# A Guide to version 4 of the Post-Apartheid Labour Market Series (PALMS)

Andrew Kerr[*] and Martin Wittenberg[†]

November 20, 2025

## 1 Introduction

This document describes the PALMS version 4 stacked cross sectional dataset created by DataFirst at the University of Cape Town (Kerr et al. n.d.b). The dataset consists of microdata from 92 household surveys conducted by Statistics South Africa between 1994 and 2025, 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 2008-2025. 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 rotating panel element to parts of the LFS and 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 120 variables in the main dataset and nearly 7.5 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 for some waves where these were asked. Not all variables from all surveys are included. The surveys are regarded as one of the more reliable sources of labour market data, including earnings. 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 PALMSv2 and later versions was the creation of a consistent earnings variable over all waves that collected earnings. Further information on earnings can be found in section 5 below, including why earnings from the QLFSs has been excluded from PALMSv4.

PALMS is available through the DataFirst data portal: https://doi.org/10.25828/gtr1-8r20

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.

[*] School of Economics and DataFirst, University of Cape Town. andrew.kerr@uct.ac.za.

[†]Martin Wittenberg was the initiator of the PALMS project while he was the director of DataFirst at the University of Cape Town. Martin passed away in July 2024, before version 4 of PALMS was completed.

[‡]Version 4 of PALMS was funded by UNU-WIDER through the SA-TIED project. Previous versions were made possible with funding from the UCT Vice Chancellor's strategic fund (the VC at the time was Max Price), REDI, funded by the South African National Treasury, and the International Labour Organisation.

[‡]In this version of the PALMS guide whenever the first person is used this reflects that this decision or opinion is Andrew Kerr's alone.

[‡]A large number of people have provided very useful feedback on prior versions of PALMS, and we thank in particular Nicola Branson, Bruce McDougall, Patrizio Piraino, Owen Crankshaw, Neil Balchin, Friedrich Kreuser, Karmen Naidoo, Rob Davies and Carlos Gradín. 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. We thank Amy Thornton who updated the cross entropy weights for PALMSv4. We thank Alex Montgomery for providing assistance with incorporating the Labour Market Dynamics earnings data into previous versions of PALMS (even though this data is not included in PALMSv4-see below). 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. I thank Raynold Runganga for alerting me to the availability of job start dates in the OHSs and for the code to incorporate these into PALMSv4. I thank Peris Wachira for her help editing this version of the PALMS guide.

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 (2025), Post-Apartheid Labour Market Series [dataset]. Version 4. Cape Town: DataFirst [producer and distributor], 2025.

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 earnings variable was created and the imputation of the QLFS earnings undertaken by Stats SA. We finally describe how the multiply imputed earnings data was created. In the Appendix we explain how users 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 4

## 2.1 New QLFS waves

For this release of PALMS the new data is for the QLFS 2019 quarters 3 and 4, QLFS 2020-2024 and QLFS 2025 quarters 1 and 2.

## 2.2 Removal of Earnings data from 2010 onwards

The main and very significant change compared to PALMSv3.3 and earlier is that I have removed all earnings data from PALMS from 2010 onwards. This decision was not taken lightly. It reflects concerns with the imputed earnings microdata provide by Statistics South Africa in the public releases of earnings up until 2019 through the "Labour Market Dynamics" publication. In a number of papers Martin and I showed that this data is of very low quality. These included Kerr and Wittenberg (2017), Kerr and Wittenberg (2019), Kerr and Wittenberg (2021) and Kerr (2025). This last paper shows most clearly the problems with the public QLFS earnings data up until 2019. My opinion based on the work I have done is that the QLFS earnings data up until 2019 that is currently publicly available is so bad that it should not be used by researchers at all and hence I have taken the decision to drop it from PALMS. The newer data for 2022 and 2023 that has been made publicly available by Stats SA is fully imputed. Whilst it looks better than the pre-2020 data it still does not contain any imputation flags, and thus is of limited value.

In Kerr (2025) I noted that the underlying non-public QLFS earnings data for 2011- 2012 and 2018-2020 that I accessed was reasonable and, as should be expected, produced fairly similar earnings distributions to the General Household Survey, also conducted by Statistics South Africa. It is thus crucial that Stats SA releases earnings data that allows a researcher to clearly see who responded in brackets and who refused or was recorded as a "don't know". Unfortunately this appears to be unlikely- I have requested a public release of this type of data multiple times and this has been rejected by Stats SA officials. The lack of earnings makes PALMS less valuable, but given the lack of unimputed earnings data or even the improved earnings data underlying the 2024 Statistics South Africa publication that was also imputed (see below) for before 2022, it seems wrong to release earnings data in PALMS that I believe to be of very low quality.

Stats SA released a report on earnings in 2024 (Statistics South Africa 2024), which showed some percentiles from the QLFS earnings distribution between 2017 and 2022. The data underlying this report was a new version of the QLFS earnings that Stats SA had applied new imputation methods to. The report states: "In 2023, Stats SA undertook an assessment of the earnings data collected through the QLFS from 2009 in order to standardise the imputation methodology for all the years. The revision of earnings is a fundamental process driven by the pursuit for improved quality, accuracy and relevance."

The 2024 report shows earnings levels that much more closely match both the GHS and the non-public QLFS data used in Kerr (2025) which did not contain the Stats SA imputations, at the 10th, 25th, median, 75th and 90th percentiles in 2017, and the trend in the median between 2017 and 2019. Unfortunately, as of May 2025, the microdata for 2017- 2021 on which this new publication is based has not been released through the new Stats SA data portal, isiBaloweb.[1]

In addition, and even more unfortunately, the report also states "it is in accordance with international best practice, to avoid inconsistencies due to key entry errors, programming, interviewer mistakes or errors due to respondent

[1] On 28 May 2025 the situation for each LMD is was follows: 2017-2019- old data. 2020- *unimputed* earnings data available (surprisingly!). 2021- NO earnings data available at all. 2022-2023- earnings data is fully imputed, likely using the new methods briefly mentioned but not explained in Statistics South Africa (2024).

reporting, that statistical data editing and imputation should be undertaken to ensure that the information provided is complete and consistent. It is against this background that Stats SA cannot make available unedited or unimputed data to the users from Stats SA repositories." This position seems to go against Statistics South Africa's own practice in the microdata releases from the October Household Surveys (1995-1999), the Labour Force Surveys (2001-2007), the Surveys of Employers and Self-Employed (2001, 2005, 2009, 2013), the National Travel Surveys (2003, 2013, 2020) and the General Household Surveys (2002-2023), since all these surveys contain earnings data that allow one to see who responded in brackets or responded with a don't know or refusal, including very recent publications. In addition, as of 28 May 2025 the 2020 LMD available on isiBaloweb contains unimputed data! My conclusion is that recent releases of Stats SA earnings microdata without imputations in other surveys, including the 2020 LMD, contradicts the stated position of Statistics South Africa in the QLFS earnings report with regards to the QLFS earnings data. The obvious solution is to release unimputed versions of the QLFS earnings data for all years in which data was collected.

## 2.3 Fixed OHS 1995 occupation error

The 3 digit occupation code in 1995 was missing for some of the self-employed in previous versions of PALMS. This was due to a coding error and has been fixed.

## 2.4 Updated QLFS 2008-2012 Q1 weights

Statistics South Africa released updated weights for QLFS 2008 Q1-2013 Q4 some time after the original release of these data sets. PALMS v3.3 and earlier releases did not incorporate the updated weights for QLFS Q1 2008- Q1 2012. We have now fixed this by using the most recent releases and weights. One implication of this error can be seen in Kerr (2025), appendix figure A1, where there is a large jump up in the number of formal sector employees at Q2 2012. This was caused because I used the older weights that were in PALMS. What the weighting adjustment does is shift the large jump in total population and total employed population back to 2007 LFS 2 abd 2008 Q1 when the change to the QLFS occurred. The LFS weights were not revised when the QLFS weights were. This highlights the need for the use of a method like the cross entropy weighting adjustments discussed in Branson and Wittenberg (2014).

## 2.5 Fixed 1993 PSLSD cross entropy weights

In PALMS v3.3 the cross entropy weights for the 1993 PSLSD were incorrect as a result of a coding error. We found this out in 2023 and replaced version 3.3 of PALMS with version 3.3.1 which had the cross entropy weights for 1993 set to missing, since at that time we did not have capacity to fix the weights. In PALMSv4 we have incorporated corrected cross entropy weights, which were created by Amy Thornton.

## 2.6 Fixed missing job start year in LFS 2002:1 and OHS 1995-1999

The old version of LFS 2002:1 worker file that was originally used in creating PALMS was missing the year each employed person's job started in. We have now used an updated version of this data from DataFirst to add in this missing variable in this wave. In addition, PALMSv4 incorporates the job start dates variables (month and year) for the OHS 1995-1999, which were not included in the previous versions of PALMS.

## 2.7 New variables

A number of new variables have been included in PALMSv4. These include whether the job is permanent or of some limited duration (jobpermanent in LFS, jobpermanent2 in QLFS), how an unemployed person is supporting themselves (variables have prefix "sup", available in LFS and QLFS, with two extra options in QLFS), the main reason someone stopped working (variable is reasonstopwork, QLFS only), the period an individual has been trying to find work or start a business (searchlength in LFS, searchlengthqlfs in QLFS) and the number of hours worked on each day of the week for those with one job (hoursmonday, hourstuesday etc. QLFS only). There are very few people with more than one job in South Africa (around 0.5% of employed) and some of the total hours worked on some days for those with more than one job are incorrect so we have not included hours worked on each day for the group of people with more than one job.

We also included an enrollment variable in the QLFS (enrolledqlfs) and the type of educational institution the enrolled individuals attended, both in LFS (educ_insti_lfs) and in QLFS from 2012:Q3 onwards when it was asked (educ_insti_qlfs). We have not included field of study, which has been asked from QLFS 2012 Q3 onwards, due to odd changes in Q2 2022 that we have not yet understood.

The LFS and QLFS asked respondents who were not working whether they had ever worked and how long it was since they last worked. We included these variables as everworked and whenlastjob (LFS) and whenlastjob_qlfs (QLFS). Individuals that had worked were also asked what they did in their previous occupation and what their previous industry was. We have now included these variables in PALMS at the one and four digit level. The variables are occup_prev, indusprev_lfs, occupprev2, indprev_qlfs (all 3/4 digit codes), previndcode and prevocccode (1 digit codes).

# 3 The Post-Apartheid Socio-Economic Series

At the same time as the release of PALMSv4, the Post-Apartheid Socio-Economic Series (PASES) is being released (Kerr et al. n.d.a). PASES uses the same idea as PALMS and applies them to the series of October Household surveys (1995-1999) and General Household Surveys (GHSs) (2002-2024), with a focus on the labour market. The GHSs are an alternative source of earnings data to the QLFS earnings data that I have noted is excluded from PALMSv4. Kerr (2025) documents that the GHS earnings appears to be of reasonable quality, and certainly much better than the publicly available imputed QLFS earnings data from 2011-2012 and 2018-2019.

# 4 Cross entropy weights

The cross entropy weights released with PALMS use a demographic model of the population that is consistent over time. This is instead of using the most recent population estimate at the time of each survey, as is done by Stats SA, which can lead to jumps up and down in the population estimates over time. For more information on the cross entropy weighting ideas, see Branson and Wittenberg (2014). As in PALMSv3.3, the cross entropy weights included in PALMSv4 are based on the latest Stats SA mid-year population estimates, which are for 2024 in PALMSv4. We did not use the 2025 MYPE since these are quite different, due to including results from the 2022 census, which may not be reliable. The MYPE only go back to 2002, so we have extended them back to 1993 using a simple exponential growth model, using the Actuarial Society of South Africa (ASSA) 2008 demographic model growth rates over the period 1993-2001. There are very minor differences for 1993-2001 between these population estimates and the ones directly from the ASSA 2008 model over the same period.

# 5 Earnings data in PALMS v2+

As noted above, earnings data are completely excluded from the QLFS. This section thus describes the processing of earnings data between 1993 and 2007, which is the same as in PALMSv3.3.

## 5.1 Comparison with pre-PALMS v2 earnings

Version 2.0 (and up) of PALMS uses a very different approach to earnings 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 "PALMSv4incomes". 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 PALMSv4 data . This can be merged into the main PALMSv4 data by the (Stata) command:

```
merge 1:1 uqnr personnr wave using PALMSv4incomes
```

The do files PALMSwages_create, PALMSwages_create2 and PALMSwages_create3 (released with this data) used to create the final earnings variables included in PALMSv4 can then be run relatively easily to replicate these variables.

### 5.2 Reweighting for Earnings Bracket Responses

PALMS v4 contains a inflation adjusted labour earnings variable called realearnings (the base period is December 2023). This is the recommended variable to undertake any analysis of earnings in PALMS, but note that these only exist for 1993-2007. 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 may 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 ceweight1 for that particular individual. The population model used to create the cross entropy weight is discussed above in section 4. In 1996 there were no actual rand amounts collected. There were no incomes collected in the first 6 waves of the QLFS and earnings data has never 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 earnings 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 v4, see section 6 below.

### 5.3 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 realearnings)
```

## 6 Multiply Imputed earnings 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 v4, and how this data was created. The imputed earnings data has been released as palmsv4miincomes.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 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 PALMSv4 data open:

```
merge 1:1 wave uqnr personnr using palmsv4miincomes.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 palmsv4 data will not be matched with the new multiply imputed data.

The realearnings variable is the one that was imputed. realearnings is set to hard missing (==.a) for those who we did not want to impute incomes (those reporting zero incomes). realearnings 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 if you have comments or suggestions.

# References


**Branson, Nicola and Martin Wittenberg**, "Reweighting South African National Household Survey Data to Create a Consistent Series Over Time: A Cross-Entropy Estimation Approach," *South African Journal of Economics*, 03 2014, *82* (1), 19–38.

**Kerr, A., A. Thornton, and M Wittenberg**, "The Post-Apartheid Socio-Economic Series v1 (PASES)."

**_ , D. Lam, and M Wittenberg**, "The Post-Apartheid Labour Market Series v4 (PALMS)."

**Kerr, Andrew**, "Earnings and Earnings Inequality in South Africa: Evidence from Household Survey and Administrative Tax Microdata from 1993 to 2020," *Review of Income and Wealth*, 2025, *71* (1).

**_ 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.

**_ and _** , "Union wage premia and wage inequality in South Africa," *Economic Modelling*, 2021, *97*, 255–271.

**Statistics South Africa**, "Monthly earnings in South Africa 2017 - 2022," Technical Report Report 02-11-20 (2017-2022), Statistics South Africa 2024.

**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.

# 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/relabeled 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 (appendlfstoohs.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 Amy Thornton, based on code from 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 ohsmerge'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 createpalmsconsistentqlfs.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_v4, PALMSwages_create2_v4, PALMSwages_create3_v4, PALMSfinal_create_v4

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 palmsv4incomes. This can then be merged into palmsv4.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 PALMSv4miincomes. These are PALMSwages_create4_v4 and PALMSwages_create6_v4.

# 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 unqr 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.

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

**year**: Year of the survey. Valid range: 1993-2025.

**wave**: Wave, with PSLSD 1993=0, OHS 1994=1 and QLFS 2025:2=92. 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 0-92.

**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, 4=Tribal areas, 5=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 (new master sample) until 2017 Q2. This variable has been used for stratification by Stats SA.

**urbrur4**: Type of area, 1=Urban, 2=traditional, 3=Farms. Covers QLFS 2017:3 onwards. This variable has been 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 2024. 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 ceweight and ceweight2 variables from older versions of PALMS to prevent confusion. Researchers wanting to replicate their results using ceweight2 should merge in the older version of PALMS into PALMSv4. For earnings analysis bracketweight should be used. See below. Valid range: 0-34331.152.

**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. Valid range: 0-17.

**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.

**enrollment3**: Enrolment information that covers the PSLSD, the OHSs and LFSs but not QLFSs. There is no direct enrolment question asked in the early QLFSs (but see enrolledqlfs below for less detailed variable that covers the QLFS 2012Q3 onwards) Valid range 1-9. 1= full-time, 2= part-time, 3=not enrolled, 9=unspecified.

**enrolledqlfs**: Enrollment in educational institution. Covers QLFS 2012:3 onwards only. 1=Yes, 2=No.

**edu_insti_lfs**: Type of educational institution the enrolled individuals attended. Valid range: 1-9. Covers LFSs only.

**edu_insti_qlfs**: Type of educational institution the enrolled individuals attended. Valid range: 0-9. Covers QLFS 2012:3 onwards only.

**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. Only asked for LFSs (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 1-99.

**jobstartyear**: The year the individual started working in their current job. Valid range 1919-2024. Asked in OHSs 1995-1999, LFSs and QLFSs.

**jobstartmonth**: The month the individual started working in their current job. Valid range 0-12. Asked in OHSs 1995-1999, LFSs and QLFSs.

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

**jobpermanent**: Whether the job is permanent or of some limited duration. Only in LFS. 1= Permanent, 2=A fixed contract, 3=Temporary, 4=Causal, 5=Seasonal, 6=Don't know, 8=Not applicable, 9=Unspecified

**jobpermanent2**: Whether the job is permanent or of some limited duration. Only in QLFS. 0= Not applicable, 1= Limited Duration, 2=Permanent nature,3= Unspecified (mostly self-employed).

**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=Yes, 2=No, 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 it why it was dropped in 2005 onwards).

**informal_emp_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, 8=Others. This variable replaces formalreg2 in PALMSv3.2 and later versions- we wanted to make it clearer that this is NOT comparable to formalreg1 or formalreg0.

**informal_sector_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 sector of the firm. 1=Formal, 2=Informal, 4=Private Households.

**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.

**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.

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

**industry**: 3 digit industry code. Defined for OHS 99 and LFSs. Valid range: 1-999 (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. Valid range: 1-999 (see release documentation, Stats SA seems to have gone back to OHS codes in QLFS)

**indust2digit**: 2 digit industry code for OHS94 and OHS 95. Valid range 0-50. 0=Not applicable.

**jobocccode**: One digit occupation code (main job) for both employees and self-employed. Valid range 1-99.

**occupation**: 4 digit occupation code. Only for OHS 96- LFS 2002:1. Valid range: 810-9999. But see occupation1 below

**occupation1**: 3 digit occupation code. Only asked in OHS 94 and OHS 95. Valid range: 0-934

**occupation2**: 4 digit occupation code. Only asked in LFS 2002:2 onwards. Valid range: 850-9999.

**jobunion**: Whether an employed individual belongs to a trade union. OHS, LFS and QLFS from 2010:3 onwards only. Valid range: 0-9. 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. 0=No, 1=Yes. Only asked in the PSLSD, LFSs and QLFSs. (See publicemp2 variable below which includes an imperfect public sector dummy created from the industry codes in the OHSs).

**publicemp2**: Public employment dummy that includes an approximation for public employment in the OHSs, where no direct question was asked. In the OHSs it probably excludes some public sector employees and probably includes some non-public sector employees. Valid range: 0-1. 0=No, 1=Yes. See Kerr and Wittenberg (2017) for further details.

**businesstype1**: The type of business an individual works for. Only asked in LFS 00:1-01:1. Valid range: 1-9.

**businesstype2**: The type of business an individual works for, different categories to businesstype1 and businesstype3. Only asked in LFS 01:2-07:2. Valid range 1-99.

**businesstype3**: The type of business an individual works for, different categories to businesstype1 and businesstype2. Only asked in QLFS. Valid range 1-9.

**hrslstwk**: 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. For tthe QLFSs this variable was created for PALMS from several variables.

**jobcontract2**: The type of contract of the employee. Valid range 1-8. QLFS only.

**hoursmonday**: Hours worked by an individual on Monday. Valid range: 0-888. This question was asked in QLFS only and it is specific for individuals who had one job only (see similar variables hours#day below).

**hourstuesday**: Hours worked by an individual on Tuesday. Valid range: 0-888. Asked in QLFSs only.

**hourswednesday**: Hours worked by an individual on Wednesday. Valid range: 0-888. Asked in QLFSs only.

**hoursthursday**: Hours worked by an individual on Thursday. Valid range: 0-888. Asked in QLFSs only.

**hoursfriday**: Hours worked by an individual on Friday. Valid range: 0-888. Asked in QLFSs only.

**hourssaturday**: Hours worked by an individual on Saturday. Valid range: 0-888. Asked in QLFSs only.

**hourssunday**: Hours worked by an individual on Sunday. Valid range: 0-888. Asked in QLFSs only.

**reasonstopwork**: Main reason and individual stopped working. Valid range: 0-99. Covers QLFSs only.

**searchlength**: Period an individual has been trying to find work or start a business. Valid range: 1-9. Asked in LFSs only.

**searchlengthqlfs**: Period an individual has been searching for work. Valid range: 0-99. Asked in QLFSs only.

**everworked**: Unemployed individual ever worked. Valid range: 0-9. Covers LFSs and QLFSs only.

**whenlastjob_qlfs**: Time since last job. Valid range: 1-99. Covers QLFSs only.

**whenlastjob**: Time since last job. Valid range: 1-9. Covers LFSs only.

**occupation_prev**: Occupation of previous job. 4 digit code. Valid range: 850-9999. Covers LFS 2000:1 to 2002:1 only.

**occupation_prev2**: Occupation of previous job. 4 digit code. Valid range: 0-9999. Covers LFS 2002:2 onwards.

**indusprev_lfs**: Industry of previous job. 3 digit code. Valid range: 10-999. Covers LFSs only.

**indprev_qlfs**: Industry of previous job. 3 digit code. Valid range: 0-999. Covers QLFSs only.

**previndcode**: Industry of previous job. 1 digit code. Valid range: 0-11. Covers LFSs and QLFSs only.

**prevocccode**: Occupation of previous job. 1 digit code. Valid range: 0-11. Covers LFSs and QLFSs only.

**dwelltype**: 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 household. 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 child support 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.

**incdepgrnt**: 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.

**incfostcrgrnt**: 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.

**supoddjobs**: Unemployed individual supports themselves by doing odd jobs. Valid range: 0-9. 1=Yes, 2=No, 8= Not applicaple, 9=Unspecified. Asked in LFSs and QLFSs only.

**suphhpersons**: Unemployed individual is supported by person in the household. Valid range: 0-9. Asked in LFSs and QLFSs only.

**supothpersons**: Unemployed individual is supported by persons not in the household. Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**supcharity**: Unemployed individual is supported by charity, church or welfare. Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**supUIF**: Unemployed individual is supported by the Unemployment Insurance Fund (UIF). Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**supsavings**: Unemployed individual is supported by savings or money previously earned. Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**suppension**: Unemployed individual is supported by old age or disability pension. Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**supother**: Unemployed individual is supported by other sources of support. e.g bursary. Valid range 0-9. 1 = Yes, 2 = No, 8 = Not applicable, 9 = Unspecified. Asked in LFSs and QLFSs only.

**supothgrant**: Unemployed individual is supported by other grant. Valid range 0-3. 1 = Yes, 2 = No, 0 = Not applicable. Asked in QLFSs only.

**supcsg**: Unemployed individual is supported by Child Support Grant. Valid range 0-3. 1 = Yes, 2 = No, 0 = Not applicable. Asked in QLFSs only.

**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. But now we exclude the QLFS earnings data. To be used in conjunction with the bracketweight variable to obtain an earnings series that takes account of bracket responses. Valid range: 0-5281000.

**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. But now we exclude the QLFS earnings data. This is the earnings variable deflated to December 2023 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- 18278044 (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 at least 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.

**bracketweight**: This is just the product of the ceweight 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 based on the 2024 Stats SA mid-year population estimates, rather than the ASSA 2008 model as in older versions of PALMS. See Wittenberg (2008b) for an explanation of the reweighting method to take account of bracket responses.

# C Earnings Variable Description

The variables in the separate data set palmsv4incomes correspond to the income variables released in versions of PALMS prior to version 2. The data can be merged in with the main PALMSv4 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 unqr 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. 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. Valid range 1-22.

**earnperiod**: The period for which the individual's income is reported. Only for OHSs and LFSs (See jobsalperiod2 selfemppayperiod2 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 MONTLY 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.

**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 earnings section for more detail.

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

**selfempexprenum**: Expenses on staff renumeration 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 earnings. See do files for more information on how this variable was created.