

# A Guide to version 3.3 of the Post-Apartheid Labour Market Series (PALMS)

Andrew Kerr\* and Martin Wittenberg†

August 29, 2019

## 1 Introduction

This document describes the PALMS version 3.3 stacked cross sectional dataset created by DataFirst at the University of Cape Town. The dataset consists of microdata from 69 household surveys conducted by Statistics South Africa between 1994 and 2019, as well as the 1993 Project for Statistics on Living Standards and Development conducted by SALDRU at UCT. The Statistics South Africa surveys include the October Household Surveys from 1994 to 1999, the bi-annual Labour Force Surveys from 2000-2007, including the smaller LFS pilot survey from February 2000, and the Quarterly Labour Force Surveys from 2008-2018. The data is at individual level, but household level variables may be created using the household id variable `uqnr`. No attempt has been made to link individuals or households across waves, although there was a panel element to the earlier rounds of the LFS, as well as the QLFS.

The data used was collected by Statistics South Africa and SALDRU (for the 1993 PSLSD) and was obtained from DataFirst at the University of Cape Town. There are currently over 120 variables in the dataset and over 5.7 million observations, including children and the elderly. The variables included are mainly those to do with the labour market, although some household variables, such as dwelling type and access to services, as well as access to government social grants, are also included. But not all variables from all surveys are included. The surveys are regarded as one of the more reliable sources of labour market data, including labour income data in South Africa. However they generally contain little other income information, except for some incomplete attempts at capturing government grants. The PSLSD and OHSs were more comprehensive but the other forms of income data collected in these surveys have not been included in PALMS.

One of the key pieces of value added in PALMSv3 and later versions is the creation of a consistent income variable over all waves that collected labour income. Further information on this variable can be found in section 4 below.

The naming of variables and value labels for the LFSs was done using code from David Lam at the University of Michigan. A similar structure was then employed to code the OHSs and QLFSs. The `do` files and `csv` files which do this naming and labelling have been made publicly available with the data.

If you use PALMS please cite the data, as well as this guide if you have used it. To cite the data please use the following citation:

Kerr, Andrew, David Lam and Martin Wittenberg (2018), Post-Apartheid Labour Market Series [dataset]. Version 3.3. Cape Town: DataFirst [producer and distributor], 2019.

In the sections below we first explain the new data and variables made available in this version of PALMS and then explain how the cross entropy weights that are included in PALMS were created. We then discuss how a consistent labour income variable was created and the imputation of the QLFS earnings undertaken by Stats SA. We finally describe how the multiply imputed labour income data was created. In the Appendix we explain how users

---

\* DataFirst, University of Cape Town. [andrew.kerr@uct.ac.za](mailto:andrew.kerr@uct.ac.za).

† DataFirst and School of Economics, University of Cape Town

‡ A large number of people have provided very useful feedback on prior versions of PALMS, and we thank in particular Nicola Branson, Patrizio Piraino, Owen Crankshaw, Neil Balchin, Friedrich Kreuser, Karmen Naidoo, Rob Davies and Carlos Gradin. We thank Takwanisa Machemedze for creating the cross entropy weights using the Stats SA mid-year population estimates model and extending this back to 1993 and Alex Montgomery for providing assistance with incorporating the Labour Market Dynamics into PALMS. We thank Gabriel Espi-Sanchis for creating the version of the 1993 PSLSD that has been incorporated into PALMS and Jacqueline Mosomi for providing the code that fixes the OHS domestic worker occupations, industries and employer codes.

can recreate the dataset themselves using the data and Stata code available on the DataFirst website. We also provide a variable list.

## **2 New in PALMS version 3.3**

### **2.1 New QLFS waves**

For this release of PALMS we have now included the QLFS 2017 data from quarters 3 and 4, QLFS 2018 and QLFS 2019 waves 1 and 2, as well as the 2016 and 2017 Labour Market Dynamics data on earnings, which contains the earnings information collected in the QLFS but which is not released by Stats SA at the same time as the quarterly releases of the rest of the QLFS data.

### **2.2 New demographic model for cross entropy weights**

The second substantial change in PALMSv3.3 is in the demographic model used to obtain the cross entropy weights. Previous versions of PALMS used the 2008 Actuarial Society of South Africa (ASSA) demographic model to generate sampling weights (cross entropy weights) that use a demographic model that is consistent over time (instead of using the most recent estimate at the time of each survey- which can lead to jumps up and down in the population estimate). The ASSA 2008 model was last updated in 2011 and did not take into account higher survival rates and the resulting population growth, mainly as a result of the large scale roll out of anti-retroviral drugs by the South African state. This meant that ASSA 2008 had population counts by 2016/7 that were several million lower than the Stats SA demography model (released as the mid-year population estimates) or the Thembisa demographic model produced by University of Cape Town academics Leigh Johnson and Rob Dorrington. We have thus decided to include cross entropy weights that are based on the new Stats SA mid-year population estimates. These estimates only go back to 2002, so we have extended back to 1993 using a simple exponential growth model, using the ASSA 2008 growth rates over the period 1993-2001. There are very minor differences for 1993-2001 between these estimates and the ones from the ASSA 2008 model over the same period.

The change to using the Stats SA mid-year population estimates also affects the bracketweights used for earnings analysis- we have constructed these bracketweights using the cross entropy weights derived from the Stats SA mid-year population estimates in PALMSv3.3. We are currently conducting more analysis to explore which demographic models to use in future versions of PALMS.

### **2.3 Warning about the QLFS earnings data**

PALMS is designed to facilitate analysis of the South African labour market. We discuss imputation of earnings in the QLFS in more detail below, but here we want to issue a warning about the QLFS earnings data, because the QLFS public release imputed data show some very worrying and unbelievable trends in earnings inequality, when measured by the Gini coefficient. Wittenberg (2016) and Finn and Ranchhod (2017) have shown some unbelievable trends in the Gini coefficient that is calculated on the QLFS public release imputed data. In Kerr and Wittenberg (2017) we described that there are two very different imputation regimes that make it hard to compare QLFS earnings over time. In Kerr and Wittenberg (2019) we extend the Gini coefficient estimation to Q4 2017 and show that it continues to fluctuate wildly, even when excluding the outliers we identified. We also show that the Gini coefficient from the General Household Survey seems more sensible, probably because it is not imputed. The GHS has run every year since 2002 but sadly does not include the job and worker questions of the QLFS and LFS for detailed labour market analysis.

We should note that Wittenberg (2016) produced sensible looking trends in various earnings percentiles using the imputed QLFS data up until 2014, so there is value in using the data for analysis. But because the earnings data seems to have a number of issues we would like to warn users to be extremely careful when using the QLFS earnings data, and also when comparing the QLFS to the LFS or OHS. This issue makes it even more crucial that users of the micro data request public releases of unimputed earnings data from the QLFS from Stats SA, something which we are continuing to do.

## 2.4 Corrected years of education variable in 2016:3 and later waves

In PALMSv3.3 we have included a corrected years of education variable. In PALMS v3.2 the years of education variable was incorrect for 2016:3 and later waves as a result of a change in the values assigned to different education levels. We thank Nicola Branson for pointing this out to us.

## 2.5 Dropping of formalreg2, replacing with informal\_derived

The variable formalreg2 has been dropped and replaced with informal\_derived, which is defined for 2009:3 onwards. We have changed the name of the variable to make it clearer that it is NOT the result of a direct question either to the self-employed or employees about whether they think the business they work in or own is in the formal sector. Instead, it is derived by Statistics South Africa from a number of questions asked to employees and the self-employed and so we have chosen to rename it informal\_derived.

informal\_derived is thus NOT comparable to the formalreg0 or formalreg1 variables which asked direct questions to both employees and the self-employed in the LFS ( and the QLFS up until 2009:2 (formalreg0) or to the self-employed only from QLFS 2009:3 onwards (formalreg1). In both the LFS and QLFS up until 2009:2 the self-employed and employees were *also* asked direct questions about whether the businesses/organisations they owned or worked for were registered for income tax or VAT. From 2009:3 onwards only the self-employed were asked these questions. Thus from 2009:3 there is NO question about formality or registration for employees about the businesses/organisations they work for.

The informal\_derived variable is an informal sector variable derived from a subset of the questions asked to both employees and employers. The Statistics South Africa QLFS Guide states that employees are defined as informal sector if they “are not registered for income tax and who work in establishments that employ less than five persons.” Later this is clarified as “Income tax deducted by employer.” This definition suggests that anyone under the tax threshold (R75750 in the 2018 tax year, way above median earnings) who did not have income tax deducted should be considered to be in the informal sector, or at least may lead to some ambiguity. Other employed individuals (Employers, own-account workers and persons helping unpaid in their household business) are considered in the informal sector if they “are not registered for either income tax or value-added tax.”

## 2.6 UIF deduction question

We have also added a variable on whether the worker’s firm deducted Unemployment Insurance Fund (UIF) contributions from their earnings. This is an alternative method of investigating informality since any UIF registered employer is required to deduct UIF, unless the worker works less than 24 hours a month). This question has been asked since the 2000 LFS, but there is a big jump up in the proportion of workers reporting UIF deductions in the switch to the QLFS in 2008. Users of this variable should investigate further.

## 2.7 3 digit industry code labels

We have now included 3 digit industry code variables for both the industry and industry2 variables. industry is the 3 digit code for OHS 99 and the LFSs, whilst industry2 is the 3 digit code for the OHS 96-98 and the QLFS. It seems strange that in the QLFS Stats SA has gone back to the old industry codes it used pre-OHS 1999, but this seems to be what the documentation suggests is correct. We would welcome feedback from users. In future versions of PALMS we would like to include consistent 4 digit occupation code labels also, but that has not been possible in this release.

# 3 A reminder of what was new in PALMS version 3.2

## 3.1 Change in base year for real earnings variable

We changed the base year for the realearnings variable in PALMS3.2 so that Dec 2016 was the base period. December 2017 is now the base period, and this is the latest year for which there is earnings data in PALMS.

## 3.2 Domestic worker occupation, industry and employer codes for the OHSs

In PALMSv3.2 we released corrected 1 digit occupation and industry codes for OHS 94 (jobocccode and jobindcode), which were not consistent in previous versions in this year, as many individuals who were recorded as domestic workers according to the industry code were then not coded as domestic workers in the occupation variable. This led to a large undercount of domestic workers if the occupation variable was used (Mosomi 2016). We used a probit to predict the probability an individual was a domestic worker in PSLSD 1993 and OHS 1995 and then used the coefficients from this model to predict domestic worker status for individuals in 1994. This procedure increased the number of domestic workers in the *sample* by around 2000. We have not adjusted the 3 or 4 digit occupation and industry codes in any wave. We are grateful to Jacqueline Mosomi, who discovered this error and who allowed us to use her code to implement the fix.

## 3.3 Employment variable excluding self-employed agricultural workers

In a recent paper Neyens and Wittenberg (2016) show that the self-employed agricultural worker series from PALMS was extremely inconsistent over time and contributed to some unbelievable changes in total employment. As a result of this work we have now included a new employed variable, *employedpreferred*, which is based on the status variable previously included in PALMS but which now excludes self-employed agricultural workers. Employment trends using this variable are much more sensible than the raw employment numbers created from the Stats SA status variables.

## 3.4 Public sector Employment variable

Questions about whether the worker worked in the public sector were asked in the Project for Statistics on Living Standards and Development (PSLSD), Labour Force Survey (LFS) and Quarterly Labour Force Survey (QLFS). In the OHSs no question was asked directly but the industry code variable for OHS 1996-1999 contains enough information that public sector employment can be inferred, although not perfectly. We thus have included a variable in PALMS v3.2 and v3.3 which is a public sector dummy that includes an imperfect measure of whether an individual worked in the public sector in the OHSs. This new variable is called *publicemp2*.

In OHS 1996-1999 we use the industry codes 910 (public administration and defense activities), 911 (central government activities), 912 (regional service council activities), 913 (Local government activities), 914 (provincial administrations), 915 (SA Defence Force), 916 (SA Police Force), 917 (Correctional service) as well as 410 (Electricity, Gas, Steam And Hot Water Supply), 411 (Production, Collection And Distribution Of Electricity) and 711 (Railway Transport). These are all unambiguously public sector employers. Employees in industries 920 (education) and 931 (human health activities) are also mostly going to be public sector employees, although private medical personnel and private teachers are also included in this category. Excluding these two industries results in estimates that are too small and so we include them, knowing that we are making errors.

Using these codes produces estimates of the size of the public sector that look too small in 1997-1999 and about right in 1996 (though as noted above we are including private medical personnel and teachers in these estimates). There are likely to be further public employees in post and telecommunications but we have chosen not to code these as public employees since a sizable fraction will be employees in the private sector.

We have also fixed the *publicemp* variable to be zero for those in self-employment in the first 3 LFSs- in earlier versions of PALMS these were mistakenly set to missing (we assume the self-employed cannot be employed in the public sector).

# 4 Labour Income data in PALMS v3+

## 4.1 Comparison with pre-PALMS v2 earnings

Version 2.0 (and up) of PALMS uses a very different approach to labour income compared to the previous versions. PALMS from version 2.0 onwards now contains only a few earnings variables. These have been cleaned and coded by Martin Wittenberg. There is a real and nominal earnings variable. The old income variables (varying substantially across waves) are now included in a separate file called "PALMSv3.3incomes". This includes all the earnings data from each wave, as it was collected, and is similar, although not identical, to the way the earnings data was released

in versions of PALMS prior to version 2.0. For those who wish to replicate or adjust the creation of the single labour earnings variable this file can be downloaded with the PALMSv3.3 data . This can be merged into the main PALMSv3.3 data by the (Stata) command:

```
merge 1:1 uqnr personnr wave using PALMSv3.3 incomes
```

The do files PALMSwages\_create, PALMSwages\_create2 and PALMSwages\_create3 (released with this data) used to create the final earnings variables included in PALMSv3.3 can then be run relatively easily to replicate these variables.

## 4.2 Imputation in the QLFS

It is important for analysts to be aware that the labour income data in the QLFS from 2010-2016 is substantially different to the OHS and LFS incomes. Firstly, every person who refused to answer or reported only a categorical value is given a rand amount for the 10 QLFSs between 2010 and Q2 2012 inclusive, ie income data is imputed for all these individuals by Stats SA. There is no way to identify who refused or only reported a categorical value in the publicly released data (The income data has been released separately in a publication called the “Labour Market Dynamics”). From Q3 2012 onwards complete refusals are no longer imputed but categorical responses do have a rand amount imputed. It is not possible to determine which individuals gave categorical answers which are then imputed and which are actual responses. The large number (449) of individuals reporting exactly R400 000 income per month in the QLFS after 2012 Q2 is still worrying, suggesting that top coding has been done by Stats SA since Q3 2012.

We have not adjusted the imputed income data in 2010-2016 but we have still flagged outliers. We hope to obtain publicly available data from Stats SA and include it in future versions of PALMS unimputed labour incomes from the QLFS. See Kerr and Wittenberg (2017) for an analysis of QLFS 2011 data without imputation.

## 4.3 Reweighting for Earnings Bracket Responses

PALMS v3.3 contains a inflation adjusted labour earnings variable called `realearnings` (the base period is December 2016). This is the recommended variable to undertake any analysis of labour incomes in PALMS. Outliers in the `realearnings` variable are flagged but not adjusted. Any analysis will be complicated by the many individuals who refused to answer this question but responded in brackets to the categorical question. DataFirst thus recommends users weight the data to account for those individuals who responded by giving an earnings bracket value but not a Rand amount. Simply ignoring the bracket responses incorrectly ignores responses that overwhelmingly come from the top end of the income distribution. The weight to do this is **bracketweight**. See Wittenberg (2008b) for a discussion of this method.

Bracketweight is a combination of the inverse of the probability of a bracket response in a particular bracket in a particular wave, multiplied by the cross entropy weight for that particular individual created from the 2008 ASSA model. In 1996 there were no actual rand amounts collected. There were no incomes collected in the first 6 waves of the QLFS and income data has not yet been released for the last 2 quarters of 2009. Below is a simple example of the use of the bracket weight (in Stata) to estimate mean real earnings over each wave of the data for which there is income data:

```
table wave [pw=bracketweight], c(m realearnings)
```

Users should note that this approach does not do anything about those who refuse to answer or who otherwise have missing data- it only corrects for bracket responses. Users who prefer to correct for this type of non-response could use the multiply imputed data file released with PALMS v3.3, see section 5 below.

## 4.4 Outliers

In the current version of PALMS we have one outlier indicator but outliers are not set to missing or imputed in the main data set. In the multiple imputation dataset outliers are set to missing and then imputed. Secondly, a studentised residual (i.e. residuals normalised against their residual standard deviation, but calculated from a regression in which that observation is left out) from a Mincerian regression of logearnings on age, population group, gender, years of education, wave, occupation and interactions between wave and pop group and wave and gender was flagged if it had an absolute value of greater than 5 (implying we should expect to see 1 observation with this value in the data.

To repeat the above analysis but excluding outliers based on a regression of earnings against education, age and occupation one would use the following command:

```
table wave if outlier==0 [pw=bracketweight], c(m rearnings)
```

## 5 Multiply Imputed Labour Income Data

DataFirst has created a set of imputed earnings data to accompany PALMS from version 2 onwards. This section describes the multiply imputed data, which have now been released with PALMS v3.3, and how this data was created. The imputed earnings data has been released as `palmsv3.3miincomes.dta`. Two new do files, `PALMSwages_create4` and `PALMSwages_create6`, accompany this imputed data and can be used to recreate or modify the imputations undertaken by DataFirst. We have included 10 replications of the imputed data, to allow for the uncertainty inherent in any imputation procedure. We expect that users of this data will use the 10 replications and correct their standard errors accordingly. For more information on using multiply imputed data in Stata, type `help mi`.

To create the imputations we first dropped all those not employed. We then imputed a bracket for those that did not even have a bracket earnings response (the “don’t know”, “refused” and “unspecified” categories) or who were classified as outliers (477 observations) based on an ordered logit for each wave of the data using province, gender, education, population group, a quadratic in age and occupation as explanatory variables. Earnings were then imputed based on the predicted bracket using predictive mean matching, a variant of hotdeck imputation.

There were no earnings amounts captured in the 1996 OHS. These responses were dealt with by “predictive mean matching” of the 1996 bracket respondents with respondents providing Rand amounts in the 1997 OHS. The real earnings figures (i.e. deflated values) were matched, so that inflation between the periods is controlled for to some extent.

The PSLSD did not ask about earnings in brackets for those who refused. So in imputing earnings for working individuals who did not report earnings information in the PSLSD we simply used predictive mean matching, a variant of hotdeck imputation, for the actual rand amounts.

Vermaak (2012) notes that imputing values for individuals providing zero incomes makes an appreciable difference to the earnings distribution. Since the pattern of people reporting zero incomes varies considerably over the surveys it is unlikely that they are providing good data. On the other hand it seems dubious to assume that someone who explicitly records a zero should really have recorded a positive amount. It seems more likely that this response reflects that the data comes from a different “data generating process” – e.g. unpaid family workers, individuals whose attachment to the job is tenuous (maybe on a waiting list). This category clearly deserves closer scrutiny. For the time being, those reporting zero incomes were set to missing and not imputed.

In Stata, one can merge the imputed data into the original data using the following code, with the `PALMSv3.3` data open:

```
merge 1:1 wave uqnr personnr using palmsv3.3miincomes.dta
```

The data is set up to be used for `mi` work but to use these imputations the data needs to be `mi` set in Stata. Type `help mi` for more information. Only those with earnings data are included so some of the `palmsv2` data will not be matched with the new multiply imputed data.

The `rearnings` variable is the one that was imputed. `rearnings` is set to hard missing (`==.a`) for those who we did not want to impute incomes (those reporting zero incomes). `rearnings` is missing (`==.`) for those who we then did impute an earnings bracket and amount for. The initial imputations are contained in the `imputed_real` and `imputed_nom` variables. Some incomes could not be imputed. In these cases the `imputed_real` variable is set to hard missing (`.a`) and the `imputed_nom` variable to missing (`.`). The 10 versions of these imputations are (for real income) then contained in the `imputed_real_v1- imputed_real_v10` variables and the imputations for nominal income are contained in the `imputed_nom_v1-imputed_nom_v10` variables.

As a result of using predictive mean matching some of the ten imputations for each individual are the same. This may be of concern for some users, but the alternative is some type of parametric imputation, which DataFirst did not wish to undertake.

This multiply imputed dataset is the only one of those released as PALMS which is set up specifically for Stata users. The other data can easily be used in other statistical programmes. We hope that users of SAS, SPSS and R can also use the multiply imputed data as it stands but we are not certain this is possible. Feedback from users of these other programmes would be useful- please email [andrew.kerr@uct.ac.za](mailto:andrew.kerr@uct.ac.za) if you have comments or suggestions.

## References

- Finn, Arden and Vimal Ranchhod**, “Short-run differences between static and dynamic measures of earnings inequality in South Africa,” Technical Report 2017.
- Kerr, Andrew and Martin Wittenberg**, “Public sector wages and employment in South Africa,” REDI Working Paper Series 42, REDI 3x3 2017.
- **and** —, “Employment and Earnings Microdata in South Africa,” UNU WIDER Working Paper 2019/47, UNU WIDER 2019.
- Mosomi, Jacqueline**, “The Role of Domestic work in Female Employment with Implications on Measurement Issues in PALMS,” 2016. AERC Biannual Research Workshop paper.
- Neyens, Liz and Martin Wittenberg**, “Changes in self-employment in the agricultural sector, South Africa: 1994-2012,” 2016. Southern Africa Labour and Development Research Unit Working Paper Number 173. Cape Town: SALDRU, University of Cape Town.
- Vermaak, Claire**, “Tracking poverty with coarse data: evidence from South Africa,” *Journal of Economic Inequality*, June 2012, 10 (2), 239–265.
- Wittenberg, Martin**, “Income in the October Household Survey 1994,” Technical Report 2008. DataFirst Technical Paper 7.
- , “Nonparametric estimation when income is reported in bands and at points,” 2008. Economic Research Southern Africa Working Paper 94.
- , “Trends in Earnings and Earnings Inequality in South Africa: 1993-2014,” Technical Report 2016. Unpublished working paper.

## A Advice for those wanting to recreate the PALMS dataset

The OHS, LFS and QLFS data have been prepared separately and then appended together. The LFS data are prepared using a method and a set of excel and do files obtained from David Lam at the University of Michigan. The OHS and QLFS data are prepared using a similar method to that used by David Lam, using a new set of do and excel files.

This section sets out how this process was done which should allow other researchers to replicate this process. This will be useful for several reasons. Researchers may wish to check the work that has been done to create PALMS. Researchers may also wish to add in the significant amounts of LFS data which have been coded by David Lam and Kendra Goostrey, but which are not included in PALMS. Those who wish to do so will need to read and understand the explanations below, and use the do files and excel files found on the PALMS sections of the DataFirst website, as well as the data files from each OHS, LFS and QLFS survey, which can be found on the DataFirst website.

### A.1 LFS Codebook

Kendra Goostrey, formerly a PhD student supervised by David Lam, wrote an unpublished explanation of the method for putting together the LFS data, which we briefly describe here. This method relies on two excel files that allow one to consistently rename variables and code values for variables across all waves of the LFS. A quote from that document is included below to help explain the process. It describes the master codebook for all waves of the LFS. In that spreadsheet, titled, "LFS master codebook.xls," you will find a column for each wave of the LFS listing out the variables matched by row with identical questions in prior waves.

“ Multiple rows often exist for certain variables (such as marital status) for several reasons:

- A change has occurred in the order of responses. For example: marstat1 (wave 00:1) lists 'Never Married' as response 1, while marstat3 (waves 04:2 to present) lists it at response 5.
- A change has occurred in the degree of information provided by the question. It now includes more/less/different information or information in a different format. For example: marstat1 & marstat list married and cohabiting responses together, but in later waves (marstat3) these responses are separated, giving a new list of value labels.
- These differences are important because each row in this spreadsheet corresponds to a set of value labels listed in LFS Master value labels.xls\*; so when the order of value labels changes, a new list of these labels is needed, thus requiring a new variable name and a new row in this sheet.

”

*David Lam & Kendra Goostrey*

### A.2 OHSs

A similar set of excel files and do files has been constructed for creating a consistent set of variables in the 94-99 October Household Surveys and we briefly outline how these work below. It should be noted that whilst the LFS files from David Lam code a very comprehensive set of variables, the OHS is currently more limited in the number of variables which are cleaned and coded consistently across waves. The OHSs also varied a lot more between waves than the LFSs. This means that there are often several rows for a similar question (popgroup2, popgroup3 and popgroup4 when looking at population group for example) in the OHSs.

#### A.2.1 OHS Master Codebook.csv

This file contains a list of some variables in the OHSs which are to be labelled and renamed. Each row represents a different variable. If a question changed between surveys then this new variable requires a new line in the csv file. As in the LFS Master Codebook, each line in "OHS Master Codebook" corresponds to a set of value labels in the "OHS Master Value Labels.csv" file.

These two .csv files are used by "OHScreatedofiles.do" to automatically generate a set of do files that rename and relabel the OHS variables in each wave. These are "OHSrename'wave'.do" (one for each wave), "OHSrename2.do", "ohslabelvars.do", "ohslabeldefine.do" and "ohslabelvalues.do". The beauty of Lam's method is that this is all automated, and that the excel files make it easy for a new user of the data to see how the variable definitions changed



across the waves (users may be intimidated by the large number of do files, but once the structure is understood it is actually not difficult to understand the method).

Like in Lam's LFS method, each round of the OHSs has its component sections appended together in OHSmerge'wave'.do, which also calls the set of renaming and labelling do files mentioned above. Each of the do files associated with an OHS survey then saves the full data set (all data from person, worker and household data, whether renamed/relabelled or not) and a smaller data set that is used to create a data set with small versions of each of the 6 OHSs. At the moment we have not automated the entire process of adding new labelled and renamed variables to the appended OHS dataset. This requires adding the variable name to the keep command at the end of the ohsmerge'wave' do file. If the variable is in every wave then each do file needs to be changed.

The OHS data from each wave is then appended together in "ohsappend.do." Unlike Lam's method for the LFSs, most of the coding that creates consistent variables across waves is done in ohsappend, rather than in the do files for each wave (OHSmerge'wave'.do). There is some coding of consistent variables done in each OHSmerge'wave'.do file. There is also some cleaning done in these do files, which Lam does not do for the LFSs. Further cleaning is done in ohsappend.do, as well as in the do file that appends the OHS and the LFS (appendlfstooohs.do).

The data used for creating this dataset come from the DataFirst Server. At the time a version of the OHS data on the DF website from Stats SA was found to have some problems, for example a large number of duplicate households in 1996. Another version of the OHS data has now been put up. Users of the OHS data or those who wish to replicate the cleaning and appending of the OHS data described above should check they have the latest version of the data from DataFirst.

### A.3 Appending the LFS and OHS

The OHS do files currently cleans and consistently codes only a limited number of variables. Thus a smaller version of the LFS data is created in "createsmallerlfswave1to16". "createohsconsistentlfs" uses this smaller data set and creates variables that are consistent with the OHS. Finally, "appendohstolfs.do" appends the OHS and smaller LFS data together and creates the data set "ohslfsdata".

### A.4 Adding in the QLFS

This has been done in the most recent version of PALMS. A similar set of do files and excel files were used to rename, relabel and append the QLFS data together, before appending it to the OHS and LFS data.

"LFS master codebook with QLFS.csv" is the main file used to create the renaming and relabeling do files. It contains LFS variables as well but these are ignored by the do file "QLFScreatedofiles" which uses the QLFS variables in "LFS master codebook with QLFS.csv" and then creates the other do files used in renaming and relabeling.

Like in Lam's LFS method and the OHS method described above, each round of the QLFSs has its component sections appended together in QLFSmerge'wave'.do, which also calls the set of renaming and labelling do files mentioned above. Each of the do files associated with a QLFS survey then saves the full data set and a smaller data set that is used to create a data set with small versions of each of the QLFSs. The QLFS data from each wave is then appended together in "qlfsappend.do." A smaller dataset is created in createsmallerqlfs.do. Some coding of variables to be consistent with PALMS is undertaken in "createpalmsconsistentqlfs.do"

Income data was not released with the quarterly QLFS data releases. However, income data has been released in the 2010 - 2017 Labour Market Dynamics (LMD). We have merged the LMD into the QLFS for 2010-2017. 2009 income data was collected in the 3rd and 4th quarters but this data has not yet been released by Stats SA and neither has the 2018 LMD.

#### A.4.1 Taking full advantage of the coded LFS

The publicly released version of PALMS data contains more than 120 variables, most of which are consistent across the entire period. There is actually a much richer amount of data coded and labelled for the LFS by David Lam and Kendra Goostrey at the University of Michigan (see the LFS master codebook.csv file), and this is useable by anyone who can run the do files and append the LFS data together (see lfsappend.do). Researchers who wish to look at the LFSs only thus can easily access a wide range of LFS data that is consistent across waves. They will still need to code and make consistent the OHS data if this is not in the final data set, however, which is more difficult because many questions in the OHSs varied across waves.

#### A.4.2 Additional variables

Cross entropy weights created by Takwanisa Machemedze are included in the data, along with the Stats SA person and household weights. These will need to be downloaded from the PALMS page at DataFirst before the do files will run.

### A.5 A Brief explanation of how the final dataset is put together

Individual wave data labelled in qlfsmarge‘wave’.do, lfsmerge‘wave’.do and ohsmmerge‘wave’.do (these do files use other renaming and labelling do files, e.g. “LFSrename2005\_1”, “lfslabelvalues” or “OHSrename2”)

OHS data appended in ohsappend.do, LFS data appended in lfsappend.do, QLFS data appended in qlfsappend.do

Smaller LFS file created in createsmallerlfswave1to16.do. Smaller QLFS file created in createsmallerqlfs.do. OHS-consistent LFS data is created in createohsconsistentlfs.do. OHS and LFS-consistent QLFS data created in createpalm-consistentqlfs.do OHS and LFS appended together in appendohstolfs.do QLFS and PSLSD appended to OHS and LFS in appendqlfspslsdtoohslfs.do

The work by Martin Wittenberg on creating the consistent income variables is then done in 4 do files: PALMSwages\_create, PALMSwages\_create2, PALMSwages\_create3, PALMSfinal\_create

The last of these drops all the other income data (similar to that released in previous versions of PALMS) from the main data set but saves it all as a separate file called palmsv3.3incomes. This can then be merged into palmsv3.3.dta if users wish to replicate/modify the work on incomes. There are then 2 more do files which create the imputations and the dataset PALMSv3.3miincomes. These are PALMSwages\_create4 and PALMSwages\_create6.

## B PALMS Variable Description

In this section we give a description of each variable in the PALMS dataset.

### B.1 Variables

**uqnr:** Household identifier, this is not unique across waves. It is the same as the original variable from Stats SA for all waves, except for 1996, where no household id was supplied and hence was created from magisterial district, enumeration area and visiting point variables.

In previous versions of the data the uqnr variable was not the same as the original hhid variable as a zero was taken out. A second variable, uqnr\_orig was added in version 1.0.8 to make merging in data easier. In the latest version there is only one variable, uqnr, and this is the original household id variable, in string format.

**perssonnr:** The number of the person in the household. Valid range: 0-85.

**year:** Year of the survey. Valid range: 1993-2017

**wave:** Wave, with PSLSD 1993=0, OHS 1994=1 and QLFS 2017:2=60. The last OHS was OHS 1999 and this was wave 6. The first LFS was a pilot survey conducted in February 2000, which is included as wave 7 in this data set. The first wave of the QLFS is March 2008, wave 23. Valid range 0-54.

**province:** Province the household is located in. Used for stratification by Stats SA in waves 1-15. Valid range: 1-9.

**metro:** QLFS metropolitan municipality identifier. Valid range: 0-76.

**urbrur:** Type of area, 1=urban, 2=rural. Only supplied by Statistics South Africa until March 2004. Used for stratification by Stats SA in waves 1-15 (March 2004).

**urbrur2:** Type of area, 1=urban formal, 2=Urban informal, 3=Tribal areas, 4=Rural formal. Only supplied by Statistics South Africa between 2008 and 2014 inclusive. Used for Stratification

**urbrur3:** Type of area, 1=Urban, 2=traditional, 3=Farms, 4=Mining Areas. Only supplied by Statistics South Africa in 2015 onwards (new master sample). Used for stratification by Stats SA.

**ea:** Enumeration area/Primary sampling unit. In OHS 94 the variable was created from the variables number1 and number2 in the house data. In OHS 1995-1998 the ea variable was created from the ea and magisterial district variables supplied by Stats SA. In the 2001 LFSs the ea variable was created from the ea variable supplied by Stats SA and the stratum variable, in both cases to obtain a variable that had roughly 10 households per ea. In some other waves the ea variable is the same as that supplied by Stats SA, whilst in others it was created using the first 7 or 8 digits of the id number (see survey do files for more information). As a result of these differences EA numbers are not always comparable over time. In principle it is possible to construct a panel of EAs for each of the master sample periods.

**stratum:** This is a variable created for PALMS for users to set up their data as complex survey data. It is a combination of wave and the actual stratum variable released in each wave of data (or the stratum variable created in each wave of data for PALMS where no stratum variable or an incorrect stratum variable was released by Stats SA).

**dc:** District Council. Only present in waves 16-22. There was explicit stratification on this variable in these waves, replacing the province and urban/rural stratification in previous waves. Valid range: 1-55.

**ceweight1:** Cross Entropy weight derived by DataFirst from Stats SA mid-year population estimates for 2018. This weight is now the recommended weight for use in PALMS, because of the ASSA 2008 model substantially underestimating the SA population (see discussion above) in more recent years. We have excluded the older ceweight and ceweight2 variables to prevent confusion. Researchers wanting to replicate their results using ceweight2 should merge in the older version of PALMS into PALMSv3.3 For earnings analysis bracketweight should be used. See below.

**pweight:** Person weight supplied by Statistics South Africa/SALDRU. Valid range: 0.07429, 53717.976. Missing for 2 individuals.

**hweight:** Household weight supplied by Statistics South Africa. Valid range: .05449,21533.18. Not supplied by Statistics South Africa in LFS 00:1, 02:1 and 03:1. From 2004 household and person weights did not differ, hence the non-missing person weights are also household weights in these years. In most LFSs after 2004 household weights are not supplied and data users should create their own household weights from the person weights if these are required. It is not clear if this also applied to 00:1, 02:1 and 03:1, where no household weights were supplied but where there was no discussion of whether person weights could be used as household weights. No hweights were supplied in QLFSs.

**inperson:** Whether the person responded in person or not. Only asked in LFS and QLFS. 1=Yes, 2= No, 8=n/a, 9=unspecified.

**popgroup:** Population group of individual. Missing for 1108 individuals. 1=African/black, 2=Coloured, 3=Indian, 4=White, 5=Other.

**gender:** Gender of the individual. 1=male, 2=female. Missing for 501 individuals. Imputed by Stats SA in the QLFSs.

**age:** Age of the individual. Missing for 2295 individuals. Valid range: 0-142 (implausible maximum, but this is in Stats SA metadata).

**marstat:** Marital status of the individual. Missing for 844 individuals. 1=married or living together as husband or wife, 2=widow/widower, 3=divorced or separated, 4=never married, 9=unspecified. In the PSLSD this variable

was created from a relationship code to the head of the house and so is not comparable with the other waves.

**yrseduc:** a derived variable showing the number of years of education. Derived from educhigh0, educhigh1, educhigh2.

**educhigh:** No longer in PALMS - see yrseduc for a derived variable giving the number of years of education or educhigh0, educhigh1 and educhigh2, giving categorical education variables for the OHS, LFS and QLFS respectively.

**educhigh0:** Highest level of education achieved in OHS. There are some categories only valid for certain waves within the OHSs. Valid range 0-26.

**educhigh1:** Highest level of education achieved, LFS only. Valid range 0-99.

**educhigh2:** Highest level of education achieved, QLFS 2008:1- 2012:2. Valid range 0-99.

**educhigh3:** Highest level of education achieved, QLFS 2012:3- 2016:2. Valid range 0-98.

**educhigh4:** Highest level of education achieved, QLFS 2016:3 onwards. Valid range 0-98.

**enrolled:** Detailed information on whether the individual was enrolled in an educational institution and the type of institution. Valid range: 1-9. Only defined for the LFS waves. (but see enrolment3 below for less detailed enrolment variable that covers the OHSs).

**enrolment3:** Enrolment information that covers the PSLSD, the OHSs and LFSs but not QLFSs. It also has less detail than enrolled variable above. (There is no direct enrolment question asked in the QLFSs). Valid range 1-9. 1= full-time, 2= part-time, 3=not enrolled, 9=unspecified.

**empstat1:** Employment status, using the strict definition of unemployment. Valid range 0-2. This is the Stats SA variable included with most of the waves of the OHS and LFS and created from the status variable in the QLFSs. For the PSLSD it is created from the employment questions. This is not comparable over time because of how Stats SA changed the definitions of what counts as work (excluding subsistence agriculture in the QLFS for example) and the criteria to be counted as searching unemployed (which became stricter in the QLFS). 0= not economically active, 1=employed, 2=unemployed.

**empstat2:** Employment status, using the expanded definition of unemployment. Valid range 0-2. This is the Stats SA variable included with most of the waves of the OHS and LFS and created from the status variable in the QLFSs. For the PSLSD it is created from the employment questions. This is not comparable over time because of how Stats SA changed the definitions of what counts as work (excluding subsistence agriculture in the QLFS for example) and the criteria to be counted as searching unemployed (which became stricter in the QLFS). 0= not economically active, 1=employed, 2=unemployed.

**employer:** Individual's type of employer. Valid range: 1-9. Missing for OHSs (but see employer1 variable below). employer=8 for those not employed.

**employer1:** The individual's employer. Valid range: 0-8888. Only asked for OHSs and created for the PSLSD (but see employer variable above). Both 0 and 8888 are not applicable codes.

**employer2:** The individual's employer. Valid range: 0-8888. Only asked for QLFSs (but see employer and employer1 variables above). Both 0 and 8888 are not applicable codes.

**numworkers:** The number of workers at the individual's place of work. It is a categorical variable. Valid for

LFSs only (but see numworkers2 below). Valid range 1-9.

**numworkers2:** The number of workers at the individual's place of work. It is a categorical variable. Valid for QLFSs only (but see numworkers above). Valid range 1-88.

**jobstartyear:** The year the individual started working in their current job. Valid range 1927-9999. Only asked in LFSs and QLFSs.

**jobstartmonth:** The month the individual started working in their current job. Valid range 0-99. Only asked in LFSs and QLFSs.

**writtencontact:** Derived variable, =1 if employee has a written contract with employer. Only asked in LFSs and QLFSs.

**selfformalreg:** Self-employed individual considers the business they operate to be formal. OHS only. 0=Missing, 1=Formal, 2=informal.

**selfvatreg:** Self-employed individual's business is VAT registered. OHS only. 0=Missing, 1=Formal, 2=informal, 3=Don't know

**selfpaidemp:** The number of paid employees in the self-employed individual's business. OHS only. Valid range: 0-2000.

**selfunpaidemp:** The number of unpaid employees in the self-employed individual's business. OHS only. Valid range: 0-1000.

**wageformalreg:** Employee considers the enterprise they work for formal. OHS only. 0=Missing/Don't know, 1=Formal, 2=Informal, including domestic work.

**formalreg0:** Employed person considers the enterprise they work for or own to be formal (with a questionnaire prompt explaining that formal businesses are registered for VAT). LFS only. 1=Formal, 2=Informal, including domestic workers, 3=Don't Know, 7=Other, 8=n/a, 9=Unspecified.

**formalreg1:** Employed person considers the enterprise they work for or own to be formal. QLFS 2008:1-2009:2 only. 1=Formal, 2=Informal, 3=private households, 4= don't know , 8=n/a.

**uif:** Individual's employer deducts UIF contributions from their earnings. LFS 2000:1- current. 1=Yes, 2=No, 3=don't know. Before LFS 2005:1 the question had 2 no options, one was for if the person was above the threshold, we have combined the two no options into one (we understand that there has never been a UIF contribution earnings threshold so this option was redundant, perhaps that is why it was dropped in 2005 onwards).

**informal\_derived:** The direct question to employees stopped in QLFS 2009:3. informal\_derived a version of a variable now created by Stats SA from the answers to a number of questions, including the size of the firm and whether various benefits are paid. 1=Formal, 2=Informal. This variable replaces formalreg2 in PALMSv3.2- we wanted to make it clearer that this is NOT comparable to formalreg1 or formalreg0.

**regisvat:** Is the Business registered for VAT? Both self-employed and employees in LFS, only self-employed in QLFS. 1=Yes, 2=No, 3=Don't know. 8=not applicable, 0= not applicable, 9= unspecified.

**jobindcode:** One digit industry code for both employees and self-employed. Valid range 1-99.

**industry:** 3 digit industry code. Defined for OHS 99 and LFSs. (see release documentation, Stats SA seems to have gone back to OHS codes in QLFS)

**industry2:** 3 digit industry code. Defined for OHS 96-98 and QLFSs. (see release documentation, Stats SA seems to have gone back to OHS codes in QLFS)

**industry2digit:** 2 digit industry code for OHS94 and OHS 95.

**jobocode:** One digit occupation code for both employees and self-employed. Valid range 1-99.

**occupation:** 4 digit occupation code. Not given in OHS94 and OHS 95. But see occupation1 below

**occupation1:** 3 digit occupation code. Only asked in OHS 94 and OHS 95.

**jobunion:** Whether an employed individual belongs to a trade union. OHS, LFS and QLFS from 2010:3 onwards only. Valid range: 1-9. 1=union member, 2=not a union member, 3=don't know, 9= unspecified, 0 = not applicable. Not asked for self-employed or those not employed.

**publicemp:** A dummy variable for whether the individual is employed in the public sector. Valid range: 0-1. Only asked in the PSLSD, LFSs and QLFSs. (although an imperfect public sector dummy can be created from the industry codes in the OHSs).

**business1:** The type of business an individual works for. Only asked in LFS 00:1-01:1. Valid range: 1-9.

**business2:** The type of business an individual works for, different categories to business1 and business3. Only asked in LFS 01:2-07:2. Valid range 1-99.

**business3:** The type of business an individual works for, different categories to business1 and business2. Only asked in QLFS. Valid range 1-8.

**hrslstk:** The number of hours worked in the last week (not hours usually worked). Valid range 0-190 (impossible since 168 should be the maximum but this is as in Stats SA metadata). The very large hours worked come mainly from OHSs, particularly OHS 1996. In the case of the QLFSs this variable was created by DataFirst for PALMS from several variables.

**jobcontract2:** The type of contract of the employee. Valid range 1-3. QLFS only.

**dweltype:** The dwelling the household lives in. Valid range: 1-7. Asked in PSLSD, all OHSs, as well as LFS 00:2, 01:2, 02:2, 03:2, 04:1, 04:2, 05:1. Other category (=7) includes several residual categories as the question changed over time, especially in the earlier waves.

**watersource:** The main source of water for the individual. Valid range: 1-12. Asked in PSLSD, all OHSs, as well as LFS 00:2, 01:2, 02:2, 03:2, 04:2. The question varied over time so there was some rationalisation of categories.

**toiletmaintype:** The type of toilet the household uses. Valid range: 1-13. Data from PSLSD, OHS 1995, 1997-1999, LFS 01:2, 02:2, 03:2, 04:2 used in creating this variable.

**hhpension:** A member of the household receives the state old age pension. Valid range 1-9. 1=yes, 2=no, 9=unspecified. Only asked for LFS 01:2, 02:2, 03:2, 04:2.

**hhdisablegrant:** A member of the household receives the state disability grant. Valid range 1-9. 1=yes, 2=no,

9=unspecified. Only asked for LFS 01:2, 02:2, 03:2, 04:2.

**hhchildsuppgrant:** A member of the household receives the state disability grant. Valid range 1-9. 1=yes, 2=no, 9=unspecified. Only asked for LFS 01:2, 02:2, 03:2, 04:2.

**hhcaredependgrant:** A member of the household receives the state care dependency grant. Valid range 1-9. 1=yes, 2=no, 9=unspecified. Only asked for LFS 01:2, 02:2, 03:2, 04:2.

**hhfostercaregrant:** A member of the household receives the state foster care grant. Valid range 1-9. 1=yes, 2=no, 9=unspecified. Only asked for LFS 01:2, 02:2, 03:2, 04:2.

**incpension:** Individual receives state old age pension. Valid range: 1-9. 1=yes, 2=no, 3=don't know, 9=unspecified. Only asked in PSLSD, OHS 1997-1999 and LFS 2000:2.

**incdisabgrnt:** Individual receives state disability grant. Valid range: 1-9. 1=yes, 2=no, 3=don't know, 9=unspecified. Only asked in PSLSD, OHS 1997-1999 and LFS 2000:2.

**inctstmaint:** Individual receives state maintenance/child support grant. Valid range: 1-9. 1=yes, 2=no, 3=don't know, 9=unspecified. Only asked in OHS 1997-1999 and LFS 2000:2.

**incdeprgrnt:** Individual receives state care dependency grant. Valid range: 1-9. 1=yes, 2=no, 3=don't know, 9=unspecified. Only asked in OHS 1997-1999 and LFS 2000:2.

**incfostcrgnt:** Individual receives state foster care grant. Valid range: 1-9. 1=yes, 2=no, 3=don't know, 9=unspecified. Only asked in OHS 1997-1999 and LFS 2000:2.

**registax:** Business is registered for income tax? Valid range: 0-9. 1=yes, 2=no, 3=don't know, 9=unspecified, 0=not applicable. Only asked in QLFS.

**publicemp2:** Public employment dummy that includes an approximation for public employment in the OHSs, where no direct question was asked. Probably excludes some public sector employees and probably includes some non-public sector employees. See Kerr and Wittenberg (2017) for further details.

**earnings:** Monthly earnings variable generated from the earnings amount data (not bracket information) across all waves where earnings amounts were asked and data have been released (all waves except OHS 1996 and QLFS waves in 2008, 2009 and 2017-2018). To be used in conjunction with the bracketweight variable to obtain an earnings series that takes account of bracket responses. Valid range 0-86450000 (but see high salary and outlier2 variables below)

**realearnings:** Monthly REAL earnings variable generated from the earnings amount data (not bracket information) across all waves where earnings amounts were asked and data have been released (all waves except OHS 1996 and QLFS waves in 2008, 2009 and 2012). This is the earnings variable deflated to 2015 Rands using the CPI. To be used in conjunction with the bracketweight variable to obtain an earnings series that takes account of bracket responses. Valid range 0- 46699440 (but see outlier variable below).

**outlier:** A flag if the studentised residual from an OLS log earnings regression (with independent variables gender, wave, popgroup, wave, yrseduc, age, age squared, jobocccode and interactions between gender and wave and pop group and wave) is more than 5.

**employerAll:** This variable expresses whether earnings were calculated from an individual's wage employment or self-employment. In the PSLSD individuals could report several jobs. In the OHSs individuals could report 2 jobs

and 2 earnings amounts. Only one of these has been used for the calculation of earnings, to make the earnings series comparable with the LFSs and QLFSs, where only 1 job and 1 earnings could be reported. Valid range 0-1. 0= wage employment, 1= self-employment.

**bracketweight1:** This is just the product of the `ceweight1` variable and the `pr_rand` variable and should be used to reweight the `realearnings` and `earnings` variables to produce a consistent earnings series that takes account of bracket responses. It is now based on the Stats SA mid-year population estimates, rather than the ASSA 2008 model. See Wittenberg (2008b) for an explanation of the reweighting method to take account of bracket responses.

## C Income Variable Description

The variables in the separate data set `palmsv3.3incomes` correspond to the income variables released in versions of PALMS prior to version 2. The data can be merged in with the main PALMSv3.3 data for those who wish to replicate or investigate the work done to create the `realearnings` variable.

### C.1 Variables

**uqnr:** Household identifier, this is not unique across waves. It is the same as the original household id variable from Stats SA/SALDRU for all waves, except for 1996, where no household id was supplied and one was created from the magisterial district, enumeration area and visiting point variables. It has been converted to a string variable.

In previous versions of the data the `uqnr` variable was not the same as the original `hhid` variable as a zero was taken out. A second variable, `uqnr_orig` was added in version 1.0.8 to make merging in data easier. In the latest version there is only one variable, `uqnr`, and this is the original household id variable.

**personnr:** The number of the person in the household. Valid range: 0-85.

**wave:** Wave, with PSLSD 1993=0, OHS 1994=1 and LFS 2015=54. The last OHS was OHS 1999 and this was wave 6. The first LFS was a pilot survey conducted in February 2000, which is included as wave 7 in this data set. The first wave of the QLFS is March 2008, wave 23. Valid range 1-54.

**earnperiod:** The period for which the individual's income is reported. Only for OHSs and LFSs (See `jobsalperiod2` selfempayperiod2 for QLFSs). Valid range 1-4. 1= per day, 2=per week, 3=per month, 4 = per year. Not all options were asked in each survey.

**earnperiodaddjob:** Earnings period for those with more than one job. Only defined for OHSs. Valid range: 1-4. 1= per day, 2=per week, 3= per month, 4=per year. Not all options were asked in each survey.

**jobsalary:** Actual gross earnings in either wage or self-employment main job, asked in LFSs only, in current prices. valid range: 0,5281000. Not converted to a fixed period, so can be daily, monthly amount etc. Extreme values set to missing only where these exceeded the range specified in the Stats SA metadata, effects one individual in wave 21.

**jobsalcat:** Bracket response for gross earnings in either wage or self-employment main job, asked in LFSs, in current prices and is an annual amount. This is for those who refused or did not know the actual earnings amount. Valid range: 1-99.

**earncatmin:** This is the minimum of the earnings bracket reported in the LFSs, using an annual figure and expressed in current prices. Valid range: 1-360001.

**earncatmax:** This is the maximum of the earnings bracket reported in the LFSs, using an annual figure and expressed in current prices. Valid range: 2400-360000.



**wageempincome:** Gross wage employment income from OHS 1995-1999 (but excluding OHS 96 where only a categorical question was asked), not adjusted for earnings period and expressed in current prices. Valid range: 1-920920. Extreme values set to missing only where these exceeded the range specified in the Stats SA metadata, effects one individual in wave 4. Includes imputed values.

**wageempincome2:** Total NET SALARY PER MONTH derived from q3.13 OHS 1994 as released by Stats SA. Expressed in current prices. Includes imputed values. See Wittenberg (2008a) for a criticism of this imputing. Valid range: 0-52500.

**empsalcat1:** Bracket response for gross earnings in wage employment for OHS 1996-1999, in current prices, using an annual figure. This is missing for those who refused or did not know the actual earnings amount. This is the ONLY wage employment income variable in OHS 96, as only a categorical question was asked. Valid range: 0-8888.

**empsalcat2:** Bracket response for gross earnings in wage employment for OHS 1995, in current prices. NOT an annual amount, but per earnings period. This is for those who refused or did not know the actual earnings amount. Valid range: 0- 30.

**empsalcat3:** Bracket response for NET MONTHLY earnings in wage employment for OHS 1994, in current prices. Includes imputed values. See Wittenberg (2008) Valid range: 0- 12.

**selfempincome1:** Self-employment income from OHSs (except OHS 96 where only a categorical question was asked), expressed in current prices, not adjusted for earnings period. Valid range: 1- 2000000. The questions on self-employment income varied slightly across the waves, see document on “Using the labour income data in PALMS.”

**selfempincome2:** Self-employment income q 3.19, calc PER MONTH by SSA, OHS 94. Includes imputes. Valid range: 0- 489742.

**imputed:** Dummy, =1 if the original net income figure is imputed in OHS 94.

**salary\_impute:** This is an improved version of the imputed employee income from OHS 94, as described in Wittenberg (2008). Valid range: 0- 52500.

**impute\_gross:** Dummy, =1 if the original self-employment income figure is imputed in OHS 94.

**emp\_impute:** This is an improved version of the imputed self-employment income from OHS 94, as described in Wittenberg (2008). Valid range: 0- 303333.

**selfempinccat1:** Bracket response for gross earnings in self-employment for OHS 1999, in current prices, using an annual figure. This is for those who refused or did not know the actual earnings amount. Valid range: 1- 8888.

**selfempinccat2:** Bracket response for gross earnings in self-employment for OHS 1996-1998, in current prices, using an annual figure. This is for those who refused or did not know the actual earnings amount. This is the ONLY self-employment income variable in OHS 96, as only a categorical question was asked. Valid range: 0-16.

**selfempinccat3:** Bracket response for gross earnings in self-employment for OHS 1995, in current prices, NOT an annual amount, but per earnings period. This is for those who refused or did not know the actual earnings amount. Valid range: 0-30.

**selfempinccat4:** Bracket response for gross earnings in self-employment for OHS 1994, in current prices, NOT an annual amount, but per earnings period. This is for those who refused or did not know the actual earnings

amount. Valid range: 0-14.

Expenses in self-employment, see the labour incomes document for more detail.

**selfempexpgoods** Expenses on goods in self-employment, asked in OHS 96-98 only. Valid range: 0-500000.

**selfempexprenum**: Expenses on staff remuneration in self-employment, OHS 96-98 only. Valid range 0-450000.

**selfempexpoth**: Other Expenses in self-employment, OHS 96-98 only. Valid range 0-500000.

**selfempexpall**: Total Expenses in self-employment, OHS 94 and 95 only. Valid range 0-90000 with some missing codes (99998 and 99999). The expenses in OHS 94 are imputed if a refused answer was given. See Wittenberg (2008).

**pslsd\_earnings**: This is the earnings variable created for PSLSD 1993. That survey asked separate questions about regular and casual employment as well as self-employment. The value present in pslsd\_earnings is regular earnings, unless a respondent was not a full-time regular worker, in which case they were assigned the highest of their three incomes. Less than 2% of the employed reported multiple sources of labour income. See do files for more information on how this variable was created.