linearly represented by one independent variable. Mathematically, it can be written as an equation of $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

Here in the simple regression equation, besides the dependent variable Y and independent variable X, we have three other items - constant (intercept) β_0 , slope β_1 and residuals ε_i . These three items are what we need to estimate with our collected data, and are what we need to interpret to make sense of our modeling results.

In other words, for simple regression, we also have four important elements to work with. And they are (1) function representation or mathematical equation that is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, (2) explanation of constant β_0 that is the Y value when X equals 0, and of slope β_1 that is the change of Y when X changes by one unit, (3) estimation of coefficients, for which Ordinary Least Square Method (OLS) is often used, and (4) assumptions.

The assumptions for simple regression include the following, which are all about residuals.

- ε_i
- zero mean
- constant variance
- independent from each other
- independent from x_i
- normally distributed

In the model building process, these assumptions are used to conduct diagnostics or model evaluations. Therefore, we actually have four Es that are necessary to deal with for any model building process: **E**quation (functional representation), **E**xplanation (Interpretation), **E**stimation, and **E**valuation (Diagnostics or Errors). We call the method of working with these four Es to build a model as the RM4Es framework.

If we use this RM4Es framework to summarize simple regression, we will get the followings:

– **Equation**

$y = a + \beta x + e$ is the equation for any simple linear regression. Here, y is often called as a dependent variable or a response, while x is often called as an independent variable or a predictor. a is called as an intercept and β is called as a slope, while e is called the error term. a and β are the equation parameters to be estimated. Adapting this equation assumes the dependent variable y is linearly related to one and only one independent variable x.

– Estimation

After specifying our equation, we need to use available data to estimate the values of the constant a and the slope β . The ordinary least squares (OLS) method is the one employed most often, but the maximum likelihood method can also be used. When conducting OLS estimation, parameters a and β are chosen to minimize a quantity called as the residual sum of squares that is $S[y - (a + \beta x)]^2$. Under the assumption errors are uncorrelated and have the same variance, the OLS estimate is the best among all linear estimation methods.

– Errors for Evaluation

e is the error term for simple linear regression that is the difference between the predicted values and the actual values of the dependent variable y . That is, $e = y - (a + \beta x)$.

Errors can be used to evaluate the goodness of fit of your simple linear regression, and can also be used to diagnose your regression model in order to improve it.

– Explanation

a , β and R^2 are what need to be explained for simple linear regression.

Here, a , the intercept, is the value of y when x equals to 0. And, β , the slope, is the rate of change in y for a unit change in x . $R^2 = 1 - \text{RSS}/\text{SYY}$ is called as coefficient of determination. R square tells us how much variability in y can be explained by our model. Simple linear regression can be represented by a straight line that graph is often used to help explanations.