Statistics/Data analysis

```
/*
NOTES;
This do-file creates the individual level dataset

*/
gl yearlist 2010 2011 2012 2013 2014 2015 2016 2017 2018

cap program drop cr1
program cr1
 syntax , irp5_ind(string) citirp5_v4(string) saveaddress_data(string)

 cap log close
 log using "$log_save\create_ind_panel", replace

foreach year in $yearlist{
    use "Z:\Master Data\IRP5\Job level\v4\IRP5_`year'_cleaned", clear
        keep amt3601 amt3805  busdistmuni_geo buslocmuni_geo busmainplc_geo busprov_ge
> o ///
 province_geo districtmunicip_geo localmunicip_geo mainplace_geo dateofbirth ///
 gender idno mainincomesourcecode natureofperson payereferenceno ///
 periodemployedfrom periodemployedto taxrefno taxyear totalperiodsinyearofassessment /
> //
 totalperiodsworked passportno certificateno revisionnumber kerr_income kerr_emp kerr_
> emp_inc

 *keep only natural persons
 tab natureofperson
 keep if natureofperson =="A"

/*****************************************************************
*create total months worked variable using employedfrom and employedto variables
use substring to create these
six variables(day from, month from, year from, day to, month to, year to)
*****************************************************************/

gen day_from = substr(periodemployedfrom,9,2)
gen month_from = substr(periodemployedfrom,6,2)
gen year_from = substr(periodemployedfrom,1,4)

gen day_to = substr(periodemployedto,9,2)
gen month_to = substr(periodemployedto,6,2)
gen year_to = substr(periodemployedto,1,4)

foreach var in day_from month_from year_from day_to month_to year_to{
destring `var', replace
}
gen year_diff = year_to - year_from
tab year_diff
keep if year_diff==1

cap drop months_worked
gen months_worked =.
replace months_worked = month_to - month_from if year_from==year_to
replace months_worked = (12-month_from) + (month_to) if year_from!=year_to


drop if months_worked >12
tab months_worked

*drop indivividuals who worked for less than a month
count if months_worked ==0
drop if months_worked==0

/*****************************************************************************
*For the earnings variable, we use the kerr income variable and amount 3601
*get the monthly wages
*****************************************************************************/
gen monthly_3601= amt3601/months_worked
gen monthly_kerr= kerr_income/months_worked if kerr_emp_inc==1
```

```
/*******************************************************************************
2013 minimum wage
hourly = 11.66
weekly = 525
monthly = 2274.82

We want to identify the proportion of individuals earning below the 2013 minimum wages
*******************************************************************************/
gen mw_worker_3601 = monthly_3601 <= 2274.82
tab mw_worker_3601
label var mw_worker_3601 "min wage worker using amount3601"

gen mw_worker_kerr = monthly_kerr <= 2274.82
label var mw_worker_kerr "min wage worker using kerr income"

/*******************************************************************************
Confirm that the data is at the individual level

count number of employees per firm
*******************************************************************************/
* create individual identifier: code from Marlies Piek: "Set_up_4_march"
gen id_new=idno
replace id_new=passportno if id_new==""
label var id_new "Unique identifer, from SA ID or passport no."
egen n_id=group(id_new)
* count how many IRP5 forms an individual had in a year
bysort n_id taxyear: egen number_certs=count(n_id)

cap drop n
bys taxyear id_new: gen n =_n
tab n
keep if n==1
drop n

gen unit=1
egen employees_n = sum(unit), by(taxrefno)
tab employees_n
label var employees_n "number of employees in a firm"

egen num_mw_3601= sum(mw_worker_3601), by(taxrefno)
label var num_mw_3601 "No. Min wage workers per firm using amt3601"

egen num_mw_kerr= sum(mw_worker_kerr), by(taxrefno)
label var num_mw_3601 "No. Min wage workers per firm using kerr income"

*average wage
egen tot_earnings_3601 = sum(amt3601) , by(taxrefno)
egen tot_earnings_kerr = sum(kerr_income) if kerr_emp_inc==1,  by(taxrefno)

gen avwage_3601 = tot_earnings_3601/employees_n
gen avwage_kerr  = tot_earnings_kerr/employees_n if kerr_emp_inc ==1


*median wage
egen medwage_360 = median(amt3601), by(taxrefno)
egen medwage_kerr= median(kerr_income) if kerr_emp_inc==1, by(taxrefno)

*proportion of minimum wage workers

gen propmin_3601 = mw_worker_3601/employees_n
label var propmin_3601 "Proportion of minimum wage waorker (a3601)"
gen propmin_kerr = mw_worker_kerr/employees_n
label var propmin_kerr "Proportion of minimum wage waorker (kerr)"

*****************************************************************
*Merge in the cit data
```

```
preserve
use if taxyear == `year' using "$citirp5_v4" , clear
tempfile cit
save `cit'
restore
merge m:1 taxrefno taxyear using `cit'

keep if _merge==3

*Keep agriculture and manufacturing sectors using CIT industry classification
keep if imp_mic_sic7_1d ==-1 | imp_mic_sic7_1d == -3

cap id_new_num
egen id_new_num = group(id_new)


save "`saveaddress_data'\merged_`year'", replace
}

use "`saveaddress_data'\merged_2018", clear
forval i = 2017(-1)2010{
append using "`saveaddress_data'\merged_`i'.dta"
}
cap drop empl_tag
gen empl_tag = 1
xtset id_new_num taxyear
tsfill , full

tab taxyear
save "`saveaddress_data'\merged_ind_panel", replace
cap log close

end
cr1, irp5_ind("$irp5_ind") citirp5_v4("$citirp5_v4") saveaddress_data("$saveaddress_da
> ta")
```