Modeling Violence against Women in India: Theories and Problems

Das, Rituparna

15. March 2010

Modeling Violence against Women in India: Theories and Problems

Rituparna Das

1. Introduction
2. What a model is
3. What a policy is
4. Link between model and policy: policy simulation
5. The issues
6. Is 'violence against women' a variable? What kind of variable is it?
7. Is it theoretically plausible to model 'violence against women'?
8. If it is theoretically plausible to model 'violence against women' then is it feasible to estimate and use such a model for policy purposes in the context of India?
9. Conclusion

1. Introduction

There are two approaches to check the occurrence of a variable (let us call it the control variable), which is not desired:

Utilitarian approach: To impose penalty whenever the control variable occurs[i].

Control approach: To control the variables, which determine the control variable.

The first approach has important legal implications, whereas the second one has numerous social and behavioral implications. Again the first approach relates to cure to a problem, whereas the second one relates to prevention of the problem[ii]. Therefore legislation relates to the first approach and policy-formulation relates to the second approach. This paper analyzes (1) the theoretical plausibility of building a statistical model of the activities falling in the purview of 'violence against women' and (2) the feasibility of working with such a model for the purpose of framing policies in the context of India.

2. What a model is

In the social context, the term 'model' implies a simple description of a system, used for explaining how something works or calculating what might happen. The term 'model' is

used as verb also. In the social context modeling a variable means to determine the nature of relationships between that variable and its determinant variables and make predictions of its value for the periods both within-sample and post-sample. A social system has an underlying 'model' in the above sense. It is a bunch of specified relationships expressed through equations involving a number of social variables. Some of the social variables are socio-economic by nature like the proportion of females in total workforce. There are two ways of classifying these variables in the context of social models:

Mode of determination: If a variable is determined within the model or the system then it is called endogenous variable. If it is determined outside the model or the system, then it is called exogenous variable. Policy variables like government expenditure on education for women are often attached the status of exogenous variable, whereas the rate of divorce, which depends on the behavior of individuals often receive the status of endogenous variable.

Type of values: If a variable takes any numerical values within a certain domain or range, then it is called a quantitative variable. If it takes limited values then it is called a qualitative variable or an attribute, e.g., the degree of discrimination in parental treatment between a boy child and a girl child.

3. What a policy is

Following the Oxford Advanced Learner's Dictionary (sixth edition), a policy can be defined as a plan of action agreed or chosen by either of a political party, a business, a government etc based on a principle that they believe in and that influences how they behave. If fixing the target value or changing the value of a social or economic variable according to the above principle is a part of policy formulation then the variable is called policy variable. A political party or a government designs policies in the context of a particular social system.

4. Link between model and policy: policy simulation

The relationships in a model are specified on the basis of theories and reports. Specification of equations is followed by estimation. Estimation is followed by simulation. There are two types of simulations:

***Ex post*, within sample or historical simulation:** In this method the extent and the direction of change in the endogenous variable following a change in an exogenous policy variable are noted and compared with the past movement of the endogenous variables. This experiment is repeated within the sample period with varying magnitudes and types of policy shocks. That policy shock, which leads to changes in the endogenous variables in a desirable manner, is finally chosen.

***Ex ante*, post sample or futuristic simulation:** The above experiment takes place for a future or post sample period with a view to producing those future values of the dependent variable, which move in the desired direction.

Both of the above two kinds of simulations are conducted with alternative models out of which that one is selected, with which, the simulation practice performs the best or produces the best result in terms of changes in the endogenous variables in a desirable manner.

5. The issues

1. Is 'violence against women' a variable? What kind of variable is it?
2. Is it theoretically plausible to model 'violence against women'?
3. If it is theoretically plausible to model 'violence against women', then is it feasible to estimate such a model and perform simulation exercises?

6. Is 'violence against women' a variable? What kind of variable is it?

'Violence against women' is defined by the United Nations as "….any act or gender-based violence that results in or is likely to result in physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion, or arbitrary deprivation of liberty, whether occurring in private or public life. This definition places 'violence against women' within the context of gender equity as acts that women suffer because of their social status with regard to men. ….. The great majority of perpetrators of violence are men; women are at the greatest risk from men they know" [iii]. Again in the context of India there is no definition of 'violence against women'. It can happen with any individual irrespective of sex and age and its form varies from one to another situation[iv].

What follows from the above is that occurrence of 'violence against women' does not result in any outcome, which is numerically measurable; rather it results in the outcomes like deprivation of liberty, which reflects some quality or attribute of life or living. If this argument is valid, then the decision to indulge in 'violence against women' can be defined as an attribute or qualitative variable of binary choice on part of the perpetrator as per the theory of choice and a value unity i.e. '1' is assigned to the outcome when the decision is 'yes' and a value zero i.e. '0' is assigned when the decision is 'no'[v].

7. Is it theoretically plausible to model 'violence against women'?

In the sociological context, a model is a miniature of an existing social system represented by a bunch of relationships between two sets of variables specified by a set of assumptions. Out of the two sets of variables, one set contains dependent variables and the other set contains independent variables. Modeling a particular variable, involves, in the beginning, to identify the variables, which are affecting it and also to specify, in what direction and to what extent they affect it. Thus one has to identify a functional relationship between the particular variable, i.e. the dependent variable, and other variables (independent variables) determining it[vi]. The existence of a relationship between the two sets of variables can be inferred on the basis of existing references including texts, reports and news-items. From the available literature, one can conjecture a relationship between educations of women (X_1) and proportion of women in aggregate workforce (X_2) on the one hand and on the other hand 'violence against women' (Y)[vii].

The second step involves estimation of the above function: $Y = f(X_1, X_2)$. In the standard literature, two techniques are available for estimation of this kind of models. They are (a) the ordinary least square (OLS) technique in most cases and (b) in a few cases the maximum likelihood (ML) technique, where the OLS technique is difficult to apply. Therefore we have to shape in the next step the form of the function $Y = f(X_1, X_2)$, in a way that is amenable to application of the either of the OLS technique and the ML technique. As per Gaus-Marcov Theorem, both of these techniques yield the most efficient, linear and unbiased estimators.

The third step involves shaping of the function $Y = f(X_1, X_2)$ in a way to make it estimable through application of the either of the OLS technique and the ML technique.

In this function the dependent variable Y is a binary variable and the independent variables X_1 and X_2 are ordinary variables, which takes only positive values. There are four alternative models, which make a function involving dependable binary variables amenable to estimation: (A) The linear probability model, (B) The logit model, (C) The probit model, and (D) The tobit model.

(A) The linear probability model (LPM)

We consider the following simple model:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, (1)

The number of observations is i = 1 to n.

Models which express the binary dependent variable as a linear function of the independent variable(s) are called linear probability models, because E ($Y_i \mid X_{1i,} X_{2i}$), the conditional expectation of Y_i, given X_i is interpreted here as the conditional probability that the event will occur, given X_i, that is Pr ($Y_i = 1 \mid X_{1i,} X_{2i}$). Thus in our case E ($Y_i \mid X_{1i,} X_{2i}$) is the probability of happening of Y when the literacy rate is X_{1i} and the proportion of women in total workforce is X_{2i}. The justification of LPM model is as follows:

Assuming E (u_i) = 0 in order to obtain unbiased estimators, we have

E ($Y_i \mid X_{1i,} X_{2i}$) = $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$ (2)

Now letting P_i = probability that $Y_i = 1$ and (1-P_i) the probability that $Y_i = 0$, the variable Y_i has the following distribution:

Table 1

Y_i	**Probability**
0	1 - P_i
1	P_i
Total	**1**

Therefore by definition of mathematical expectation we have

E (Y_i) = 0(1 – P_i) + 1(P_i) = P_i (3)

Comparing (2) and (3) we equate E ($Y_i \mid X_{1i,} X_{2i}$) = $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} = P_i$.

This means the conditional expectation of the model can be interpreted as the conditional probability of Y_i. Since the probability P_i must lie between 0 and 1, we have the restriction $0 \leq E(Y_i \mid X_{1i,} X_{2i}) \leq 1$.

Problems in estimation of LPM

Since (1) looks like a regression model, one can estimate it by standard ordinary least square (OLS) method. But doing this leads to the following problems:

(i). Violation of normality assumption in small sample cases

(ii). Heteroskedasticity of variances of u_i

(iii). Possibility of $\hat{Y}_i$ falling outside the range 0-1

(iv). Low value of R^2

(i). Violation of normality assumption in small sample cases

The assumption of normality of u_i is no more tenable because u_i takes only two values depending on the value of Y_i as follows:

Table 2

Y_i	u_i	Probability
1	$1 - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}$	P_i
0	$- \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}$	$1 - P_i$

Here u_i does not follow normal distribution. Rather, it follows binomial distribution. However, on the basis of central limit theorem, one can prove that as sample size increases the OLS estimators tend to be distributed normally.

(ii). Heteroskedasticity of variances of u_i

Homoskedasticity of variances of u_i terms can no longer be maintained even if $E(u_i) = 0$ and $E(u_i u_j) = 0$, for $i \neq j$, i.e. no serial correlation. On the basis of table 2, we calculate the variance of u_i. By definition var $(u_i) = E[u_i - E(u_i)]^2 = E(u^2_i)$, for $E(u_i) = 0$ by assumption.

Now var $(u_i) = E(u^2_i) = (- \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 (1 - Pi) + (1 - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 P_i = (- P_i)^2 (1 - P_i) + (1 - P_i)^2 P_i = P_i (1 - P_i)$, (4)

where $E(Y_i \mid X_{1i,} X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} = P_i$

Equation (4) tells that u_i is heteroskedastic. Again, since P_i is a function of X_{1i} and X_{2i}, var (u_i) is dependent on these independent variables and not homoskedastic. In presence

of heteroskedasticity the OLS estimators, though unbiased, are not efficient, i.e. they do not have minimum variance. Here the cure to the problem is dividing both sides of the model by $\sqrt{\{P_i (1 - P_i)\}} = \sqrt{w_i}$. Then the disturbance term would be homoskedastic. Now we may estimate $(Y_i/\sqrt{w_i}) = (\beta_0/\sqrt{w_i}) + \beta_1 (X_{1i}/\sqrt{w_i}) + \beta_2 (X_{2i}/\sqrt{w_i}) + (u_i/\sqrt{w_i})$ (5)

But the true $E (Y_i \mid X_{1i,} X_{2i})$ is unknown and hence w_i, the weights are unknown. In order to estimate w_i we use the following method:

We apply the OLS technique to estimate (1) in spite of the heteroskedasticity problem and get $\hat{Y}_i$ = Estimated $E (Y_i \mid X_{1i,} X_{2i}) = \hat{E} (Y_i \mid X_{1i,} X_{2i})$, and then get $\hat{w}_i = \hat{Y}_i (1 - \hat{Y}_i)$, the estimated w_i. We use $\hat{w}_i$ to transform the data like (5) and run the OLS regression on the transformed data.

(iii). Possibility of $\hat{Y}_i$ falling outside the range 0-1

Since $E (Y_i \mid X_{1i,} X_{2i})$ in the linear probability model measures the conditional probability of the event Y occurring given X, it must necessarily lie between 0 and 1. But there is no guarantee that $\hat{Y}_i = \hat{E} (Y_i \mid X_{1i,} X_{2i})$ will satisfy this restriction[viii]. There are two alternative ways of finding out whether $0 \leq \hat{Y}_i = \hat{E} (Y_i \mid X_{1i,} X_{2i}) \leq 1$ as follows

(a). To estimate the LPM by the usual OLS method and find out whether $0 \leq \hat{Y}_i \leq 1$. If some $\hat{Y}_i < 0$, then they are assumed to be zero. If they are greater than one, then they are assumed to be one.

(b). To devise an estimating technique that will guarantee that $0 \leq \hat{Y}_i \leq 1$. The estimating techniques here may be logit and probit models, which can guarantee that $0 \leq \hat{Y}_i \leq 1$.

(iv). Low value of R^2

The conventional R^2 is not useful in case of binary dependent variable(s), because conventionally computed R^2 would be much lower than unity owing to the fact that Y has two values 0 and 1, corresponding to any pair of X_{1i} and X_{2i}. In most of the cases $0.2 \leq R^2 \leq 0.6$. Therefore we should avoid use of R^2 as summary statistics[ix].

(B) Logit model

The insurmountable problem with LPM is that it is not a logically attractive model, because it assumes that P_i increases linearly with $E (Y_i \mid X_{1i,} X_{2i})$ and LPM does not guarantee a P_i within the range 0 - 1. In reality P_i may increase non-linearly with X and fall outside the range 0 - 1. So, we need an alternative probability model, which is free of these defects. Logit model serves this purpose.

Shape of logit model

Logit model looks like $P_i = E(Y_i \mid X_{1i}, X_{2i}) = 1/(1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}) = 1/(1+e^{-Zi})$, where $Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. As Z_i ranges from $-\infty$ to ∞, P_i ranges between 0 and 1 and is non-linearly related to X_{1i} and X_{2i}.

In order to make P_i amenable to OLS technique we construct a linear relationship

$L_i = ln\ (P_i/(1 - P_i)) = Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, where $1 - P_i = 1/(1 + e^{Zi})$; $P_i/(1 - P_i)$ is odds in favor of occurrence of Y; *L* is called logit. As Z_i ranges from $-\infty$ to ∞, P_i ranges between 0 and 1 and the *L* ranges from $-\infty$ to ∞.

Estimation of logit model

For applying OLS technique, we need data on L, X_{1i} and X_{2i}. Data on X_{1i} and X_{2i} are available from published reports, but we have to generate the data on L by calculating P_i from a reasonably large sample.

If sample size is fairly large and each observation in each of X_{1i} and X_{2i} follows independently a binomial distribution with mean equal to true P_i and variance equal P_i (1 - P_i), then $u_i \sim N\ [0, 1/(N_i P_i (1 - P_i)]^x$. Here u_i is heteroskedastic. So we apply weighted least square technique, where the weight is $w_i = \sqrt{(P_i (1 - P_i))}$.

(C) Probit (normit) model

If the occurrence of Y in the i^{th} family depends on an unobservable utility index I_i, which is determined by X_{1i} and X_{2i}, then the index I_i can expressed as $I_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. Here we assume that for each family there is a critical or threshold level of the index, called I^*_i such that, Y occurs when $I_i > I^*_i$, and vice-versa. Though observations on I_i are not available, information are available that distinguish between two categories of observations: (1). High values of I_i and (2). Low values of I_i. Probit analysis solves the problem of how to obtain estimates for the parameters β_0, β_1, and β_2 and at the same time obtaining information about the underlying index I_i. To focus on this problem let us consider an analysis of the exposure of a typical woman to 'violence against women'. I_i represents the degree of her exposure to 'violence against women', which may be a linear function of X_1 and X_2. Probit model may provide suitable means of estimating β_0, β_1, and β_2. Given that Y represents a binary variable taking values 1 and zero as per occurrence and non-occurrence of 'violence against women' respectively, I_i represents the critical cut off value, which translates the underlying index into facing 'violence against women'.

I_i follows normal distribution. Now $P_i = Pr\ (Y = 1) = Pr\ (I^*_i \leq I_i) = F\ (I_i) = 1/\sqrt{(2\pi)} \int_{t = -\infty \text{ to } T_i} \exp\ (-t^2/2)\ dt = 1/\sqrt{(2\pi)} \int_{t = -\infty \text{ to } \beta 0 + \beta 1\ X1i + \beta 2\ X2i} \exp(-t^2/2)\ dt$, where t is the standard normal variate ~ N(0,1).

From $P_i = F\ (I_i)$, we have $I_i = F^{-1}\ (P_i) = \beta_0 + \beta_1\ X_{1i} + \beta_2\ X_{2i}$, where F^{-1} is the inverse of the normal cumulative distribution function (CDF). A CDF is defined as having as its value the probability that an observed value of a variable X will be less than or equal to a particular X. The range of the cumulative probability function is (0, 1) interval, since all probabilities lie between zero and one. I_i is here known as normal equivalent deviate (n.e.d) or normit. The probability P_i resulting from the probit model has an estimate of the conditional probability that the typical woman would face 'violence against women' given some measures of women literacy and womens' share in total employment.

(D) Tobit model

Tobit model is an extension of probit model developed by Tobin. In this model the families are divided on two groups. We have information on X_{1i} and X_{2i} of one group of size n_1 and we do not have information on X_{1i} and X_{2i} of one group of size n_2. If we run OLS to $Y_i = \beta_0 + \beta_1\ X_{1i} + \beta_2\ X_{2i} + u_i$, using only n_1 observations, the parameter estimates would be biased and inconsistent. So here the model is estimated by using maximum likelihood technique (ML)[xi].

So it is theoretically plausible to model 'violence against women'.

8. What is the specific use of above models? How should we get use of these models?

The above models can provide the following utilities when applied in the way described in each case:

8.1. Whether occurrence of 'violence against women' (Y_i) with a particular woman can be explained by her level of education (N_i) and her level of income (M_i) together or individually can be examined by the linear probability model. Here we can use the linear probability model (LPM). We have to collect information on Y_i, M_i and N_i. Following the established sampling techniques, we can select, say 30 ladies[xii]. Then i = 1 to 30. In each case we note the values of Y_i, M_i and N_i. If the i^{th} individual lady has faced 'violence against women' in her life then $Y_i = 1$, and if she has not faced it then $Y_i = 0$ for

her. Similarly we can assign values 1, 2,…etc to N_i in the ascending order to different levels of education like below:

Level of education (N_i)	**Value**
Did not go to school	0
Studied in primary school only	1
Above primary, but did not pass matriculation	2
Passed matriculation	3
Passed higher secondary	4
Graduate	5
Post graduate	6

Each of above categories can further be divided into a number of sub-categories according to the need of the researcher. Then we note the monthly/annual income of the lady for M_i. M_i can take any positive figure starting from zero. Then we can apply the weighted least square technique to estimate the following LPM:

$Y_i = \alpha + \beta N_i + \gamma M_i + u_i$

Then we test the following hypotheses:

(i). $\alpha = 0$. If α proves to be zero or less than zero, this means 'violence against women' cannot occur in absence of the causes like N and M and we need to drop the intercept α from the above model. If α proves greater than zero, this means 'violence against women' can occur independent of these causes. We reject or do not reject this hypothesis according to whether the calculated *t* value of the parameter α exceeds or does not exceed the table value.

(ii). $\beta = 0$. If β proves zero, this means the level of education is not a determining factor of 'violence against women' and we need to drop the variable N from the above model. On the other hand if β is different than zero, whether positive or negative, this means the level of education is a determining factor of 'violence against women'. Accordingly the government can frame women-education policies. We reject or do not reject this hypothesis according to whether the calculated *t* value of the parameter β exceeds or does not exceed the table value.

(iii). $\gamma = 0$. If γ proves zero, this means the level of income is not a determining factor of 'violence against women' and we need to drop the variable M from the above model. On

the other hand if γ is different than zero, whether positive or negative, this means the level of income is a determining factor of 'violence against women'. Accordingly the government can frame women-employment policies. We reject or do not reject this hypothesis according to whether the calculated *t* value of the parameter γ exceeds or does not exceed the table value.

If all the above hypotheses are rejected then we can use the model $Y_i = \alpha_1 + \beta_1 N_i + \gamma_1 M_i + u_i$, α_1, β_1 and γ_1 are the estimated values of α, β and γ for forecasting occurrence or non-occurrence of 'violence against women'. Using the 'Analysis Tool Pack' software package we accomplish this kind of estimation. We feed the program the data on M and N and set the linear regression of Y on these variables and we get the results under CLRM assumptions. Let us suppose we get $Y_i = 0.76 + 2.5 M_i + 0.9 N_i$. Now for any given pair of values of M and N, we can get the value of Y. If it is negative or zero, we decide non-occurrence of 'violence against women'. If it is more than one, we decide sure occurrence of 'violence against women'. If it is positive but less than one, we decide that there is positive probability of 'violence against women' to the extent of the value of Y.

8.2. Whether the education level alone affects occurrence of 'violence against women' or does it together with income level can be examined with the help of following model:

(a). $Y_i = \alpha + \beta N_i + \gamma M_i + u_{1i}$

(b). $Y_i = \alpha + \beta N_i + u_{2i}$

(c). $Y_i = \alpha + \gamma M_i + u_{3i}$

We estimate equations (a) through (c) applying OLS technique. Let us suppose that the sums of squared residuals are SSR_a, SSR_b and SSR_c respectively computed from equations (a) through (c) and the degrees of freedom of these equations are d_a, d_b and d_c respectively. Now we can construct the follwing *F* statistics:

(i). $F = \{(SSR_b - SSR_a)/(d_b - d_a)\}/(SSR_a/d_a)$

(ii). $F = \{(SSR_c - SSR_a)/(d_c - d_a)\}/(SSR_a/d_a)$

in order to test the following hypotheses respectively:

(i). $\beta = 0$ in (a). This hypothesis is rejected if the computed *F* value exceeds the table value. This means M alone does not cause Y. On the other hand, if this hypothesis is not rejected then we choose the model $Y_i = \alpha + \gamma M_i + u_{3i}$ for forecasting and policy making purpose.

(ii). $\gamma = 0$ in (b). This hypothesis is rejected if the computed *F* value exceeds the table value. This means N alone does not cause Y. On the other hand, if this hypothesis is not rejected then we choose the model $Y_i = \alpha + \beta N_i + u_{2i}$ for forecasting and policy making purpose[xiii].

8.3. Once the model $Y_i = \alpha + \beta N_i + \gamma M_i + u_i$ is chosen on the basis of the exercise described in 8.1, what proportion of a particular number of women with a specified combination of the levels of M and N, selected in course of the standard sampling process faces 'violence against women' can be inferred by the logit model.

Similarly, once the model $Y_i = \alpha + \beta N_i + u_i$ is chosen on the basis of exercise 8.2, what proportion of a particular number of women with a specified level of N, selected in course of the standard sampling process faces 'violence against women' can be inferred by the logit model.

Similarly, once the model $Y_i = \alpha + \gamma M_i + u_i$ is chosen on the basis of exercise 8.2, what proportion of a particular number of women with a specified level of M, selected in course of the standard sampling process faces 'violence against women' can be inferred by the logit model[xiv].

Let us suppose that we have chosen the simplest model: $Y_i = \alpha + \gamma M_i + u_i$. Corresponding to every M_i, we select a sample of size N_i, out of which n_i ($\leq N_i$) number of women, such that the calculated sample probability of any women with income M_i facing 'violence against women' is $P^*_i = n_i/N_i$. If N_i is fairly large (≥ 30), P^*_i is reasonably a good estimate of population P_i. Using P^*_i, we get the estimated logit $L^*_i = ln \{P^*_i/(1-P^*_i)\} = \alpha^* + \gamma^* M_i$. In this case, as already stated, u_i follows binomial distribution with mean zero and variance $1/\{N_i P_i (1-P_i)\}$. So we multiply both sides of the above equation by $w_i = \sqrt{\{N_i P^*_i (1-P^*_i)\}}$ and estimate $L^{**}_i = \alpha^{**} + \gamma^{**} M_i + u^*_i$ applying OLS technique, where $L^{**}_i = wL^*_i$, $\alpha^{**} = \alpha^* w_i$, $\gamma^{**} = \gamma^* w_i$. Now, for any level of M_i, we can get an estimated $L^{**}_i = k$.

L^{**}_i is again a function of P^*_i. For a given value of N_i, we can solve the equation $L^{**}_i = k$ for P^*_i and get the probability of occurrence of 'violence against women' corresponding to some particular income level.

8.4. Once one out of the following models

(a). $Y_i = \alpha + \beta N_i + \gamma M_i + u_{1i}$

(b). $Y_i = \alpha + \beta N_i + u_{2i}$

(c). $Y_i = \alpha + \gamma M_i + u_{3i}$

are chosen, the degree of exposure to 'violence against women' of a woman with a particular level education N and (or) income M can be estimated with help of the probit model.

Let us suppose, as before, that we have chosen the simplest model: $Y_i = \alpha + \gamma M_i + u_i$. Corresponding to every M_i, we select a sample of size N_i, out of which n_i ($\leq N_i$) number of women, such that the calculated sample probability of any women with income M_i facing 'violence against women' is $P^*_i = n_i/N_i$. Corresponding to every P^*_i we can compute an I^*_i, using the standard normal cumulative distribution function (CDF) table[xv]. Then we can apply OLS technique to estimate the model $I_i = \alpha + \gamma M_i + u_i$. With the help of the estimated model, $I^{**}_i = \alpha^* + \gamma^* M_i$ we can estimate I^{**}_i for any value of M_i. The higher the value of I^{**}_i, the greater the probability that a woman with a specified income level would face 'violence against women'[xvi].

8.5. In a cross section sample of women with different levels of income and/or education, where some of the selected units (women) have not faced 'violence against women', application of OLS technique does not help find any meaningful relationship between occurrences of 'violence against women' as pointed out by James Tobin. Here we can use tobit model to estimate the relationship(s) between occurrence of 'violence against women', and education of level or/and income level of a woman.

For simplicity, let us consider the model $Y_i = a + b N_i + e_i$, Y and N are as defined earlier, e is the disturbance term, 'a' is intercept parameter, 'b' is the slope coefficient reflecting the impact of rise in female education level by one step on the occurrence of 'violence against women' and the number of observations is n, i.e. i = 1 to n. For $Y_i = 1$, the

associated level of N_i is proposed to be compatible with occurrence of the event; but for $Y_i = 0$, we do not know, what level of N_i would have been compatible with occurrence of 'violence against women'. This model is called tobit model or censored regression model. Estimates of 'a' and 'b' obtained from application of OLS technique would be obviously biased and inconsistent in this case, because $e_i = - a - b N_i$ for $Y_i = 0$, and $E(e_i) \neq 0$. So we require application of maximum likelihood method here. Here e_i, called the censored regression error term.

The probability density function of e_i is

$f(e_i) = f(e_i \mid e_i = - a - b N_i) = f(e_i) / \int_{\text{from } (- a - b Ni) \text{ to } \infty} f(\theta)\, d\theta$

Now $E(e_i \mid e_i = - a - b N_i) = \sigma f(a + b N_i) / F(a + b N_i) = \sigma \lambda_i$, σ is the standard deviation of the true error term e_i, f is the probability density function of a standard normal variable and F is the associated CDF[xvii]. λ_i is called the rate of social hazard. We use the estimates of λ_i to normalize the mean of e_i to zero and hence obtain consistent estimators of the parameters. Here we have to use a two-stage estimation process that can yield consistent estimates of the parameters. First, we estimate λ_i by utilizing the probit model $P_i = F(a + b N_i) = F(Y_i)$. This model can be estimated by using the ML technique by distinguishing those observations with Y = 1 from those with Y = 0. Now we calculate λ_i by using the normal distribution table. In the second stage, we estimate the model $Y_i = a + b N_i + \sigma \lambda^*_i + v_i$, where estimated $\lambda_i = \lambda^*_i$, v_i is the random disturbance term; Y, N, a, b and σ are defined as earlier. As sample size approaches infinity, λ^*_i approaches λ_i, $E(v_i)$ approaches zero and ML estimation of the above model gives consistent estimates of 'a' and 'b' parameters[xviii].

9. Is it feasible to estimate and use such a model for policy purposes in India?

Once the job of model construction is accomplished, estimation of the model depends crucially on availability of data. For the purpose of policy simulation both of static and dynamic time series data are needed, whereas for interspatial comparison cross section data is needed. Gauging the success or the failure of a policy measure in terms of its temporal impact on the control variable requires availability of time series data. In the context of India, Census conducted by the Registrar General of India is the source of cross section data on women-literacy and employed women in all the states. In Census,

the number of figures in a series is equal to the number of states. But, for model-estimation purpose one needs sizeable data for having sufficient degree of freedom. Cross section data as they exist in India do not provide this facility[xix].

Regarding employment data, we note that the term 'employment' covers all employment in primary (agriculture), secondary (industry) and tertiary (service) sectors, which are again broadly divided into organized/formal and unorganized/informal sectors. Again the factor income approach to gross domestic product (GDP) accounting has to take into account employment in primary (agriculture), secondary (industry) and tertiary (service) sectors[xx]. The difficulty of obtaining data from the unorganized sector including agriculture and many areas of industrial and service sectors proves formidable[xxi]. If the cross section data were collected at the grass root level, i.e. block level, then sufficient degrees of freedom would have been available. The standard sources of information like Economic Survey and India Development Report do not provide time series data on female-literacy and female-employment. On the other hand absence of definition of 'violence against women' reflects the fact that there has not taken place sufficient research on 'violence against women'. The logical corollary is that data on 'violence against women' is not systematically compiled in India and consequently the kind of 'data explosion' that has happened in England and Wales in the context of social and socio-economic variables has not happened in India. The reason may be that there does not exist enough demand for these data so as to give one incentive to compile them either in the government level or in the private level or in the NGO level. In other words there are not perhaps sufficient buyers of such data[xxii]. Otherwise data on 'violence against women' could have been procured from the records of police station and family courts. If such data were readily available, then on the basis of cross section data one could estimate the probability of occurrence of 'violence against women' in a typical household of India as well as other countries and make a cross-country comparison. Further, the index of exposure of a typical woman to 'violence against women' could be included as the fourth indicator of human development index, because it relates to the safety of women in the society. But, there is no planned effort on part of the Ministry of Statistics and Program Implementation. The Report of The National Statistical Commission,

though has overlooked this issue, has admitted the serious deficiencies of the Indian Statistical System.

10. Conclusion

So we conclude the following:

1. The decision to perpetrate 'violence against women' is a binary variable, which takes value unity (1) when the decision is 'yes' and zero (0) when the decision is 'no'.

2. It is theoretically plausible to construct the models of estimating and forecasting the probability of occurrence of 'violence against women' facing a typical woman in a particular society on the basis of necessary information.

3. It is not feasible in practice to apply above models for the purposes of policy-formulation and policy-simulation in India because of absence of compilation or systematic compilation of the data on 'violence against women' and the variables determining 'violence against women'.

Endnotes

[i] This approach has its origin in Bentham. See more in Bentham J (1986): *Theory of Legislation*, N. M. Tripathy Pvt Ltd, Delhi, (Foreword: P. M. Bakshi)

[ii] "From the beginning in 1829 the constables of the new Metropolitan Police were informed that their first duty was the prevention of crime." - Emsley C (2002): 'The History of Crime and Crime Control Institutions', in Maguire M, Morgan R and Reiner R ed. (2002): *The Oxford Handbook of Criminology,* The Oxford University Press, Oxford, p 213.

[iii] United Nations Declaration on the Elimination of Violence Against Women - United Nations General Assembly, 1993

[iv] Jaising I. ed. (2001): *Law of Domestic Violence*, Universal Law Publishing, Delhi, Chapter 1

[v] Mcfadden D (1973): 'Conditional Logit Analysis of Qualitative Choice Behavior' in P. Zarembka (ed.) *Frontiers in Econometrics*, Academic Press, New York

[vi] Hanushek E and Jackson J (1977): *Statistical Methods for Social Scientists*, Academic Press, New York, Chapter 2

[vii] Majumdar S (2003): 'In India Domestic Violence Rises with Education' *Womenenews*, June 11, 2003

[viii] $\hat{Y}_i$ is the estimated value of Y.

[ix] Aldrich J and Nelson F (1984): *Linear Probability, Logit and Tobit Models*, Sage Publications, California, 1st Edition, p 15

[x] Theil H (1970): 'On the Relationships involving Qualitative Variables', *American Journal of Sociology*, July

[xi] Amemiya T (1984): 'Tobit Models: A Survey', *Journal of Econometrics*, January

[xii] There are ten assumptions of the classical linear regression model (CLRM) of application of the ordinary least square technique (OLS) on the basis of which the linear probability model, a modified version of CLRM, is built up. The seventh assumption is that the sample size is greater than the number of parameters in the model. In our case the there are three parameters. So the sample size can be at least four. See more in *http://www.statsoftinc.com/textbook/sttimser.html*

[xiii] Ibid.

[xiv] Ibid.

[xv] Ibid.

[xvi] Supra no vi, p 189

[xvii] Heckman J (1979), 'Sample Selection Bias as a Specification Error', *Econometrica*, January

[xviii] Supra no x

[xix] On socio-economic statistics, the Secretary of the Department of Statistics (DoS) points out to the failures of the (Administrative Statistical System) AdSS, and the lack of manpower resources devoted to education statistics by states, the unsatisfactory state of the civil registration system, of vital statistics and, of course, of gender statistics. The secretary mentions the collapse of AdSS only in the case of statistics of agriculture, industries and labor. The other failures of the (Indian Statistical System) ISS he points out are on the counts of data gaps, conflicting statistics, lack of timeliness, non-adoption of statistical standards, insufficient periodicity or frequency, questionable reliability, and the (catch-all count of) quality, and duplication. No statistics is found without a defect.

Again there are serious gaps in education statistics as per the Department of Statistics. Till recently the Annual Estimates of Literacy rates for the country and states/UTs were not available. In this sector lot of investments are made but in respect of literacy rate, only the 1991 census figures are being quoted. The

DoS has taken the initiative to process the data available from NSSO surveys to generate Annual Estimates of Literacy. Other data on education statistics are those based on the administrative returns collected from schools. These statistics show unrealistic figures of gross enrolment rate above 100 per cent. The data collected in the Sixth All India Educational Survey were made available only after a gap of six years. While the Department of Education decided to participate in the OECD/UNESCO pilot project of World Education Indicators requiring supply of data on 43 indicators, data on most of them are not available.

Factories are also required under the act to file every year certain annual returns with the CIF, providing data that are relevant to the administration of the Factory Act and other related acts such as the Payment of Wages Act, Maternity Benefit Act, Trade Unions Act, and the Workmen's Compensation Act. These returns are also the main source of data about employment and conditions of labor in the important sector of organized manufacturing. The position about the collection of these returns is also very unsatisfactory affecting adversely the quality and timeliness of labor statistics in this important sector based on them (NSC Report Paragraph 9.2.20) - Vidwans S. M (2002): 'Indian Statistical System at the Crossroads', *Economic and Political Weekly*, September 14

[xx] Experts had pointed out the conceptual and operational problems in the generation of quarterly estimates of the GDP and of employment and unemployment, and the lack of relevance of the quarterly estimates of unemployment to SDDS (Special Data Dissemination System) requirements of IMF - Ibid.

[xxi] The Secretary of the DoS points to the unsatisfactory data collected through sample surveys (NSSs) of the unorganized manufacturing sector, taken as a follow-up of economic censuses. The story is the same for the unorganized services sector, mainly constituting the informal sector. The secretary points to the "absence of legal provisions, education, and lack of importance attached to statistics by the informants and by the unincorporated private sector" as reasons for the unrealistic estimates from the large-scale regular surveys (NSSs). The data from the registered companies is also unsatisfactory. The non-availability of data for non-profit institutions serving households (NPISHs) or of even a frame to plan a sample survey for them does not escape the secretary's criticism. After being convinced about the present sorry state of affairs of statistics in production sectors for annual GDP estimation, the DoS has embarked on a program of quarterly GDP estimation - Ibid.

[xxii] A diametrically opposite situations exists in the developed countries like England. "The rapid development of electronic data storage in the last quarter of the twentieth century has fuelled what might be called 'data explosion in almost every are of public and private life. This is certainly true of crime and justice… What is most interesting …is …the extensive range of 'crime problem' that are now being carefully measured… The focus here is upon systematically gathered (and mainly quantitative data)….However, the more systematic kinds of data directly inform policy-making…." Maguire M (2002): 'Crime Statistics: 'The Data Explosion' and Its Implications', in Maguire M, Morgan R and Reiner R. ed. (2002): *The Oxford Handbook of Criminology,* The Oxford University Press, Oxford, p 324.
