DataFirst Technical Papers

Income in the October Household Survey 1994

by
Martin Wittenberg

Technical Paper Series
Number 7

About the Author(s) and Acknowledgments

Recommended citation

Wittenberg, M. (2008). Income in the October Household Survey 1994. A DataFirst Technical Paper Number 7. Cape Town: DataFirst, University of Cape Town

---

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Income in the October Household Survey 1994

From DataFirst

Martin Wittenberg School of Economics and SALDRU, University of Cape Town January 2008

Contents

Introduction

Income dynamics in the post-apartheid era are of particular interest to economists, since they relate to the incidence of poverty and inequality and throw light on the operation of the labour market. The 1994 October Household Survey is strategically placed at the beginning of the post-apartheid era. Nevertheless the income information in the survey needs to be treated with considerable caution since much of it is imputed in ways which are questionable.

This document sets out what we have managed to discover about the process of imputation and ways of correcting for these imputations.

The structure of the income data

There are two key places in the questionnaire where income information is solicited. In Question 3.13 the “income from main job” is asked for; while in Question 3.19 the “gross income/turnover for all own account activities” is the subject of enquiry. In both cases provision is made for a Rand amount to be supplied, a series of categories are given and then the respondent is prompted to indicate whether the reporting period is per day, week, month or year.

This “raw” information, however, is not provided in the dataset. Instead the following four variables are given: “Net income of employee (Rand) (calculated per month)”, “Net income of employee (Code) (calculated per month)”, both derived from Question 3.13 and “Gross income of employer (Rand) (calculated per month)” and “Gross income of employer (Code) (calculated per month)” both derived from Question 3.19. The calculations seem to have involved at least the following: conversion of daily, weekly and annual data to monthly; imputing information given only in brackets to point values; and correcting gross values to net values. The effect of the imputations has led to a particular structure of the data as supplied, in particular there are pronounced “spikes” within the data. We will discuss the “net income of employee” variable first, before turning to the income from own account activities.

Spikes in the “net income of employee” variable

The consequence of the imputations can be seen in the “spike plot” shown in Figure 1, which gives the information for the “net income of employees” variable. The placement of these pronounced spikes is odd, since they are not located at “nice” round numbers. Indeed the placement is bizarre to say the least: one would not expect around 30% of all individuals who reported being paid daily to have received precisely the same net monthly income of R1364. This is not an artefact of small numbers either: this spike corresponds to 172 individuals.

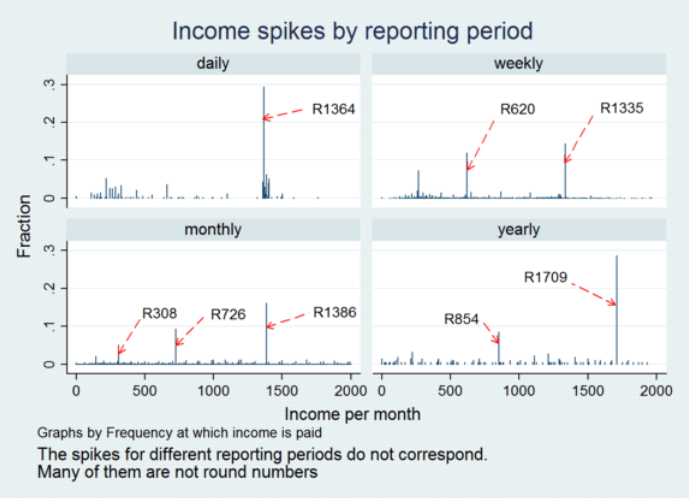

Figure 1 Spikes in the "Net income of employee calculated per month" variable

One potential culprit for these strange numbers might be the conversion from gross income to net income. Individuals that reported round numbers might have had these spikes shifted by corrections in the conversion process. Figure 2 shows, however, that this is not the case. Individuals that claimed to have reported **net income**, i.e. where no conversion was necessary, were the ones were the data is spiky. The conversion from gross to net income seems to have smoothed over the spikes instead of creating them.

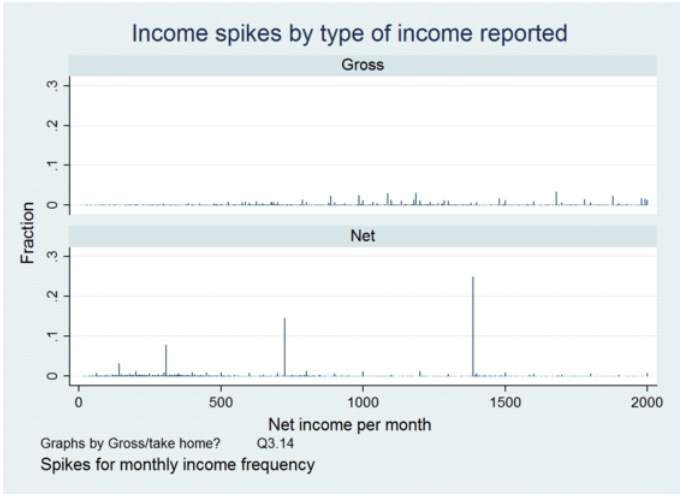

Figure 2 The income spikes are evident only for people claiming to report net income

Looking at the spikes of the individuals that reported monthly income data (bottom panel of Figure 2) is quite suggestive – particularly when the income brackets are borne in mind. It seems as if there is a "spike" corresponding to each income bracket. The most logical explanation is that the spike corresponds to individuals that reported an income category only and that were then placed at these particular values. The "spiky" nature of the data therefore potentially allows us to separate out the point information from the categorical data: with the caveat that the process of correcting the data among those reporting "gross income" seems to have removed any spikes that might have been there initially.

We can see the impact of the conversion in Figure 3, where we have added the "income deductions" back to the net income variable to create an estimate of what the individuals would have reported originally. The spikes magically reappear. In fact, as Figure 4 shows (when compared to Figure 1) the spikes are precisely at the same points.

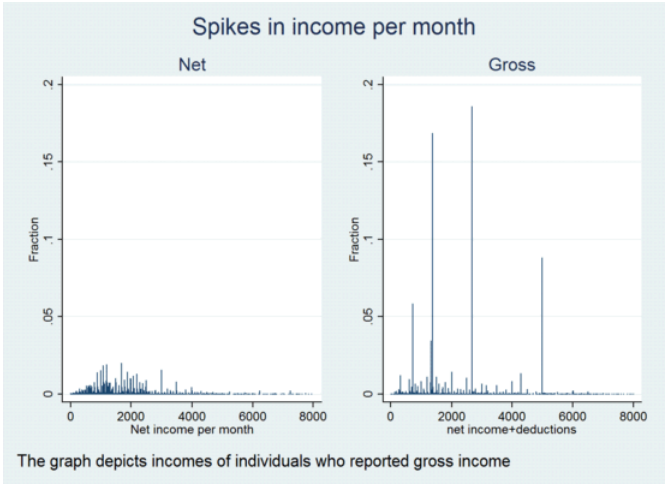

Figure 3 The spikes re-emerge when the deductions are added back to the "net income" figures.

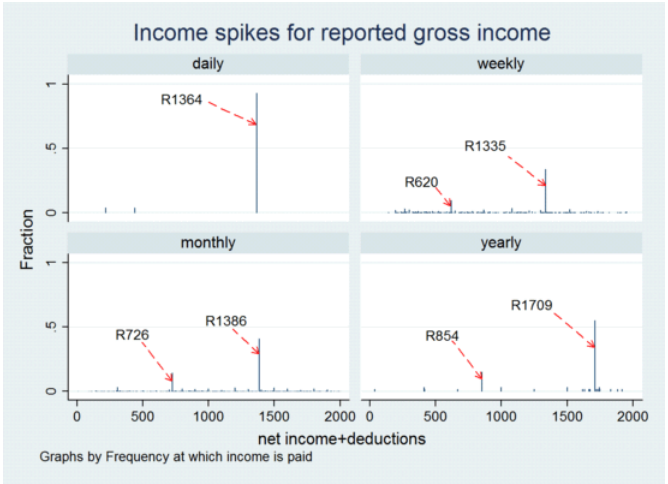

Figure 4 The spikes for individuals reporting "gross income" are at the same points as for those reporting net income

Conversion rates between reporting periods

Before we analyse the location of the spikes a bit further, it is important to consider the way in which Statistics South Africa seems to have converted between reporting periods. In the case of daily data the pattern emerges when one looks at a tabulation of the actual values. Ignoring spikes at strange values, there is a periodic pattern of spikes at multiples of 110. This kind of pattern would be produced if the original "round" Rand incomes (i.e. spikes at multiples of five and ten) had been multiplied by 22. This multiplication factor in turn makes sense, since there are roughly twenty-two working days in every month.

. tab salary_r if income_i==1

Income per Month	Freq.	Percent	Cum.
0	15	2.38	2.38
33	1	0.16	2.54
70	1	0.16	2.7
110	9	1.43	4.13
132	5	0.79	4.92
147	1	0.16	5.08
150	1	0.16	5.24
154	7	1.11	6.35
170	1	0.16	6.51
172	2	0.32	6.83
176	8	1.27	8.1
186	1	0.16	8.25
210	1	0.16	8.41
220	31	4.92	13.33
223	1	0.16	13.49
242	16	2.54	16.03
264	15	2.38	18.41
280	1	0.16	18.57
286	17	2.7	21.27
300	1	0.16	21.43
308	7	1.11	22.54
330	20	3.17	25.71
336	1	0.16	25.87
360	1	0.16	26.03
374	1	0.16	26.19
396	2	0.32	26.51
400	1	0.16	26.67
418	1	0.16	26.83
440	13	2.06	28.89
453	1	0.16	29.05
484	3	0.48	29.52
528	2	0.32	29.84
550	5	0.79	30.63
640	1	0.16	30.79
660	21	3.33	34.13
682	2	0.32	34.44

In the case of the weekly data the pattern is harder to pick up, but there is a pattern nonetheless with spikes at 130, 260, 390, 520 and so on. This pattern would be produced if round incomes had been multiplied by 4⅓. This multiplier makes sense if one remembers that there are 52 weeks in a year and also 12 months. But 52/12 is equal to 13/3.

The location of the spikes

The strongest confirmation that these numbers indeed give the conversion rates comes by regraphing the spikes in terms of the original reporting periods, i.e. dividing the daily income figure by 22 and the weekly one by 4⅓. In the resulting spike plot (see Figure 5) the spikes are now at precisely the same position.

Figure 5 Converting the monthly data back to the "original" form produces spikes at the same location

Taking the spikes in the monthly data one can create all the other spikes, provided that one uses the appropriate conversion rates (dividing by 12 in the case of annual data) and rounding the resulting number to the nearest integer. In Table 1 we have shown the relationship between these spikes. Literally **all** the "odd" spikes in the data are in this table.

Table 1 Spikes in the "net income of employee, calculated per month" variable

Bracket:	Monthly	Daily	Weekly	Annual
R1-R99	62	1364	269	5
R100-R199	143	3146	620	12
R200-R499	308	6776	1335	26
R500-R999	726	15972	3146	61
R1000-R1999	1386	30492	6006	116
R2000-R3999	2679	58938	11609	223
R4000-R7999	4993	109846	21636	416
R8000-R15999	10244	225368	44391	854
R16000-R32999	20509	451198	88872	1709
R33000+	51500	1133000	223167	4292

This evidence suggests fairly strongly that individuals that merely ticked a category without providing a Rand income were placed at these values within the brackets and this amount was then converted to a monthly figure, using the appropriate conversion rate. That begs, of course, the question why these particular values should have been used for the imputation in the first place.

That question turns out not to have any easy answers. It seems natural to suppose that these numbers would be based on the means or medians of those individuals actually supplying point estimates within each category. Nevertheless trying all sorts of permutations of these (even geometric means) we could not reproduce these numbers.

The impact of the imputations

Since the spikes are well defined it is straightforward to remove them. This may remove also the odd individual who coincidentally gave precisely that figure, but this error is likely to be negligible. Taking out the spikes has a dramatic effect. Table 2 shows fully 54% of all the observations at one or other of the spikes! The same table also shows that there is a systematic negative relationship between reporting actual information and income received. Note that category 12 is not the highest income category. It is a category for individuals that did not supply categorical data, i.e. only a Rand amount.

Table 2 Proportion of data imputed by income bracket

	Net income per month (category) - calculated per month												
Variable	1	2	3	4	5	6	7	8	9	10	11	12	Total
not imputed	71	171	836	3,395	2,755	3,356	1,576	369	71	11	1	1,167	13,779
(percent)	100	61.07	67.69	69.02	44.8	36.02	31.08	24.67	29.83	29.73	14.29	100	45.95
imputed	0	109	399	1,524	3,395	5,961	3,495	1,127	167	26	6	0	16,209
(percent)	0	38.93	32.31	30.98	55.2	63.98	68.92	75.33	70.17	70.27	85.71	0	54.05
Total	71	280	1,235	4,919	6,150	9,317	5,071	1,496	238	37	7	1,167	29,988
	100	100	100	100	100	100	100	100	100	100	100	100	100

Given that there are so many missing values, one might assume that it would be beneficial to impute. Nevertheless the procedure actually adopted in this data set is unlikely to be useful. It is inconceivable that the distribution of daily, weekly, monthly or annual data will be the same within each bracket. Indeed the empirical (non-imputed) figures shown in Table 3 show significant differences. For instance the mean income among daily paid workers that fell into the lowest bracket was R30 and the median only R20. Converted to monthly figures (by the factor of 22) this would give R660 or R440 as the appropriate imputed value, rather than the R1364 actually assigned in the data set. While the daily and weekly values are over-imputed, the annual data is likely to be underimputed. Individuals with annual income in the top bracket (R33 000+) had a mean monthly income of R6231, instead of the R4292 assigned them by the imputation algorithm. The net effect of these different biases for the different categories is difficult to predict. Redoing the imputation but using the category means separately for each reporting period leads to a slight increase in mean monthly "net income" from R1484 to R1529.

Table 3 Distribution of reported income within the bracket, by reporting period

	Income reported					Income reported			
Income category	daily	weekly	monthly	yearly	Income category	daily	weekly	monthly	year
2 Mean	30.2	70.9	64.55		7 Mean		3028	2694	281
Median	20	73.62	70		Median		2829	2696	252
Imputed	0.371	0.242	0.379		Imputed		0.654	0.7	0.63
n	582	1,617	203		n		26	5,320	11
3 Mean	124.9	149	150		8 Mean		5025	5166	573
Median	120.5	150	152		Median		4818	5043	600
Imputed	0.5	0.5	0.369		Imputed		0.6	0.695	0.45
n	18	1,357	929		n		5	2,126	11
4 Mean	278.4	297.7	329.1	395	9 Mean			11040	100
Median	200	296.1	336	360	Median			10744	105
Imputed	0.571	0.529	0.312	0	Imputed			0.696	0.85
n	7	1,964	2,985	5	n			345	35
5 Mean	500	653.6	749.5	899	10 Mean			20756	239
Median	500	600	780	984	Median			20809	240
Imputed	0	0.595	0.555	0	Impute			0.686	0.8
n	1	289	3,734	2	n			35	130
6 Mean	1400	1368	1441	1214	11 Mean			52500	747
Median	1400	1389	1428	1200	Median			52500	563
Imputed	0.8	0.6	0.654	0.5	Imputed			0.857	0.75
n	5	70	6,453	8	n			7	488

Notes:

1. The statistics are weighted using the Statistics South Africa released person weights.

1. The income categories are as per reporting period, e.g. category 2 is R1 to R99 either per day, per week, per month or per year.
2. The statistics are calculated over individuals who were not located at one of the spikes, i.e. pre-imputation.
3. The statistics provide the mean and median of the point data, proportion of data set located at spikes and sample size of the cell as a whole (i.e. imputed and non-ir

An increase of 3% in mean income (among those reporting) does not seem an awful lot, although the discrepancy is sufficiently large that naive hypothesis tests (not controlling for clustering) would reject the idea that the two means are equal. Table 4 shows that there is a noticeable impact of the imputation process on average incomes among people who report daily incomes and those reporting annual income. Indeed the difference is statistically significant in the latter case.

Table 4 Impact of category specific imputations on average income

	Mean	Robust Std. Err.	[95% Conf. Interval]			Mean	Robust Std. Err.
Net income (as calculated)					**Salary imputed per reporting period**		
daily	1437.4	236.7	972.8	1901.9	daily	1216.4	235.5
weekly	1174.0	40.5	1094.4	1253.5	weekly	1163.3	40.2
monthly	1505.1	51.6	1403.8	1606.4	monthly	1539.4	53.6
yearly	3080.6	151.5	2783.2	3377.9	yearly	4156.6	209.6

Notes:

1. The left panel uses the StatsSA net income figures. The right hand panel imputes income using different points within each bracket for individuals giving daily, we
2. The statistics are weighted using the Statistics South Africa released person weights.
3. The standard errors are robust to clustering on PSU.

Imputations in the "gross income from own activity" variable

The general points from the discussion on the "net income of employee" variable carries forward to the "gross income from own activity" variable. There are still spikes which seem to be due to imputations from the bracket data, and there are still conversions from daily, weekly and annual data that need to be taken into account. The situation is, however, somewhat more messy as Figure 6 indicates. Here we have again converted the "calculated" values back to the original ones, using the same conversion rates as for the "net income of employee" variable. These conversion rates seem to be the right ones, because the spikes again line up precisely.

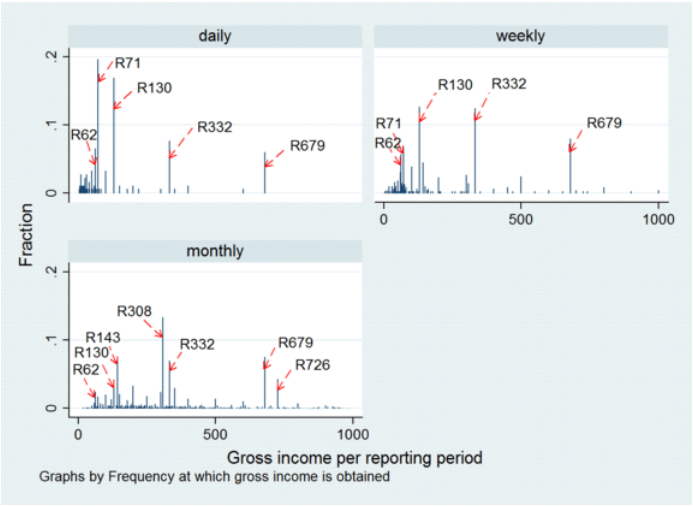

Figure 6 In the "gross income from own activity" variable there seem to be two spikes per bracket

In this case, however, there seem to be **two** imputation spikes per bracket. For instance in the lowest category there are spikes at R62 as well as at R71. The first of these is identical to the spike in the "net income of employee" variable while the latter seems specific to this variable. In the next bracket there are spikes at R130 and R143. The former is new whereas the latter is identical to a spike we have seen before.

In Table 5 we give a listing of the "spikes" we observe in the data set. As before all the strange spikes can be obtained by simple transformations from the monthly spikes. Not all these potential spikes are actually observed in the data. For the higher income categories (above R1000 per reporting period) there is generally only one spike per interval.

Table 5 Income spikes in the "gross income from own account activities" variable

	Daily		Weekly		Monthly
Reporting period bracket:	common	new	common	new	common
R1-R99	1364	1562	269	308	62
R100-R199	3146	2860	620	563	143
R200-R499	6776	7304	1335	1439	308
R500-R999	15972	14938	3146	2942	726
R1000-R1999	30492	29150	6006	5742	1386
R2000-R3999	58938	57332	11609	11293	2679
R4000-R7999	109846	112266	21636	22113	4993
R8000-R15999	225368	230186	44391	45340	10244
R16000-R31999	451198	489742	88872	96464	20509
R32000-R63999	1133000	999526	223167	196876	51500
R64000-127999		1805562		355641	
R128k+		6545000		1289167	

Notes:

1. The columns labelled "common" list spikes that are common with the "net income of employee" spikes. The columns labelled "new" indicate spikes that seem spe
2. Cells with dark shading indicate that this value is not observed in the data at all.
3. Cells with light shading indicate that this value is observed, but the "spike" is not a local maximum, or it represents a negligible fraction of the values in that brack

It is difficult to make sense of this pattern. It seems clear why there would be a "new" set of imputations. We wouldn't expect the point estimates from the "net income of employee" variable to be good approximations to the values of the typical "gross income from own account activities". It is hard to see why there are **any** spikes in common with the other set of imputations, let alone how there could be two different imputations for values in the same bracket.

Although it is difficult to understand the process that generated these imputations, it is fairly easy to identify the imputations themselves. It is therefore also easy to remove the imputed values and so separate the categorical information from the proper point estimates.

Conclusion

It is evident that there are different levels of manipulation that led to the creation of the two income variables. However it is dubious to put daily and annually paid individuals at the same point within each bracket as monthly paid workers. The existence of the spikes, however, alerts us to the existence of the problem and enables us to correct appropriately.

Files to accompany this document

OHSIncome.do [1] This creates imputation flags for the two income variables plus it generates new imputations, different by reporting period.

Category: OHS 1994

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708