Online Access available from: http://www.healthprospect.org

View Point

Health Prospect
Journal of Public Health

Open Access

Estimating population relation- method and assumptions

Sapkota VP[1]

Abstract

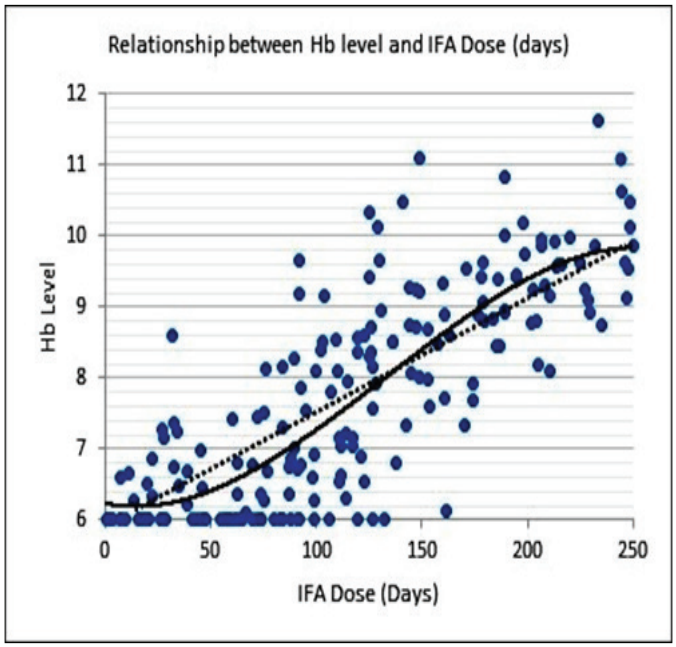
Researchers often try to estimate the relationship among the variables in a population using regression modelling. In public health statistical literature, a clear theoretical introduction about the theory underlying the estimation techniques and assumptions is often lacking. In empirical papers, many do not discuss the violation of regression assumptions, and quite often, emphasis has been laid only on the aspects that are less important. In this theoretical review, I have described a theory behind the estimation of population relation, preliminaries of regression methods and various assumptions one has to take care of while estimating the population relation in an unbiased manner.

Keywords: population relation; empirical relation; regression assumptions.

Submitted: 22 June 2014; **Revised:** 30 July 2014; **Accepted:** 2 August 2014.

Introduction

Introductory statistics course covered theory behind population mean or proportion estimation. Especially, how the sample mean (or proportion) is an attractive estimator of population mean (or proportion). In empirical studies, population parameters are estimated using a sample of population units. The objective of this article is to explain the ways to carry over this concept to estimate a relationship among two or more variables at population level. Here, a systematic introduction of underlying theory and assumptions has been included with some practical examples.



Consider a variable Y = haemoglobin level of a women during pregnancy and another variable X = Iron Folate tablet (IFA) consumption (in terms of doses). When the relationship at the population level is discussed, we tend to describe average value of Hb level (a dependent variable) for a given dose of IFA consumed by a women (independent variable) (1, 2). In the scatter diagram, each point shows the dose of IFA and corresponding Hb level for each population unit at time t. The black curve shows average value of Hb in population for the women consuming various doses of IFA. Each point lying on this black curve maps the average value of Hb at different doses of IFA. This average value when discussed at population level is called expected value, denoted by E(Y|X=x). This expression is read as the expected value of Y (Hb level) when the variable-IFA dose (X) takes on the specific values of x (say, 120 days). The vertical distance between each point in the scatter diagram and the expected value (points on the curve) is called "error" or "residual" generally represented by u. We can predict the average level of Hb in the population using the level of IFA dose once the curve (E(Y | X)) has been estimated. In general, the curve is represented by function f(x). It is an algebraic expression which shows that when Xs are combined in some mathematical expression, it gives expected value of Y i.e. E(Y | X). Estimation of any population relationship is, therefore,estimation of f(x) that best describes the average population level relationship between the X and Y. Mathematically, it is written as:

$$E(Y \mid X) = f(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

[1]*The data was generated hypothetically for the purpose of this article.*

Correspondence:
[1]Vishnu Prasad Sapkota
Health Economist
Institute for Nepal Environment and Health System Development
Shantinagar Gate, Kathmandu
Email: visapkota@gmail.com

Adding the expected value of Y for a particular value of X and the error u, It should give the original value of Y. It is shown in the expression below
$Y= E(Y|X) + u$…….(2)

From (1) and (2), we can write the following expression.
$Y= f(X) + u$……. (3)

It means that actual value of Hb level (Y) is a sum of expected value of Y at given value of X i.e. E(Y|X) and the error u. Our purpose is to estimate f(X) in unbiased manner. Regression methods can be used in order to estimate f(X). The simple form of f(X) will be linear as shown below.
$f(x)= \beta_0+\beta_1X$……...(4)

From equation (3) and (4), the regression equation takes the following form.
$Y= \beta_0+\beta_1X+ u$……(5)

Assumptions

Following assumptions should be fulfilled, once we identify the equation to be estimated (equation 5). These assumptions are generic and are applicable to most of regressions used in practice.

Linearity assumption: The "$\beta_0+\beta_1X$" part of equation 5 is expected to represent average value of Y at given level of X i.e. $E(Y|X)$. This assumption is called linearity assumption (3). It means that the conditional expectation $E(Y|X)$ is best estimated (in this example) by the linear combination of X. "$\beta_0+\beta_1X$" is shown in the above figure by the dotted line. It can be observed that the population level relationship $E(Y|X)$ is shown by the nonlinear curve and, from equation 5, this nonlinear relationship is being estimated using linear expression. There are some other assumptions that need to be fulfilled in order to best estimate the average relationship.

Zero conditional mean of error component (u). The error should have zero expected value given any value of explanatory variable (X) i. e. $E(u|X)= 0$ (3). In the other words, average value of error u should be zero for any value of independent variable X. This is the most important assumption while estimating population level relationship. It means that u is uncorrelated with X and f(x) has been properly specified such that it represents true $E(Y|X)$. In our case, it means that "$\beta_0+\beta_1X$"and error term (u) will be uncorrelated. Lets consider equation 5, X involves a list of variables that has been observed and can explain the Y significantly. The error(u) is the part of Y that is not explained by the observed list of variables (X). Error, therefore, includes unobserved list of independent variables that can strongly explain Y. So, a question arises in the mind, how many variables should be measured and included in the equation in order to properly specify the model? There are a variety of modelling techniques in order to properly specify equation 5. This is out of scope of present article. However, one guiding principle for such decision is that variables that go in u (or the one that are not observed) should be uncorrelated with the variables that are contained in X (i.e. the variables that are observed). It ensures accurate estimation of population conditional expectation function $E(Y|X)$. In case we failed to fulfil this condition, the problem is called endogeneity. This phenomenon can be practically explained in connection with our example. In this example, only one independent variable (IFA dose) is used to explain Hb level. However, Hb level is also determined by many other factors, for example, initial Hb level before starting IFA consumption, number and spacing of parity, present or recent past illness such as malaria that drastically reduces Hb etc. In the present condition let us suppose, we couldn't measure these variables. The linear combination of these variables becomes a part of residuals (u). The rule is that the observed Xs and those that could not be observed due to study limitations should be uncorrelated at population level. In this case, it can be observed that initial level of Hb strongly determines the future level of Hb even after initiation of IFA consumption. At the same time, those women who are anaemic in the beginning are more likely to get the prescription of IFA dose. If we didn't include this variable (initial level of Hb) in the regression equation then the assumption of zero conditional mean for error term u will be violated. And, the estimated f(x) will be biased. In the more familiar term, we can say that our results are confounded and such variables are often called confounding variables. Therefore, it is always recommended to measure the variables and include in the regression equation that are correlated with independent variables and are also a major predictor of dependent variable(s). Quite often, unobserved factors strongly determine the Hb level and are also correlated with the IFA compliance. Sometimes it is very hard to quantify the unobserved factors and results are biased. There are techniques to control such unobserved factors.

Homoscedasticity assumption states that for each value of X, the variance for u should be same (3). In terms of notation, it is represented as $Var(u|X)=\sigma^2$. This phenomenon can be observed. For example, if the dots around the E(Y|X) curve are homogenously distributed, we can say that the assumption has been met. There are some other ways to test this assumption, it is called Lagrange Multiplier (LM) test (1). If this assumption doesn't hold, this situation is called Heteroscedasticity. In this situation, though the point estimates of beta coefficient are unbiased, standard errors are not correct which affects the p-values (1). In such circumstances we need to calculate White's standard error or Heteroscedasticity robust standard errors (1, 3). This technique is available in almost all the statistical packages (R, STATA etc).

Random sample/Independent sample is a basic requirement while estimating E(Y|X) based on a cross-sectional data. Regression requires a random sample where study units are independent and identically distributed (also called iid assumption). This assumption is ensured by randomly selecting of population units. One example of violation of this assumption is that when study units (for example pregnant women) are selected from hospitals/

[2] *Here f(X) is linear because our dependent variable is continuous. If it were categorical or count type, other types of cumulative probability functions for f(X) would be used. For example, when Y takes the binary form, logistic probability function is mostly used for f(X).*

health centres, it is likely that the study units are clustered. This has impact on the u i.e. it will be correlated across the observations which affects model specification and hence standard errors. In order to get rid of this problem, we need to run a fixed effect regression model that adjusts hospital specific clustering effect and hence provides an estimate of unbiased relationship (1).

Fulfilment of the four assumptions ensures that our estimated function f(x) for E(Y|X) is unbiased and so are the beta coefficients (often called regression parameters) (1). Another important feature about regression is the value of R^2. It is a statistical measure of how well a regression line approximates real data points. It is a descriptive measure between zero and one, and signifies the predictive ability of regression equation. In the simplest sense, R^2 is equivalent to the squared correlation coefficient between Y for each observation (the scattered points) and fitted values ($\hat{Y}$) (the points on the dotted line in figure 1) (3). If all the observed data points lie on the dotted line, R^2 value would be one showing absolute fitness. A value of R^2 that is nearly equal to zero indicates a poor fit of the regression line. In the process of estimating population relation, low R^2 are not uncommon, especially for cross-sectional analysis. It is worth emphasizing that a low R-squared does not necessarily mean that an estimated population relation is useless. In fact the most important thing about R^2 is that it is not important when we have fulfilled the above four assumptions to estimate the population conditional expectation function $E(Y \mid X)^2$. On the other hand, if our objective of analysis is to estimate a function and use it for prediction purpose, then the value of R^2 does play vital role.

Conclusion

The assumptions on error term (u) are the most important aspect of estimating relationship using regression methods. While designing any study to estimate the relationship among the variables, one should carefully select the independent variables. If there are any variables that the researcher could not measure but they are likely to be correlated with the observed Xs, one should report or discuss in limitations. As discussed above, it is necessary to use the value of R^2 as required. Quite often it is the case that researchers evaluate the model based on the value of R^2 alone. It is far better to evaluate the estimated relationship based on the fulfilment of the four assumptions discussed above.

References

1. Wooldridge JM. Econometric analysis of cross section and panel data: MIT press; 2010.
2. Goldberger AS. A course in econometrics: Harvard University Press; 1991.
3. Wooldridge J. Introductory econometrics: A modern approach: Cengage Learning; 2012.