August 8, 1996

HOUSEHOLD EXPENDITURE AND INCOME DATA FOR TRANSITIONAL ECONOMIES (HEIDE): DATA CLEANING AND RENT IMPUTATION

Appendix 1 of RAD project
"Poverty and Targeting of Social Assistance
in Eastern Europe and the Former Soviet Union"

by

Robert Ackland
Jeanine Braithwaite
Mark Foley
Thesia Garner
Christiaan Grootaert
Branko Milanovic
Gi-Taik Oh
Sandor Sipos
Sasoun Tsirounian
Yvonne Ying

1. Data Cleaning

This section summarizes the techniques used to clean the Household Expenditure and Income Data for Transitional Economies (HEIDE). The two stages in the data cleaning process involved the treatment of outliers and missing values (see below for a precise definition of these terms). The procedures used in the data cleaning are 'mechanical' in nature in that they involved the use of SAS macros which were applied uniformly across the variables. It was deemed necessary to use mechanical procedures first because conformity of procedures across data sets is desirable and second because the large number of variables and observations involved (the HEIDE contains over 3 million observations) made data cleaning using visual or 'variable-by-variable' techniques infeasible.[1]

Before describing the data cleaning procedures, some definitions are necessary. The HEIDE is divided into four types of variables: expenditure, income, asset and descriptive variables. Data cleaning was only performed on the expenditure and income variables (XX - say why). Within the expenditure and income variables, there are three 'classes' or 'levels'. The first class of variable contains aggregate variables and there are two such variables: total expenditure (TOTHHX) and total disposable income (TOTHHY).[2] Sub-aggregates are those variables that directly sum up to the aggregates. For example, the sub-aggregate income variables are wages (WAGEY), self-employment income (SELFEMY), self-consumption (SELFCNY) etc. (see HEIDE variable list for further details).

Component variables are those variables that directly sum up to the sub-aggregates. For example, the expenditure sub-aggregate variable housing expenditure (HOUSEX) is the sum of two component variables rent (RENTX) and other housing expenditures (OTHOUSEX). In the expenditure data there is only one sub-aggregate with component variables (HOUSEX), while in the income data there are four sub-aggregates which consist of component variables (WAGEY, SELFEMY, TOTPENY and TAXESY). Note that for an individual country a particular sub-aggregate or component variable may not exist i.e. data may not have been collected for that particular variable or may not be available in the data file.

1.1 Outliers

For the purposes of the construction of HEIDE, outliers were defined as those observations which deviated by more than 5 standard deviations from the mean. In all of the datasets (except Armenia) outliers were replaced with means using a two-stage

[1] The design of the Armenia survey was very different to that of the other countries in the HEIDE (for example, in the Armenia survey food purchases were recorded only for the previous day while in the other datasets the reference period was generally one month). For this reason, it was decided that the mechanical data cleaning techniques were not appropriate for the Armenia data. See separate appendix for details.

[2] In the HEIDE expenditure variables have the suffix 'X' while income variables have the suffix 'Y'.

procedure contained in a SAS macro. It is generally desirable to minimize the influence of outliers as their presence will affect both means and regression coefficients. It was considered particularly necessary to adjust for the presence of outliers in the HEIDE because some datasets in the database may be more prone to outliers and adjustment was therefore necessary before valid cross-country comparisons could be made.

Before detailing the exact procedure of outlier adjustment, it is necessary to establish exactly which variables were adjusted. If a sub-aggregate variable did not consist of component variables, then outliers were adjusted for in that sub-aggregate. If a sub-aggregate consisted of component variables then the outlier procedure was applied to the components and the sub-aggregate was calculated as the sum of the outlier adjusted components. It is important to note that a sub-aggregate which consisted of component variables did not have the outlier adjustment procedure subsequently applied to it.

Once the sub-aggregates were all adjusted for outliers (either through their components or directly), they were summed to the aggregates. There was no further outlier adjustment in the aggregates. There were practical reasons for this: if an aggregate was found to have outliers and they were adjusted for, then the sub-aggregates would no longer sum up to the aggregate. It is also for this reason that for those sub-aggregates with components, the outlier procedure was only applied to the components.

The following steps were involved in the outlier adjustment procedure:

1. All variables were measured at the household level and in monthly per capita terms. Since income and in particular expenditure variables are generally increasing with household size, using the total value of the variables would lead to large households being more often identified as having outliers.

2. Outlier identification and adjustment were performed only over the positive observations for each variable. One reason for this is that for some variables many households record zero values (the tax variable for countries from the FSU is a good example) and therefore the mean for such variables will be so low that too many observations will be identified as outliers. Further, it is appropriate to replace an outlier with a mean which has been calculated only over the positive observations and not all observations; by replacing an outlier with a mean calculated over *all* observations we would be implicitly assuming that the true value of the variable could have been zero, while in fact we know that it was positive.

3. Certain steps were taken to ensure that total disposable income (TOTHHY) was always positive -- a practical reason for ensuring this is the fact that the existence of zero or negative TOTHHY would complicate data manipulation and analysis (e.g. log transformations and the calculation of gini coefficients). For self-employment income (SELFEMY), it was possible for a particular household to record a negative value. Negative values were transformed into zeros during the data cleaning process, and remained at zero for the subsequent analysis. A justification for this is that there is a tendency to overstate self-employment expenses (often households include capital

expenditures in self-employment expenses, when only current expenses should be reported). Hence a reported negative self-employment income is probably not an accurate indication of the household's financial situation. There were also a few instances where TOTHHY was zero or negative because of high direct taxes (PITAXY). In these cases PITAXY was reduced so as to make TOTHHY slightly positive. However, in order to preserve rankings, PITAXY was reduced so that the new TOTHHY was less than the smallest reported positive TOTHHY. There were also some cases where TOTHHY was zero, yet both SELFEMY and PITAXY were also zero (as well as all other income sub-aggregates). In such cases, TOTHHY was made positive by increasing other income (OTHERY), but again, the increase was such that the new TOTHHY was less than the smallest reported positive TOTHHY.

4. Outlier adjustment was conducted on unweighted-data[3] for two reasons. First, given that zero valued observations are excluded from the calculation of means, and it is to be expected that these will not be distributed randomly, the weights are unlikely to work as intended (i.e. the weights were designed to be used on the entire sample and not a subset of positive observations). Second, the weights *per se* do not add any more information as to whether a particular observation is an outlier or not and therefore they need not be applied at this stage.

5. The sample was divided into three localities (capital city, other urban and rural) and outlier identification and replacement was done within these localities (also see point 6 below). This division of the sample was necessary as otherwise many households residing in the capital (where income and expenditure are usually higher) would be falsely identified as outliers.

6. The outlier replacement routine was only run on a variable in a particular locality if there were more than 100 positive observations in that locality. It was felt that a minimum of 100 observations was necessary for the distribution to reflect the population distribution and thus be suitable for the calculation of means and standard deviations. If a particular locality had less than 100 positive observations then, if possible, localities were merged and the outlier procedure was run on the merged data. Localities were merged in the following way.

If either the capital or other urban localities or both had fewer than 100 positive observations (but the rural locality had more than 100 positive observations) these were merged if doing so would result in the pooled observations being greater than 100.

If either the other urban or rural localities or both had fewer than 100 positive observations (but the capital had more than 100 positive observations) these were merged if doing so would result in the pooled observations being greater than 100.

If the capital and rural localities both had fewer than 100 positive observations, then observations from all localities were pooled.

[3] Where 'weighting' here refers to the use of statistical weights to ensure the sample is representative of the population.

1.2. Missing Values

A missing value is defined here as where, for example, a household stated that it purchased a particular item (or received a particular type of income), but either did not know the value or refused to state this information. Some surveys provide special codes to distinguish missing values from 'legitimate skips' (also known as non-applicable or legitimate missing values). A legitimate skip is where the household did not purchase the item and therefore either a '.' was recorded for the value or else a record describing the purchase simply does not exist. Other surveys may not provide special codes for missing values; in such cases either they may not be able to be distinguished from legitimate skips or else it may be possible to identify missing values by using the 'leader' questions (i.e. 'Did you purchase item X over the last 30 days?'). Legitimate skips were set to zero.

If it was possible to distinguish missing values from legitimate skips, then missing values were adjusted for by replacement with per capita means calculated over localities. Thus all of the information available in the survey is used to give a measure of expenditure or income as close as possible to the household's 'true' expenditure or income. By not adjusting for missing values we would be implicitly setting the household's expenditure for that item to zero, even though we know that it was in fact positive.

In only two of the surveys (Kyrgyz Republic and Russia) was there enough information to distinguish missing values from legitimate skips,[4] and hence missing value replacement was conducted for only these two datasets. For all the other countries there was either not sufficient information to distinguish missing values from legitimate skips or else the data was provided already cleaned of all missing values (XX - specify countries?). For these countries, in the absence of further information, there was no option but to believe that these are all legitimate skips and hence, as mentioned above, the value was set to zero.

The only exception to this rule pertained to the consumption of food. There were some households that reported zero consumption of food (FOODX=0 and zero self-consumption of food).[5] Depending on the reference period[6],it is generally safe to assume that this is erroneous. Such instances of zero food expenditures were therefore regarded as outliers and were thus replaced using the outlier procedure described above.

The following summarizes the steps involved in the replacement of missing values:

1. Missing value replacement was conducted after the outlier adjustment.

---

[4] The KMPS, for example, records missing values with either a 997 (household did not know) or 998 (household refused to state).

[5] This note will identify the datasets where this was the case (maybe only Kyrgyz Republic).

[6] As the reference period for the Armenia survey was 1 day, it is quite possible that zero food consumption is the true value.

2. As with the outlier adjustment, missing value replacement was only conducted over sub-aggregates or, if they exist, the components of these sub-aggregates. Missing value replacement was done by locality.

3. All variables were measured at the household level and in monthly per capita terms and the data was unweighted. By replacing missing values using per capita data there is an implicit assumption that larger households have both larger expenditures and incomes. While such an assumption may be defensible for expenditure variables, it is less so for income variables. For the income variables which can be described as 'person-bound' (for example wages), a more appropriate method may have been to replace missing values on a case-by-case basis, for example by using means calculated over the same gender, education level, locality etc. However, such a procedure was not feasible given the size of the database.[7] An implication of performing missing value replacement only at the household level is that a situation may arise where for a particular household only one member's wage income, for example, is missing while there other members have valid observations. It was left to the individual researcher to decide what to do in such situations.[8]

2. Rent Imputation

The imputation of rental expenditures is an important step in the estimation of a household's standard of living. Rent imputation is especially important when one is wanting to make accurate welfare comparisons between households that own their housing ('owner-occupiers') and those who rent. For example, an income comparison of two households having the same income but with one household renting and the other being an owner-occupier would, in absence of imputation, conclude that their position is the same; in reality the owner household is better-off because it enjoys housing services for free. If the two households had moreover the same expenditures on all goods and services except that the renter household had to pay rent, an expenditure comparison would conclude that the renter household is better off while in reality their welfare is the same. This simple example shows the essential outlines of how rent should be imputed: for those who receive housing services without paying, an imputed value of these services must be added to both income and expenditures. For those who do pay rent, rent is treated as any other expenditure, and nothing is imputed to income. This section outlines the rent imputation techniques used in the construction of the HEIDE database.

The rent imputation procedure involved the estimation of 'hedonic' regression models which aimed to quantify the impact of different housing characteristics on the actual rent paid. The estimated coefficients from these regressions were then used to impute rent

[7] These points are equally valid for the outlier adjustment procedure.

[8] The data cleaning rules only imposed uniformity on adjustment of sub-aggregate and component variables, while individual household members' wages can be seen as sub-components.

expenditures for those households that did not report paying any rent (either because they were owner-occupiers or else because their housing was rent-free).[9] The hedonic rent regressions were estimated for each country (and locality) separately, subject to the following guidelines.

1. The key decision is what households to include in the hedonic regression. Once this is determined the rest follows directly: those who are included in the regression have their reported payments treated as expenditures and nothing is imputed to their income; the rest have both their expenditures and income increased by the amount of imputed rent. For a household to be included in the regression, it must have reported paying some rent. This is generally (but not always; see below) the case with renter, private or public, households (TENANCA=2,3). In addition, in most countries owner-occupier households also report positive values for RENTX . Such payments could be mortgage payments, co-op dues or the like. Based on (1) what owner-occupied dwelling really implies in each country, i..e. whether the reported payments were likely to correspond to rental services, and (2) whether the mean payment for those owners reporting RENTX>0 was sufficiently similar to the mean rent reported by households with TENANCA=2,3, it was decided to include such households in the regression. (If the two conditions were not met, such households were excluded from the regression and their reported RENTX was set to 0; see Table 1). The regression thus included households reporting positive rent: all such rentor households and, depending on the country, a subset of owner households. All other households, including rentor households who do **not** report paying rent, were left out of the regression.

Table 1. Rent imputation procedure

	'renter' households		*'non-renter' (owner and other) households*		
Reported rent	Positive	Zero	Positive		Zero
Included in regresssion	YES	NO	YES	NO	NO
RENTX	as reported in survey	imputed	as reported in survey	reported rent set=0 and imputed	imputed
IMPRENTY	=0	=0	=0	=RENTX	=RENTX

Note: Other households are those living (for free or not) with relatives or friends, or in their dwellings.

[9] Some of the surveys included 'self-imputed rent' where either the household itself or the enumerator was asked to estimate what the house could be rented to a third party for. As such information may be unreliable, it was decided to not use self-imputed rent information (except for the case of Bulgaria 95 where self-imputed rent was treated as actual rent payment - XXWhy?).

2. The hedonic regressions had either outlier-adjusted rent expenses (RENTX) or the log of this variable on the left hand side. The rent regressions were ideally run separately for each locality (capital city, other urban, rural). However, a regression was only run for a particular locality if the number of observations for that locality was either 100 or else 10% percent of the entire sample. If a particular locality had less than 100 households then localities were merged in the same manner as described above in point 6 of the outlier adjustment procedure. If it was the case that a valid regression could only be run over households from all localities, then a locality dummy variable was included on the right hand side.

3. The right hand side variables in the regressions were housing characteristics. One subset of housing characteristics variables, household amenities, was modeled in two ways -- as individual dummy variables (the 'dummy variable approach'), and by using a single variable indexing the number of types of amenities present in the house (the 'index approach'). The argument against the index approach is that it implicitly assumes that each amenity is worth the same. The argument against the dummy variable approach is that some amenities can be seen to come in 'packages'; for example, having a bathroom may be of value only if the household has access to running water. Thus the regression coefficient for WC or bathroom may be conditional upon having running water. Also included on the right hand side of the regression was the size of the house, or if this was not available then the number of rooms. The regressions for some countries also included a dummy variable reflecting whether the house was rented from the private or public rental market.

4. A total of four specifications were estimated for each country: linear and semi-log (log of RENTX on the left hand side) versions of both the dummy variable and index models. Of the four, the 'best' (in terms of adjusted R^2 and significance of coefficients) regression was selected and its coefficients (including those coefficients found to be statistically insignificant from zero) were then used to calculate imputed rent. In the case of the Kyrgyz Republic the hedonic regressions were not adequately specified (adjusted R^2 was less than 0.1 and the F-statistic was insignificant) for their coefficients to be validly used in the rent imputation. In this case imputed rent was the mean rent (by locality) of the households selected for the hedonic regression.

5. As explained before, imputed rent expenditure was assigned to all households who were not in the hedonic regression. A new RENTX variable was then constructed: this is equal to actual rent paid by the households in the regression group and imputed rent for all other households. The outlier procedure was applied for a second time to the new RENTX variable.

6. The final stage of the rent imputation procedure was to calculate imputed rent income (IMPRENTY). Positive IMPRENTY was only assigned to households who had their rent imputed (i.e. were left out of the regression), and for them IMPRENTY was set equal to RENTX. For all other households, IMPRENTY=0.