# Income Imputations for CAPS*

Jonathan Argent†
SALDRU

December 15, 2010

## 1 Introduction

This report documents the construction and, where necessary, imputation of monthly income variables for the Cape Area Panel Study (CAPS). The construction of income variables concerns waves 1, 3, 4, and 5 of CAPS, as wave 2 did not contain questions of this sort. The Stata do-files for this process are available online, and this should be the first source of interest for more detail on the imputations.

Ordinary least squares regression imputation is used in all cases where imputation is applied. The available data varies by waves, and by consequence both the raw income data available and the variables used in the imputation of missing income data varies across waves. Note that this form of imputation strengthens patterns already in the data, as well as reducing variability. Its use as a regressor on the right hand side of a model is thus inappropriate. For this purpose, multiple imputation or an alternative technique should be applied. For a treatment of the issues relating to the measurement of income in household surveys see Deaton (1997).

The wave 1 data in particular, and certain subsamples[1] of the other waves, provide very good quality cross-sectional income data. However, given inconsistencies in measurement across waves, as well as substantial differences in non-response, it would be ill-advised to use these income variables in longitudinal applications without further work.

*I acknowledge preliminary work from Jesse Naidoo towards the completion of this document and the imputation process in general. I also acknowledge the work by Meredith Sparks on wave 1 and 3, and Brendan Maughan-Brown on wave 5.

†Comments and questions are welcome. Contact me by email at jtargent@gmail.com.

[1]For example, the African strata (CAPS was stratified on major population group of PSU) has much lower levels of non-response across all waves. Since each strata is actually a separately drawn sample, their use in isolation does not create a selection problem. However, this does mean that the inferences drawn from such analysis would apply only within this strata of the population.

# 2 Household level measures

## 2.1 Household income

The CAPS household level questionnaire included a 'one-shot' total household income question in waves 1, 3, and 5. In each wave the respondent was asked to indicate how much money the household receives in total in a typical month. If the respondent refused or professed that they did not know, a follow up question asked the respondent to indicate an income bracket from a showcard. For the purpose of constructing a household income variable, I take the point estimate if it is available. If it is not, I use the midpoint of the bracket answer if that is available.

In wave 5, the household level questions were given to each young adult, as opposed to a 'knowledgable member of the household', which means that for some households we have more than one observation per household, complicating our hierarchy of information somewhat. However wave 5 also had very high non-response and only 4% of households have more than one point estimate. A further 4% have more than one bracket response. In these (few) cases where this occurs, I take the upper estimate[2] to be the correct one and follow the same hierarchy of data sources as with the other waves.

Table 1
*One-shot income data across waves*

| Data source | Wave 1 | Wave 3 | Wave 5 | Total |
|---|---|---|---|---|
| Point | 74% | 56% | 34% | 60% |
| Bracket | 20% | 28% | 25% | 23% |
| Missing | 6% | 16% | 42% | 16% |
| Total | 5,255 | 2,549 | 2,313 | 10,117 |

All comments.

Table 1 shows the distribution of data sources for the construction of the one-shot household income variable across waves. The reason why wave 1 has a much larger total number of households is because waves 3 and 5 only revisited households where young adults (aged 14-22 in wave 1) resided. The number of young adult households for wave 1 is 3,479. However, the distribution across data sources is almost identical[3] between those with young adults and those without in wave 1.

Wave 1 experienced only 6% non-response, with 20% of estimates coming from brackets. It is clear from table 1, that response rates were not as healthy in the other three waves. Beyond the loss in precision from more bracket responses, the rise in missing data is substantial. The particularly large deterioration in wave 5

[2]I choose the upper estimates because recent data from the National Income Dynamics Study (NIDS) showed that on average, one-shot income questions tended to under-report income relative to measures that aggregated across individuals and sources of income.

[3]The difference in each category is less than 1% of the total. Point estimates (73.99%), Brackets (20.55%) and Missing (5.46%).

may well be due to the question being posed to the young adults themselves as opposed to a knowledgable household member. Certainly we would expect them to know less about the financial affairs of the household.

For the missing data, I use a regression[4] imputation strategy. Since the set of explanatory variable available differs across waves, the regression model used is not identical across waves. The actual specifications used in the regressions can be found in the Stata do-files, available online. The variables used in the models include: month of interview; household size; measures of household age distribution; measures of demographics on household head; an asset ownership index; indicators for access to sanitation and quality of housing; and subjective measures of food adequacy and financial status.

The high non-response rate wave 3, and wave 5 in particular, mean that any imputed income variable should be treated with caution. With wave 1 being the only measure on household income where the data quality is really defensible, changes in income across waves as derived from this data must be considered highly suspect. However, certainly for wave 1, this is not a bad cross-sectional measure at all.

## 2.2 Household expenditure

Waves 3 and 4 of CAPS included a battery of questions at the household level that asked for total household expenditure on certain items. Since wave 4 did not include a one-shot income question, the expenditure data provides an alternative. However, substantial differences in the size of expenditures measured in wave 3 relative to incomes in the same wave suggests that some caution should be employed.

There are three questions asked per item in the expenditure module. The first asks if the household spends on that item. The second asks how much is spent on the item (assuming an affirmative to the first). The last question identifies if the answer given to the second question was in monthly or yearly form. I assume that those that refuse or don't know if they spend money on a particular item do not spend money on that item. Where a household identifies itself as spending on an item, but does not respond to the question on the amount (whether a refusal or a don't know) I impute the missing value. All yearly values[5] are divided by 12 to give monthly values.

Non-response is very low in the answers to these questions across both waves. There are some apparent differences in data quality between the two waves, with wave 4 having better response rates to both questions. Wave 3 has on average of roughly double the proportion of 'don't know' answers to the question of whether the household receives income from a particular source (the first question). Also, for those that answer yes to the first question, roughly double the proportion in wave 3 an-

[4]The dependent variable in these regressions is the log of one-shot household income.

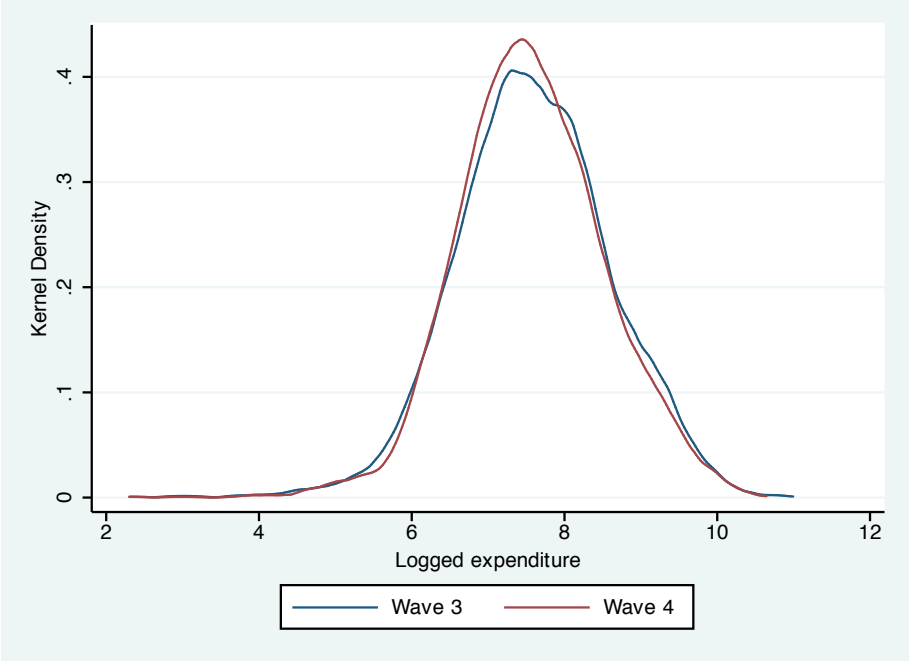
[5]I rely on positive identification here. Only where we have a positive (non-missing) identification of a yearly value, I divide by 12. Very few answers are given in yearly form which means this assumption does not make a great difference.

swer 'don't know' when probed for the amount (second question). Wave 3 also had comparatively more refusals to the first question, although total refusals are still less than 1% on average. There was little difference in refusals in response to the second question. While it certainly appears that the data quality in wave 4 is better, even that of wave 3 still looks very good in terms of response rates.

There are several other encouraging signs that the expenditure data are comparable across waves. For example, the distribution of the choice to answer in monthly/yearly format looks very similar across the two waves. Also, the cumulative distribution of the number of values imputed per household (out of 12) is almost identical across the two waves.

Figure 1 shows the kernel density estimates of the distribution of logged sum of the expenditures measured in waves 3 and 4. The figure does not include imputed values, but the same plots including imputed values show no material difference (which is unsurprising given how little missing data there is). Interestingly, wave 4 lies almost perfectly over wave 3 rather than shifting out to the right in the manner of the total household income figures calculated by the aggregation method (as we will see later on).

Figure 1
*Comparing logged total household expenditures*




This figure does not include imputed values.

There is one major concern with the measurement of expenditures. Mean total expenditure is only 54% of mean total household income in wave 3. The ratio for the medians is identical. This is really not particularly surprising given the amount of work it generally takes to measure expenditures accurately[6].

[6]See for example the Income and Expenditure Survey in 2000.

# 3 Individual component measures

An alternative way to measure household income, is the aggregation of individually measured incomes. One of the advantages of the use of this data in CAPS is that it was measured in all four relevant waves. This section goes through the construction of the individual components (wages, government grants and remittances) and discusses the use of these in an aggregate household measure.

## 3.1 Wages

When constructing wage data for individuals in CAPS, there are two sources of information. Firstly we have the household roster data. All waves of CAPS have individual wage data in the household roster. However, only the first wave of CAPS asked respondents that refused or professed not to know their wage income point estimate to allocate themselves to an income bracket. The second source of information is for the young adults only, which comes from the labour market module of the young adult questionnaire. The methodology for this module varies across waves in terms of the number of jobs asked about, but all waves include the bracket question as a follow up to a failure to obtain a point estimate.

Table 2
*Wage data across waves*

Pane 1: Young adult wage data

| Data source | Wave 1 | Wave 3 | Wave 4 | Wave 5 | Total |
|---|---|---|---|---|---|
| Youth | 59.7% | 79.0% | 82.2% | 71.8% | 73.7% |
| Roster | 21.8% | 7.7 % | 6.0% | 6.9% | 10.0% |
| Youth-brackets | 3.2% | 1.3% | 0.6% | 2.2% | 1.8% |
| Roster-brackets | 5.0% | 0.0% | 0.0% | 0.0% | 1.1% |
| Missing | 10.3% | 12.0% | 11.2% | 19.1% | 13.4% |
| Numbers | 1,359 | 1,492 | 1,738 | 1,806 | 6,395 |

Pane 2: Adult (not young adults) wage data

| Data source | Wave 1 | Wave 3 | Wave 4 | Wave 5 | Total |
|---|---|---|---|---|---|
| Roster | 73.0 | 61.6 | 71.0 | 33.0 | 64.3 |
| Roster-brackets | 17.3 | 0.0 | 0.0 | 0.0 | 7.4 |
| Missing | 9.7 | 38.4 | 29.0 | 67.0 | 28.3 |
| Numbers | 7,241 | 3,196 | 3,815 | 2,568 | 16,820 |

In many cases there are multiple available data sources.
The hierachy is given by order above.

The existence of two sets of estimates for youth wages presents a problem. What sort of hierarchy do we apply to this data? This would not be such a large problem if there were not large disparities between the two sources of data. As an example, just over 20% of those those who claim to be currently employed in the young

adult questionnaire, and give non-zero wage responses are missing in the household roster. While reliability is certainly a debatable issue, it would seem tough to argue that young adults do not have better information about their own employment that other household members. For the construction of wage variables I adopt the following hierarchy. Young adult answers are preferred to roster answers, and point estimates are preferred to bracket data. Thus the final order applied is: young adult point estimate; roster point estimate; young adult bracket; roster bracket. Where a bracket is used, I assign the individual to the midpoint of the bracket. Where the top bracket (which is open ended) is chosen, I assign the respondent to twice the lower bound of the top bracket. However, this occurs only a handful of times in total across all waves.

Table 2 shows the distribution of data sources for the construction of individual wages. The first pane of table 2 shows the source of wage data for the young adults in the sample. The second pane, shows the breakdown for all other adults. The striking rise in the percentage of answers that come from young adult questionnaire point estimates is perhaps not too surprising given that these young adults are in the midst of a transition into the labour market. The sudden drop in the quality of adult wage data in wave 5 is likely caused by the roster having been completed by young adults, as opposed to knowledgeable household members. This presents something of a problem for imputation, with only a 34% response rate among (not-young) adults in the fifth wave.

Table 3
*Aggregated (household level) wage data across waves*

| Data source | Wave 1 | Wave 3 | Wave 4 | Wave 5 | Total |
|---|---|---|---|---|---|
| Survey | 89.2% | 66.3% | 74.8% | 46.3% | 73.9% |
| Some imputed | 6.2 % | 17.9% | 12.9% | 28.6% | 13.9% |
| All imputed | 4.7 % | 15.8% | 12.3% | 25.1% | 12.2% |
| Total | 5255 | 2549 | 3312 | 2313 | 13429 |

All comments.

On the whole, response rates from young adults to the wage questions look good enough to create panel wage data. In contrast, waves 3, 4 and 5 have such high non-response to the adult wage data (from the roster), that any imputation of this data must be viewed with hostility. It may be that certain sub-sections of the sample (across strata for example) may have sufficient response rates to be a convincing base for imputation of remaining missing values.

Notwithstanding the problems with the data, regression imputation has been applied to complete the individual wage data where it is missing. For the actual specifications used, see the do-files. The explanatory variables used include: interview month, age, sex, race, education and full/part-time employment status[7].

[7]Household level explanatory variables were not employed as this may lead to systematic bias. Consider where person A and B live together and A earns a large salary and B earns very little. Where B has missing wage data, using household assets to impute this would likely overstate wages, as they will incorporate information relating to A.

Table 3 presents the data sources for household level wage data. This presents a slightly more positive picture than the individual data in complete wage records, suggesting that individuals for which we have poor wage data are concentrated within certain households[8].

## 3.2 Government grants

There is no uncertainly as to the value of government social grants, since they are fixed each year on the 1st April (and sometimes increased on the 1st October). These values are public knowledge. Where any individual acknowledges the receipt of a particular social grant, I allocate the appropriate value for that grant at that point in time. Of course this means that people interviewed on either side of the day of increase may be allocated different values. But this is an accurate picture of their reality at the time of interview, and consistent with the rest of the income data being in nominal form. The incidence of government grants in the CAPS data is presented in Table 4.

Table 4
*Incidence of government grants across waves*

Pane 1: All grants across waves

| wave | Obs | Y.A obs | Pension | Disability | Child | Foster | Care.Dep. | Other |
|---|---|---|---|---|---|---|---|---|
| 1 | 22,629 | 5,282 | 724 | 438 | 587 | 0 | 0 | 117 |
| 3 | 12,994 | 3,560 | 464 | 284 | 772 | 72 | 0 | 0 |
| 4 | 15,636 | 3,647 | 775 | 419 | 950 | 67 | 11 | 0 |
| 5 | 12,350 | 3,261 | 490 | 287 | 934 | 24 | 11 | 0 |

Pane 2: Alternative measurements

| wave | Child | Foster | Care.Dep. |
|---|---|---|---|
| 4 | 1293 | 92 | 28 |
| 5 | 1321 | 29 | 10 |

Pane 2 shows incidence of grants calculated from summing *children for whom grants are received.*

There are some minor methodological complications. In the first wave of CAPS, there was a single question in the household roster of the receipt of household grants. This question did not allow recipients to indicate that they receive multiple grants, and did not allow precise identification of all grants since it explicitly contained categories for state old age pensions, disability grant and child support grant only. All other grants were allocated to "other grants". There are only two other government grants with significant numbers, which are the care dependency grant and the foster care grant. Given the very low incidence of the care dependency grant (as seen in the other waves), I allocate all those in "other grants" the value of the foster care grant in wave 1. This is a conservative choice, as the care dependency grant is significantly larger.

[8]This is not necessarily a good thing from the analyst's perspective

In waves 3, 4 and 5 a multiple question approach was adopted for the measurement of social grants. In wave 3, there were individual categories for all except the care dependency grant. As noted before, the extremely low incidence of this grant suggests that its exclusion is unlikely to make a great difference. Waves 4 and 5 adopted improved methodology for measurement of grants. In these two waves, the roster required the household respondent to allocate both the recipient and *those for whom the grant(s) are received*. This means that in waves 4 and 5 we are able to detect the number of these grants received by a household. The second pane of Table 4 presents the incidence of social grants as calculated by the improved methodology. Comparison with the top pane suggests that the prior two wave of CAPS understate the incidence of these three social grants due to the imprecision on the number received per recipient.

Note that the jump in wave 4 of government pension and disability grants is due to a change in sample. In wave 4, CAPS revisited *all* of the wave 1 households, including those that did not include young adults. That we find more of these grants relative to total number of achieved sample in comparison with wave 1, may be driven by a rise in incidence of these grants over time.

## 3.3 Remittances

Waves 3 and 4 of CAPS asked questions at the household level on remittances (inter-household transfers). This includes transfers of both money and goods[9]. Wave 5 included questions on remittances in the young adult questionnaire. However, these questions only asked about remittances received, not remittances paid. Also, there were not separate questions as to whether they receive remittances and if so, how much. So it is impossible to separate those who receive and refuse the amount from those who do not receive at all. The fact that in this data we find that only 117 households receive remittances - about 12% of the numbers in the previous wave - suggests that the non-response here is very high. For these reasons, I exclude the wave 5 remittances data from income aggregates and analysis.

Table 5
*Household level net remittance data across waves*

| Data source | Wave 3 | Wave 4 | Total |
|---|---|---|---|
| Survey | 94.9% | 95.4% | 95.2% |
| Part missing | 0.3% | 0.1% | 0.2% |
| All missing | 4.8% | 4.5% | 4.6% |
| Number | 706 | 961 | 1,667 |

Net remittances are remittances received, less remittances paid.

Table 5 shows the distribution of data source for waves 3 and 4. The data displayed is from net remittances calculated as the difference between remittances received

[9]There is of course a theoretical issue regarding the addition of goods as income. Treatment of this issue is beyond the scope of this report.

and remittances paid. Where households claim multiple remittances, but only give values from some of these, this entry is considered a complete response. Only where a household identifies itself as having remittances (paid or received) and there are no values given are they considered to have missing data. A small number of households have missing data for the number of times in the past year that they received such a transfer. In these cases, I make the assumption across both waves the amount is received six months out of twelve. This is a conservative assumption as the median answer to this question was above six in all cases.

Net remittances are not particularly large. In wave 3 (4) about 28% (30%) of households claim remittances. Mean (net) remittances in wave 3 are only R28 (R17) a month. These low means do however hide the large variance, with one household paying net remittances of over R13,000 per month and another receiving net remittances of R43,000. Certainly this will have a large effect for some households in terms of their position in the income distribution calculated by the sum of individual income components.

There is no imputation done on the missing remittances data. It is difficult to imagine how from the characteristics of households we would reasonably be able to impute the value of net remittances, in contrast to wages for example. However, given the very high response rates to these questions, it seems hard to imagine that this missing data is particularly problematic.

## 3.4 Aggregating to form a household income measure

For each wave I generate a household income measure from the aggregation of available parts. The aim here is to generate the best measure that we can for each individual wave. This should serve as a further warning against using this data in longitudinal applications - it is not comparable over time.

All waves include wage data for all individuals, imputed where missing. All waves include government grant data, although this data does differ in methodology as previously noted. Most important of these differences in government grants is that waves 4 and 5 have much more accurate measurement in child, foster and care dependency grants as they measure then number received, as opposed to just receipt at all. Waves 3 and 4 include remittances data. While the wage and goverment grant data is in monthly form, the net remittance data is in expected monthly value form[10].

Comparing the measurement of total monthly from the one-shot household level question, and an aggregation approach, there is very little difference between the two. It is not *a priori* clear which of the two is to be preferred. Work on the National Income Dynamics Study (NIDS) suggests that one-shot measures tend to understate household income compared to aggregation methods (Argent, 2009). But this does not appear to be the case in CAPS. Because the individual wage and grant data

[10]I calculate the expected value of a remittance (paid or received) as the typical monthly payment multiplied by the fraction of months in the year in which the payment (or receipt) is made

is collected from the person that answers the household roster (with exception of young adults wage data), there is less informational advantage to the aggregated method than there would be in a context where individual data was collected from individual questionnaires.

Figure 2
*Comparing logged income estimates across waves*

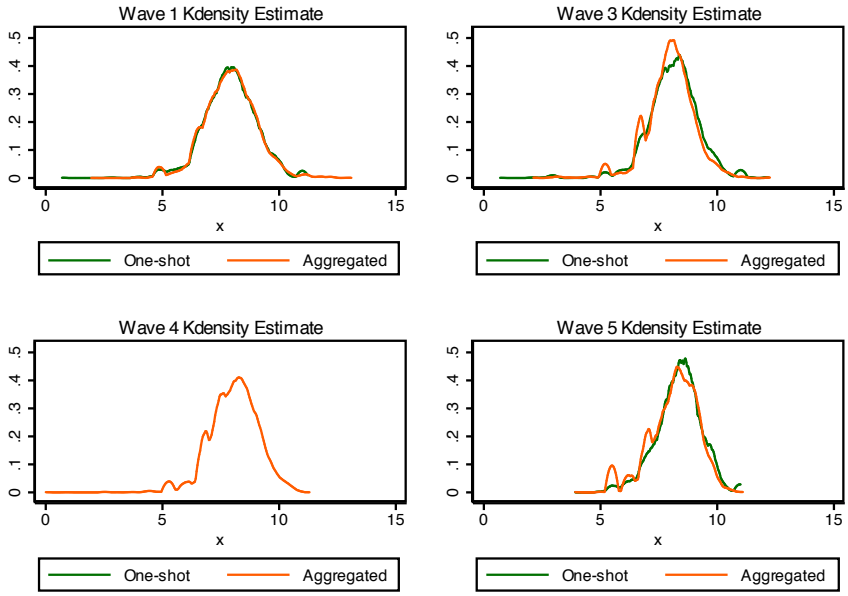



Figure 2 shows the kernel density estimates of the log of each of the two measures (before imputation), across waves. For wave 4, where there is no one-shot income data, only the aggregating method is shown. Looking at Figure 2, it is clear that the differences between methods are not large on average[11]. The fact that the difference is small provides some encouragement for the use of the aggregated measure in wave 4, in which there is no one-shot income question. Of course an alternative is to use the expenditure data, but the differences between that and the two income measures is much more substantial. It is beyond the scope of this report to fully explore the patterns of difference between these three measures.

# 4 Concluding remarks

In concluding I want to reiterate the caution about the use of the income data from CAPS in a longitudinal manner. Certainly there may be applications, or particular subsamples for which this data can be used in a longitudinal manner. However, the burden of proof lies with the analyst to show that their use of the data in this manner is defensible. Within a particular wave (cross-sectional analysis), the use of the income data is more defensible. However, the missing data and applied

[11]The amount of shuffling of position within the distribution is not clear.

imputation techniques documented here should still be taken seriously even in this case.

# References

J. Argent (2009). 'Household Income : Report on NIDS Wave 1'.

A. Deaton (1997). *The Analysis of Household Surveys*. Baltimore, Maryland: The Johns Hopkins University Press.