Who benefits from being self-employed?

Adeola Oyenubi
University of Cape Town South Africa

Second draft 3rd Dec 2014

ABSTRACT

We describe the income gap between self-employed individuals and wage earners. Since there is reasonable overlap between self-employment activities and the informal economy in developing countries we hypothesize that heterogeneity in the informal sector will have implications for the income gap between self-employed individuals and wage earners. The mean income gap will not portray the true picture as it mask the heterogeneity in the income gap along the income distribution i.e. one should expect income gap to between the two categories of workers to vary over the income distribution. We therefore estimate the conditional and the unconditional quantile effect of being self-employed over the income distribution and further decompose the income gap at each quantile into its explained (endowment) and unexplained (returns) components. Our hypothesis is that heterogeneity in the wage gap should lead to differential effect of returns and characteristics at different portions of the income distribution.

Our result show self-employed individuals enjoying a positive premium at the upper end of the income distribution while the result is mixed at the lower end. Our quantile effect result suggests that the positive premium observed at the mean for self-employed individuals after controlling for human capital characteristics can be attributed to superstar entrepreneurs (Rosen, 1981). We also show by decomposing the earnings gap at each quantile that this heterogeneity implies that of the portion of the earning differential explained by human capital characteristics and returns to these characteristics differs at different parts of the earnings distribution. Furthermore, the result of our quantile decomposition shows that characteristics explain a higher percentage of the earnings gap at the lower quantiles while its effect dwindles at the upper end. This result agrees with the finding of Tannuri-Pianto & Donald, (2002) that self-employment has more benefits only at the upper end of the earnings distribution.

This research is funded under the Network for Jobs and Development (NJD) programme.
The authors are grateful for the assistance of Andre Kerr in providing valuable guidance on the data used and the methodology.
The authors are also grateful to Esfandiar Maasoumi, Daniel Millet and Dipanwita Sakars for sharing their code with us

Introduction

The term self-employed and the informal economy are closely related with considerable overlap especially in developing economies. Informal economy (or sector) refers to Jobs or employment activities that are unregulated by the government. The term self-employment was coined by a British anthropologist while studying low income communities in Ghana. Self-employment on the other hand refers to people working on their own as against being formally employed. It turns out that in developing economies these definitions roughly refers to the same concept as those who work for themselves generally work in the unregulated sector.

This paper looks at the wage gap between self-employed individual and wage employees in Ghana. Since most self-employed individuals work in the informal economy and informal sector is known to be highly heterogeneous (Bargain & Kwenda, 2014) one should expect this heterogeneity to translate to heterogeneity in the self-employment income gap. So that at each point of the income distribution the income gap between a self-employed individual and a wage earner would vary. Specifically we are interested in the portion of the income distribution where the income gap is highest/lowest and what explains the difference.

Looking at the mean income gap will conceal the information we are interested in, therefore, we take a quantile approach by estimating the wage gap at quantiles of the income distribution. This will reveal the heterogeneous nature of the income gap. Furthermore we are interested in decomposing this income gap along the income distribution. What we expect is that the decomposition too will display heterogeneity so that what explains the wage gap at the lower end of the distribution may not have significant influence at the higher quantiles. For example is the difference between self-employed and wage earners at upper quantiles better explained by their set of endowments or the return to those endowments? In other words what is the difference between super star entrepreneurs (a term coined by Rosen (1981)) and a comparable wage earner? The answers to these questions may have important policy implications similar to the implications that can be derived from typical Oaxaca-Blinder Decomposition.

Econometric Approach

The self-employed and wage earners can be thought of as being in two different sectors. We are therefore interested in the income difference between the two sectors and what explains it. One immediate problem in isolating the effect of being in one of the two sectors is endogeniety. Selection into the self-employment group is not likely to be exogenous. One way to deal with this is to use the Heckman selection model, however this requires an exclusion restriction i.e. a variable that is correlated with participation in the self-employment sector but is uncorrelated with earnings. Due to data restrictions we cannot pursue this method. Instead we follow Falco, *et al.*(2011) and use a fixed effects estimator. This allows us to control for time invariant unobservables that may affect the choice of sector. The framework of Abowd, *et al.,* (1999) adopted by Falco, *et al.* (2011) assumes that the movement between sectors are exogenous for the identification of size and sectorial effects. However Falco, *et al.* (2011) showed using a System GMM estimator (Blundell & Bond, 1998) that there is no

evidence that the fixed effect estimator that is relied upon for identification of size and sector effects are biased. Since our sample is not significantly different from theirs, we rely on their result for now as far as exogenous movement is concerned[1].

We utilize a battery of methods in order to isolate the effect of being self-employed and to check the robustness of the effect found. We start at the mean by using a fixed effects estimator. Then we look at the effect of being self-employed on the entire distribution of wages. To achieve this we calculated the conditional Quantile Treatment Effect (QTE). Unconditional Quantile effect (UQE) was also calculated to check the robustness of the QTE estimates. Furthermore the UQE calculated using the Regression Influence Function (RIF) methodology allows the decomposition of the income gap at quantiles of the income distribution. We could therefore decompose the wage gap at the quantiles into returns and endowment effect.

Fixed Effect estimator

We start by replicating the result of Falco, *et al.* (2011) for our variable of interest. The authors looked at four sectors namely Public Enterprise, Civil service, Private wage earners and self-employment. In our analysis we collapse these groups into wage earners (comprising Public Enterprise, Civil service and Private wage earners[2]) and the self-employed (Self-employed with no employee and Self-employed with at least one employee). Following their lead we adopted the framework of Abowd, et al., (1999) and consider the following model

$$y_{it} = x_{it}\beta + \theta_i + \psi_{j(i,t)} + \delta time + \varepsilon_{it}$$

Where y_{it} is the natural log of earnings, x_{it} is a vector of time varying observables, θ_i is the time invariant individual characteristics , $\psi_{j(i,t)}$ is the sector j in which individual i works at time $t, j = s \; or \; w$ representing self-employed and wage earners respectively. Time dummy captures the effect of time. Sector variation $\psi_{j(i,t)}$ is modelled as a dummy variable that takes the value one when a worker is observed as a self-employed individual and zero otherwise. This means that unlike Falco, et al.(2011) there is no need to construct a sector index but similar to them, a fixed effect estimator is used to extract θ_i which contains all observable and unobservable time invariant variations. The variation due to time invariant observables is adjusted for by modelling the fixed effect variation as follows:

$$\theta_i = \alpha_i + \mu_i\eta$$

Where μ_i is a vector of time invariant observables like education and gender (estimated without a constant term), α_i is then isolated as the residual of this regression. Falco, et al.(2011) called it market ability: it contains each individual's time invariant unobservable variation. This term is then used in a pooled OLS regression as an additional control. This means that for this OLS regression time invariant unobservables are controlled for. Following the authors we address the problem of bias due to selection by assumption, that is: our ability to control for unobservables with the panel also controls for selection. In our subsequent

[1] We intend to look at this issue in more detail in subsequent drafts.
[2] Wage earners employed by private entities

analysis we exploit the above methodology and continue to control for the time invariant unobservable even when we looked at distributional effects. This approach is similar to the one introduced by Canay (2011) where the author assumed that θ_i are pure location shifters and estimated it using the first difference method. This extracted fixed effect was then used to calculate corrected income $\hat{y}_i = y_i - \theta_i$. Doing this Canay (2011) controlled for time invariant unobservables by eliminating them. Under this approach just like the traditional fixed effects estimator we cannot have an inference on any time invariant observable. On the other hand the method used by Falco, *et al.* (2011) which we follow control for the fixed effect in such a way that inference can still be made on time invariant observables.

We also follow Falco, *et al.* (2011) in modelling of the number of employees and firm size. The former refers to the number of employment generated by the firm in which an individual works. This value is unity for wage earners and own self-employed. The firm size captures variation in the number of people working in the firm in which an individual works, for wage earners this will be the number of people employed by their firm while for own self-employed it will be the number of people they employed plus one.

Conditional and Unconditional Quantile treatment Effect (QTE & UQTE)

As stated earlier it is reasonable to expect that the first moment (mean income gap) between the two sectors will not paint a complete picture of the dynamics of the income gap at different point in the distribution. To tell a more complete story we consider the quantile effects. Let Y_w and Y_s denote two earnings variables to be compared. Y_w (Y_s) Represent potential log earnings for wage earners (self-employed). We have N_w (N_s) individuals observed in wage earning (self-employment) sector. The treatment here is being self-employed so that $D_w(D_s)$ denote that an individual is observed in the wage earner (self-employed) group. Let $F_w(y) \equiv \Pr[Y_w < y]$ represent the cumulative density function (CDF) of Y_w and $F_s(y) \equiv \Pr[Y_s < y]$ be the CDF for Y_w

To explore the return dynamics at different parts of the distribution we consider the QTE and the UQE. For the QTE we use the approach of Maasoumi, et al., (2009) this involves calculating $\hat{\Delta}_p = y_w^p - y_s^p$ for $p = 0.001 \ldots \ldots 0.99$[3]. The estimates will be consistent if $Y_w, Y_s \perp D$ (i.e. if unconfoundedness assumption holds) and the CDFs of the potential earnings are continuous and monotonically increasing at the quantiles. The dependence between sector choice and other determinants of earnings suggests that the identification assumption might not hold. Therefore estimates for QTE adjusted for covariates is obtained using the inverse propensity score weighting. Such estimates will be consistent under the Conditional independence assumption (CIA) and Common Support Assumption in addition to the requirement that potential earnings should be continuous and increasing at quantiles for which QTE is estimated. To do this we follow Bitler, et al., (2006) and Maasoumi, et al., (2009) and estimate the empirical CDF

[3] It is important to note that the QTEs do not correspond to quantiles of the distribution of treatment effect unless the rank preservation assumption holds (Firpo, 2007).

$$\hat{F}_{j\ N_j}\ (y) = \frac{\sum_{i=1}^{N_j} \hat{\eta}_j (Y_j \leq y)}{\sum_{i=1}^{N_j} \hat{\eta}_j} \quad j = s\ or\ w$$

Where the weights $\hat{\eta}_j$ are given by

$$\hat{\eta}_j = \frac{D_i}{\hat{p}(x)} + \frac{1 - D_i}{1 - \hat{p}(x)}$$

$\hat{p}(x)$ is the estimated propensity score. The propensity scores were calculated controlling for observed characteristics and variation due time invariant unobservables α_i extracted from the fixed effect estimation. Maasoumi, *et al.,*(2009) controlled for time invariant unobservables by considering ΔY_w ,$\Delta Y_s \perp D$ (i.e. using the difference of the outcome variable). We follow a different route to achieve an analytically similar construct. Since the recovered α_i extracted from the panel analysis controls for time invariant unobservables in our pooled OLS regression, we also use it directly to control for time invariant unobservables in QTE estimation. This will make our result from the fixed effects analysis more comparable with the one in this section.

The QTE explored above is a conditional quantile effect. It is given by the difference

$$\Delta_p = F_s^{-1}(p|X = s) - F_w^{-1}(p|X = w)$$

and is generally different from

$$\frac{d\Delta_p(p)}{dp} = \Pr\left[Y > \Delta_p \middle| X = s\right] - \Pr\left[Y > \Delta_p \middle| X = w\right] / f_y(\Delta_p)$$

Which is the effect of increasing the proportion of self-employed individuals on the p^{th} quantile of the unconditional distribution of Y (Firpo, *et al.*, 2009). This unconditional quantile regression can be calculated using the Re-centered influence function (RIF) of the unconditional quantile on the explanatory variables. Conditional QTE (or quantile regression) may generate results that are often not generalizable in terms of policy or population context (Firpo, *et al.,* 2009). On the other hand Unconditional quantile effects can be interpreted in policy or population context as it marginalizes the effect over the distribution of other covariates in the model (Borah & Basu, 2013).

Borah & Basu (2013) investigated conditions under which QTE and UQTE results might differ, their investigation suggest that when there are multiple covariates UQTE is more appropriate than QTE. Their simulation shows that QTE is sensitive to the specification of the model. Furthermore, when the DGP is such that the effect of the variable of interest varies over the levels of other covariates UQTE is more appropriate. Therefore comparing our QTE result to the UQTE result will serve as a robustness check and also enable the decomposition of the income gap at quantiles of the income distribution.

Brief background on Ghanaian Labour market and the Data used

Self-employment makes up a large part of the African urban economy (Glick & Sahn, 1999) of which Ghana is not an exception. Estimates from the GLSS (2005/2006)[4] data suggests that women and young workers dominate the informal sector while the few available formal sector jobs are dominated by men (Heintz & Lynda, 2012). The gender and age dimension of the divide between wage earners and self-employed individuals is such that the demography that is worst hit by unemployment is the one that dominates the self-employment (or informal) sector. This is in line with the overlap between the informal sector and self-employment in developing countries. These coupled with a hiring freeze in the public sector due to the Structural Adjustment Programme in the 1980s in part explain the distribution of unemployment in the between the young and the old.
Previous research suggests that the urban labour market in Ghana is heterogeneous in that determinants of entry into different segments of the labour market are not the same (Glick & Sahn, 1999). The authors also found that education reduces the likelihood of being self-employed while in the formal or wage earning sector education is a key attribute that predicts entry.

The data used for analysis is the longitudinal labour market survey conducted by the Centre for the Study of African Economies (CSAE) at Oxford University. We used the Ghanaian data set (data was also collected for Tanzania). The data spans three years covering 2004 to 2006. The survey collects information on income, education, labour market experience, household characteristics and various other modules for labour force participants (respondents are between 15 to 60 years old) in urban areas. These areas include the four largest urban centres in the country: Accra (and neighbouring Tema), Kumasi, Takoradi and Cape Coast.

The samples were based on a stratified random sample of urban households from the 2000 census in Ghana. While the initial sample was household based, interviews were conducted on an individual basis, and the unit of analysis in what follows will be at the individual level. A total of 830 individuals were interviewed in the first round of the survey in Ghana. Information on earnings for self-employed individuals in low income developing countries is controversial. The argument is that self-employed individuals in the informal sector rarely keep accounts and their self-reported income data may be too noisy. For household based businesses the distinction between business and personal expenditure may be unclear to respondents. We acknowledge that this creates some limitations. However like Falco, *et al.*(2011) we believe that this aspect of the labour market is too important to be overlooked.

To mitigate some of the effect of measurement error the income measure is based on self-reported profits net of operating expenses and fixed capital expenditure. The concepts of revenue, cost and expenses were explained to respondents by enumerators and a simple mechanical check (hand held computers were used to collect the data) forces the enumerator to go through the numbers again for inconsistencies.

Table 1 shows the summary statistics and the pattern is roughly along the lines of what has been observed in the literature. Wage earners on average earn more and are more educated. They are however less experienced, work less hours and younger.

[4] GLSS Ghana Living Standard Survey

Table 1: Summary Statistics

variable	Wage	Self
Earnings (97 USD)	20.46	16.37
	(21.21)	(15.21)
Experience	16.18	22.94
	(10.34)	(11.95)
hours	46.63	48.63
	(12.21)	(13.74)
Firm size	33.51	1
	(50.15)	(NA)
Employees	1	1.36
	(NA)	(1.14)
Education	9.8	7.44
	(3.29)	(4.1)
Age	31.98	36.38
	(10.4)	(10.65)
Tenure	5.85	10.77
	(7.32)	(9.58)
Male	0.65	0.35
N	909	1297

Analysis and Results

Tables 2 and 3 present the results from our OLS and fixed effect estimator. For all estimation clustering is done at individual level. The results are similar to what Falco*, et al*.(2011) found so we have kept the discussion brief. Column (1) of table 1 shows the OLS earnings gap without controls, the result is as expected: a negative premium of about 22% is associated with being self-employed. The result in Column (2) shows that once we introduce a set of controls the magnitude of this premium becomes positive but insignificant. The fixed effect estimation in column (3) shows that once time invariant unobservables are controlled for, only firm size and experience are significant in explaining variation in earnings. The result suggests that at the mean there is no significant difference in the observed earnings profile of self-employed individuals and wage earners. Column (4) is an OLS regression that includes the probability of being observed in 2005 and 2006. Similar to Falco, *et al*.(2011) the result suggests that attrition is not a big concern.

The next set of results is shown in table 2. Again the result is similar to what is found by Falco, *et al.* (2011). Column (1) is a replica of column (2) in table 1 (included to facilitate comparison); Column (2) shows the result once variation due to time invariant unobservable is controlled for in the OLS regression. In contrast to table 1, self-employed individuals actually earn significantly higher than their wage earning counterparts. This result provides a mixed signal which we address in the next section.

Table 2: OLS and FE Regression of log earnings

log earnings	(1)	(2)	(3)	(4)
	OLS	OLS	FE	OLS (Attrition)
Self	-0.223	0.056	0.159	0.060
	(0.048)***	(0.061)	(0.117)	(0.211)
Total Exp		0.044	0.164	0.046
		(0.007)***	(0.047)***	(0.007)***
exp^2		-0.065	0.012	-0.060
		(0.013)***	(0.105)	(0.014)***
Ln(Hours)		0.142	0.090	0.166
		(0.059)**	(0.080)	(0.299)
Ln(firsiz)		0.101	0.053	0.082
		(0.017)***	(0.019)***	(0.030)***
Ln(emps)		0.256	0.090	0.210
		(0.058)***	(0.074)	(0.079)***
Tenure (years)		0.022	-0.015	0.003
		(0.008)***	(0.017)	(0.022)
Tenure^2		-0.030	0.063	0.019
		(0.026)	(0.057)	(0.064)
Male		0.276		0.299
		(0.044)***		(0.058)***
Education (years)		-0.058		-0.051
		(0.016)***		(0.020)**
Educ^2		0.835		0.799
		(0.105)***		(0.137)***
2005		0.082	-0.077	0.084
		(0.040)**	(0.031)**	(0.040)**
2006		0.314		0.316
		(0.043)***		(0.043)***
Probo5				-0.331
				(0.668)
Prob06				-0.444
				(0.453)
Constant	2.624	0.590	-1.321	0.293
	(0.039)***	(0.245)**	(0.631)**	(1.530)
Adjusted R-squared	0.01	0.24	0.08	0.24
N	1,294	1,294	1,294	1,294
F	22.00	45.34	7.99	39.95

* $p<0.1$; ** $p<0.05$; *** $p<0.01$ (1) Regression of log earnings on self-employment indicator without controls (2) Regression one with controls (3) Fixed effect regression with controls (4) Regression in (1) including probability of being observed in 2005 and 2006 to check the effect of attrition. Clustering was done at the individual level.

Table 3: Pooled Regression with and without time invariant unobserves

log earnings	(1)	(2)
	OLS	FE
Self	0.056	0.161
	(0.061)	(0.016)***
Total Exp	0.044	0.069
	(0.007)***	(0.001)***
exp^2	-0.065	0.012
	(0.013)***	(0.001)***
Ln(Hours)	0.142	0.089
	(0.059)**	(0.027)***
Ln(firsiz)	0.101	0.053
	(0.017)***	(0.004)***
Ln(emps)	0.256	0.089
	(0.058)***	(0.024)***
Tenure (years)	0.022	-0.015
	(0.008)***	(0.002)***
Tenure^2	-0.030	0.063
	(0.026)	(0.005)***
Male	0.276	0.500
	(0.044)***	(0.006)***
Education (years)	-0.058	0.341
	(0.016)***	(0.001)***
Educ^2	0.835	-0.886
	(0.105)***	(0.004)***
2005	0.082	-0.077
	(0.040)**	(0.025)***
2006	0.314	0.001
	(0.043)***	(0.028)
Ability Index		1.000
		(0.002)***
Constant	0.590	-1.892
	(0.245)**	(0.110)***
Adjusted R-squared	0.24	0.83
N	1,294	1,294

* $p<0.1$; ** $p<0.05$; *** $p<0.01$
With ability and sector indices

Table 4: Oaxaca-Blinder Decomposition of earning gap

lrearn97	Coef.	Robust Std. Err.	z	P>z	[95% Conf.	Interval]
Differential						
Wage	2.623233	0.030489	86.04	0.00	2.563476	2.682991
Self	2.408008	0.0257466	93.53	0.00	2.357546	2.458471
Difference	0.215225	0.0399058	5.39	0.00	0.137011	0.293439
Explained	0.375855	0.0406705	9.24	0.00	0.2961422	0.455568
Unexplained	-0.16063	0.0227714	-7.05	0.00	-0.2052611	-0.116

Using Oaxaca-Blinder decomposition the wage gap at the mean was decomposed into its components. The mean decomposition show that more of this gap is explained by differences in human capital characteristics. Table 4 shows this result. Although both returns and characteristics explain a significant portion of the total income gap, differences in characteristics explain a larger share of the difference. The full decomposition result is presented in the appendix (table 4A). The negative unexplained part suggests that self-employed individuals are actually being compensated in terms of returns (the negative unexplained portion of the difference).

This mixed signal may be as a result of an uninformative mean due to the possible heterogeneous nature of the income gap. There is therefore need to exercise caution in the interpretation of this result because the positive premium might be driven by few individuals in the self-employment sector i.e. "superstars" among the self-employed Rosen (1981). This is the proposition that mean earnings may be biased upwards by a few successful entrepreneurs such that the result is not representative. Heterogeneity in the informal sector means that it could be the case that a handful of skilled professionals like doctors and lawyers are in the same group with street traders, tailors and barbers. This disparity in expertise and hence earnings will be masked at the mean.
These individuals are not comparable in terms of human capital characteristics. This may explain the positive returns to self-employment in table 2. To get a clearer picture we explore a methodology that allows us to look beyond the mean impact of these variables.

Distributional Analysis

In this section we ask if the income gap observed at the mean is uniform over the income distribution. We start by simply comparing the empirical CDF in the two groups with and without adjusting for observables and time invariant unobservables by using the inverse probability weighting described earlier (Maasoumi, *et al.,* 2009). Figure 1 (A) and (C) present the result of the comparison between the unconditional and conditional earning profiles (or CDFs) respectively. The unconditional result show that the CDF for wage earners is "better" than the one for the self-employed. Fig 1 (A) shows that at any quantile τ for the unconditional distribution the probability of observing a worker earning less than τ is lower under the wage earners CDF than under the self-employed CDF. In technical parlance the CDF of Y_w first-order stochastically dominates (FSD) Y_s i.e. $F_w(y) \leq F_s(y)$[5]. However once we

[5] This test was actually conducted. The result is presented in the appendix (Table 6A). The important point here is that a social planner will prefer the earning profile for the wage earners because it is the dominant profile.

adjust for covariates the first order dominance is lost. Fig 1 (C) shows that the comparison yields different conclusion at different parts of the distribution. At the lower quantiles the wage earning profile still dominates but at higher quantiles the role is reversed. This is in line with the expected heterogeneity in the income gap[6].

To explore the dynamics more closely Figure 1 panel (B) and (D) show the QTE with and without controls respectively. In panel (B) there is a negative premium to being self-employed. This effect is close to uniform across the earnings distribution aligning perfectly with the OLS regression in column 1 of table 1 and previous distributional results. Being self-employed is associated with around 22% earnings penalty. This result is not surprising since when an effect is uniform throughout the distribution, the mean will give a valid description of the effect.
However one would observe that there are still some heterogeneity, the effect is larger at the upper end of the distribution with earnings penalty magnitudes close to 45%.

Once the controls are introduced, the shape of the QTE graph changes as shown in panel (D). At the lower quantiles the negative premium for being self-employed is larger but so is the standard error. However since it is not bounded away from zero this effect is insignificant which suggests that given the characteristics, the negative premium at the lower and middle quantiles are not significant. We note that this insignificant negative premium is fairly uniform below the 40th quantile. Between the 40th and the 60th quantile there is a significant negative premium of about 50%, beyond the 70th quantile a significant positive premium is observed.

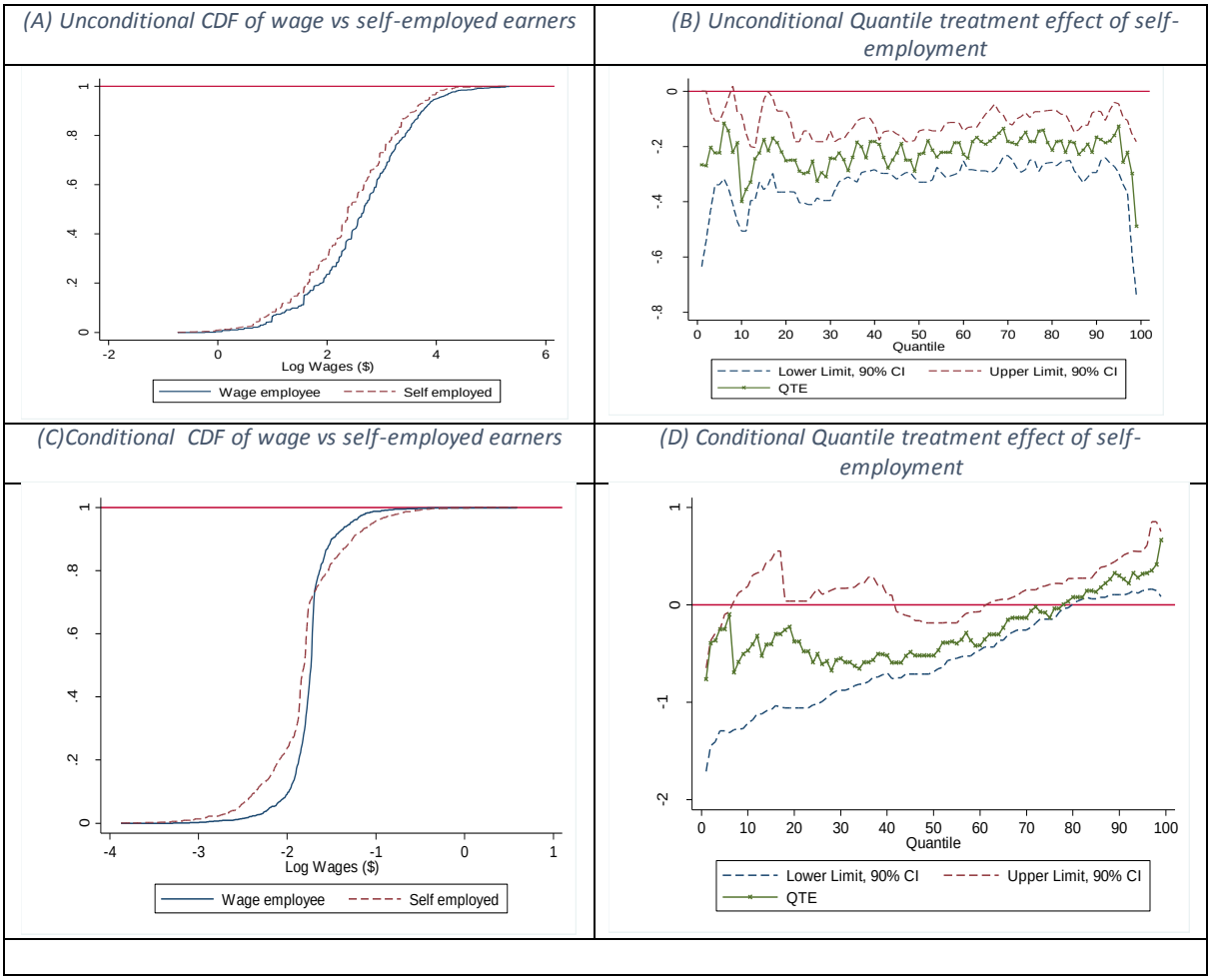
At the upper end of the earnings distribution self-employed individuals are earning up to 68% more than their wage earning counterparts. This is in contrast to the lower end which confirms the suspicion that the effect might be different at different parts of the distribution. Specifically, the result in table 2, column 2; which suggests that self-employed individuals are earning a positive premium of around 16%, can be better interpreted in light of the result from the QTE. This positive premium can be attributed to the "superstar" effect Rosen (1981). Self-employed individuals in the upper quantiles are more likely to be professionals like doctors or lawyers with private practices as compared to the lower parts of the distribution where self-employed individuals are more likely to be involved in less lucrative businesses. The high earners are more likely to be working in the formal sector even though they are self-employed. The unconditional QTE suggests that the negative premium should be higher at the upper end Figure 4(B) but after adjusting for the controls the mean is pulled up by the self-employed at the upper end of the earnings distribution that are doing better on average compared to wage earners with similar characteristics. The net effect is the 16% premium observed at the mean.

Unconditional Quantile Effect

The unconditional version of the quantile effect is calculated for two reasons. First it serves as a robustness check to the QTE explored in the previous section since it can be interpreted in a policy context.

[6] In terms of stochastic Dominance wage earners CDF second order stochastically dominate the self-employed income CDF.

Figure1: Conditional and Unconditional CDFs and Quantile Treatment Effects

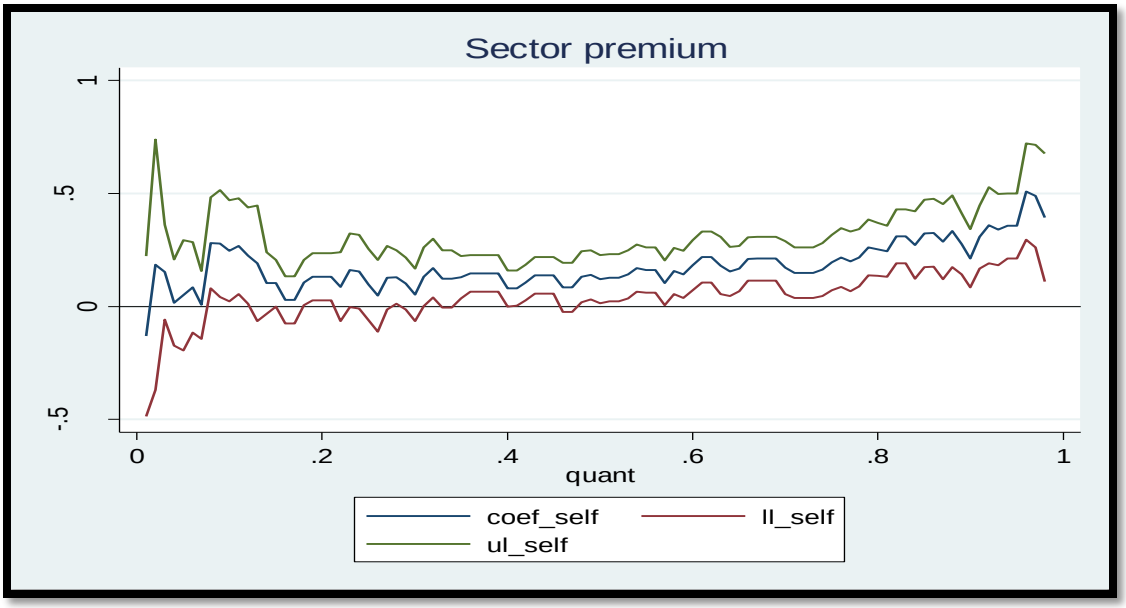


Second, the RIF methodology which is used to calculate the UQE also allows us to decompose the wage gap at different quantiles of the income distribution. This method generalizes the Oaxaca-Blinder decomposition method by extending the decomposition to distributional measures (besides the mean) (Firpo, *et al.,* 2009). Our distributional functional of interest are the quantiles of the income distribution. We are interested in how the influence of returns and characteristics (or endownments) affect the income gap througout the distribution.

Figure 2 below shows how sector premium vary over the entire distribution. Comparing Figure 2 with Figure 1 (D) the QTE and the UQE agree in pattern. Holding all other factors constant being self-employed is not associated with a significant difference in earning at the lower part of the income distribution, but at the upper end (>50 percentile) there is a positive premium of about 50%. The premium to being self-employed increases as we move to the upper part of the income distribution. However there are some specific differences. We suspect this difference is as a result of the sensitivity of the QTE result to the specification of the quantile regression equation (Borah & Basu, 2013). As noted earlier QTE result may vary in the presence of multiple conditioning variables especially when the impact of one variable vary over the levels of the other (Borah & Basu, 2013).

Both method however confirm the heterogeneous nature of the income gap. Furthermore both result suggest the super star entrepreneur effect Rosen (1981) is consistent with our data.

Figure 2: Unconditional Quantile Regression



Other Unconditional quantile regression results (see the appendix for these results Fig2A) further show that the effect of education and experience attenuate as we move to the upper quantiles. This is to be expected since at upper quantiles more and more people have higher levels of education therefore reducing its marginal effect. Similarly, the effects of gender and Ability index[7] decline as we move to the upper part of the distribution. This suggests that the gender gap is more pronounced at the lower end of the distribution (Although the effect is likely to be unstable).

The results for hours worked and tenure have an inverted U shape meaning that long hours and experience on the job increases earnings at the lower end of the distribution but the effect is reversed as we move to the upper end. The effect of firm size mimics that of self-employment in that it increases as we move along the distribution from the lower to the upper part (note that this affect only wage earners).

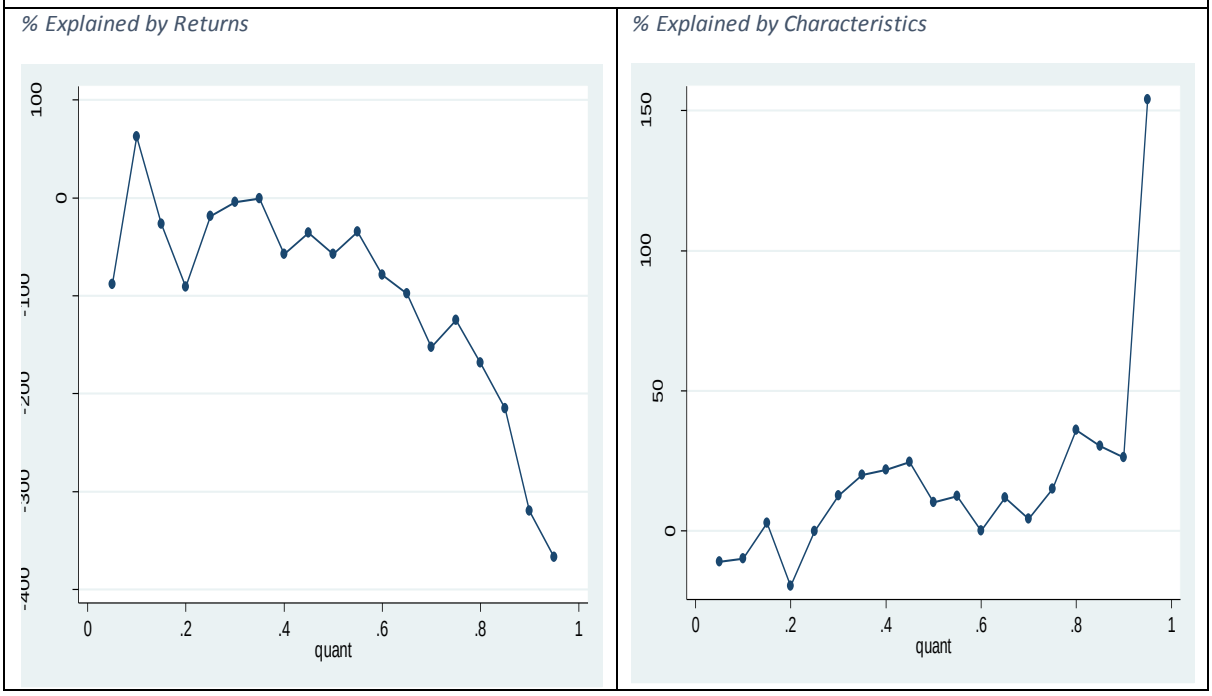
Income Gap Decomposition at the Quantiles

Lastly we decompose the wage gap along the distribution of earnings using the RIF decomposition. First a probit model is used to recover the probability of being observed in the self-employment sector. This probability is then used to reweight the observation in the subsequent RIF decomposition. The result of the probit model is shown in table 5A of the appendix. Males and educated individuals are less likely to be observed as self-employed. Experience and tenure on the other hand increases the likelihood of being self-employed.

[7] There is a striking similarity between the effect of Ability index and Education throughout the distribution. If the unobservables represented by ability index is linked to how workers match with employment opportunities then the result suggests that education has a lot to do with these matching.

Figure 3 shows the percentage of the earnings gap that is explained by returns and characteristics respectively. The effects of returns and characteristics contrast throughout the distribution. The result shows that percentage of the wage gap explained by both returns and characteristics increases relative to its initial level as we move to the upper part of the distribution[8]. As we move from the lower to the upper part self-employed individuals are being increasingly compensated (negative returns effect). Similarly, endowment matters more and more as we move to the upper part of the distribution. This joint returns compensating factor and importance of endowment at higher quantiles explains why the income gap on that portion of the distribution is in favour of the self-employed.

Figure 3: RIF Decomposition



*note that the differences do not add up to the total difference, this is because the RIF regression is a linear approximation to what may very well be a non-linear relationship.

One way to interpret this result is that characteristics are homogeneous at different portions of the distribution i.e. individuals are sorted into different portion of the income distribution based on their characteristics. Both return and characteristics become more important as we move to the upper part of the distribution. The suggestion is that at the upper end the self-employed have double advantage i.e. the gap between them and the wage earners is due to both returns and characteristic effects. While at the lower end the effect of both factors are low.

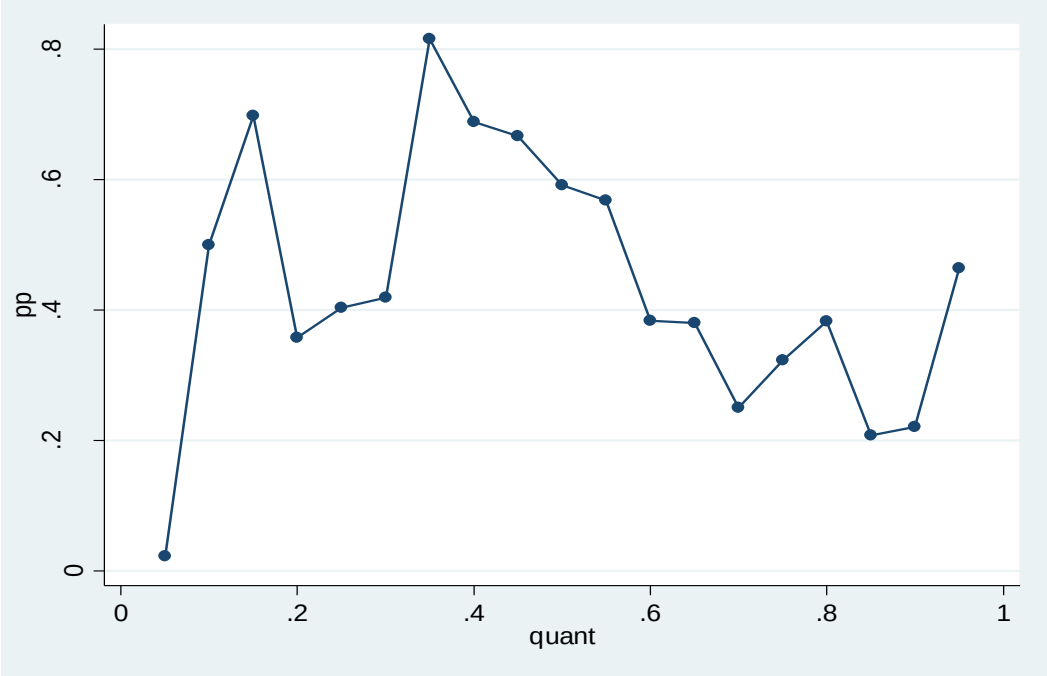
A plausible explanation for this is that wage earners may have less incentives to improve on their human capital characteristics compared to self-employed professionals which relates to the investment and agency model prediction. The investment model suggests that self-employment earnings may be steeper than the wage employment one because human and

[8] The endowment effect is straightforward, for the unexplained part or the returns effect the negative value indicate that self-employed individual are being compensated more and more as we move to the upper part of the income distribution.

physical capital is not shared with the employer for self-employed individuals. Therefore the self-employed individuals may tend to invest more in human capital which increases their return relative to their wage earning counterparts.

To compare the relative effect at each part of the distribution, we add the returns and the characteristic effect together and plot the percentage of the difference explained by characteristics. It should be noted that this addition do not add up to the total difference (because the RIF regression is a linear approximation to what might very well be a non-linear relationship). However this analysis tells us something about the effect of characteristics relative to returns at different parts of the distribution.

Figure 4: Relative effect of characteristics



The result is shown in table 4. We exclude the first and the last quantiles in our explanation because of possible violation of common support (we go into details of this in the next section). Excluding the first and last quantiles the result suggests that characteristics explain the bigger fraction of income variation at the lower end of the income distribution while returns become more and more important at the upper end.

One logical explanation of this result is that at the lower end those with better human capital characteristics (majorly education) are able to get into the formal sector where they attract better wages. The self-employed at this part of the distribution are more likely small business owners that could not get into the formal sector because of their human capital constraints. These will include tailors, hawkers etc. At the upper end both wage earners and self-employed have comparable human capital characteristics, however the self-employment premium separate them from the wage earners in terms of income.

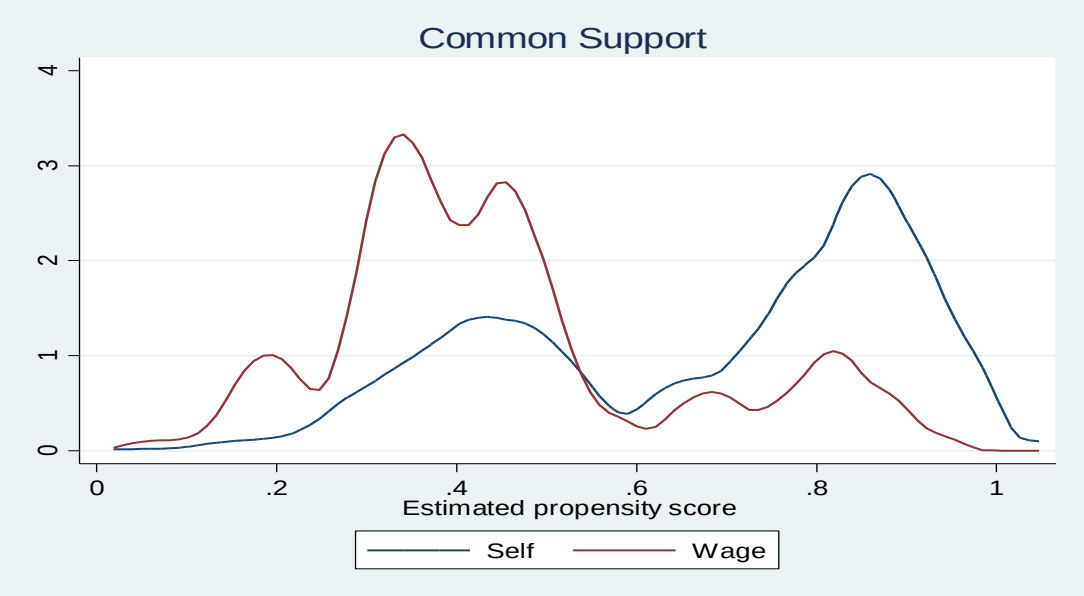
One policy implication of this result is that human capital characteristics are important for welfare consideration. Young people in developing countries are often encouraged to develop

entrepreneurial spirit as an alternative means of mitigating the effect of unemployment. Our result suggest that education and other human capital characteristics still matters even for self-employment activity.

Further Robustness Check: Checking the Plausibility the Common Support Assumption holds

Our result for the QTE and the RIF decomposition require common support assumption. One might wonder if the distribution of the population is such that there are only self-employed individuals at the upper end of the income distribution i.e. there is a support difference so that at the uppermost quantile only high-earning self-employed individuals exists with no comparable wage earner and the convers happens at the lower end. If this is the case comparison at the upper and lower ends will be invalid.

Figure 5: Propensity Score Distributions (Common Support Assumption)



Since the QTE method and the RIF makes use of propensity scores we check the common support assumption with the distribution of propensity scores. This is similar to what is done under Propensity Score Matching method. Fig 5 above shows the kernel density estimate of propensity score distributions in both groups. There seem to be two distinct populations. There are more wage earners on one side of the distributions (units with propensity score close to zero) and more self-employed on the other side (units with high propensity scores).

This pattern is not strange once one recall that individuals with low propensity scores are those that are less likely to be self-employed and the converse is also true at the upper end of the distribution. This sorting at different parts of the distribution suggests that individuals are matching themselves to sectors based on their characteristics but it also explain why excluding the first and the last quantiles in our previous result is the right thing to do. The result also give some comfort about the common support situation.

Conclusion

We showed that the income gap between self-employed and wage earners are heterogeneous over the income distribution after controlling for time invariant unobservables. The consequence of this is that the mean impact is misleading as it mask interesting differences in the wage gap along the income distribution. Decomposing the wage gap reveals that there is significant gain in being self-employed especially at the upper end of the distribution. Our result agrees with Tannuri-Pianto & Donald, (2002) in that self-employment is beneficial at the upper end of the income distribution.

Abowd, J. M., Kramarz, F. & Margolis, D. N., 1999. High wage workers and high wage firms. *Econometrica 67.2,* pp. 251-333.

Bargain, O. & Kwenda, P., 2014. The Informal Sector Wage Gap: New Evidence Using Quantile Estimations on Panel Data. *Economic Development and Cultural Change,* 1(63), pp. 117-153.

Bitler, M. P., Gelbach, J. B. & Hoynes, H., 2006. What mean impacts miss: distributional effects of welfare reform experiments. *American Economic Review Vol. 96,* p. 988–1012.

Blundell, R. & Bond, S., 1998. Initial Condition and Moment Restriction in Dynamic Panel Data Models. *Jornal of Econometrics 87,* pp. 115-143.

Borah, B. J. & Basu, A., 2013. Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health economics,* Issue 22.9, pp. 1052-1070.

Canay, I. A., 2011. A simple approach to quantile regression for panel data. *The Econometrics Journal,* 3(14), pp. 368-386.

Falco, P. et al., 2011. The returns to formality and informality in urban Africa. *Labour Economics 18,* pp. S23-S31.

Firpo, S., Fortin, N. & Lemieux, T., 2009. Unconditional quantile regressions. *Econometrica 77.3,* pp. 953-973.

Firpo, S., Fortin, N. M. & Lemieux, T., 2009. Unconditional quantile regressions. *Econometrica 77.3 ,* pp. 953-973.

Firpo, S., Nicole, F., Thomas & Lemieux, 1995. Labor market Institutions and the distribution of wages, 1973-1992. A semiparametric approach. *National Bureau of Economic Research.*

Firpo, S. P., 2007. Efficient semiparametric estimation of quantile treatment effects. *Econometrica Vol. 75,* p. 259–276.

Glick, P. & Sahn, D. E., 1999. Gender and education impacts on employment and earnings in West Africa: Evidence from Guinea. *Economic Development and Cultural Change 45.4,* pp. 793-823.

Hamilton, B. H., 2000. Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political economy 108.3,* pp. 604-631.

Heintz, J. & Lynda, P., 2012. The Determinants of Selection into Non-agricultural Self-employment in Ghana. *Margin: The Journal of Applied Economic Research 6.2,* pp. 181-209.

Kuepie, M., Nordman, C. J. & Roubaud, F., 2009. Education and earnings in urban West Africa. *Journal of Comparative Economics 37.3 ,* pp. 491-515.

Lazear, E. P. & Moore, R. L., 1984. Incentives, Productivity, and Labor Contracts. *The Quarterly Journal of Economics,* pp. 75-296.

Loayza, J. & Norman, R., 2006. *Informality trends and cycles.* Vol. 4078: World Bank Publications.

Maasoumi, E., Millimet, D. L. & Sarkar, D., 2009. Who Benefits from Marriage?. *Oxford Bulletin of Economics and Statistics 71.1,* pp. 1-33..

Maloney, W. F., 2004. Informality revisited. *World development 32.7,* pp. 1159-1178.

Rosen, S., 1981. The economics of superstars. *The American economic review,* pp. 845-858.

Tannuri-Pianto, M. & Donald, P., 2002. Informal employment in Brazil-a choice at the top and segmentation at the bottom: a quantile regression approach. *Anais do XXIV Encontro Brasileiro de Econometria 2.*

APPENDIX

Table 5A: Full result of Oaxaca-Blinder Decomposition of earning gap

lrearn97	Coef.	RobustStd. Err.	z	P>z	[95% Conf.	Interval]
Differential						
Prediction_1	2.623233	0.030489	86.04	0	2.563476	2.682991
Prediction_2	2.408008	0.025747	93.53	0	2.357546	2.458471
Difference	0.215225	0.039906	5.39	0	0.137011	0.293439
Explained						
exp	-0.46789	0.037912	-12.34	0	-0.5422	-0.39359
expsq100	-0.03592	0.018964	-1.89	0.058	-0.07309	0.001245
lhours	-0.00301	0.001674	-1.8	0.072	-0.00629	0.000274
lfirmsize	0.122199	0.013236	9.23	0	0.096257	0.148142
lemployees	-0.01545	0.005259	-2.94	0.003	-0.02575	-0.00514
tenure	0.071589	0.018012	3.97	0	0.036286	0.106893
tenuresq100	-0.07634	0.016025	-4.76	0	-0.10775	-0.04493
male	0.17373	0.01188	14.62	0	0.150446	0.197015
educ	0.805451	0.057098	14.11	0	0.693541	0.917362
educsq100	-0.3112	0.026057	-11.94	0	-0.36227	-0.26013
abind	0.104091	0.043574	2.39	0.017	0.018687	0.189494
year05	0.008548	0.003057	2.8	0.005	0.002557	0.014538
year06	5.72E-05	0.001787	0.03	0.974	-0.00345	0.00356
Total	0.375855	0.040671	9.24	0	0.296142	0.455568
Unexplained						
exp	-0.00079	0.096955	-0.01	0.993	-0.19082	0.189236
expsq100	0.026506	0.050031	0.53	0.596	-0.07155	0.124566
lhours	-0.24432	0.20714	-1.18	0.238	-0.6503	0.161671
lfirmsize	-0.00129	0.004777	-0.27	0.787	-0.01065	0.00807
lemployees	-0.00018	0.000718	-0.25	0.801	-0.00159	0.001226
tenure	0.012179	0.045639	0.27	0.79	-0.07727	0.10163
tenuresq100	-0.01388	0.023997	-0.58	0.563	-0.06091	0.033154
male	-0.01012	0.01683	-0.6	0.548	-0.0431	0.022869
educ	0.034667	0.156155	0.22	0.824	-0.27139	0.340725
educsq100	0.002442	0.08722	0.03	0.978	-0.16851	0.17339
abind	0.000875	0.003922	0.22	0.823	-0.00681	0.008562
year05	-0.01779	0.015291	-1.16	0.245	-0.04776	0.012181
year06	-0.03619	0.020669	-1.75	0.08	-0.0767	0.004324
_cons	0.087252	0.252415	0.35	0.73	-0.40747	0.581977
Total	-0.16063	0.022771	-7.05	0	-0.20526	-0.116

Figure 2A: Unconditional Quantile Regression

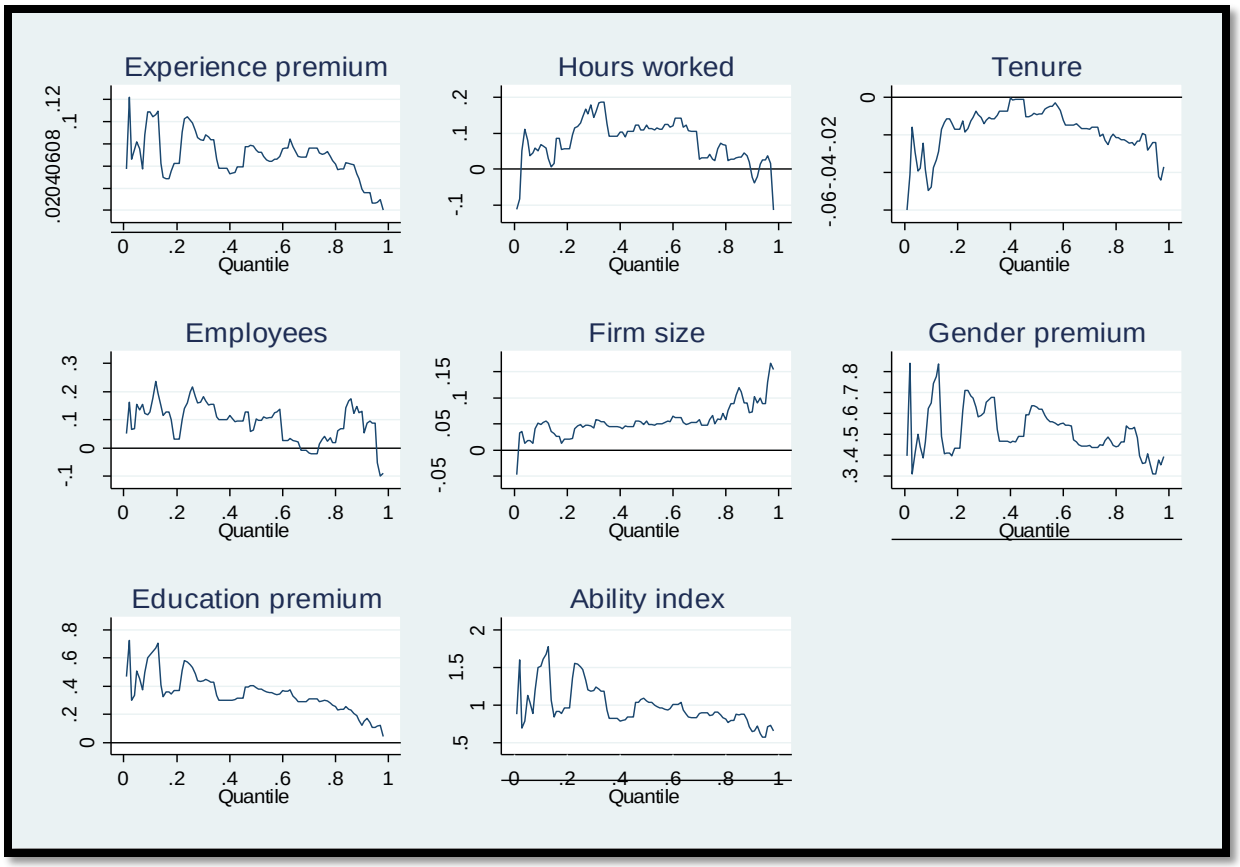


Table 5A: Probit model for determinants of self-employment

self	Male	-0.788
		(0.075)***
	educ_cat2	-0.463
		(0.167)***
	educ_cat3	-1.014
		(0.180)***
	educ_cat4	-0.716
		(0.314)**
	exp_cat2	0.589
		(0.519)
	exp_cat3	1.172
		(0.251)***
	exp_cat4	1.121
		(0.416)***
	exp_cat5	4.737
		(87.724)
	o.exp_cat6	0.000
		(0.000)
	tenure_cat2	0.516
		(0.086)***
	tenure_cat3	0.669

		(0.124)***
	tenure_cat4	0.662
		(0.222)***
	age_cat4	-0.162
		(0.138)
	age_cat5	-0.279
		(0.230)
	age_cat6	-0.687
		(0.545)
	exp_male2	-0.449
		(0.155)***
	exp_male3	-0.480
		(0.187)**
	exp_male4	0.420
		(0.358)
	exp_male5	-3.869
		(87.725)
	educ_exp2	0.008
		(0.561)
	educ_exp3	-0.121
		(0.321)
	educ_exp4	-0.437
		(0.423)
	educ_exp7	0.515
		(0.104)***
	educ_exp8	0.682
		(0.534)
	educ_exp9	-0.031
		(0.313)
	educ_exp10	-0.839
		(0.525)
	educ_exp13	0.842
		(0.126)***
	educ_exp14	0.734
		(0.500)
	educ_exp19	-0.391
		(0.438)
	2005	0.348
		(0.083)***
	2006	0.022
		(0.081)
	Constant	0.442
		(0.163)***
Pseudo R-squared		0.21
N		1,294
LR		689

* $p<0.1$; ** $p<0.05$; *** $p<0.01$
Probability of being Self Employment

Stochastic Dominance Results

Table 6A: Distributional test for Conditional and unconditional earnings

Distributions		First order Stochastic Dominance					First order Stochastic Dominance				
w	s	d_1, (MAX)	d_2, (MAX)	d	pr{d≤0} Simple bootstrap	pr{d*≥d} equal bootstrap	S_1, (MAX)	S_2, (MAX)	s	pr{s≤0} Simple bootstrap	pr{s*≥s} equal bootstrap
(A) W SSD S		0	3.186	0.00	0.64	**0.998**	0.00	443.6	0.00	**0.976**	**0.962**
(B)W SSD S		2.16	6.63	2.84	0.00	0.608	0.00	1131.5	0.00	**0.974**	**0.902**