Introduction to Linear Regression

Shivank Goel, Email : sg2359@cornell.edu

INTRODUCTION

Regression is an approach to model the relationship between dependent variables (often a single dependent variable), and a set of independent variables, often to understand the cause and effect relationship between them.[1] When this relationship is linear, and we have a single explanatory variable we call it as a simple linear regression model. When we have more than one explanatory variables, the approach is called multivariate linear regression.[2]

DESCRIPTION

Given any random variable y, we can write it as, $y = E(y|x) + \epsilon$, where (y, x, ϵ) are random variables and $E(\epsilon|x) = 0$. $E(y|x)$ is called *Conditional Expectation Function (CEF)*. We can observe that, for any given x, distribution of y is governed by ϵ, and is centered around CEF. We can estimate the CEF, using the regression equation, $y = \alpha + \beta x + u$. Note that in this specification, it is safe to assume that, $E(u) = 0$, because otherwise, u can be broken into, $u = k + u'$, where k is a constant and can be merged with α. Also, we can say that β represents the true causal effect of x on y iff x and u are not correlated. A zero correlation between x and y is also necessary, so that y can be a consistent estimator of CEF. Further, for y to be an unbiased estimator of CEF, i.e. $E(y|x) = \alpha + \beta x$, it is required that $E(u|x) = E(u) = 0$. This is known as the *Conditional Mean Independence (CMI)* assumption. We note that, CMI condition will automatically hold if x and y will be uncorrelated. There can be many ways in which CMI can be violated, and can create a bias in the estimation of CEF, such as *Omitted Variable Bias*, *Measurement Error Bias* and *Simultaneity Bias*.

MULTIVARIATE ESTIMATES AND PARTIAL EFFECTS

We can describe the basic multivariate model as, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$. For unbiased estimates, apart from the assumptions made in simple linear regression, it is required that there does not exist a perfect collinearity among any two covariates. Intuitively, multivariate regression is equivalent to finding effect of each independent variable after partialing out effect of other variables. In other words, we can also get the coefficient, $\hat{\beta}_i$, by first collecting residuals obtained by regressing y on all other variables except x_i (say $\tilde{y}$). Intuitively, $\tilde{y}$ represents the variation left in the dependent variable y after removing the variation caused by other variables. Similarly, remove variation in x_i caused by other variables, by regressing x_i on all other dependent variables and collect the residuals, say $\tilde{x}$. Now, we can regress $\tilde{y}$ on $\tilde{x}$, and get the same estimate for β_i, which we can get from the multivariate regression model.

GOODNESS OF FIT (R^2)

Let, N be the number of observations in the sample. Define, SST as the total variation in y in our dataset, i.e., $SST = \sum_i (y_i - \bar{y})^2$, where $\bar{y}$ is the mean of dependent variable in our data. Let, SSE be defined as the total variation in the predicted y values, i.e., $SSE = \sum_i (\hat{y}_i - \bar{y})^2$. Let, SSR denote the variation in residuals, i.e., $SSR = \sum_i \hat{u_i}^2$. Mathematically, one can show that, SST = SSE + SSR. R^2 is defined as the proportion of total variation that can be explained by our estimated CEF, i.e, $R^2 = SSE/SST$. Mathematically, it can also be shown that $R^2 = Corr(y, \hat{y})^2$. Hence, if we will add more independent variables, R^2 can never go down. Due to this fact, we use *Adjusted* $R^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-1-k}\right)$, where k is the number of independent variables. Hence, as k increases and if R^2 stays constant, then *Adjusted* R^2 will decrease. In order to make up for that R^2 must also increase with increasing k.

---

[1] This pdf has been summarized from *Todd A. Gormley's* notes, Lecture 1, http://www.gormley.info/phd-notes.html

[2] Generally, we refer to multivariate regression as simply regression, since having multiple explanatory variables is quite common