SCHOOL OF DEVELOPMENT STUDIES
RESEARCH REPORT No. 84

ANALYSIS OF UNMATCHED DATA USING PROPENSITY SCORES

PART 1: CROSS-SECTION ANALYSIS

Louis Munyakazi, Vaughan Dutton
and Julian May

October 2010

Analysis of unmatched data using propensity scores
Part 1: Cross-section analysis

First published by the School of Development Studies in 2010
ISBN 978-1-86840-697-8

Available from the website: www.sds.ukzn.ac.za/
Or
The Librarian
School of Development Studies
University of KwaZulu-Natal
Howard College Campus
Durban 4041
SOUTH AFRICA

Tel: +27 31 260-1031

The School of Development Studies is one of the world's leading centres for the study of the political economy of development. Its research and graduate teaching programmes in economic development, social policy and population studies, as well as the projects, public seminars and activism around issues of civil society and social justice, organised through its affiliated Centre for Civil Society place it among the most well-respected and innovative inter-disciplinary schools of its type in the world

We specialise in the following research areas: civil society; demographic research; globalisation, industry and urban development; macroeconomics, trade and finance; poverty and inequality; reproductive health; social aspects of HIV/AIDS; social policy; work and informal economy.

Rationale

The analysis of observational data usually lacks balance in the observable (and non observable) variables. Such unbalance-ness often results in the incorrect results when testing for the differences between a *treatment* group and a *control* group (our *primary factors* of interest) especially when such a comparison is solely based on their observed responses. The reason is that the responses are to some degree "*contaminated*" by the so-called baseline covariates. It is therefore imperative to find an alternative validation approach analogue to the *randomization* technique required in the classical experimental design. The randomization balances for the *secondary factors*-observable and unobservable variables- and prevents bias in subsequent analysis. Without randomization, there is no guarantee that the results are unbiased.

In many practical situations randomization is not feasible (for ex. when testing for the effects of smoking, exposure to chemicals, left eye and right eye measurements, couples preferences, before and after measurements on the same individual etc…). To help alleviate the problem, fundamental work in the field was done by Cochran (1953), Cochran (1968), Althauser and Rubin (1970), Cochran and Rubin (1973), to name the few. To date, a satisfactory approach that leads to reasonable results is the one based on *matching* the primary variables of interest on propensity scores. The idea is to remove (or minimize) the effects due to the secondary variables of age, race, social status, gender and other demographic characteristics. The expectation is that the distribution of the variables is ultimately similar in both groups and therefore the remaining *signal* in the data is primarily due to the treatment effect alone.

The statistical analyses in the cross section enable us to get insight knowledge of each wave and provide, at the same time, a better understanding on how to construct the model applicable to the *combined* (entire panel) data. *At first,* a *logistic regression model* is used to select the most appropriate model and obtain the propensity scores. *Second* we determine the groups defining the *stratification* based on the quantile approach. *Third* a series of t-test is used to evaluate the effectiveness of the stratification via, among other things, the *percent reduction in bias*. *Fourth,* a combination of stratification and an *ANalysis of COVAriance* (ANCOVA) is used along the *geometric mean* regression method to derive the relationship between one dimension of ICT (ownership) and one dimension of poverty (income). *Two alternatives* to the geometric regression are introduced. In addition, we provide a simple intuitive method to deal with missing data typical in this type of studies. *The last* section of the report provides the SAS® codes used to implement the above and the SAS® log that shows the program ran successfully. Some of the SAS codes

provided in the text assumes the new version of SAS® 9.2. However the analyses were done in SAS® 9.1.3 currently available at UKZN.

Fitting the logistic regression to obtain propensity scores

For *binary response* (i.e. response with outcome equals to 0 or 1), the propensity scores are obtained from a *logistic regression* model (SAS, 2009). This calls for a transformation to the probabilities because the relationship between the probabilities and the covariates is *not linear*. The logistic regression model formulation specifies that the probability of the "*event*" ("1") is related to the associate secondary (or exposure) variables $x_1, x_2, \ldots, x_m$ via a link function:

$$\log\{p/(1-p)\} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_m\beta_m + \text{selected interactions and quadratics} \quad (1)$$

and

$$p = 1/e^{(\beta 0 + x1\beta_1 + x2\,\beta 2 + \ldots + xm\beta m + \text{selected interactions and quadratics})} \quad (2)$$

The estimates of the treatment are adjusted by means of a model relating the dependent variable *g(p)=log {p/(1-p)}* to the confounding variables $x_1, x_2, \ldots, x_m$ and selected interactions plus some quadratics (Joffe and Rosenbam, 1999). The logit *g(p)* is the log of the odds, *log {p/(1-p)*. The log odds are written as an intercept (β_0) plus a combination of exploratory variables multiplied (x_s) by the appropriate parameter values (β_s). The propensity score are useful in the reduction of bias and increase of precision because they create a "*quasi-randomized*" experiment (D'Agostino, 1998). In other words, if two subjects have the same propensity score, they could be thought as if they have been assigned randomly to either the treatment or the control. This is a very important property one needs to exploit prior to matching or stratification.

The logistic regression fit the selected model by maximum likelihood (ML) assuming the underlying assumptions are satisfied (Pregibon, 1984; SAS, 2009; Allison, 2005). The estimates β_0 and β_S are such that all the values of ($X\beta$) in ($-\infty$, $+\infty$) map into (0, 1) for p. In essence the predicted probabilities are made to only be between 0 and 1. Therefore there are no possible predicted values that are either negative or greater than 1. When fitting the model, SAS (2009) uses Fisher scoring method. This method is equivalent to model fitting with iteratively weighted least squares (Stokes et al, 1995).

Model Selection for Propensity Estimation

The implementation of the above model (1) can be done in many ways. The following statements in SAS (2009) can be applied:

```
proc glmselect;
      effect MyPoly = polynomial(x1-x3/degree=2);
      model y = MyPoly;
run;
```

An identical analysis is obtained by

```
 proc glmselect;
   model y = x1 x2 x3 x1*x1 x1*x2 x1*x3 x2*x2 x2*x3 x3*x3;
  run;
```

In our study however, the logistic model defined in (1) used at least 11covariates together with selected interactions and quadratics. This step removes not only bias in the *original 11 covariates* but also most of the bias in their *squares* and *paired-wise interactions* (Joffe and Rosenbaun, 1999). The selection of interactions and the quadratics can be specific (by intelligent choice) or can be selected on the basis of their contribution to the overall fit thru modeling. At this stage, the main objective of computing the propensity scores is to create *balance* between the interviewed and the non-interviewed individuals, not to make any *inferential statement* about the two groups.

The PROC GLMSELECT is called to perform effect selection in the framework of general linear models.

```
proc glmselect data=one plots=all outdesign=---;
          class Country EnumerationArea;
       effect MyPoly = polynomial(x1-x11/degree=2);
model y  = MyPoly Country EnumerationArea
    /details=all stats=all selection=stepwise(choose= adjrsq);
run;
```

The following PROC REG produces a useful set of regression diagnostics corresponding to the model selected by PROC GLMSELECT above

```
ods graphics on;
  proc reg data=----;
    model improvement = &_GLSMOD;
  quit;
   ods graphics off;
```

The Statistical Analysis under the No Matching

Stratification (or sub-classification) on the Propensity scores

The propensity scores *e(X)=prob(Z=1|X)* are estimated using the logit regression model on the 11 so-called *baseline covariates* plus some selected interactions and quadratics (see model 1). We have several ways to use these scores in the analysis without having to spit them into "treated" and "control" group and do the matching. We present below some of the available methods to deal with the analysis of unmatched data. At first, one can take advantage of the selection model mentioned above using PROC GLMSELECT to retain variables of *significant* interest. Failing this important step can result in a model with 11 main effects, 11 quadratics effects and 55 first order interactions effects for a total of 77 terms. To reduce such a large number of items and bring it to a *manageable* size, we use the model selection procedure described before (see **Model Selection for Propensity Estimation**).

The following SAS codes is then applied to run the final model defined in (1) and save the propensity scores for further use.

```
proc logistic;
  class country EnumerationArea / param=ref;
  model stay_home(event="1") = (selected among the 12 covariates
                    and possibly their interactions and quadratics);
output out=preds predprobs=individual;
run;
```

The selected variables are not necessarily the ones that are "*statistically significant"* but rather *"practically significant"*. This suggests that, in addition to the above statistical approach to model selection, one may consider and retain meaningful terms. It is recommended to categorize the propensity scores using (1) *quintiles*, (2) *spline*, or (3) a locally weighted scatterplot smoothing or *loess* smooth of the scores is an alternative allowing several *degrees of freedom* (SAS, unknown). Finally one may use a model of *covariate adjustment* using propensity scores.

1. Quintile approach to stratification

The Individuals scores can be divided into in five strata each stratum containing *20% of the individuals*. This method was suggested by

Miettinen (1976) and Marshall and Rosenbaum (1999) and subsequently adopted by many researchers (Austin, 2009; Stone et al, 1995; Rosenbaum and Rubin, 1984). The process is expected to generate strata that are homogeneous *within the same stratum;* meaning the interviewed and non-interviewed groups are represented and should have similar (overlay) distributions of the covariates. Therefore, the two groups within such defined stratum are deemed directly comparable. In fact Rosenbaum and Rubin (1983) demonstrated that for a *perfect* stratification based on the propensity score, the average treatment effect within stratum is an *unbiased* estimate of the true treatment effect.

2. The spline approach to stratification

Another approach to sub-classification is to find homogenous strata thru the use of the *spline function* (Harvey Goldstein and Huiqi Pan, 1992). A spline function is a piecewise polynomial functions made of individual polynomials. These polynomials connect smoothly at join points known as *knots*. The basic idea for using the *spline* is to categorize the individual propensity scores and use the categories in the manner similar to stratification described above. The following SAS statements can be used to identify strata (SAS, 2009).

```
proc glmselect data=---;
     effect spl = spline(x / knotmethod=equal 4)
                          split details);
model propensity scores = spl;
output out=out1 p=pBumps;
run;
```

or within the generalized mixed model framework,

```
proc glimmix;
     class ---- ;
     effect spl = spline(propensity_scores);
     model () = spl (and other factors);
run;
```

In the above procedures, the columns of *spl* are formed from the variable "*propensity_ scores"* as a cubic B-spline basis with *four equally spaced interior knots* (SAS, 2009). For consistency, *w*e propose, if possible, to maintain the same number of strata (5) as we did in the partitioning of the propensity scores based on quantiles. A plot of the propensity scores can help identify a useful number of interior knots. Since the knots must be *pre-specified*, a visual aid is useful thru the following SAS statements:

```
proc sgplot data=---;
     scatter y= propensity_scores     x=ordered_prop;
run;
```

3. Using the LOcally Weighted Scatterplot Smoothing (LOESS or LOWESS):

The LOESS method is useful for situations in which we don't know the parametric form of the regression surface. For our purpose, we use the procedure solely as a visual tool to identify the number of knots (if any). We "*smooth*" the propensity scores as function of created variable called *ordered_prop* (in the data set, the responses are pre-ordered from *max* to *min* and an ID is given). From there we want to capture the *periodic pattern* in the propensity scores and use it for stratification. For consistency with the previous approaches, a smoothing parameter that gives five strata is preferable. However, we should strive to create, as much as possible, homogeneous strata that are suggested by the data.

```
ods graphics on;
  proc loess data=----;
    model  propensity_scores =ordered_prop/ smooth=0.1
0.25 0.4 0.6 residual;
  run;
```

Statistical analysis of Stratified data

The analysis is the same for all the three stratification methods described above (quintile, spline and lowess). In order to estimate the average difference Δ_k, in the outcome of interest (i.e. "treated *vs.* untreated in ICT"), one must first calculate the difference *within* each stratum and then *sum* them over the strata:

$$\Delta = \frac{1}{k}\sum_{k=1}^{5}(mean\ of\ \text{treated} - mean\ of\ control)$$

k i indexes the propensity score stratum. The standard error of Δ, namely SE_Δ is obtained from the pooled stratum specific variances

$$SE_\Delta = \sqrt{\sum_{k=1}^{5}\frac{1}{K^2}\left(\frac{s^2_{k\_trt}}{n_{k\_trt}} + \frac{s^2_{k\_cont}}{n_{k\_cont}}\right)}$$

where n_{k_trt} and n_{k_cont} are the number of treated and control individuals in each stratum respectively. The numbers of observations in each group are not expected to be equal neither between nor within stratum. The ratio of Δ to its SE follows a *t*-distribution and therefore can be used to test the hypothesis whether Δ equals zero at a given α. In general, for a given *estimate* y_i *(the estimate could be a mean, a proportion, a slope, or a difference between means or proportions), the weights* are the reciprocal of its *standard error*

$$w_i = \frac{1}{(SE(y_i)^2}$$

The weighted mean is then

$$\bar{\bar{y}} = \frac{\sum_1^k w_i y_i}{\sum_1^k w_i} \qquad (5)$$

The standard error of $\bar{\bar{y}}$ is given by

$$SE_{\bar{\bar{y}}} = 1/\sqrt{\sum_1^k w_i} \qquad (6)$$

The ratio between (4) and (5) is distributed as a standard normal.

Covariance adjustment using the Propensity scores

Another popular method that makes use of propensity scores is the *covariate adjustment method*. In this approach the treatment effect on a particular outcome is evaluated thru a regression analysis *with dummy variable*. With a treatment indicator variable, Z=1 for the treated and Z=0 for untreated groups respectively, one can fit a *multiple regression model* that includes both Z and the propensity score as follow:

$$y_i = \alpha_0 + \beta Z_i + \alpha_1 e_i + \epsilon_i \qquad (7)$$

where y_i is the outcome, Z and e_i denote the treatment and the propensity score respectively. The parameters α_0,β , and α_1 are the coefficients and ϵ_i are residuals assumed to be normally distributed with mean 0 and variance equals to σ^2 . The coefficient β can be interpreted as a measure of change in the outcome *due to treatment*. One direct application of this formulation is the effect of the variable *ICT* usage on *income*. In this case, equation (7) models the *ICT* as the *treatment effect* Z, and y_i represents *income*. The parameter estimates of α_0,β , and α_1 are obtained thru *Ordinary Least Squares* (OLS) fit. For *binary outcome*, one may use the approximately equivalent *weighted least squares* regression model on the *logit*. The appropriate weights, given by

$$w \approx np(1-p), \qquad (8)$$

allow to perform a non-iterative weighted OLS fit to the logit. The relevant hypothesis is H_{05}: $\beta = 0$. If one rejects the null hypothesis at a predetermined alpha level (usually $\alpha = 0.05$), the effect of ICT on the income is accepted. Note that the *covariate adjustment method* can be applied within each stratum assuming they are enough observations *and* both groups are represented. In this case, average treatment effect is the average of the within stratum effects. Rosenbaum (1994) showed that this approach "appears to be a more efficient estimator (of treatment effect) that one based on matching alone". Eq (7) can actually include an interaction term $Z * e_i$ (test for parallelism) and if significant, it would suggest that either (1) individuals who had high propensity to be interviewed *but were not* were more likely to have

different income (higher or lower) than those who were interviewed assuming the income is the dependent variable, or (2) assuming the ICT is the dependent variable, the individuals with high probability of being interviewed *but were not* have different (high or lower) ICT ownership. When the group variances are different, propensity score methods for matching and stratification are preferable (D"Agostino, 1998).

The Geometric Mean Regression

The geometric mean functional relationship approach, or GMFR, is applied to linear regression problem when both axes are subject to errors. The usual regression method used to derive coefficients of the line assumes that one variable (the predictor) is measured without error and that only the response is subject to variability. The ordinary least squares method (OLS) is fitted to data by minimizing horizontal deviations to the line.

1. The Usual Regression Model:

Classical linear regression theory requires that for a given set of ***n*** data values (X_i,Y_i), these measurements are such that only Y_i is subject to error whereas X_i is correctly defined without error. In that case, the vector of residuals, ε_i is said to be independently and normally distributed with mean equals to zero and a common variance set to σ^2. The classical representation of the above is IID N $(0, \sigma^2)$. The usual model formulation is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \ldots, n \qquad (9)$$

In matrix form, the above equation is

$$Y = X\beta + \varepsilon \qquad (10)$$

where

Y is the vector of observed measurements (error is inherent to experimental responses and is random by definition)

X is the vector of a *known* form thus is without error

β is the vector of *unknown* parameters (β_0 and β_1). They must be estimated from the data.

Solution to (9) or (10) is uniquely and easily obtainable:

$$\beta = [\beta_0 \ \beta_1]`$$

where

$$b = (X`X)^{-1} X`Y$$

b is an estimate of β that minimizes the error sums of squares $\varepsilon`\varepsilon$ $(= \sigma^2)$. The solution vector b is software *independent* since it is based a *full rank* matrix X from which $(X`X)^{-1}$ is unique.

2. When both Y and X are Subject to Error:

For situations in which both Y and X are subject to error, a different approach must be implemented to derive the solution vector b. In order to describe this particular case, equation (1) is transformed into two equations:

$$Y_i = \eta_i + \varepsilon_I \qquad (11)$$

$$Y_i = \xi_i + \delta_i$$

each illustrate the errors associated with the measurement methods (in our case, *Income* and *ICT poverty*). In addition, we can assume a true relationship exists between η_i and ξ_i:

$$\eta_i = \beta_0 + \beta_1 \xi_i$$

There are no easy solutions to the issue above when both variables X and Y are measured with error. However, with additional assumption on the unknown parameterλ, some practical solutions, among which, the use of the maximum likelihood solution is the best:

$$b_1 = [S_{YY} - \lambda S_{XX} + \{(S_{YY} - \lambda S_{XX})^2 + 4\lambda\, S^2_{XY}\}^{1/2}\ /(2S_{XY}) \qquad (12)$$

$$b_0 = Y_{mean} - b_1 X_{mean} \qquad (13)$$

In the (somewhat unrealistic) assumption that $\lambda = S_{YY}/S_{XX}$, the above equations are equal to:

$$b_1 = S_{XY}/S_{XX},$$

$$a^{-1} = (S_{XY}/S_{YY})^{-1}$$

where b_1 is the *least squares fit of Y vs. X*

$$Y = b_0 + b_1 X \qquad (14)$$

and a_1 is the *least squares fit of X vs. Y*

$$X = a_0 + a_1 Y \qquad (15)$$

We can then invert (4) to have the *same* form as (3),

$$Y = a_0/a_1 + a^{-1} X \qquad (16)$$

Once both equations are in the same form, an appealing solution is to use the *Geometric Mean Functional Relationship***.** The slope estimate from the geometric mean functional relationship represents a compromise value lying between the two "Y and X" slopes (eqs. 14 and 15).

It is proposed that the geometric mean functional relationship of both the slopes and the intercepts estimated from (14) and (16) *be used in lieu of* either (14) or (15). Such a procedure results in a *unique* solution vector regardless what is used as X or Y (Draper and Smith, ; Barker et al, 1988). Therefore, if the roles of X and Y are reversed, *exactly the same equation* is found. The geometric mean functional relationship has its slope equals to

$$Slope = (b_1 * a^{-1})^{0.5} = sign(S_{XY})\sqrt{\beta_{OLS\,Y\,on\,X}/(\beta_{OLS\,X\,on\,Y})} \quad (17)$$

and its intercept equals to

$$Intercept = \bar{Y} - Slope * \bar{X} \quad (18)$$

The fitted line Y = *intercept* + *Slope*X* is uniquely defined. Moreover, given the difficulties associated with the computation of the confidence intervals and/or test of hypotheses on the parameters (*intercept* and *Slope*), the geometric mean regression has a *natural symmetry* that we can exploit. This symmetry provides information whether the two dimensions of poverty (Income and ICT poverty) are *related to one another*. This is only possible after both equations are written in Y on X form (eqs. 14 and 16 above).

3. The Geometric Mean Functional Relationship within ANCOVA-stratum combination

In order to implement the above, we employed the analyses of covariance adjustment method performed within each stratum following the equation:

$$Y = X\beta + \epsilon \quad (19)$$

Where Y is the *income* and X is the design matrix of *Interviewed* (interviewed vs. non-interviewed), *country* (Uganda, Rwanda, Kenya, and Tanzania), *environment* (rural vs. urban *nested* within country). *gender* (male vs female), *ownership of ICT* (as a 3-level covariate), the interaction between *ICT and environment nested within country,* and propensity score. The result is a set of 11 slopes of income on ICT per stratum for a total of 55 slopes. The following SAS codes are used:

```
proc glm data=stratglm&m;
        class Interv country rural_urban gender;
model &dep = Interv country rural_urban (country) gender
&indep*rural_urban (country) propensity&n/ss3 solution;
estimate "Slope TZN Major Urban" &indep*rural_urban(country) 1;
estimate "Slope TZN Other Urban" &indep*rural_urban(country) 0 1;
estimate "Slope TZN Rurual" &indep*rural_urban(country)      0 0 1;
estimate "Slope KNY Major Urban" &indep*rural_urban(country) 0 0 0 1;
estimate "Slope KNY Rural" &indep*rural_urban(country)       0 0 0 0 1;
estimate "Slope RWD Major Urban" &indep*rural_urban(country) 0 0 0 0 0 1;
estimate "Slope RWD Other Urban" &indep*rural_urban(country) 0 0 0 0 0 0
1;
estimate "Slope RWD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0 0
1;
estimate "Slope UGD Major Urban" &indep*rural_urban(country) 0 0 0 0 0 0
0 0 1;
estimate "Slope UGD Other Urban" &indep*rural_urban(country) 0 0 0 0 0 0
0 0 0 1;
estimate "Slope UGD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0 0 0
0 0 1;
ods output estimates=Slope&indep;
```

```
title Geometric Mean Regression;
run;
```

In the above, a macro is used to reverse the regressors in X once as ICT and then as Income (see &dep and &indep macro specs) resulting in a total of 109 number of slopes. The vector β is the solution vector from set-to-zero restrictions on the parameters. The above "L "vectors are easily estimable functions. They were used to recover the slopes in each combination of country-environment. However, the combined estimates did not follow the procedure described in eqs. (5) and (6) because the $[SE(\beta]_i)^{\square}$ are not available (at the moment). These will be implemented latter in line of the recommendations by Gillard, (2003). A note for caution is that the income variable was not check for normality as it is quite known that this variable is skewed and a log transformation should always be applied. Such a transformation will be used in the combined wave analyses (panel data). Note that the GMFR is on two variables (income and ICT). However each of the variables is part of a multi-dimensional approach to the issue of ICT and its relation with the reduction of poverty, i.e. ICT includes *access to, and use of, number of episodes, expenditure on ICT, number of applications* whereas poverty includes *financial, education, physical, vulnerability, capability, exclusion, and services.*

There are potentially two approaches that one may investigate: (1) the most obvious is to use *Principal Component scores* leading to PC_{ICT} , and $PC_{poverty}$ and proceed as above or (2) use a *generalization* of the geometric mean functional relationship (Draper and Yang, 1997) and consider the measurement error model for multi-variable regression of Y on many X_s of the other variable.

4. Useful Alternatives to the Geometric Mean Regression

When dealing with cross-section studies, one of the objectives is to measure the relationship between two or more variables at a particular point in time (Levy and Lemeshow, 1999). Because the measurements in variables are always prone to errors, the ordinary least squares solution is not appropriate. In the PICTURE survey, this is true especially in regard to *essential* variables such as income, all the proxies of multi-dimensions poverty and/or their principal component scores. Fortunately, alternative models exist to accommodate this situation. Here are two in addition to the geometric mean regression (Carroll and Ruppert (1996):

1- Classical Orthogonal Regression
2- Method of Moment estimator

The classical Orthogonal Regression (and the Method of Moment Estimator) results in a slope that lies between the slope of the regression of *Y on X* and the inverse slope of the regression of *X on Y*. However, the attracting feature of the geometric regression (GMR) is its *unique solution regardless whether Y is regressed on X or vice versa*. Referring to GMR, Ricker (1973) used the term the "geometric mean estimate of the functional regression of Y on X". Weisberg (1985) provided the major sources of random component of the errors.

5. Missing data in the computation of propensity scores

In many observational studies, missing data are found in one or more covariates /dimensions. As a consequence, a large number of subjects are eliminated from the analyses. In those instances, regression analysis provide *biased* estimates of the coefficients when missing data are present in any of the baseline covariates especially if they are not missing at random. The logistic regression uses the sets of covariates with no missing values to compute the corresponding propensity scores. In our study, more than 50% of the data would be lost if no method of replacing missing data is implemented. The model information (SAS, 2009) shows that out of the 8055 observations from the data, only 3016 were used. It is imperative to implement a method to replace missing values based on rigorous statistical theory prior to computing the propensity scores or principal components on the proxies of multi-dimensional poverty.

Implementation

Figure 1 Proxies of multi-dimensional poverty

Dimension	Proxy	Unit
Financial	1) Per capita monthly expenditure normalised to the poverty line 2) Assets	Multiples of the poverty line Number of durables owned by the household
Physical	Access to services and housing	Index based on the number of services and housing attributes
Vulnerability	Shocks	Number of negative events in the previous two years
Capability	Human capital	Index based on mean education of household members and the proportion of literate household members
Exclusion	Participation in local institutions	Index of group membership and participation in local decision making structures
Digital	Access to, and use of ICT	Index based on the type of ICT used by household members

For more information, see Julian et al (2010)

Model Selection for Propensity Estimation

1. The *Full model* includes the following variables:

Linear:

Hhsize
Maristatus
Actualage
Education
Assets
Vulnerability
Capabilities
Physical
Gender
Services
Exclusion

Quadratics:

Education
Assets
Vulnerability
Capabilities
Physical
Exclusion

Interactions:

education*gender
Assets*gender
Assets*education
Vulnerability*gender
Vulnerability*education
Vulnerability*Assets
Capabilities*gender
Capabilities*education
Capabilities*Assets
Capabilities*Vulnerability
Income*gender
Income*education
Income*Assets
Income*Vulnerability
Income*Capabilities
Physical*gender
Physical*education
Physical*Assets
Physical*Vulnerability
Physical*Capabilities
Exclusion*gender
Exclusion*education
Exclusion*Assets
Exclusion*Vulnerability
Exclusion*Capabilities
Exclusion*Physical
Services*gender
Services*education
Services*Assets
Services*Vulnerability
Services*Capabilities
Services*Physical
Services*Exclusion
maristatus*gender
maristatus*education
maristatus*assets
maristatus*vulnerability
maristatus*Capabilities
maristatus*Physical
maristatus*Exclusion
hhsize*maristatus
hhsize*gender
hhsize*education

HHsize*Physical
HHsize*exclusion
hhsize*assets
hhsize*vulnerability
hhsize*Capabilities

2. The *Reduced model*:

Base on a combination of backward elimination and forward selection method within *the logistic regression* and within the *response surface regression* model (cutoff point of α =10% was applied), two reduced models were constructed. However, we gave more weight to the logistic regression results than the non-iterative regression model in the choice of the final model.

1. Selection using *logistic regression*

Note that all the *linear* dimensions were *deliberately* selected to be included in the final model.

Linear:

hhsize
maristatus,gender
education
Assets
Vulnerability
Capabilities
Physical
Exclusion

Quadratics:

Education

Cross-products:

Gender*education
Assets*education
Gender*Capabilities
Education*Capabilities

2. Selection using *Response Surface*

Alternatively we used a *Weighted Regression Response Surface*. The weights are given by $w \approx np(1-p)$. At this stage, the primary objective was to verify whether the selected variables are same or approximately the same in both methods (logistic and weighted regression) recognizing that the regression are designed to model continuous variables. The weighted regression is a non-iterative fit that takes into account the fact that the logit

transformation produces a linear relationship, but the residual variances are unequal (SAS technical report,). From this run, a reduced model was found to be:

Linear:

Hhsize
Maristatus,
Gender
Education
Assets
Vulnerability
Capabilities
Services
Physical
Exclusion

Quadratics:

Education
Actualage
Physical
HHsize

Cross-products:

Gender*HHsize
Gender* Maristatus
Actualage*Gender
Gender*Education
Capacity*Actualage
Assets*Education
Exclusion*Actualage
Exclusion*Assets
Services*Physical

3. The *final model* combines the results from the above and contains all the linear, one quadratic (on education) and the cross-products Hhsize*gender, Gender*education, Education*assets, Gender*capabilities, Education*capabilities

The LOGISTIC Procedure for the Final Model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-1.6464	0.3777	19.0026	<.0001	
hhsize	1	-0.0939	0.0284	10.9019	0.0010	
maristatus	1	0.0251	0.0395	0.4053	0.5244	
gender	1	2.0416	0.3407	35.9074	<.0001	
education	1	0.5499	0.1665	10.9091	0.0010	
Assets	1	0.0321	0.0813	0.1560	0.6928	
Vulnerability	1	0.0184	0.0853	0.0466	0.8291	
Capabilities	1	-1.3330	0.3409	15.2925	<.0001	
Physical	1	-0.1701	0.0841	4.0865	0.0432	
Exclusion	1	0.1686	0.0982	2.9449	0.0861	
hhsize*gender	1	-0.1249	0.0375	11.1045	0.0009	<--
gender*education	1	-0.3296	0.0997	10.9213	0.0010	<--
education*education	1	-0.0662	0.0197	11.2957	0.0008	<--
gender*Capabilities	1	0.4672	0.2341	3.9828	0.0460	<--
education*Capabiliti	1	0.2503	0.0754	11.0037	0.0009	<--

The model above produced a scalar called "*propensity score*" that is function of 14 covariates, 5 of which are interaction and quadratics. The resulting score summarizes the information required to balance for the distribution of the covariates (Rosenbaum and Rubin, 1984).

Quintile approach to stratification

Measuring the effectiveness of Sub-classification

1-Intra-class correlation

The estimate variance component of Stratum equals 0.01458. It represents the variance *among* the strata means whereas the estimate for Residual represents the variance of the propensity scores *within* the same stratum. Obviously the strata means are very different since the *intra-class correlation coefficient* $\frac{\sigma^2_{strata}}{\sigma^2_{strata} + \sigma^2_{residual}}$ is very close to one (=**0.92** with a confidence interval **0.82** and **0.99**). This provides clear evidence to the dominance of σ^2_{strata} over the total variance $\sigma^2_{strata} + \sigma^2_{residual}$.

Table . Covariance Parameter Estimates Between and Within Subclasses

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
Strata	0.01458	0.01031	1.41	0.0786	0.05	0.005236	0.1204
Residual	0.001232	0.000032	38.71	<.0001	0.05	0.001172	0.001297
Intra-Corr	0.922					0.817104	0.98934

The intra-class correlation is generally used to measure the homogeneity of elements in the five created strata (Steel and Torrie, 1980). As such it verifies that the stratification based on propensity scores was effective in advance of further analyses (Snedecor and Cochran, 1989) and. More importantly it is shown (see the table below) that subclassification on the propensity score balances for the

baseline covariates. It also provide evidence that in the *within* subclasses the distribution is small relative to between subclasses.

2-Remove initial differences

2.1. Examination of whether the groups are balanced BEFORE stratification:

T-Tests for Difference in Group Means-Variables of interest
Before any Stratification on Propensity Scores

Variable	Method	Variances	DF	t Value	Pr > |t|	
hhsize	Satterthwaite	Unequal	2412	16.87	<.0001	←--
maristatus	Satterthwaite	Unequal	1640	0.40	0.6903	
gender	Satterthwaite	Unequal	2253	-10.95	<.0001	←--
Actualage	Satterthwaite	Unequal	2324	-17.73	<.0001	←--
education	Satterthwaite	Unequal	2251	-7.51	<.0001	←--
Assets	Satterthwaite	Unequal	2181	3.10	0.0020	←--
Vulnerabilit	Satterthwaite	Unequal	2197	-3.23	0.0012	←--
Capabilities	Satterthwaite	Unequal	2060	1.17	0.2402	
Physical	Satterthwaite	Unequal	1513	1.44	0.1515	
Exclusion	Satterthwaite	Unequal	2213	0.31	0.7599	
Services	Satterthwaite	Unequal	1512	1.44	0.1512	

$\alpha = 0.05$ The Satterthwaite is more general since it doesn't assume equal variance

Equality of Variances before stratification

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	6581	1472	1.31	<.0001	←--
maristatus	Folded F	3355	986	1.05	0.3937	
gender	Folded F	6575	1471	1.10	0.0231	←--
Actualage	Folded F	6547	1463	1.21	<.0001	←--
education	Folded F	6570	1470	1.10	0.0248	←--
Assets	Folded F	6581	1472	1.00	0.9779	
Vulnerabilit	Folded F	6581	1472	1.02	0.5824	
Capabilities	Folded F	1470	6577	1.19	<.0001	←--
Physical	Folded F	4523	1018	1.00	0.9573	
Exclusion	Folded F	6576	1472	1.04	0.3112	
Services	Folded F	4523	1018	1.00	0.9621	

$\alpha = 0.05$

2.2 Examination of whether the groups are balanced AFTER stratification:
After stratification, there are practically no differences among means and variances in (almost) all strata.

T-Tests Stratum=4

Variable	Method	Variances	DF	t Value	Pr > |t|
hhsize	Pooled	Equal	601	1.58	0.1152
maristatus	Pooled	Equal	601	-0.24	0.8136
gender	Pooled	Equal	601	0.46	0.6438
Actualage	Pooled	Equal	600	-1.51	0.1313
education	Pooled	Equal	601	-0.10	0.9222
Assets	Pooled	Equal	601	0.11	0.9149
Vulnerabilit	Pooled	Equal	601	-0.42	0.6728
Capabilities	Pooled	Equal	601	-0.44	0.6594
Physical	Pooled	Equal	601	-0.22	0.8271
Exclusion	Pooled	Equal	601	-0.09	0.9273
Services	Pooled	Equal	601	-0.22	0.8250

Test using pooled variances are justified.

Equality of Variances Stratum=4

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	157	444	1.18	0.2055	
maristatus	Folded F	157	444	1.11	0.3965	
gender	Folded F	157	444	1.05	0.6775	
Actualage	Folded F	444	156	1.39	0.0169	<--
education	Folded F	444	157	1.30	0.0520	<--
Assets	Folded F	444	157	1.17	0.2538	
Vulnerabilit	Folded F	157	444	1.00	0.9712	
Capabilities	Folded F	444	157	1.05	0.7275	
Physical	Folded F	444	157	1.05	0.7214	
Exclusion	Folded F	157	444	1.12	0.3810	
Services	Folded F	444	157	1.05	0.7174	

$\alpha = 0.05$

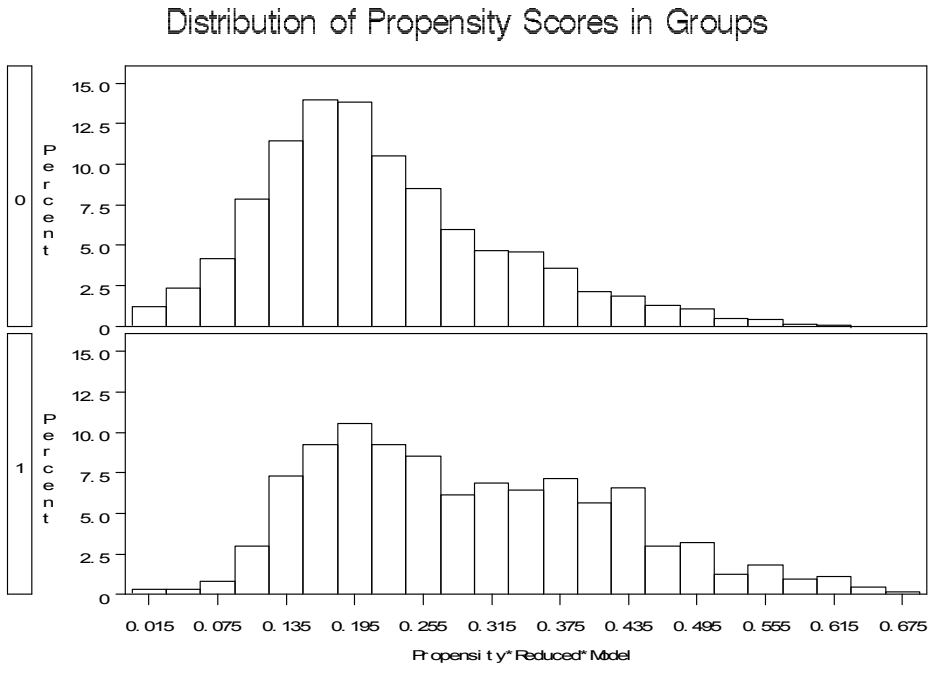
3- Percent Bias Reduction

Percent Reduction Bias From Reduced Model Before and After Stratification

Variable	Abs(Mean) After	Abs(Mean) Before	Percent Bias Reduction
Actualage	0.87586	8.6080	89.82
Assets	0.00578	0.0588	90.12
Capability	0.00924	0.0167	45.91
Physical	0.00600	0.0341	82.40
Services	0.01762	0.0990	82.22
Vulnerabity	0.00232	0.0500	96.00
Education	0.05510	0.2720	79.74
Hhsize	0.08474	1.2563	93.26

4-Graphical display

The graph for the *overall* distribution of the propensity scores is as follows:

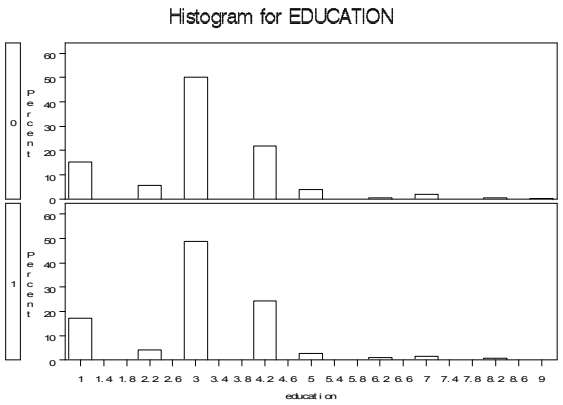


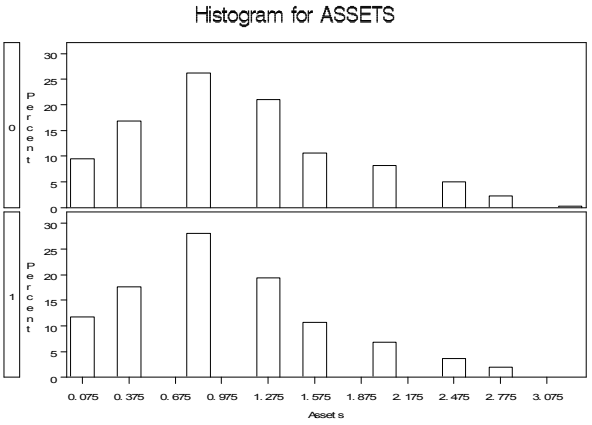
The histogram suggests a reasonable overlap in the propensity scores between interviewed and non interviewed individuals. However we may want to remove some of the scores that are greater than 0.60 for there may be a risk of extrapolating outside the range of the data when we adjust (Kleiman, 2010). We can then consider the two groups as *comparable*.

The Table below also shows that the maximum propensity score for the interviewed group is larger (0.30) than that of the non-interviewed group (0.22) by approximately 8%. This may suggest removing some of the extra values to avoid extrapolation during the covariance adjustment (see **Covariance Adjustment using Propensity score**s below). As expected from the logistic fit, the mean of the interviewed is larger.

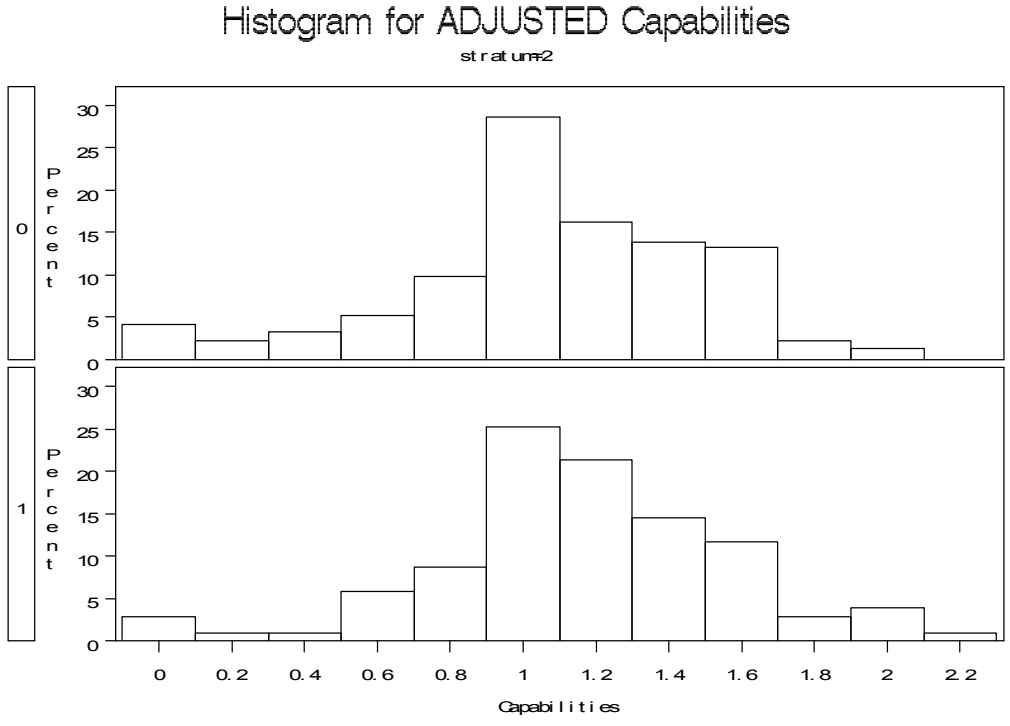
5-Graphs of selected Poverty Dimensions

The following (partial) series of graphs display the distribution of the dimensions of poverty. The distribution *does not differ* from one group to another. Once again the effectiveness of the adjustment method by sub-classification is apparent. This is especially clear for the two graphs on the capacity dimension of poverty. The adjusted distributions are much closer than when the variable is not adjusted. All the graphs are for combined across strata.





Similar graphical representation is performed for selected dimensions within *each stratum separately*. They clearly show that within subclasses that are homogeneous in the propensity scores, the distribution of the covariates is the same for treated and control groups.



The Geometric Mean Regression

The geometric mean (slope) of the least squares regression coefficient for the regression of income on ICT and the reciprocal of ICT on income are as follow. When drawing this slope thru the mean (income,

ICT) one obtains the *same line* whether *income* is regressed on *ICT* or vice versa.

Income and ICT_total as the Dependents *across* strata

Income is the *Dependent*

Obs	Dependent	Parameter	Estimate	StdErr
1	Income	Slope ALLTZN Major Urban	0.12091150	0.01283163
2	Income	Slope ALLTZN Other Urban	0.07471073	0.02042262
3	Income	Slope ALLTZN Rurual	0.16766947	0.03439367
4	Income	Slope ALLKNY Major Urban	0.05804188	0.01526563
5	Income	Slope ALLKNY Rural	0.10427444	0.01058406
6	Income	Slope ALLRWD Major Urban	0.11697288	0.00835514
7	Income	Slope ALLRWD Other Urban	0.14210371	0.01672936
8	Income	Slope ALLRWD Rural	0.14813052	0.02084141
9	Income	Slope ALLUGD Major Urban	0.05813257	0.01807319
10	Income	Slope ALLUGD Other Urban	0.09611823	0.02499468
11	Income	Slope ALLUGD Rural	-0.04322515	0.03482878

ICT-Total is the *Dependent*

Obs	Dependent	Parameter	Estimate	StdErr
1	ICT_total	Slope ALLTZN Major Urban	1.77578964	0.17787609
2	ICT_total	Slope ALLTZN Other Urban	0.84824054	0.24882731
3	ICT_total	Slope ALLTZN Rurual	0.44724948	0.22158084
4	ICT_total	Slope ALLKNY Major Urban	1.45777849	0.27460696
5	ICT_total	Slope ALLKNY Rural	1.29214362	0.13554002
6	ICT_total	Slope ALLRWD Major Urban	2.74226549	0.14713334
7	ICT_total	Slope ALLRWD Other Urban	2.07981779	0.23174829
8	ICT_total	Slope ALLRWD Rural	0.80270462	0.17528863
9	ICT_total	Slope ALLUGD Major Urban	0.56915244	0.20237452
10	ICT_total	Slope ALLUGD Other Urban	2.92303673	0.49318397
11	ICT_total	Slope ALLUGD Rural	-0.33754369	0.39499426

Equation (18) is applied to the above slopes to compute GMFR. The results are as follow:

Geometric Means Relationship

Relationship between Income and ICT by country-environment

COUNTRY	Average Geometric Parameter	Stand Error Geometric MeanSlope	MeanSlope	t_obs	Testing Null Hypoth
Kenya	Maj_Urb	0.12659	0.07255	1.7448	NS
Kenya	Rural	0.30158	0.02972	10.1479	***
Rwanda	Maj_Urb	0.21719	0.02353	9.2298	***
Rwanda	Oth_Urb	0.26955	0.02592	10.4006	***
Rwanda	Rural	0.45181	0.04605	9.8112	***
Tanzania	Maj_Urb	0.26225	0.00964	27.1982	***
Tanzania	Oth_Urb	0.30561	0.01776	17.2092	***
Tanzania	Rural	0.87395	0.37580	2.3256	*
Uganda	Maj_Urb	-0.72761	1.09054	-0.6672	NS
Uganda	Oth_Urb	0.16060	0.02598	6.1817	***
Uganda	Rural	-0.21827	0.10527	-2.0736	NS

Relationship between Income and ICT by country

COUNTRY	Average Geometric MeanSlope	Stand Error Geometric MeanSlope	t_obs	Testing Null Hypoth
Kenya	0.21408	0.04708	4.54721	**
Rwanda	0.31285	0.03226	9.6968	***
Tanzania	0.48060	0.13798	3.4832	***
Uganda	-0.26487	0.37736	-0.70190	NS

Across countries Relationship

Country	Average Geometric MeanSlope	Standard Error Geometric MeanSlope	t_obs	Testing Null Hypoth/Slope
All (**)	0.35106	0.05591	6.27866	**
All	0.19138	0.11015	1.73736	NS

The standard errors are underestimated. (**) Excluding Uganda

Geometric Means Relationship
Intercept across Strata

Obs	COUNTRY	Parameter	Average Intercept	Average Geometric MeanSlope	Testing Null Hypoth Intercept	
1	Kenya	Maj_Urb	1.09940	0.12659	**	
2	Kenya	Rural	0.93521	0.30158	**	
3	Rwanda	Maj_Urb	0.93888	0.21719	**	
4	Rwanda	Oth_Urb	0.92390	0.26955	**	
5	Rwanda	Rural	0.83499	0.45181	**	
6	Tanzania	Maj_Urb	0.90885	0.26225	**	
7	Tanzania	Oth_Urb	0.88162	0.30561	**	
8	Tanzania	Rural	0.75967	0.87395	**	
9	Uganda	Maj_Urb	0.46492	-0.72761	NS	← suspect
10	Uganda	Oth_Urb	0.85350	0.16060	**	
11	Uganda	Rural	0.88612	-0.21827	**	← suspect

Geometric Means Relationship
Intercept Across Environments

Obs	COUNTRY	Average Intercept	Average Geometric MeanSlope	Testing Null Hypoth Intercept	
1	Kenya	1.01731	0.21408	***	
2	Rwanda	0.89925	0.31285	***	
3	Tanzania	0.85005	0.48060	***	
4	Uganda	0.72404	***-0.26487***	***	← very suspect
Overall (**)		0.91031	0.35106	**	
Overall		0.86202	0.19138	**	

The standard errors are underestimated. (**) Excluding Uganda

Alternatives to the Geometric Mean Relationship

In this report, the above alternatives are not implemented, but can be considered in future exploitation of the study.

Dealing with Missing Data

Regression analysis provides *biased* estimates of the coefficients when missing data are present in any of the baseline covariates. In that situation, the logistic regression uses the sets of covariates with no missing values to compute the corresponding propensity scores. In our study, more than 50% of the data would be lost if no method of replacing missing data is implemented. The model information below shows that out of the **8055** observations from the data, only **3016** were used.

The LOGISTIC Procedure

Model Information

Data Set	WORK.ONE2
Response Variable	interv
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	**8055**
Number of Observations Used	**3016**

For simplicity, missing values of a particular variable are replaced by their median. The logistic regression is then applied giving more weights to the observed data. The weights, in this case, are the *number of non-missing values* for that covariate. The model information shows that of all the **8055** observations were used in the logistic regression.

The LOGISTIC Procedure

Model Information

Data Set	WORK.ONE1
Response Variable	interv
Number of Response Levels	2
Weight Variable	Weight
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	**8055**
Number of Observations Used	**8055**

Under this scenario, it appears that the *full model* is the model of choice to compute the propensity scores since almost all of the

individual factors (linear), the pair-wise interactions and the quadratics are significant.

The logistic model output is as follow:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhsize	1	25.4782	<.0001
maristatus	4	332.1276	<.0001
gender	1	1521.6515	<.0001
education	1	1041.7687	<.0001
Assets	1	12.8127	0.0003
Vulnerabilit	1	40.0798	<.0001
Capabilities	1	831.9038	<.0001
Physical	1	114.2053	<.0001
Exclusion	1	18.1157	<.0001
gender*education	1	270.8726	<.0001
education*education	1	2456.3819	<.0001
gender*Assets	1	24.4978	<.0001
education*Assets	1	297.7107	<.0001
Assets*Assets	1	31.2326	<.0001
gender*Vulnerabilit	1	1.1541	**0.2827**
education*Vulnerabil	1	41.7326	<.0001
Assets*Vulnerabilit	1	23.5368	<.0001
Vulnerabi*Vulnerabil	1	4.5326	0.0333
gender*Capabilities	1	21.4837	<.0001
education*Capabiliti	1	1312.3781	<.0001
Assets*Capabilities	1	54.9882	<.0001
Vulnerabi*Capabiliti	1	10.0109	0.0016
Capabilit*Capabiliti	1	97.6829	<.0001
gender*Physical	1	61.8608	<.0001
education*Physical	1	758.0622	<.0001
Assets*Physical	1	20.6919	<.0001
Vulnerabili*Physical	1	12.7383	0.0004
Capabilitie*Physical	1	182.3212	<.0001
Physical*Physical	1	2.8396	0.0920
gender*Exclusion	1	261.6051	<.0001
education*Exclusion	1	50.2341	<.0001
Assets*Exclusion	1	0.4676	**0.4941**
Vulnerabil*Exclusion	1	1.6829	**0.1945**
Capabiliti*Exclusion	1	35.6716	<.0001
Physical*Exclusion	1	54.9062	<.0001
Exclusion*Exclusion	1	20.7089	<.0001
gender*Services	1	21.6408	<.0001
education*Services	1	705.2728	<.0001
Assets*Services	1	21.1309	<.0001
Vulnerabili*Services	1	13.1187	0.0003
Capabilitie*Services	1	184.6352	<.0001
Physical*Services	1	2.8259	0.0928
Exclusion*Services	1	52.5468	<.0001
gender*maristatus	4	60.9809	<.0001
education*maristatus	4	339.9464	<.0001
Assets*maristatus	4	100.4337	<.0001
Vulnerabi*maristatus	4	81.7924	<.0001
Capabilit*maristatus	4	373.6822	<.0001
Physical*maristatus	4	323.9538	<.0001
Exclusion*maristatus	4	452.8140	<.0001
hhsize*maristatus	4	755.8879	<.0001
hhsize*gender	1	743.0503	<.0001
hhsize*education	1	830.7118	<.0001
hhsize*Assets	1	5.7112	0.0169
hhsize*Vulnerabilit	1	193.2105	<.0001
hhsize*Capabilities	1	403.0994	<.0001
hhsize*Physical	1	125.7056	<.0001
hhsize*Exclusion	1	14.1608	0.0002

The verification steps to the "road to a quasi-randomized experiment" remain the same. The new intra-class correlation is still very satisfactory. The table below shows that the coefficient is 91% with a confidence interval of 78%-98%.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
stratum	0.01490	0.01053	1.41	0.0787	0.05	0.005346	0.1230
Residual	0.001535	0.000024	63.44	<.0001	0.05	0.001488	0.001583

For each stratum, the t-test comparisons between the interviewed and the non-interviewed group are shown below. For completeness, we also include the *test for equal variance*. The statistics of interest are under the heading "***Mean***" (indicated in bold) *which reflects the mean difference between interviewed and non-interviewed group* in each stratum**.**

---------- stratum=1 ----------

The TEST Procedure

Variable	interv	Lower CL Mean	***Mean***	Upper CL Mean	Lower CL Std Dev	Upper CL Std Dev	Std Dev	Std Err
Hhsize	Diff (1-2)	-0.665	0.3765	1.4183	3.8493	4.0668	4.3105	0.5305
maristatus	Diff (1-2)	-0.452	-0.142	0.1689	1.1473	1.2121	1.2847	0.1581
Actualage	Diff (1-2)	-14.44	-9.765	-5.091	17.019	17.987	19.073	2.38
Education	Diff (1-2)	-0.35	-0.012	0.3249	1.2466	1.317	1.396	0.1718
Assets	Diff (1-2)	-0.065	0.1174	0.3001	0.6751	0.7132	0.756	0.093
Vulnerabil	Diff (1-2)	-0.074	0.0664	0.2065	0.5176	0.5468	0.5796	0.0713
Capabiliti	Diff (1-2)	-0.039	0.0694	0.1776	0.3995	0.4221	0.4474	0.0551
Physical	Diff (1-2)	-0.098	0.0749	0.2483	0.6406	0.6768	0.7174	0.0883
Exclusion	Diff (1-2)	-0.079	0.0185	0.1156	0.3588	0.379	0.4017	0.0494
Services	Diff (1-2)	-0.285	0.2173	0.7192	1.8545	1.9593	2.0767	0.2556

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
hhsize	Folded F	536	65	1.40	0.0921
maristatus	Folded F	65	536	1.01	0.9232
gender	Folded F	65	536	1.08	0.6370
Actualage	Folded F	63	530	1.23	0.2363
education	Folded F	536	65	1.11	0.6116
Assets	Folded F	536	65	1.13	0.5362
Vulnerabilit	Folded F	65	536	1.03	0.8316
Capabilities	Folded F	65	536	1.15	0.4290
Physical	Folded F	536	65	1.05	0.8125
Exclusion	Folded F	536	65	1.43	0.0724
Services	Folded F	536	65	1.05	0.8142

---------- stratum=2 ----------

The TTEST Procedure

Variable	interv	N	Mean	Lower CL ***Mean***	Upper CL Mean	Lower CL Std Dev	Upper CL Std Dev	Std Dev	Std Err
hhsize	Diff (1-2)	-0.487	-0.058	0.3705	1.9107	2.0186	2.1396	0.2184
maristatus	Diff (1-2)	-0.175	0.0739	0.3223	1.1066	1.1691	1.2392	0.1265
Actualage	Diff (1-2)	-8.795	-5.305	-1.816	15.472	16.348	17.33	1.7767
education	Diff (1-2)	-0.314	-0.081	0.1517	1.0376	1.0962	1.1619	0.1186
Assets	Diff (1-2)	-0.152	-0.008	0.1363	0.643	0.6793	0.72	0.0735
Vulnerabil	Diff (1-2)	-0.116	-0.004	0.1092	0.5025	0.5308	0.5627	0.0574
Capabiliti	Diff (1-2)	-0.165	-0.077	0.0115	0.3941	0.4164	0.4413	0.0451
Physical	Diff (1-2)	-0.146	-0.004	0.1381	0.6331	0.6689	0.709	0.0724
Exclusion	Diff (1-2)	-0.077	0.0161	0.1094	0.4157	0.4392	0.4655	0.0475

Services	Diff (1-2)	-0.424	-0.013	0.3991	1.8341	1.9377	2.0539	0.2097

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	102	499	1.12	0.4505	
maristatus	Folded F	499	102	1.03	0.8699	
gender	Folded F	102	499	1.29	0.0841	← Suggest
Actualage	Folded F	497	101	1.02	0.9295	
education	Folded F	102	499	1.03	0.8378	
Assets	Folded F	102	499	1.10	0.4939	
Vulnerabilit	Folded F	499	102	1.02	0.9477	
Capabilities	Folded F	499	102	1.05	0.7993	
Physical	Folded F	102	499	1.01	0.9223	
Exclusion	Folded F	499	102	1.19	0.2911	
Services	Folded F	102	499	1.01	0.9232	

---------- stratum=3 ----------

The TTEST Procedure

Variable	interv	N	Mean	Lower CL ***Mean***	Mean	Upper CL Std Dev	Lower CL Std Dev	Std Dev	Upper CL Std Err
hhsize	Diff (1-2)		-0.329	0.0713	0.472	1.9226	2.0312	2.1529	0.2041
maristatus	Diff (1-2)		-0.254	-0.028	0.1968	1.0808	1.1419	1.2103	0.1147
Actualage	Diff (1-2)		-8.722	-5.593	-2.465	15.006	15.854	16.805	1.593
Education	Diff (1-2)		-0.332	-0.087	0.1581	1.1758	1.2422	1.3166	0.1248
Assets	Diff (1-2)		-0.122	0.0107	0.1434	0.6368	0.6727	0.7131	0.0676
Vulnerabil	Diff (1-2)		-0.082	0.0219	0.1259	0.499	0.5272	0.5588	0.053
Capabiliti	Diff (1-2)		-0.051	0.0376	0.1267	0.4274	0.4516	0.4786	0.0454
Physical	Diff (1-2)		-0.077	0.0595	0.1964	0.6567	0.6938	0.7354	0.0697
Exclusion	Diff (1-2)		-0.133	-0.041	0.0507	0.4409	0.4658	0.4937	0.0468
Services	Diff (1-2)		-0.224	0.1726	0.5693	1.9032	2.0107	2.1312	0.202

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	477	124	1.02	0.9137	
maristatus	Folded F	477	124	1.18	0.2761	
gender	Folded F	477	124	1.08	0.6142	
Actualage	Folded F	476	124	1.00	0.9984	
education	Folded F	124	477	1.37	0.0213	<-- Significant
Assets	Folded F	124	477	1.00	0.9704	
Vulnerabilit	Folded F	477	124	1.07	0.6398	
Capabilities	Folded F	124	477	1.06	0.6696	
Physical	Folded F	124	477	1.05	0.7010	
Exclusion	Folded F	477	124	1.13	0.4008	
Services	Folded F	124	477	1.05	0.6969	

---------- stratum=4 ----------

The TTEST Procedure

Variable	interv	N	Lower CL Mean	***Mean***	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper Std Dev	Std Err
hhsize	Diff (1-2)		-0.054	0.2198	0.4934	1.4241	1.5045	1.5947	0.1393
maristatus	Diff (1-2)		-0.222	-0.024	0.1743	1.0306	1.0889	1.1541	0.1008
Actualage	Diff (1-2)		-5.443	-2.367	0.7095	15.972	16.875	17.887	1.5664
education	Diff (1-2)		-0.235	-0.011	0.213	1.1664	1.2323	1.3061	0.1141
Assets	Diff (1-2)		-0.116	0.0067	0.1297	0.6401	0.6763	0.7168	0.0626
Vulnerabil	Diff (1-2)		-0.117	-0.021	0.0757	0.502	0.5303	0.5621	0.0491
Capabiliti	Diff (1-2)		-0.106	-0.019	0.0672	0.451	0.4765	0.5051	0.0441
Physical	Diff (1-2)		-0.139	-0.014	0.1114	0.6522	0.689	0.7303	0.0638
Exclusion	Diff (1-2)		-0.088	-0.004	0.0801	0.4372	0.4619	0.4896	0.0428
Services	Diff (1-2)		-0.404	-0.041	0.3223	1.8903	1.9971	2.1168	0.1849

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	157	444	1.18	0.2055	
maristatus	Folded F	157	444	1.11	0.3965	
gender	Folded F	157	444	1.05	0.6775	
Actualage	Folded F	444	156	1.39	0.0169	**<-- Significant**
education	Folded F	444	157	1.30	0.0520	**<--Suggest**
Assets	Folded F	444	157	1.17	0.2538	
Vulnerabilit	Folded F	157	444	1.00	0.9712	
Capabilities	Folded F	444	157	1.05	0.7275	
Physical	Folded F	444	157	1.05	0.7214	
Exclusion	Folded F	157	444	1.12	0.3810	
Services	Folded F	444	157	1.05	0.7174	

---------- stratum=5 ----------

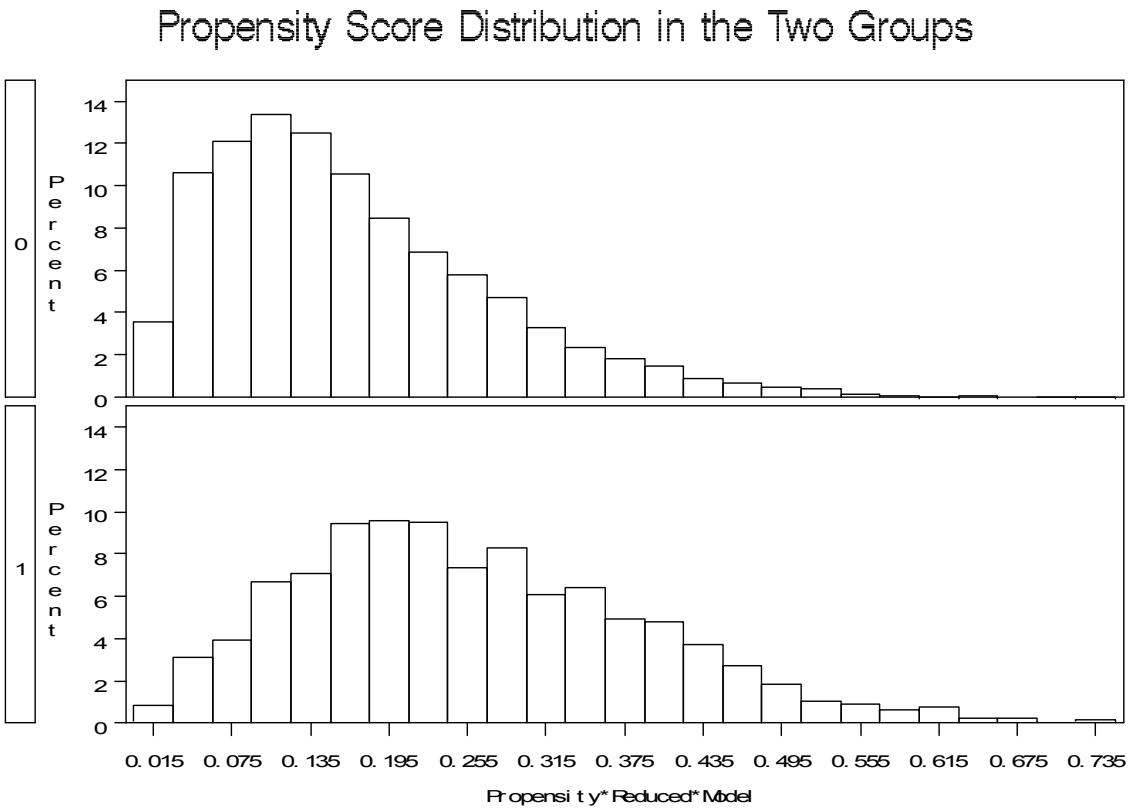
The TTEST Procedure

Variable	interv	Lower CL Mean	***Mean***	Upper CL Mean	Lower CL Std Dev	Upper CL Std Dev	Std Dev	Std Err
Hhsize	Diff (1-2)	0.2137	0.4025	0.5912	1.1075	1.17	1.2401	0.0961
maristatus	Diff (1-2)	-0.121	0.0543	0.2302	1.0317	1.09	1.1552	0.0895
Actualage	Diff (1-2)	-4.927	-2.05	0.8267	16.881	17.834	18.902	1.4649
education	Diff (1-2)	-0.014	0.182	0.3777	1.1485	1.2133	1.286	0.0997
Assets	Diff (1-2)	-0.107	0.0018	0.1105	0.6379	0.6739	0.7142	0.0554
Vulnerabil	Diff (1-2)	-0.139	-0.053	0.0333	0.5061	0.5346	0.5666	0.0439
Capabiliti	Diff (1-2)	-0.032	0.0616	0.1556	0.5516	0.5828	0.6177	0.0479
Physical	Diff (1-2)	-0.09	0.0177	0.1257	0.634	0.6697	0.7098	0.055
Exclusion	Diff (1-2)	-0.085	-0.007	0.0704	0.4551	0.4808	0.5096	0.0395
Services	Diff (1-2)	-0.262	0.0514	0.3646	1.8375	1.9412	2.0575	0.1595

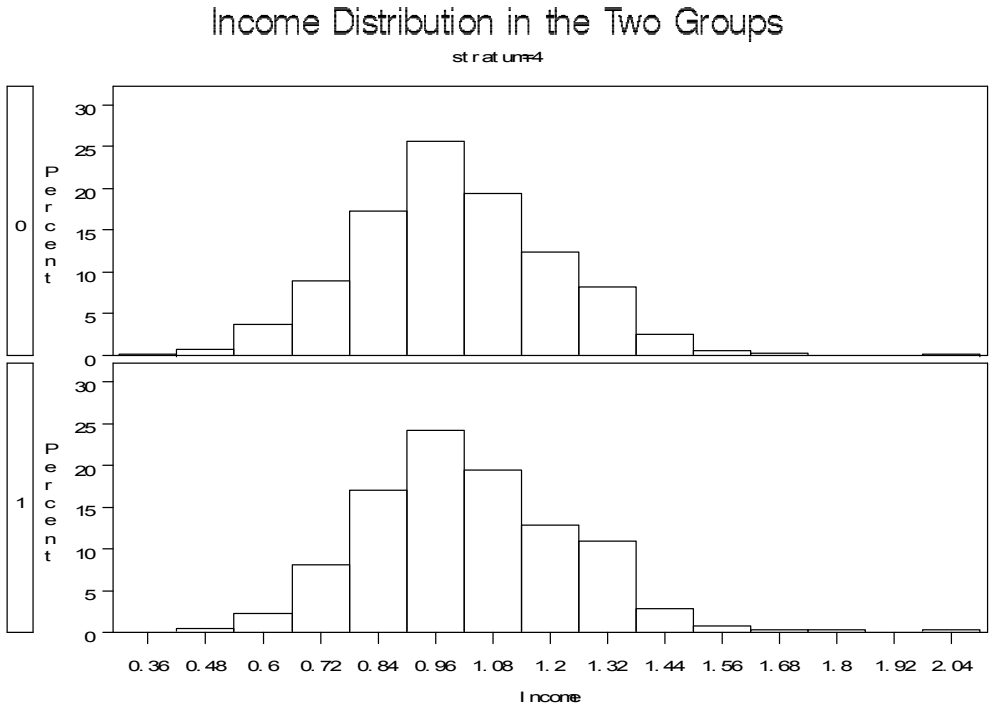
Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F	
hhsize	Folded F	260	342	1.25	0.0507	<- Suggest
maristatus	Folded F	260	342	1.01	0.9389	
gender	Folded F	342	260	2.26	<.0001	<-- Significant
Actualage	Folded F	342	260	1.20	0.1286	
education	Folded F	342	260	1.07	0.5838	
Assets	Folded F	342	260	1.03	0.7882	
Vulnerabilit	Folded F	342	260	1.08	0.5106	
Capabilities	Folded F	260	342	1.04	0.7396	
Physical	Folded F	260	342	1.04	0.7224	
Exclusion	Folded F	342	260	1.09	0.4722	
Services	Folded F	260	342	1.04	0.7164	

In addition, the overlap of the interviewed vs. non-interviewed individuals are



An example of within stratum distribution of income variable is as follow:



The percent *reductions in bias* on the same *selected baseline variables* are substantial especially for the variable *assets* for which the bias is *completely removed*:

Percent Bias Reduction When Missing Data are Replaced

Variable	Abs(Mean) After	Abs(Mean) Before	Percent Bias Reduction
Assets	0.00022	0.0588	99.99
Capabili	0.00434	0.0174	75.06
Services	0.01570	0.0591	71.74
Vulnerab	0.00495	0.0500	90.00
Educatio	0.05766	0.2720	82.46
Hhsize	0.08938	1.2563	92.87
Maristat	0.02722	0.2100	87.04

The geometric mean regression coefficients are combined following eqs. 5 & 6.The estimates of slopes themselves are obtained from an ANCOVA that includes *Stratum –Country and Areas* combination. The individual slopes are listed below: 1) *Income* as the *dependent* variable and 2) *ICT* is the *dependent* variable. In the first case, the slopes represent changes in income by an additional unit ownership in ICT and in the latter case the slopes represent change in ICT ownership by unit change in the income. As expected both estimates are different.

Dependent Variable: Income

---------- stratum=1 ----------

Parameter	ICT slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Other Urban	0.08337581	0.01199475	6.95	<.0001
Slope TZN Rurual	-0.03827191	0.05741271	-0.67	0.5051
Slope KNY Major Urban	0.01364155	0.02533744	0.54	0.5904
Slope KNY Rural	-0.01113262	0.00876413	-1.27	0.2042
Slope RWD Major Urban	-0.01229389	0.00537214	-2.29	0.0222
Slope RWD Other Urban	0.06700798	0.00867513	7.72	<.0001
Slope RWD Rural	0.08229939	0.02452823	3.36	0.0008
Slope UGD Major Urban	0.03174382	0.02072558	1.53	0.1258
Slope UGD Other Urban	-0.00021982	0.01114738	-0.02	0.9843
Slope UGD Rural	0.00783557	0.01598425	0.49	0.6241

---------- stratum=2 ----------

Parameter	ICT slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	0.07612730	0.18234909	0.42	0.6764
Slope TZN Other Urban	0.07054693	0.01407132	5.01	<.0001
Slope TZN Rurual	0.01677615	0.02141080	0.78	0.4334
Slope KNY Major Urban	0.03613413	0.02589955	1.40	0.1632
Slope KNY Rural	-0.04058566	0.01180348	-3.44	0.0006
Slope RWD Major Urban	-0.02109313	0.00604812	-3.49	0.0005
Slope RWD Other Urban	0.06621127	0.00972627	6.81	<.0001
Slope RWD Rural	0.04961682	0.01966179	2.52	0.0117
Slope UGD Major Urban	0.00944654	0.02634848	0.36	0.7200
Slope UGD Other Urban	-0.00335227	0.01148872	-0.29	0.7705
Slope UGD Rural	-0.02658577	0.01659538	-1.60	0.1094

---------- stratum=3 ----------

Parameter	ICT slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Other Urban	0.05483002	0.01581597	3.47	0.0005
Slope TZN Rurual	0.07363660	0.03223550	2.28	0.0225
Slope KNY Major Urban	0.16051151	0.07215771	2.22	0.0263
Slope KNY Rural	-0.03268645	0.01842497	-1.77	0.0763
Slope RWD Major Urban	0.00372274	0.00690691	0.54	0.5900
Slope RWD Other Urban	0.08298754	0.01401023	5.92	<.0001
Slope RWD Rural	0.03495218	0.01641869	2.13	0.0334
Slope UGD Major Urban	0.05148055	0.02085714	2.47	0.0137
Slope UGD Other Urban	0.03583219	0.01437924	2.49	0.0128
Slope UGD Rural	-0.00298308	0.01841807	-0.16	0.8714

---------- stratum=4 ----------

Parameter	ICT slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	0.05006127	0.01716192	2.92	0.0036
Slope TZN Other Urban	0.02525265	0.02499337	1.01	0.3125
Slope TZN Rurual	0.10317721	0.04269017	2.42	0.0158
Slope KNY Major Urban	-0.02017227	0.01075940	-1.87	0.0610
Slope KNY Rural	-0.02382588	0.00678908	-3.51	0.0005
Slope RWD Major Urban	0.09876199	0.01348374	7.32	<.0001
Slope RWD Other Urban	0.05151710	0.01484505	3.47	0.0005
Slope RWD Rural	0.04002595	0.01801575	2.22	0.0264
Slope UGD Major Urban	0.02778854	0.02098508	1.32	0.1856
Slope UGD Other Urban	0.01275492	0.01995307	0.64	0.5228
Slope UGD Rural	0.00209248	0.00693313	0.30	0.7628

---------- stratum=5 ----------

Parameter	ICT slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	0.08848325	0.02169504	4.08	<.0001
Slope TZN Other Urban	0.12621845	0.04202433	3.00	0.0027
Slope TZN Rurual	0.01517380	0.02443448	0.62	0.5347
Slope KNY Major Urban	-0.01328044	0.01198836	-1.11	0.2681
Slope KNY Rural	-0.01102887	0.00611583	-1.80	0.0715
Slope RWD Major Urban	0.11348172	0.01196183	9.49	<.0001
Slope RWD Other Urban	0.12812089	0.01427720	8.97	<.0001
Slope RWD Rural	0.07053733	0.02200023	3.21	0.0014
Slope UGD Major Urban	0.00050134	0.01134395	0.04	0.9648
Slope UGD Other Urban	0.02291501	0.02125393	1.08	0.2811
Slope UGD Rural	0.01729752	0.00766666	2.26	0.0242

Dependent Variable: ICT_total

---------- stratum=1 ----------

Parameter	Income Slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	0.68887304	13.5782344	0.05	0.9595
Slope TZN Other Urban	2.13036563	0.5028019	4.24	<.0001
Slope TZN Rurual	-0.22866665	0.9036406	-0.25	0.8003
Slope KNY Major Urban	0.13332040	0.6939655	0.19	0.8477
Slope KNY Rural	-1.94190309	1.0086138	-1.93	0.0544
Slope RWD Major Urban	-1.76783540	0.5482982	-3.22	0.0013
Slope RWD Other Urban	3.53732310	0.5401355	6.55	<.0001

Slope RWD Rural	0.74144124	0.6235628	1.19	0.2346
Slope UGD Major Urban	0.26795468	0.5068814	0.53	0.5971
Slope UGD Other Urban	0.08180400	1.0141468	0.08	0.9357
Slope UGD Rural	4.73944162	2.8632988	1.66	0.0981

---------- stratum=2 ----------

Parameter	Income Slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	1.68340484	7.30471717	0.23	0.8178
Slope TZN Other Urban	1.84616508	0.56033569	3.29	0.0010
Slope TZN Rurual	0.49006549	0.87381988	0.56	0.5750
Slope KNY Major Urban	0.56433240	0.76197745	0.74	0.4590
Slope KNY Rural	-3.88135069	0.87431541	-4.44	<.0001
Slope RWD Major Urban	-1.92679332	0.44490363	-4.33	<.0001
Slope RWD Other Urban	2.69179226	0.46886866	5.74	<.0001
Slope RWD Rural	0.92713642	0.65005664	1.43	0.1540
Slope UGD Major Urban	0.11530283	0.51121349	0.23	0.8216
Slope UGD Other Urban	-0.68376264	1.19708156	-0.57	0.5680
Slope UGD Rural	-1.50035154	0.94496660	-1.59	0.1125

---------- stratum=3 ----------

Parameter	Income Slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	-0.41878077	13.7898414	-0.03	0.9758
Slope TZN Other Urban	1.26954205	0.4967153	2.56	0.0107
Slope TZN Rurual	0.70752515	0.6977951	1.01	0.3108
Slope KNY Major Urban	0.23768344	0.6349089	0.37	0.7082
Slope KNY Rural	-1.37103096	0.8133468	-1.69	0.0921
Slope RWD Major Urban	0.28233252	0.4019533	0.70	0.4825
Slope RWD Other Urban	2.29158110	0.4976958	4.60	<.0001
Slope RWD Rural	1.18634711	0.6347505	1.87	0.0618
Slope UGD Major Urban	0.65851951	0.5079060	1.30	0.1950
Slope UGD Other Urban	1.30343805	0.6074532	2.15	0.0320
Slope UGD Rural	-0.14705501	0.9957496	-0.15	0.8826

---------- stratum=4 ----------

Parameter	Income Slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	1.44337542	0.61946978	2.33	0.0199
Slope TZN Other Urban	0.60952124	0.69680269	0.87	0.3818
Slope TZN Rurual	0.51252512	0.64297371	0.80	0.4255
Slope KNY Major Urban	-1.67518671	0.65217865	-2.57	0.0103
Slope KNY Rural	-2.32412568	0.43264484	-5.37	<.0001
Slope RWD Major Urban	3.22633892	0.50599615	6.38	<.0001
Slope RWD Other Urban	2.09260832	0.64383283	3.25	0.0012
Slope RWD Rural	0.61215787	0.46041535	1.33	0.1838
Slope UGD Major Urban	0.58579723	0.56867459	1.03	0.3031
Slope UGD Other Urban	0.43637806	0.77108027	0.57	0.5715
Slope UGD Rural	0.11974744	0.35084291	0.34	0.7329

---------- stratum=5 ----------

Parameter	Income Slope Estimate	Standard Error	t Value	Pr > |t|
Slope TZN Major Urban	3.10106040	0.83631699	3.71	0.0002
Slope TZN Other Urban	1.45676602	0.93748483	1.55	0.1204
Slope TZN Rurual	0.20744249	0.57327431	0.36	0.7175
Slope KNY Major Urban	-0.77568331	0.64481901	-1.20	0.2292
Slope KNY Rural	-1.02941609	0.35674073	-2.89	0.0040
Slope RWD Major Urban	3.30805691	0.42775894	7.73	<.0001
Slope RWD Other Urban	3.15799922	0.46394104	6.81	<.0001
Slope RWD Rural	0.58934735	0.42894705	1.37	0.1697
Slope UGD Major Urban	0.00386137	0.59987337	0.01	0.9949
Slope UGD Other Urban	1.21440105	1.02217721	1.19	0.2350
Slope UGD Rural	1.34321455	0.43763646	3.07	0.0022

---------- Overall ----------

Obs	Dependent	Parameter	Income Slope Estimate	StdErr
1	ICT_total	Slope ALLTZN Major Urban	1.12775412	5.41583384
2	ICT_total	Slope ALLTZN Other Urban	1.86485246	0.24943535
3	ICT_total	Slope ALLTZN Rurual	0.66761202	0.35324493
4	ICT_total	Slope ALLKNY Major Urban	0.29168522	0.29517420
5	ICT_total	Slope ALLKNY Rural	-1.69298700	0.34958790
6	ICT_total	Slope ALLRWD Major Urban	-1.25036921	0.19287535
7	ICT_total	Slope ALLRWD Other Urban	2.97260451	0.21619907
8	ICT_total	Slope ALLRWD Rural	1.90693847	0.26100967
9	ICT_total	Slope ALLUGD Major Urban	0.41875100	0.21727457
10	ICT_total	Slope ALLUGD Other Urban	0.35197654	0.32385060
11	ICT_total	Slope ALLUGD Rural	0.11179713	0.46786332

Geometric Means Relationship
The missing data are replaced by their Median

Obs	COUNTRY	Parameter	Average Geometric MeanSlope	Standard Error Geometric MeanSlope	t_obs	Testing Null Hypoth
1	Kenya	Major_Ur	0.23082	0.17384	1.3278	NS
2	Kenya	Rural	-0.10743	0.01283	-8.3724	NS
3	Rwanda	Major_Ur	0.05740	0.06306	0.9102	NS
4	Rwanda	Other_Ur	0.16862	0.01179	14.2964	**
5	Rwanda	Rural	0.26756	0.03248	8.2370	**
6	Tanzania	Major_Ur	0.18927	0.01272	14.8835	**
7	Tanzania	Other_Ur	0.21981	0.01876	11.7158	**
8	Tanzania	Rural	0.16353	0.14940	1.0946	NS
9	Uganda	Major_Ur	0.29763	0.02542	11.7072	**
10	Uganda	Other_Ur	0.10103	0.05749	1.7573	NS
11	Uganda	Rural	0.00216	0.05916	0.0365	NS

Geometric Means Relationship
The missing data are replaced by their Median

Obs	COUNTRY	Average Geometric MeanSlope	Standard Error Geometric MeanSlope	t_obs	Testing Null Hypoth
1	Kenya	0.06170	0.099653	0.61913	NS
2	Rwanda	0.16453	0.031920	5.15433	**
3	Tanzania	0.19111	0.054442	3.51045	**
4	Uganda	0.13593	0.043706	3.11015	**
	Overall	0.14370	0.027609	5.20474	**

Geometric Means Slopes and Intercepts
By Country and Urban Rural Environments
The Missing Data are Replaced by Their Median

Obs	COUNTRY	Parameter	Average Intercept	Average Geometric MeanSlope	Testing Null Hypoth Intercept/Slope
1	Kenya	Major_Ur	0.86327	0.23082	**
2	Kenya	Rural	0.92719	-0.10743	**
3	Rwanda	Major_Ur	1.00938	0.05740	**
4	Rwanda	Other_Ur	0.94160	0.16862	**
5	Rwanda	Rural	0.84450	0.26756	**
6	Tanzania	Major_Ur	0.94173	0.18927	**
7	Tanzania	Other_Ur	0.89213	0.21981	**
8	Tanzania	Rural	0.77251	0.16353	**
9	Uganda	Major_Ur	0.78909	0.29763	**
10	Uganda	Other_Ur	0.84168	0.10103	**
11	Uganda	Rural	0.85073	0.00216	**

Per country Geometric Means Slopes
and Intercepts
The Missing Data are Replaced by Their Median

Obs	COUNTRY	Average Intercept	Average Geometric MeanSlope	Testing Null Hypoth Intercept
1	Kenya	0.89523	0.06170	**
2	Rwanda	0.93183	0.16453	**
3	Tanzania	0.85757	0.19111	**
4	Uganda	0.82613	0.13593	**
	Overall	0.87777	0.14370	**

Final comments and conclusion

Some concluding remarks are warranted. The *Poverty and ICT in Urban and Rural East Africa* (PICTURE) survey was initially designed to interview randomly selected individual family members. The population of interest was defined to represent all the 20 poorest enumeration areas in each one of the four countries (Kenya, Rwanda, Tanzania and Uganda). Subsequently 20 households were randomly selected in each of the enumeration areas. However, such a scenario was not strictly followed. As a consequence, a convenient sampling was actually done since, in practice, the interviewed individuals were the one who happen to “be at home” at the time of the interview.

*I*t was *first* important to correct for such a departure using available statistical means. The method based on *stratification* of propensity scores was successful in reducing bias in the baseline covariates, allowing for a fair assessment of the *partial effect* of ICT ownership on the income defined as a multiple of the poverty line. *Second*, at least in this first wave, the geometric mean regression seems to provide, to a certain degree, a directional free approach to the causal relationship between the two primary factors of interest (income and ICT) *conditional* to the baseline covariates.

Missing data creates *additional* problems in the analysis of the cross-section data. We opted for a *simple approach* of replacing the missing observation by the *median* value of the corresponding covariate with exhibited any missing information. Implicit in the analysis is that data are missing at random, meaning that if we were to re-sample the same population and such sample includes part of the same points, the values would not necessary be missing. This *may* be an unrealistic assumption since no relevant test was conducted. Therefore we propose that in future exploitation of the study, we strive for a more rigorous and scientifically stronger approach to missing data.

Both in the full non-missing data and in the analyses of data with missing values, the slopes must be interpreted to represent the *expected differences* in response of two individuals that *differ by one unit* on the predictor. Therefore an individual who owns an ICT tool (computer, scanner, etc..) has *on the average* a 14% greater income (full data) and 19% (partial data) than one who doesn't. Inversely, two individuals, who differ by one unit in income, are also the likely to exhibit differential own*ership in* ICT appliance. In the absence of additional income, the individual cannot afford an ICT; and conversely those individuals with no ICT remain below the poverty line (less that 1.00). On the average, the income is 0.88 and 0.86 for the full data and data with missing values respectively. These values are expected given the poor areas from which they are derived.

The data from Uganda behave *differently* than those collected in the rest of the East African countries. This is obvious in the estimates of the corresponding and unexpectedly *negative* geometric coefficients in both *Major Urban* and *Rural* environments. When removing the Uganda data from the analysis, both the overall intercept and the slope increased from 0.86 to 0.91 (intercepts), and from 0.19 to 0.35 (slopes). When the medians are used as substitute of missing values, estimates from all the countries appear to remain reasonable. The overall average intercept is 0.88 whereas the overall slope is 0.14. Surprisingly, the estimate slope for Kenya is 3 times less in the full data than it was in the partial data (0.06 vs. 0.21). Kenyan result is perhaps due to many things including the substitution that took place during the fieldwork as well as the manner in which the national poverty line has been calculated. This was reflected on the negative coefficient in *Rural* environment (-0.107).

Finally the *generalization* of the results can only be made relative to the (poor) areas included in the study or relative to areas of similar characteristics (i.e. comparable household and individual traits). *The aim of the study was to provide evidence of the impact of ICTs on poverty for a deliberately selected sample of sites from the poorest area*s. This means that the data is not representative of the national state in the four countries and generalisations at the national level cannot and should not be made. Moreover, the assessment of the relationship between income and ICT was made *conditional* to the

selected controls (assets, education, actual age, capability, their quadratics and interactions, etc…). We strive to remove their influence when studying the relationship of interest. However, using different controls may lead to different conclusions about the causal relationship between income and ICT. From this standpoint, caution must be exercised in the interpretation of the above partial effects. What can be inferred is whether ICTs have a *positive impact* on reduction of household poverty. It is clear that the relationship is definitely positive but its *magnitude* is subject to change depending *on the controls that were used*. It is suggested that, if possible, a *meta-analysis* of *all* the studies that share the same or equivalent structure be conducted. In the meantime, we intend to verify whether the values obtained in the first wave are repeated in the second wave, and in the analysis of the entire panel.

References

1. Austin. P. C (2010): The performance of different propensity score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2,

2. Zanuto E. L. (2006). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science* **4**:67-91

3. May. J. et al (2010). Poverty and Information Communication Technologies in Urban and Rural Eastern Africa (PICTURE-Africa). Case studies from Kenya, Rwanda, Tanzania, and Uganda Summary Report

4. Joffe. M.M and Rosenbaum P. R (1999). Invited Commentary: Propensity Scores. *Amer. J. Epidemiol* 150 327-333

5. Akter A and Awudu A (2019): The adoption of Genetically Modified Cotton and Poverty Reduction in Pakistan. *Journal of Agriculture Economics*. **61**:175-192

6. Milliken A.G and Johnson D.E (2002). Analysis of Messy Data. Vol III. Analysis of Covariance. Chapman Hall/CRC 605 pp

7. Fleiss J. L., Levin B, and Paik M. C (2003). Statistical Methods for Rates and Proportions. 3rd Ed. Willey 760 pp

8. Sheikh K (2007) Investigation of selection bias using inverse probability weighing. *Eur J Epidemiol* **22**:349-350.

9. Stokes M. E, Davis C. S, and Koch G. G (1995) Categorical Analysis Using the SAS System. Cary NC: SAS Institute Inc 499pp

10. Allison Paul .D. (1999). Logistic Regression Using SAS System. Theory and Application. Cary NC SAS Institute Inc 287pp

11. Miettinen. O.S. (1976). Stratification by a Multivariate Confounder score. Am. J. Epidemiol. **104**, 609-620

12. Cochran W.G. (1968). The effectiveness of Adjustment by Stratification in Removing Bias in Observational Studies. *Biometrics* **24** 2 295-313

13. Cochran W.G., Rubin (1973). Controlling Bias in Observational Studies: A Review, Synkya, Ser A, **35**, 417-446

14. Smith J.A., Todd P.E (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, **125** 305-353

15. Yafeee R. (2003). A Prime for Panel Data Analysis. New York University. Information Technology Services

16. Bruderl J. (2005). Panel Data Analysis. http://www2.sowi.uni-mannheim.de/lsssm/veranst/ Panelanalyse.pdf

17. Cheng Hsiao (2003). Analysis of Panel Data. Cambridge University Press 366pp

18. Wooldridge J.M (2005). Violating ignorability of treatment by controlling for too much factors. Econometric Theory **21**, 1026-1028.

19. Wooldridge J.M (2002). Econometric Analysis of Cross section and Panel Data. MIT Press Cambridge, Massachusetts. 752pp

20. Rosenbaum R.P, D.B. Rubin (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70** 41-55

21. Rosenbaum P, D. Rubin (1984). Reducing Bias in Observational Studies Using Subclassification on the propensity Score. *Journ.l of the Am. Statistical Association* **79**, 387 516-524

22. SAS Technical Report (19--). Introduction to Logistic Regression. Handout. SAS Institute Inc. SAS Campus Drive, Cary NC 27513

23. SAS Institute Inc (1995). *Logistic regression Examples Using the SAS System*, Version 6, first edition, cary NC: SAS Institute inc. 163pp.

24. Pregibon, D (19840. Data Analytical Methods for Matched Case-Control Studies. *Biometrics*, **40**, 639-651

25. Levy P.S, S. Lemeshow (1999). Sampling of Populations. Methods and Applications 3rd Ed Wiley Series in probability Statistics 525pp

26. Steel, R.G, and T.H. Torrie (1980). Principles and Procedures of Statistics. A Biometric Approach. 2nd ed. 633pp

27. Snedecor G.W and W.G Cochran, 1989. Statistical Methods, 8th Ed.503pp

28. Rubin D. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**, 185-203

29. Draper, N.R and Y Yang (1996). Generalization of the geometric mean functional relationship. *Computational Statistics and Data Analysis* **23** 355-372.

30. Barker F, Soh, Y.C, and R.J. Evans (1988). Properties of the Geometric Mean Functional Relationship. *Biometrics*, 44, **1** 279-281

31. Leng L, T. Zhang, L Kleinman, and W. Zhu (2007). Ordinary Least Square Regression, Orthogonal Regression, Geometric Mean Regression and their applications in Aerosol Science. *Journal of Physics* Conference Series **78** 012084

32. Curtis, L, H et al (2007). Using Inverse Probability-Weighted Estimators in Comparative Effectiveness Analysis with Observational Databases. Medical care 45, 10 (2) S103-S107.

33. Halfon, E. (1085). Regression Method in Ecotoxicology: A better formulation using the Geometric Mean Functional Regression. Notes. Environ. Sci. Technol. **19**, 747-749.

34. Gillard J.W. (2006) An historical review of linear regression with errors in both variables. School of Mathematics, Senghenydd Rd Cardiff University

35. Gillard J.W. and T.C. Illes (2006). Variance Covariance matrices for Linear regression with Errors in both variables. School of Mathematics, Senghenydd Rd Cardiff University

36. Carroll J. R (1998). Measurement Error in Epidemiologic Studies

37. Sprent P. and G.R. Dolby (1980). Query: the geometric mean functional relationship. Biometrics, **36** (3) 547-550

38. Mahlon S. Wilson and Shinichi Ichikawa (1989) Comparison between the Geometric and Harmonic Mean Electronegativity Equilibration Techniques. *J. Phys. Chem.* **93**, 3087-3089

39. Derr R.E (). Performing Exact Logistic Regression with the SAS System *SUGI* paper 254-25

40. Little R.C (2007). Repeated Measures Analyses with clustered subjects *SUGI* paper 178-2007

41. Leslie S. R and H. Ghomrawi (2008). The use of propensity scores and Instrumental Variable Methods to adjust for treatment selection bias. *SUGI* paper 366-2008

42. Beck, C. A (2009). Selection Bias in observational studies. Out of control? *Neurology* 72 108-109

43. Kopec J. A and J.M. Esdaile (1990). Bias in case-control studies. A review. Journal of Epidemiology and Community Health **44**: 179-186

44. Pasta D. J (2000). Using the propensity scores to adjust for group differences: Examples comparing alternative surgical methods. SUGI paper 261-265

45. D'Agostino R.B (1998). Tutorial in Biostatistics Propensity Score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat, Med **17**, 2265-2281

46. Rosenbaum P. R and Rubin D.B (1984). Reducing bias in observational studies using sub classification on the propensity score. Journal of Amer. Stat. Assoc. **79**, 516-524

47. Ricker W. E. (1973). Linear regressions in Fishery research. *J Fish. Res. Board Can.*, **30**: 409-434

48. Weisberg, S. (1985). *Applied Linear Regression* (2nd ed). New York: John Wiley

49. Fuller, W. A (1987). *Measurement error models.* New York: John Wiley

50. Carroll R. J and D. Ruppert (1996). The use and Misuse of Orthogonal Regression in Linear Error-in-variables Models. The American Statistician, **50**, 1 1-6

IMPLEMENTATION THROUGH SAS® Software

1. The SAS program

```
%macro prop(n,indep,titl);
data one&n;
        set one;

******* Response Surface model **********;

proc rsreg data=one&n noprint;
        weight weit;
model interv = hhsize maristatus gender Actualage education Assets
Vulnerabilit
                        Capabilities Physical Exclusion Services;
run;

*************************************************************************;
****** Logistic Regression ***************************************;
***** Is used to predict probabilities of having one ICT appliance ****;
*************************************************************************;
**** In fitting the logistic regression, we choose not to *************;
**** include INCOME and its INTERACTIONS with others since ************;
**** we assume that TOTAL ICT can be a response to levels of **********;

proc logistic descending;
        class maristatus country;
 model interv = &indep;
output out=probs&n predicted=propensity&n;

data probs&n;
        set probs&n;
        if propensity&n ne .;
label
        propensity&n="Propensity*Reduced*Model";
run;

************************************************************;
**** The probabilities are sorted and grouped ***************;
**** So that ALL SIMILAR values can be grouped together ******;
**** These groups formed a homogenoue groups of **************;
**** similar baseline covariate *****************************;
************************************************************;

proc sort data=probs&n;
        by propensity&n;
run;
```

```
        proc print data=probs&n(obs=20);
        title1 Sorted Prpensities all;
                                        var Education Assets Vulnerabilit
propensity&n;
        run;

************************************************************;
*** Distribution of propensity scores in the two groups *******;
************************************************************;

        proc univariate data=probs&n;
                class interv;
                var propensity&n;
        histogram propensity&n;
        run;

        proc means;
                class interv;
                var propensity&n;
        run;
************************************************;

data quantal1 quantal2 quantal3 quantal4 quantal5;
        set probs&n;
        if _n_ <= 603 then output quantal1;
        if 603 < _n_ <= 1206 then output quantal2;
        if 1206 < _n_ <= 1809 then output quantal3;
        if 1809 < _n_ <= 2412 then output quantal4;
        if _n_ > 2412 then output quantal5;

************************ Quantal 1 ************;
                        data quantal1;
                                set quantal1;
                        stratum=1;

                                proc print data=quantal1(obs=20);
                                        title1 Sorted Prpensities stratum 1;
                                        var ICT_email ICT_mobile Education
Assets Vulnerabilit propensity&n;
                                run;

                        proc means noprint;
                                var Education Assets Vulnerabilit
                                                Capabilities Physical
Exclusion;id Stratum;
                                output out=m1 mean= meducation mAssets
mVulnerabilit
                                                mCapabilities mPhysical
mExclusion
                                                        cv = CVeducation
CVAssets CVVulnerabilit
                                                CVCapabilities CVPhysical
CVExclusion;
                        run;
************************ Esimate of the difference Treated vs Control *****;
```

```
                proc sort data=quantal1;
                        by interv;

                        proc means data=quantal1;
                                by interv;
                        var ICT_email ICT_mobile ICT_computer ICT_printer
ICT_Scanner
                                ICT_Internet ICT_total;
                        output out=diff1 mean=MeanEmail MeanMobile
MeanComputer MeanPrinter MeanScanner

        MeanInternet Mean
                                                        var = VAREmail
VARMobile VARComputer VARPrinter VARScanner
                                                                n = nEmail
nMobile nComputer nPrinter nScanner;
                run;

*********************************;
 proc transpose data=diff1
                        out=design1(rename=(col1=MControl col2=MIntervwed

        _name_=VariableInterest));
                                                        var MeanEmail
MeanMobile MeanComputer MeanPrinter MeanScanner;
 run;

data design1;
        set design1;
        quantal=1;
        Code=_n_;
 proc print; run;

 proc sort;
        by code;
 run;

*****************************;
proc transpose data=diff1
                        out=design2(rename=(col1=VARControl
col2=VARIntervwed

        _name_=VariableInterest));
                                                        var VAREmail VARMobile
VARComputer VARPrinter VARScanner;
 run;

 data design2;
        set design2;
        quantal=1;
        Code=_n_;

 proc sort;
        by code;
```

```
 run;
 proc print;run;

**************************;
proc transpose data=diff1
                        out=design3(rename=(col1=nControl col2=nIntervwed

        _name_=VariableInterest));
                                                        var nEmail nMobile
nComputer nPrinter nScanner;
 run;

 data design3;
        set design3;
        quantal=1;
        Code=_n_;
proc print;run;

 proc sort;
        by code;
 run;
 proc print;
run;

 *************************;

data design123(drop=VariableInterest);
        merge design1 design2 design3;
        by code;
 run;
 proc print; run;

************************ Quantal 2 *********************;
                        data quantal2;
                                set quantal2;
                        stratum=2;

                                proc print data=quantal2(obs=20);
                                        title1 Sorted Prpensities stratum 2;
                                        var Education Assets Vulnerabilit
propensity&n;
                                run;

                        proc means noprint;
                                var Education Assets Vulnerabilit
                                                Capabilities Physical
Exclusion;id Stratum;
                                output out=m2 mean= meducation mAssets
mVulnerabilit
                                                mCapabilities mPhysical
mExclusion
                                                cv= CVeducation CVAssets
CVVulnerabilit
                                                CVCapabilities CVPhysical
CVExclusion;
```

```
                        run;

**************************Quantal 3 ************************;
                        data quantal3;
                                set quantal3;
                        stratum=3;

                                proc print data=quantal3(obs=20);
                                        title1 Sorted Prpensities stratum 3;
                                        var Education Assets Vulnerabilit
propensity&n;
                                run;

                        proc means noprint;
                                var Education Assets Vulnerabilit
                                                Capabilities Physical
Exclusion;id Stratum;
                                output out=m3 mean= meducation mAssets
mVulnerabilit
                                                mCapabilities mPhysical
mExclusion
                                                cv=CVeducation CVAssets
CVVulnerabilit
                                                CVCapabilities CVPhysical
CVExclusion;
                        run;


********************** Quantal 4 *********************;
                        data quantal4;
                                set quantal4;
                        stratum=4;

                                proc print data=quantal4(obs=20);
                                        title1 Sorted Prpensities stratum 4;
                                        var Education Assets Vulnerabilit
propensity&n;
                                run;

                        proc means noprint;
                                var Education Assets Vulnerabilit
                                                Capabilities Physical
Exclusion;id Stratum;
                                output out=m4 mean= meducation mAssets
mVulnerabilit
                                                mCapabilities mPhysical
mExclusion
                                                cv=CVeducation CVAssets
CVVulnerabilit
                                                CVCapabilities CVPhysical
CVExclusion;;
                        run;

                ************************** Quantile 5 ******************;
                        data quantal5;
```

```
                        set quantal5;
                stratum=5;

                        proc print data=quantal5(obs=20);
                                title1 Sorted Prpensities stratum 5;
                                var Education Assets Vulnerabilit
propensity&n;
                        run;

                proc means noprint;
                        var Education Assets Vulnerabilit
                                        Capabilities Physical
Exclusion;id Stratum;
                        output out=m5 mean= meducation mAssets
mVulnerabilit
                                        mCapabilities mPhysical
mExclusion
                                        cv=CVeducation CVAssets
CVVulnerabilit
                                        CVCapabilities CVPhysical
CVExclusion;
                run;

data allstratam;
        set m1
                m2
                m3
                m4
                m5;

proc print;
        var stratum meducation mAssets mVulnerabilit mCapabilities
mPhysical mExclusion;
        title1 Sorted Mean Prpensity Scores ALL strata;
run;

proc print;
        var stratum CVeducation CVAssets CVVulnerabilit
CVCapabilities CVPhysical CVExclusion;
        title1 Sorted Coefficients Prpensity Scores ALL strata;
run;

title1 Model Selection in Logistic Regression;
title2 Create a data set that show PROPENSITY SCORES;
title3 "Model &n &titl";
run;

data stratall;
        set quantal1 quantal2 quantal3 quantal4 quantal5;
        if ICT_total ge 3 then ICT_total=3;
 label Interv='Stay Home and Interviewed';
        proc format;
          value country 1='Uganda'
                                2='Tanzania'
                                3='Rwanda'
```

```
                                                    4='Kenya'
                                                    ;
                    value intervfmt 1='Stay Home and Was Interviewed'
                                    0='Away from Home Was not Interviewed';
                    value IntpFmt    1='Yes'
                                     0='No';
                    value EmailFmt   1='Yes'
                                     0='No';
                    value mobileFmt     1='Yes'
                                     0='No';
                    value  computerFmt          1='Yes'
                                     0='No';
                    value  printerFmt           1='Yes'
                                     0='No';
                    value scannerFmt            1='Yes'
                                     0='No';
                    value internetFmt   1='Yes'
                                     0='No';
                          value totalFmt 1='Have At Least One ICT'
                                     0='Have Not';
                run;

                proc sort ;
                        by ICT_total;
                  by descending interv descending
                                ICT_email descending ICT_mobile descending
                                ICT_computer descending ICT_printer
descending
                                 ICT_scanner descending ICT_internet;
                run;

        proc freq order=data;
                    format interv intervfmt.
                                                ICT_total totalFmt.
                        ;
                *tables interv*(ICT_email ICT_mobile ICT_computer
ICT_printer
                                 ICT_scanner ICT_internet)/chisq relrisk;
                tables interv*ICT_email/chisq relrisk;
                output out=ChiSqData n pchi lrchi;
                exact pchi or;
                run;

        proc print
data=ChiSqData(rename=(XP_PCHI=Two_sided_Fisher_PValue
                                P_PCHI=Asymptotic_Pr_ChiSq
P_LRCHI=Likelihood_Ratio_Chi_Square
                                    )) noobs;
                title1 'Chi-Square Statistics for Association Treatment and ICT
Poverty';
                title2 'Two Sided Fisher Exact Test';
run;

*****************************************************************;
*** Estimate of treatment (interview) effect within each stratum ****;
```

```
*** Using PROC CATMOD **********;
*****************************************************************;

data stratall;
        set stratall;
if propensity&n le .60;
proc sort; by stratum;

        proc reg;
                A: model ICT_total=income;
                B: model Income=ICT_total;
        run;

        proc means;
                class interv;
                var ICT_email;
        run;

        proc sort data=stratall;
                        by stratum;
        run;

        proc means data=stratall;
                by stratum;
                var Assets Vulnerabilit Capabilities Physical Exclusion
         ICT_computer ICT_printer ICT_internet ICT_total;
        output out=mm var=VARAssets VARVulnerabilit VARCapabilities
VARPhysical VARExclusion
         VARICT_computer VarICT_printer
                                VARICT_internet VARICT_total;

        proc print data=mm(drop=_type_ _freq_) noobs;
        title1 Check for Equal Variance ;
        title2 Variances are AFTER stratification in FIVE Strata;
        title3 ICT Individual Dimensions;
        run;

%macro covmod(m,dep,indep,titl);
data stratglm&m;
        set stratall;
        if &dep ne .;

proc glm NOPRINT;
        class ICT_email;
                by stratum;
model interv=ICT_email;
lsmeans ICT_email;
estimate 'Treatment Difference' ICT_email -1 1;
run;
*****************************************************************;
**** In the second stage, we corrects for self-selection *************;
**** by incorporating a transformation of these predicted individual**;
****** probabilities as an additional explanatory variable ***********;
*****************************************************************;
*** The analysis is done ONLY FOR THE TREATED GROUP *****************;
```

```
*** to correspond to the conditional expectation of ICT USAGE given **;
*** GIVEN the person STAYS AT HOME *********************************;
************************************************;
****** Covariance Adjustment method ***************;
************************************************;
**** Using General Model for Binary Outcome *******;
**** This model does NOT constraint the probabilities ***********;
**** to be between 0 and 1 *********;
************************************************;
******* Within Stratum *******;
***** General Linear Model ***;

proc glm NOPRINT;
        class Interv country rural_urban gender;
        by stratum;
model &dep = Interv country rural_urban(country) gender
&indep*rural_urban(country)
                                propensity&n/ss3;

estimate "Slope TZN Major Urban" &indep*rural_urban(country) 1;
estimate "Slope TZN Other Urban" &indep*rural_urban(country) 0 1;
estimate "Slope TZN Rurual" &indep*rural_urban(country)      0 0 1;
estimate "Slope KNY Major Urban" &indep*rural_urban(country) 0 0 0 1;
estimate "Slope KNY Rural" &indep*rural_urban(country)       0 0 0 0 1;
estimate "Slope RWD Major Urban" &indep*rural_urban(country) 0 0 0 0 0 1;
estimate "Slope RWD Other Urban" &indep*rural_urban(country) 0 0 0 0 0 0
1;
estimate "Slope RWD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0 0
1;
estimate "Slope UGD Major Urban" &indep*rural_urban(country) 0 0 0 0 0 0
0 0 1;
estimate "Slope UGD Other Urban" &indep*rural_urban(country) 0 0 0 0 0 0
0 0 0 1;
estimate "Slope UGD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0 0 0
0 0 1;
ods output estimates=Slope&indep;

title1 Geometric Mean Functional Relationship;
title2 Variables of Interest are: Digital Poverty Dimensions;
title3 and ICT Poverty Levels;
run;

******* Across Strata ************;
***** General Linear Model *******;
****** UnWeighted Analysis *******;

proc glm data=stratglm&m;
        class Interv country rural_urban gender;
model &dep = Interv country rural_urban(country) gender
&indep*rural_urban(country)

        propensity&n/ss3 SOLUTION;
estimate "Slope ALLTZN Major Urban" &indep*rural_urban(country) 1;
estimate "Slope ALLTZN Other Urban" &indep*rural_urban(country) 0 1;
estimate "Slope ALLTZN Rurual" &indep*rural_urban(country)      0 0 1;
```

```
estimate "Slope ALLKNY Major Urban" &indep*rural_urban(country) 0 0 0 1;
estimate "Slope ALLKNY Rural" &indep*rural_urban(country)       0 0 0 0 1;
estimate "Slope ALLRWD Major Urban" &indep*rural_urban(country) 0 0 0 0
0 1;
estimate "Slope ALLRWD Other Urban" &indep*rural_urban(country) 0 0 0 0
0 0 1;
estimate "Slope ALLRWD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0
0 1;
estimate "Slope ALLUGD Major Urban" &indep*rural_urban(country) 0 0 0 0
0 0 0 0 1;
estimate "Slope ALLUGD Other Urban" &indep*rural_urban(country) 0 0 0 0
0 0 0 0 0 1;
estimate "Slope ALLUGD Rural"       &indep*rural_urban(country) 0 0 0 0 0 0
0 0 0 0 1;
ods output estimates=Slop&indep;
title1 Geometric Mean Functional Relationship;
title2 Variables of Interest are: Digital Poverty Dimensions;
title3 and ICT Poverty Levels;
run;

data slop;
        set Slop&indep(drop=tValue Probt);
proc print;
run;

**********************************;
**** Intercept *****************;
**********************************;

proc sort data=stratglm&m;
by stratum country rural_urban;
run;

proc means noprint data=stratglm&m;
by stratum country rural_urban;
        var income ICT_total;
output out=lm(drop=_freq_ _type_) mean=MeanInterceptIncome
MeanInterceptICT_total;
run;

data lm;
        set lm;
        if rural_urban ne .;
run;
proc print;
run;

%mend;
%covmod(m=1, dep=Income, indep=ICT_total);run;
%covmod(m=2, dep=ICT_total, indep=Income);run;
quit;

%mend;
*%prop(n=1,indep=hhsize maristatus gender education Assets Vulnerabilit
Capabilities Physical
```

```
  Exclusion education*gender education*education Assets*gender
  Assets*education Assets*Assets  Vulnerabilit*gender Vulnerabilit*education
  Vulnerabilit*Assets Vulnerabilit*Vulnerabilit Capabilities*gender
Capabilities*education
  Capabilities*Assets Capabilities*Vulnerabilit Capabilities*Capabilities
  Physical*gender Physical*education Physical*Assets
  Physical*Vulnerabilit Physical*Capabilities Physical*Physical
  Exclusion*gender Exclusion*education Exclusion*Assets
Exclusion*Vulnerabilit
  Exclusion*Capabilities Exclusion*Physical Exclusion*Exclusion
  Services*gender Services*education Services*Assets Services*Vulnerabilit
  Services*Capabilities Services*Physical Services*Exclusion
  maristatus*gender maristatus*education maristatus*assets
maristatus*vulnerabilit
  maristatus*Capabilities maristatus*Physical maristatus*Exclusion
  hhsize*maristatus hhsize*gender hhsize*education hhsize*assets
hhsize*vulnerabilit
  hhsize*Capabilities hhsize*Physical hhsize*Exclusion,titl=FULL MODEL);

*%prop(n=2,indep=hhsize maristatus gender,titl=Very REDUCED MODEL);

%prop(n=2,indep=hhsize maristatus gender education Assets Vulnerabilit
                                  Capabilities Physical Exclusion
                                  hhsize*gender gender*education
education*education
                                  gender*capabilities education*capabilities
 ,titl=REDUCED MODEL);

run;

*****************************************************************;
*** Because AT THE MOMENT we don't have the right PROC to perform **
*** the variable selection sas VS 9.1.1 does not have PROC GLMSELECT ;
**** We temporarily use PROC RSREG to do the job, even though it is ;
*** modeling a binary response ***;
*****************************************************************;

quit;
```

2. SAS log

The portion below verifies that the program ran correctly

```
NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M3)
      Licensed to Univ of Kwazulu Natal, Site 0084768003.
NOTE: This session is executing on the XP_PRO  platform.


NOTE: SAS 9.1.3 Service Pack 2

NOTE: SAS initialization used:
      real time           1.42 seconds
      cpu time            0.35 seconds

NOTE: There were 8055 observations read from the data set WORK.ONE.
NOTE: The data set WORK.ONE2 has 8055 observations and 75 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


NOTE: PROCEDURE RSREG used (Total process time):
      real time           0.12 seconds
      cpu time            0.12 seconds

NOTE: PROC LOGISTIC is modeling the probability that interv=1.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 8055 observations read from the data set WORK.ONE2.
NOTE: The data set WORK.PROBS2 has 8055 observations and 77 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.07 seconds
      cpu time            0.06 seconds

NOTE: There were 8055 observations read from the data set WORK.PROBS2.
NOTE: The data set WORK.PROBS2 has 3016 observations and 77 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds

NOTE: There were 3016 observations read from the data set WORK.PROBS2.
NOTE: The data set WORK.PROBS2 has 3016 observations and 77 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds

NOTE: There were 20 observations read from the data set WORK.PROBS2.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.12 seconds
      cpu time            0.12 seconds


NOTE: There were 3016 observations read from the data set WORK.PROBS2.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      cpu time            0.03 seconds
```

```
NOTE: There were 3016 observations read from the data set WORK.PROBS2.
NOTE: The data set WORK.QUANTAL1 has 603 observations and 77 variables.
NOTE: The data set WORK.QUANTAL2 has 603 observations and 77 variables.
NOTE: The data set WORK.QUANTAL3 has 603 observations and 77 variables.
NOTE: The data set WORK.QUANTAL4 has 603 observations and 77 variables.
NOTE: The data set WORK.QUANTAL5 has 604 observations and 77 variables.
NOTE: DATA statement used (Total process time):
      real time          0.03 seconds
      cpu time           0.01 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL1.
NOTE: The data set WORK.QUANTAL1 has 603 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 20 observations read from the data set WORK.QUANTAL1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL1.
NOTE: The data set WORK.M1 has 1 observations and 15 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL1.
NOTE: The data set WORK.QUANTAL1 has 603 observations and 78 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL1.
NOTE: The data set WORK.DIFF1 has 2 observations and 20 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 2 observations read from the data set WORK.DIFF1.
NOTE: The data set WORK.DESIGN1 has 5 observations and 3 variables.
NOTE: PROCEDURE TRANSPOSE used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN1.
NOTE: The data set WORK.DESIGN1 has 5 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds



NOTE: There were 5 observations read from the data set WORK.DESIGN1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN1.
NOTE: The data set WORK.DESIGN1 has 5 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 2 observations read from the data set WORK.DIFF1.
NOTE: The data set WORK.DESIGN2 has 5 observations and 3 variables.
NOTE: PROCEDURE TRANSPOSE used (Total process time):
      real time          0.00 seconds
```

```
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN2.
NOTE: The data set WORK.DESIGN2 has 5 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN2.
NOTE: The data set WORK.DESIGN2 has 5 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN2.
NOTE: PROCEDURE PRINT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 2 observations read from the data set WORK.DIFF1.
NOTE: The data set WORK.DESIGN3 has 5 observations and 3 variables.
NOTE: PROCEDURE TRANSPOSE used (Total process time):
      real time         0.01 seconds
      cpu time          0.01 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN3.
NOTE: The data set WORK.DESIGN3 has 5 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN3.
NOTE: PROCEDURE PRINT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN3.
NOTE: The data set WORK.DESIGN3 has 5 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds
NOTE: There were 5 observations read from the data set WORK.DESIGN3.
NOTE: PROCEDURE PRINT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN1.
NOTE: There were 5 observations read from the data set WORK.DESIGN2.
NOTE: There were 5 observations read from the data set WORK.DESIGN3.
NOTE: The data set WORK.DESIGN123 has 5 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 5 observations read from the data set WORK.DESIGN123.
NOTE: PROCEDURE PRINT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL2.
NOTE: The data set WORK.QUANTAL2 has 603 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds

NOTE: There were 20 observations read from the data set WORK.QUANTAL2.
NOTE: PROCEDURE PRINT used (Total process time):
      real time         0.00 seconds
      cpu time          0.00 seconds
```

```
NOTE: There were 603 observations read from the data set WORK.QUANTAL2.
NOTE: The data set WORK.M2 has 1 observations and 15 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.03 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL3.
NOTE: The data set WORK.QUANTAL3 has 603 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 20 observations read from the data set WORK.QUANTAL3.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL3.
NOTE: The data set WORK.M3 has 1 observations and 15 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.00 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL4.
NOTE: The data set WORK.QUANTAL4 has 603 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds
NOTE: There were 20 observations read from the data set WORK.QUANTAL4.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds

NOTE: There were 603 observations read from the data set WORK.QUANTAL4.
NOTE: The data set WORK.M4 has 1 observations and 15 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 604 observations read from the data set WORK.QUANTAL5.
NOTE: The data set WORK.QUANTAL5 has 604 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 20 observations read from the data set WORK.QUANTAL5.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 604 observations read from the data set WORK.QUANTAL5.
NOTE: The data set WORK.M5 has 1 observations and 15 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 1 observations read from the data set WORK.M1.
NOTE: There were 1 observations read from the data set WORK.M2.
NOTE: There were 1 observations read from the data set WORK.M3.
NOTE: There were 1 observations read from the data set WORK.M4.
```

```
NOTE: There were 1 observations read from the data set WORK.M5.
NOTE: The data set WORK.ALLSTRATAM has 5 observations and 15 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 5 observations read from the data set WORK.ALLSTRATAM.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 5 observations read from the data set WORK.ALLSTRATAM.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 603 observations read from the data set WORK.QUANTAL1.
NOTE: There were 603 observations read from the data set WORK.QUANTAL2.
NOTE: There were 603 observations read from the data set WORK.QUANTAL3.
NOTE: There were 603 observations read from the data set WORK.QUANTAL4.
NOTE: There were 604 observations read from the data set WORK.QUANTAL5.
NOTE: The data set WORK.STRATALL has 3016 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: Format COUNTRY is already on the library.
NOTE: Format COUNTRY has been output.
NOTE: Format INTERVFMT is already on the library.
NOTE: Format INTERVFMT has been output.
NOTE: Format INTPFMT is already on the library.
NOTE: Format INTPFMT has been output.
NOTE: Format EMAILFMT is already on the library.
NOTE: Format EMAILFMT has been output.
NOTE: Format MOBILEFMT is already on the library.
NOTE: Format MOBILEFMT has been output.
NOTE: Format COMPUTERFMT is already on the library.
NOTE: Format COMPUTERFMT has been output.
NOTE: Format PRINTERFMT is already on the library.
NOTE: Format PRINTERFMT has been output.
NOTE: Format SCANNERFMT is already on the library.
NOTE: Format SCANNERFMT has been output.
NOTE: Format INTERNETFMT is already on the library.
NOTE: Format INTERNETFMT has been output.
NOTE: Format TOTALFMT is already on the library.
NOTE: Format TOTALFMT has been output.

NOTE: PROCEDURE FORMAT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3016 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.STRATALL has 3016 observations and 78 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 3016 observations read from the data set WORK.STRATALL.
```

```
NOTE: The data set WORK.CHISQDATA has 1 observations and 8 variables.
NOTE: PROCEDURE FREQ used (Total process time):
      real time          0.01 seconds
      cpu time           0.00 seconds


NOTE: There were 1 observations read from the data set WORK.CHISQDATA.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3016 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.STRATALL has 3002 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.STRATALL has 3002 observations and 78 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: PROCEDURE REG used (Total process time):
      real time          0.04 seconds
      cpu time           0.03 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATALL.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.MM has 5 observations and 12 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 5 observations read from the data set WORK.MM.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.STRATGLM1 has 2950 observations and 78 variables.
NOTE: DATA statement used (Total process time):
```

```
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: Interactivity disabled with BY processing.

NOTE: PROCEDURE GLM used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds



NOTE: Interactivity disabled with BY processing.
NOTE: Slope UGD Rural is not estimable.
NOTE: The above message was for the following by-group:
      stratum=1

NOTE: PROCEDURE GLM used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

WARNING: Output 'estimates' was not created.  Make sure that the output object name, label, or
         path is spelled correctly.  Also, verify that the appropriate procedure options are used
         to produce the requested output object.  For example, verify that the NOPRINT option is
         not used.
WARNING: The current ODS SELECT/EXCLUDE/OUTPUT statement was cleared because the
end of a
         procedure step was detected. Probable causes for this include the non-termination of an
         interactive procedure (type quit; to end the procedure) and a run group with no output.

NOTE: The data set WORK.SLOPICT_TOTAL has 11 observations and 6 variables.
NOTE: PROCEDURE GLM used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds



NOTE: There were 11 observations read from the data set WORK.SLOPICT_TOTAL.
NOTE: The data set WORK.SLOP has 11 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds



NOTE: There were 11 observations read from the data set WORK.SLOP.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds



NOTE: There were 2950 observations read from the data set WORK.STRATGLM1.
NOTE: The data set WORK.STRATGLM1 has 2950 observations and 78 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds



NOTE: There were 2950 observations read from the data set WORK.STRATGLM1.
NOTE: The data set WORK.LM has 59 observations and 5 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds



NOTE: There were 59 observations read from the data set WORK.LM.
```

```
NOTE: The data set WORK.LM has 55 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 55 observations read from the data set WORK.LM.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATALL.
NOTE: The data set WORK.STRATGLM2 has 3002 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: Interactivity disabled with BY processing.

NOTE: PROCEDURE GLM used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: Interactivity disabled with BY processing.

NOTE: PROCEDURE GLM used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

WARNING: Output 'estimates' was not created.  Make sure that the output object name, label, or
         path is spelled correctly.  Also, verify that the appropriate procedure options are used
         to produce the requested output object.  For example, verify that the NOPRINT option is
         not used.
WARNING: The current ODS SELECT/EXCLUDE/OUTPUT statement was cleared because the
end of a
         procedure step was detected. Probable causes for this include the non-termination of an
         interactive procedure (type quit; to end the procedure) and a run group with no output.

NOTE: The data set WORK.SLOPINCOME has 11 observations and 6 variables.
NOTE: PROCEDURE GLM used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 11 observations read from the data set WORK.SLOPINCOME.
NOTE: The data set WORK.SLOP has 11 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 11 observations read from the data set WORK.SLOP.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATGLM2.
NOTE: The data set WORK.STRATGLM2 has 3002 observations and 78 variables.
NOTE: PROCEDURE SORT used (Total process time):
```

```
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 3002 observations read from the data set WORK.STRATGLM2.
NOTE: The data set WORK.LM has 59 observations and 5 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds


NOTE: There were 59 observations read from the data set WORK.LM.
NOTE: The data set WORK.LM has 55 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


NOTE: There were 55 observations read from the data set WORK.LM.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds


77412
77413  run;
77414
77415  ************************************************************************;
77416  *** Because AT THE MOMENT we don't have the right PROC to perform **
77417  *** the variable selection sas VS 9.1.1 does not have PROC GLMSELECT ;
77418  **** We temporarily use PROC RSREG to do the job, even though it is ;
77419  *** modeling a binary response ***;
77420  ************************************************************************;
77421
```