QLFS earnings data quality

Andrew Kerr and Martin Wittenberg

DataFirst, UCT

DataFirst data quality workshop, July 2017

Outline

- Changes in earnings imputation and sampling
- Illustration of strange changes over the QLFS from work on union wage gap, inequality and public sector work for REDI.
- Way forward?

LFS and QLFS earnings basics

- We want to know what people to earn to understand how inequality is changing, who is benefitting from economic growth, how South Africa is changing and many other types of questions.
- Labour force surveys ask about people’s earnings.
- Both LFS and QLFS ask about gross income amount in main job.
- Refusals are asked about their income brackets.
- Most papers using the Stats SA data are not very transparent about how they deal with a lot of the measurement issues and choices about how to deal with them that are faced by researchers.

Changes in data and sampling

- Differences between LFS and QLFS in imputations, how this has changed since 2012 Q3.
- DF also has 1 year of unimputed earnings data for 2011 from Stats SA which gives us a sense check on what causes some of the differences.
- Weird trends identified- some start around mid 2012 when earnings data changed to partially imputed- eg returns to education and white premium.
- Other trends look odd only in later periods eg U premium drops off in Q4 2014.

New 2013 master sample

- In the new 2013 Master Sample, which was used from Q1 2015 onwards in the QLFS, about 3300 PSUs used so the sample size has increased by 10%.
- This new sample also changed the composition of the QLFS by increasing the number of PSUs sampled in Gauteng by 60%, in KZN by 16% and in the Eastern Cape by 21% (3 largest provinces) and decreasing sample in others.
- Surprisingly, despite the extra 300 PSUs and 3000 dwelling units in the sample, the realised sample in the publicly released data decreased by about 2500 households.

Different types of earnings data over time

- LFS- not imputed- both complete refusals and brackets with no amounts in the data.
- QLFS until Q2 2012 is almost fully imputed- 99.1% of those employed have incomes.
- QLFS after Q2 2012 is partially imputed, with those who give bracket amounts given rand amts and no earnings for those with complete refusals.
- This means Stats SA has given us **3 quite different** types of earnings data in the 2000-2015 period.
- Refusal rates are higher in QLFS than they were in LFS (next slide)- but we only know this for 1 year.

LFS and QLFS earnings refusal rates

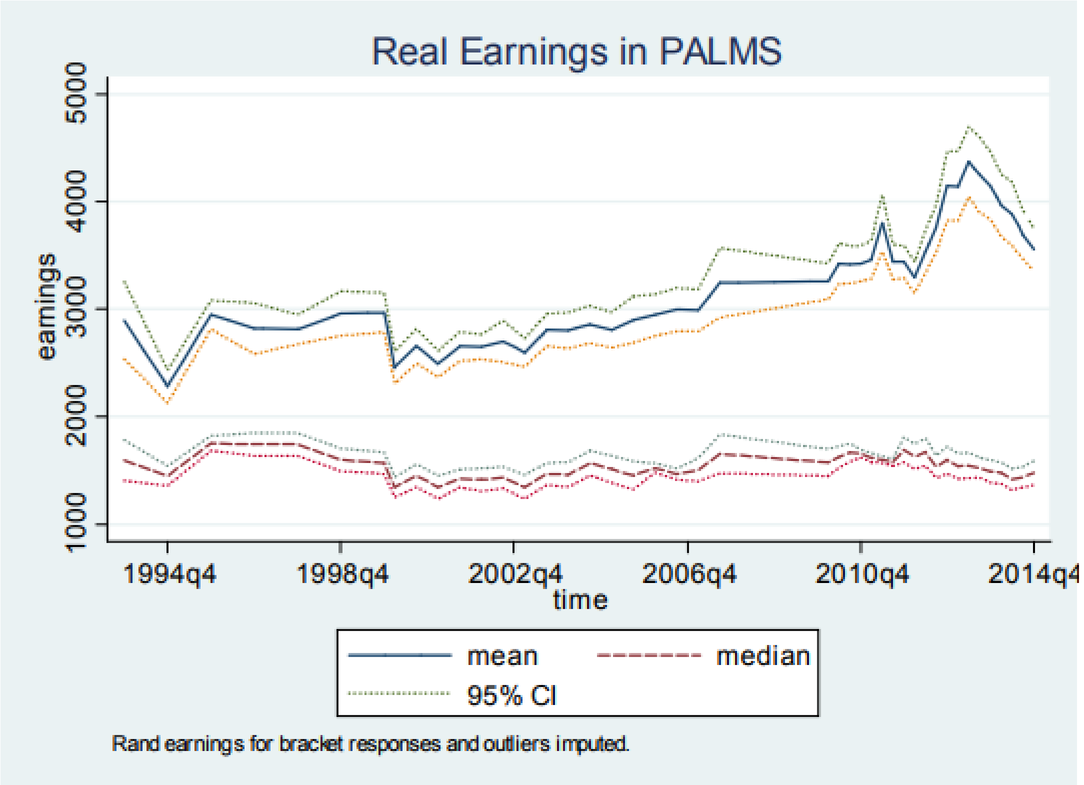
We have QLFS refusals for 2011 because of access to unimputed data from Stats SA in this year only.

		Complete refusal	Amt refusal	Don't know	Zero income	1+3+4
QLFS 2011	QLFS 2011:1	0.11	0.41	0.09	0	0.2
	QLFS 2011:2	0.11	0.42	0.1	0	0.21
	QLFS 2011:3	0.12	0.44	0.1	0.01	0.22
	QLFS 2011:4	0.11	0.43	0.09	0.01	0.21
LFS average		0.04	0.31	0.02	0.05	0.11

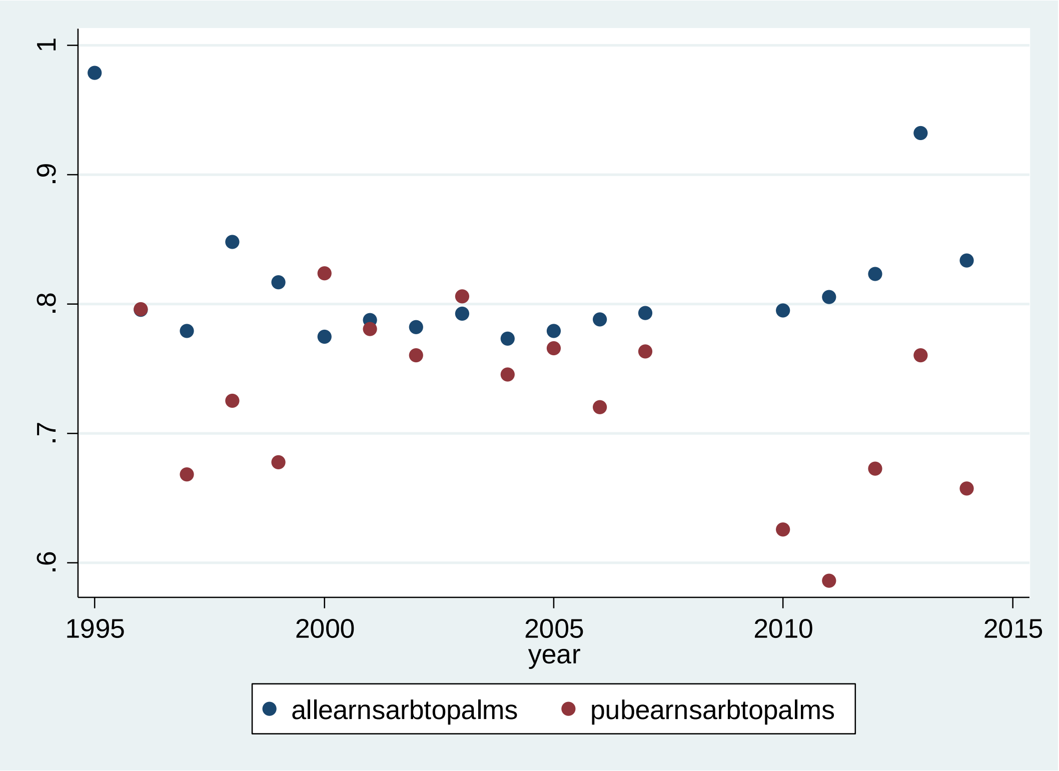
QLFS earnings imputations

- Stats SA documentation (not public) says that Hotdeck imputation is used, although no further detail is given.
- 572 obs of 1275 **employees** that earn more than 100k per month in QLFS are supposed to be earning exactly 400k
  - All 572 report monthly period of earnings, compared to only 65% for other employees with earnings >100k pm.
  - If hot deck impute why so many at 400k? Topcoding? Nothing comparable for self employed.
- For the self-employed there are quite a number of earners (.2% of sample) earning more than 700k per month. All above 1 300 000pm are flagged as outliers by the detection procedure done in PALMS.
  - Many of these are clearly cases where monthly figures have been reported but the period is reported as hourly- max is 85m pm!

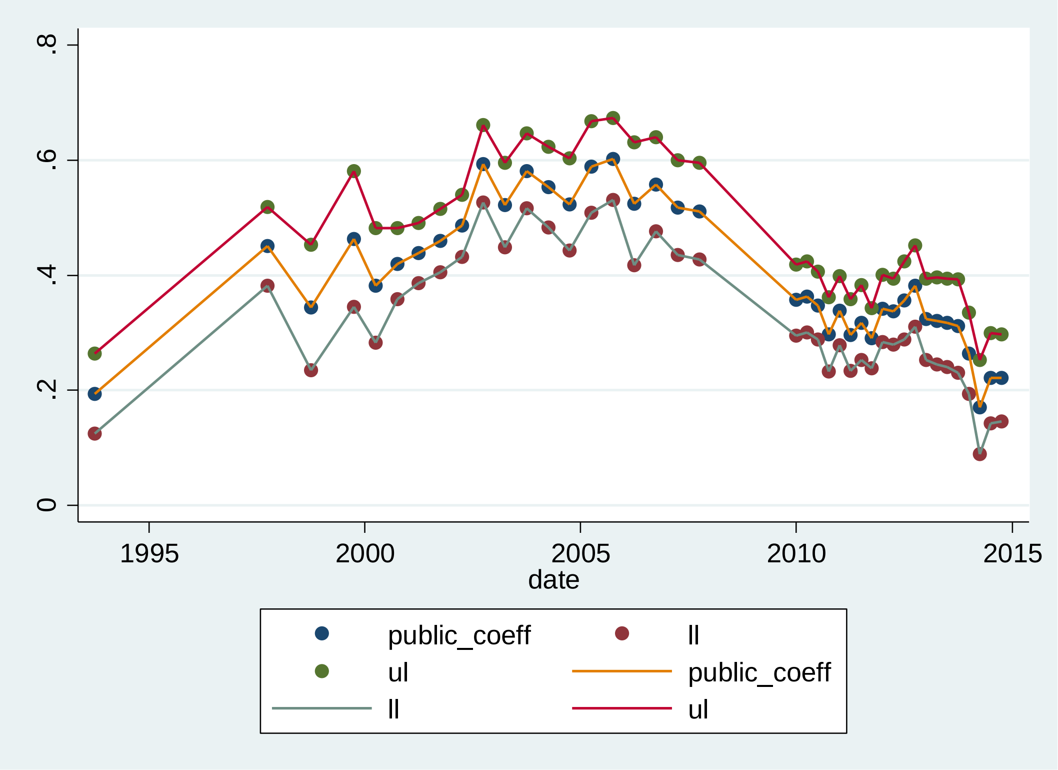
Means and medians in private formal and public sectors over time- Wittenberg (2016).



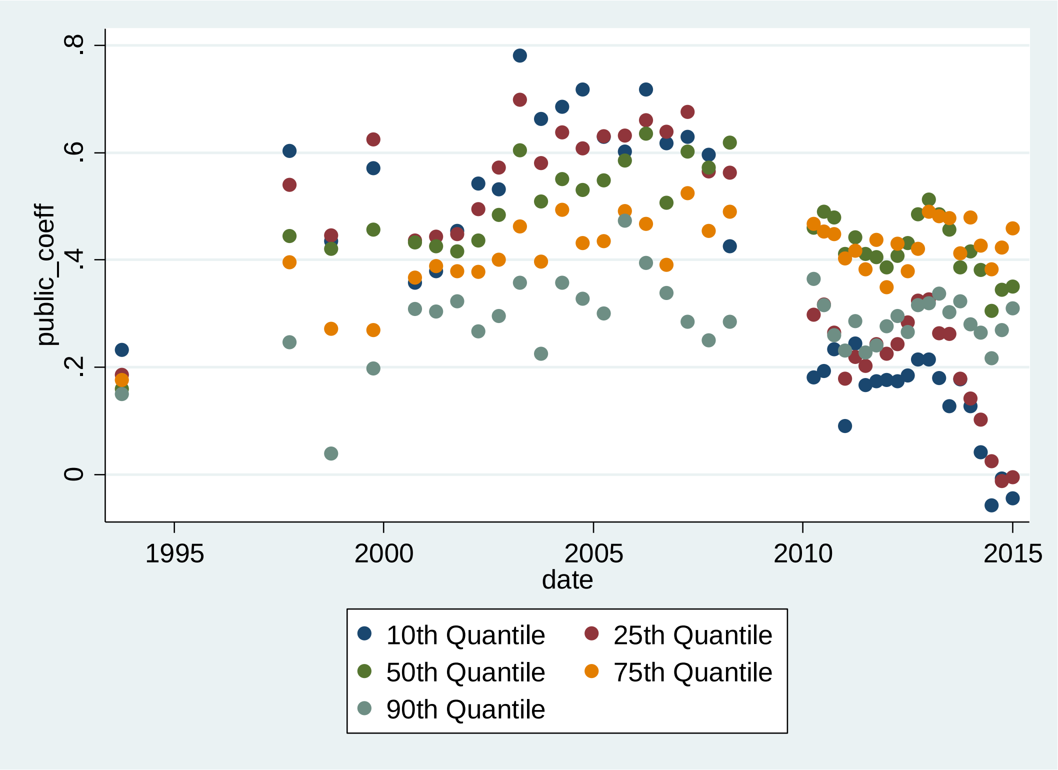
Seekings and Nattrass (2015) critique about worsening of labour surveys earnings totals relative to National Accounts not correct?

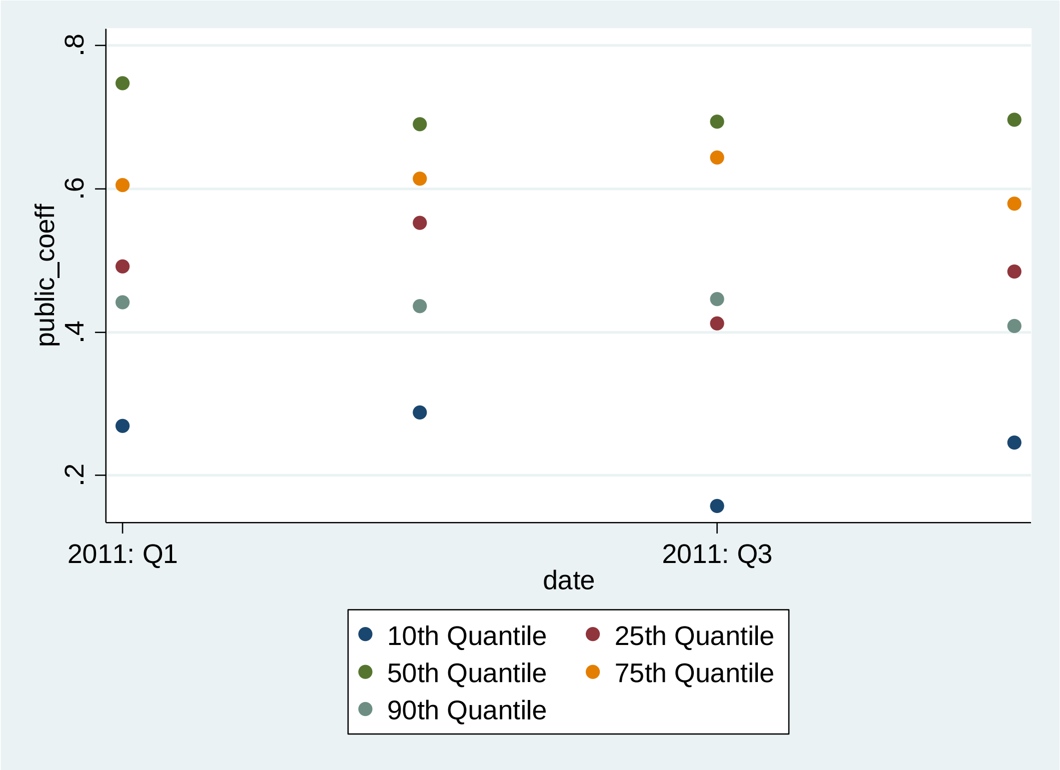


Public sector dummy regression coefficients (Kerr and Wittenberg 2016)

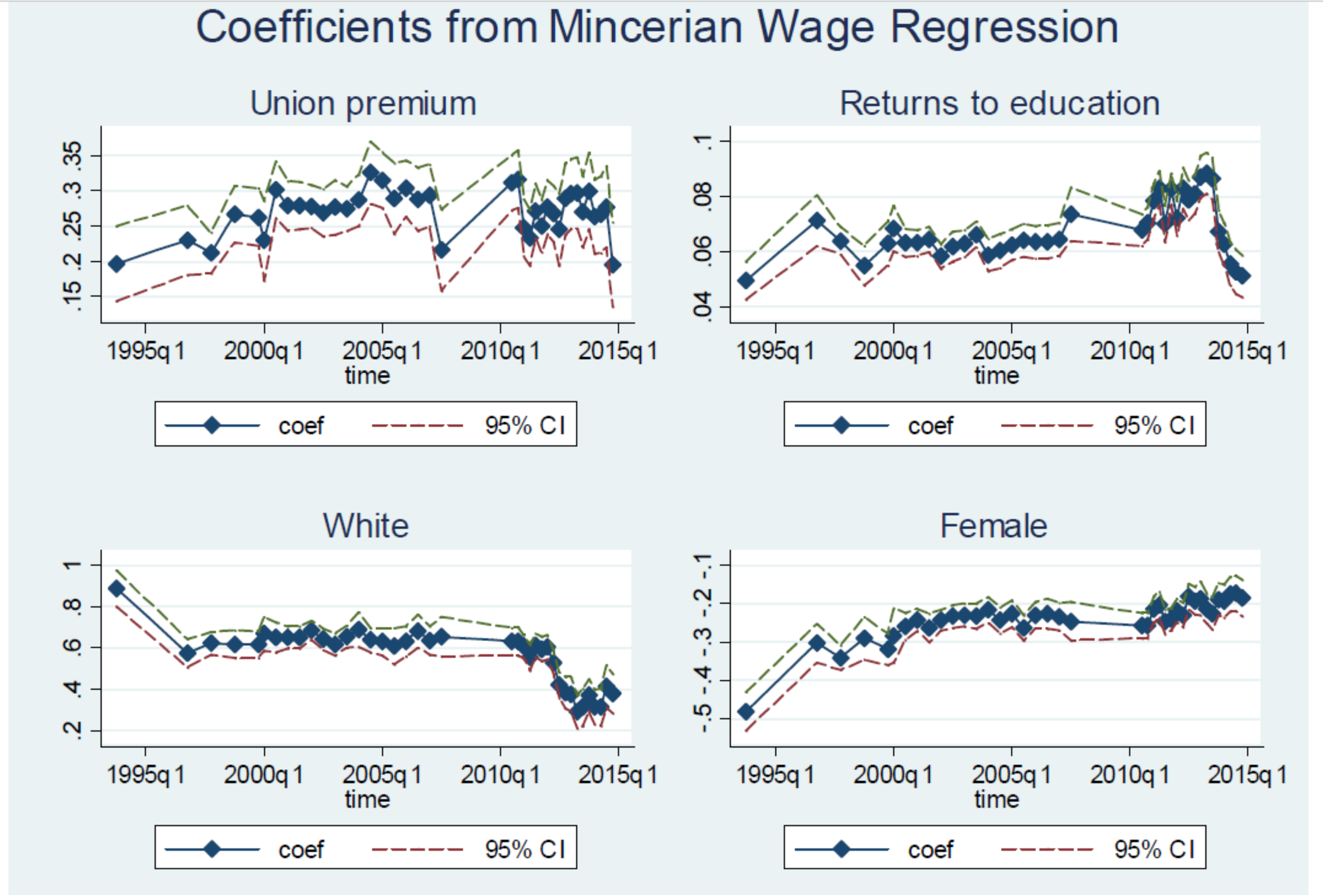


Quantile Regression estimates of a Public sector Dummy Coefficient from earnings regression + then using unimputed 2011 data (Kerr and Wittenberg 2016)

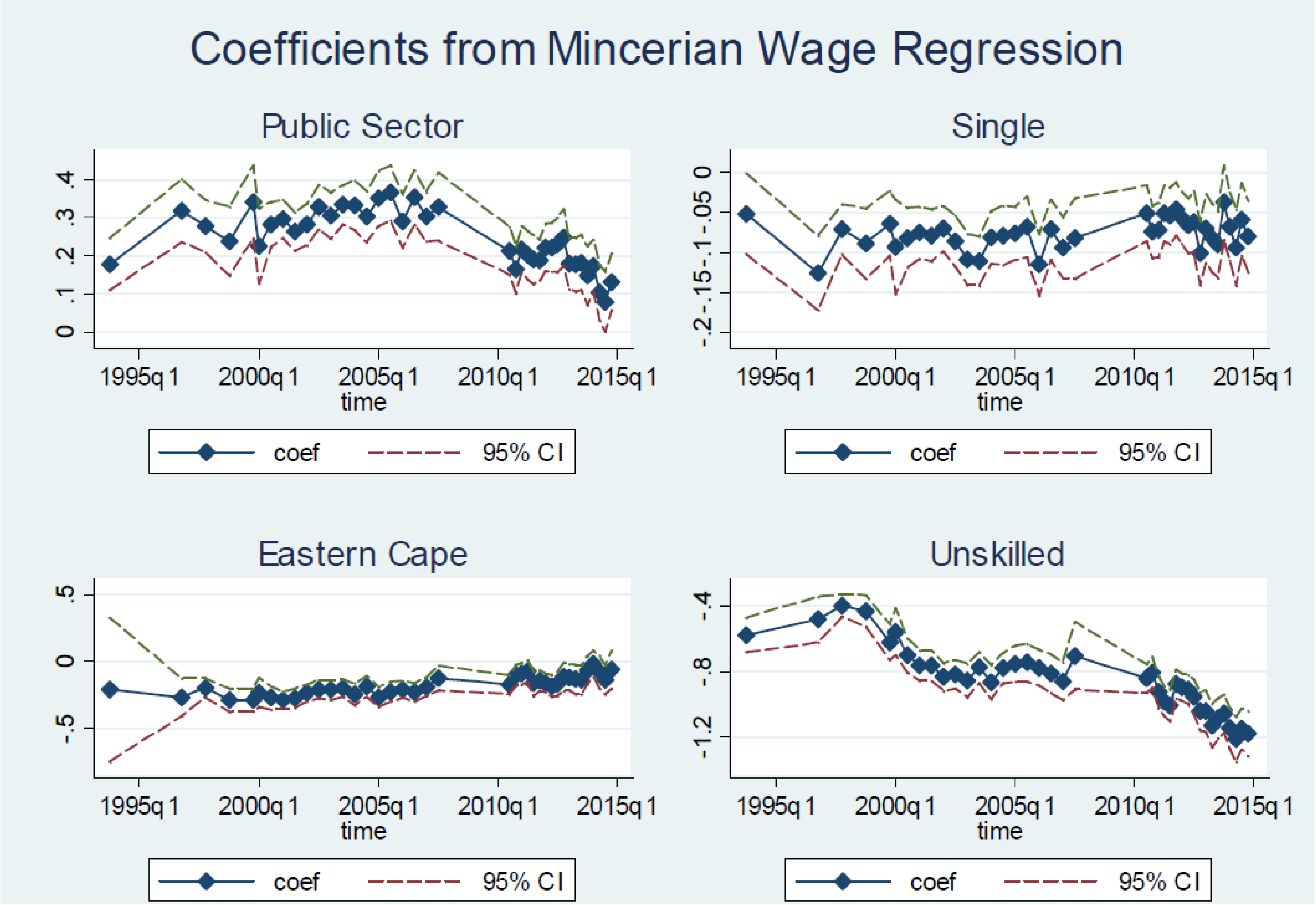




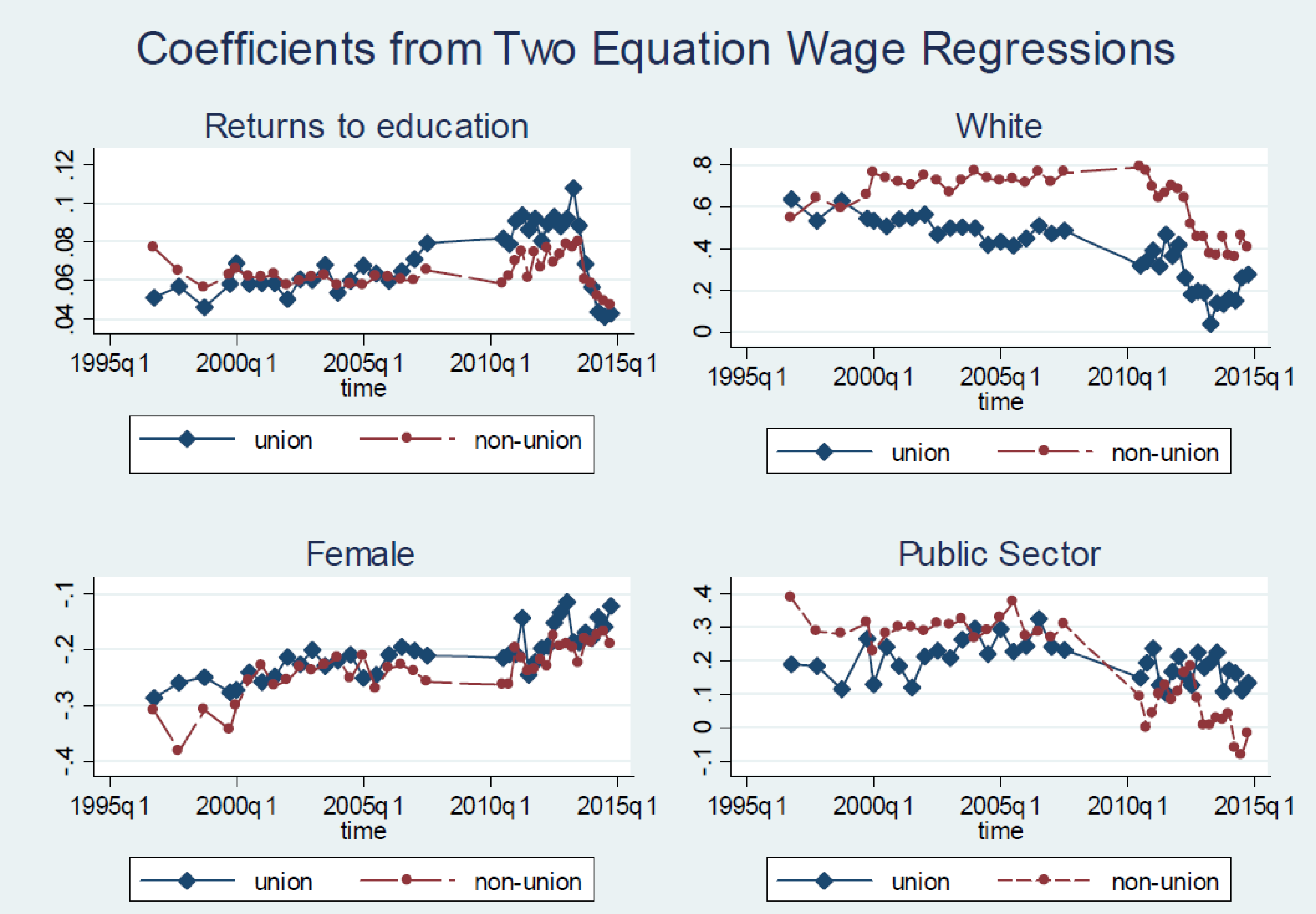
Wittenberg and Kerr (2016) on unions- odd changes in 2014 and 2012



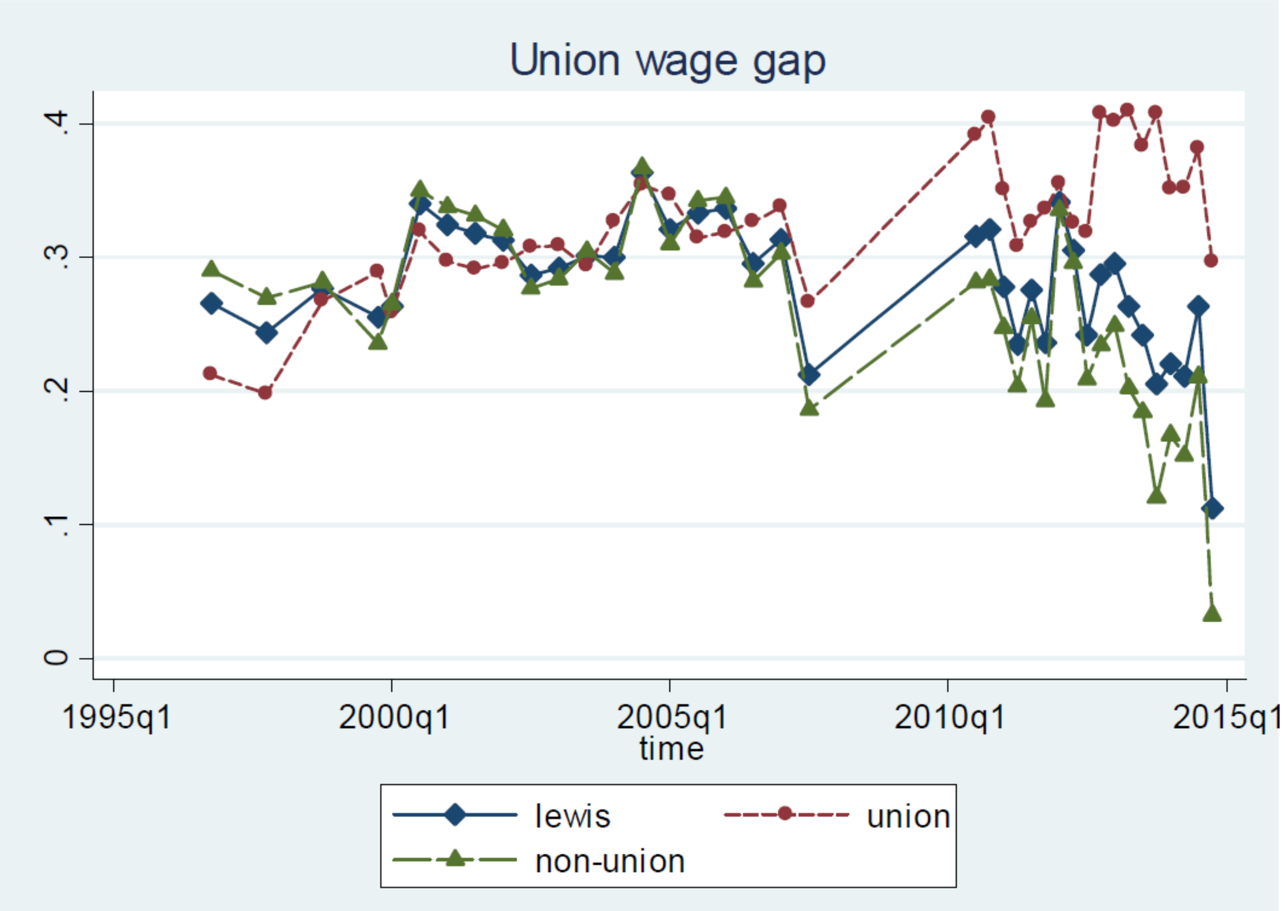
Wittenberg and Kerr (2016)



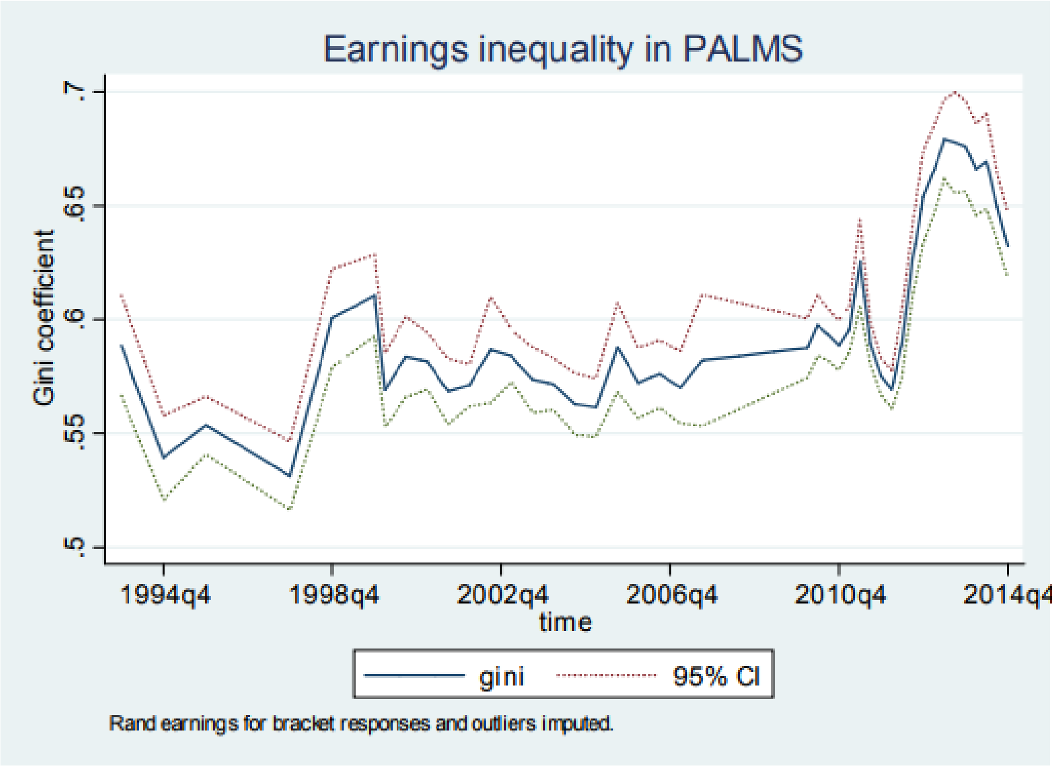
Separate Union and non union regressions



Union wage gap from Kerr and Wittenberg (2016)



Inequality changes from Wittenberg (2016)



Rand earnings for bracket responses and outliers imputed.

Figure 11: Earnings inequality as measured by the Gini coefficient 1993-2014

Conclusions

- Lots of important changes in QLFS data over time
  - Sampling changes
  - Imputation changes
- Giving out unimputed data may solve the change in imputation issue and allow us to see how odd things look without that source of oddness.
- But the state of the QLFS is still concerning – DF plans to engage with Stats SA about these issues.