THE GRADUATE INSTITUTE | GENEVA

INSTITUT DE HAUTES ETUDES
INTERNATIONALES ET DU DÉVELOPPEMENT

GRADUATE INSTITUTE OF INTERNATIONAL
AND DEVELOPMENT STUDIES

Graduate Institute of
International and Development Studies Working Paper
No: 10/2012

Regional Analysis of Eastern Province Feeder Road Project District level estimation of the Poverty Alleviation Effects of Rural Roads Improvements in Zambia's Eastern Province

Christian K.M. Kingombe

Overseas Development Institute / Graduate Institute of International Studies

Abstract

Remarkably little is known about the long-term impacts of project aid to lagging poor areas (Chen, Mu et al. 2006, 2008). This paper contributes to the debate about the role of rural transport infrastructure development in explaining the long-term rural development. In line with Grimm and Klasen (2008) we agree that there is value-added to consider this debate at the micro level within a country as particularly questions of parameter heterogeneity and unobserved heterogeneity are likely to be smaller than between countries. Moreover, at the micro level it is possible to identify more precise transmission mechanisms from rural transport infrastructure to socio-economic development outcomes. This is done empirically by analyzing a UNDP&UNCDF financed rural development project in Zambia's Eastern Province running from 1997-2002.

The secondary datasets consist of respectively a series of repeated cross-sectional living conditions monitoring surveys (LCMSs). The LCMSs were collected in 1998 (baseline) and 2004 (follow-up), that is both prior, during and after the project implementation. Our aim is to assess the ability of the parametric and semi-parametric models as well as using a time- series of cross-sections to provide an adequate description of the logarithm of per adult equivalent consumption of rural household conditional on few covariates, including an infrastructure treatment dummy variable. Although, the mean cotton sales share of household income has more than doubled despite the fact that the mean distance to the input market remained unchanged from 1998 to 2004, the parametric and semi-parametric estimation results are only small and statistically insignificant in terms of gains to mean consumption emerged in the longer-term. The main results are robust to corrections for various sources of selection bias.

Regional Analysis of Eastern Province Feeder Road Project

District level estimation of the Poverty Alleviation Effects of Rural Roads Improvements in Zambia's Eastern Province

Christian K.M. Kingombe

Overseas Development Institute / Graduate Institute of International Studies

Abstract

Remarkably little is known about the long-term impacts of project aid to lagging poor areas (Chen, Mu et al. 2006, 2008). This paper contributes to the debate about the role of rural transport infrastructure development in explaining the long-term rural development. In line with Grimm and Klasen (2008) we agree that there is value-added to consider this debate at the micro level within a country as particularly questions of parameter heterogeneity and unobserved heterogeneity are likely to be smaller than between countries. Moreover, at the micro level it is possible to identify more precise transmission mechanisms from rural transport infrastructure to socio-economic development outcomes. This is done empirically by analyzing a UNDP&UNCDF financed rural development project in Zambia's Eastern Province running from 1997-2002.

The secondary datasets consist of respectively a series of repeated cross-sectional living conditions monitoring surveys (LCMSs). The LCMSs were collected in 1998 (baseline) and 2004 (follow-up), that is both prior, during and after the project implementation. Our aim is to assess the ability of the parametric and semi-parametric models as well as using a time- series of cross-sections to provide an adequate description of the logarithm of per adult equivalent consumption of rural household conditional on few covariates, including an infrastructure treatment dummy variable. Although, the mean cotton sales share of household income has more than doubled despite the fact that the mean distance to the input market remained unchanged from 1998 to 2004, the parametric and semi-parametric estimation results are only small and statistically insignificant in terms of gains to mean consumption emerged in the longer-term. The main results are robust to corrections for various sources of selection bias.

Regional Analysis of Eastern Province Feeder Road Project
District level estimation of the Poverty Alleviation Effects of Rural Roads Improvements in Zambia's Eastern Province

Christian K.M. Kingombe[1]

Abstract:

Remarkably little is known about the long-term impacts of project aid to lagging poor areas (Chen, Mu et al. 2006, 2008). This paper contributes to the debate about the role of rural transport infrastructure development in explaining the long-term rural development. In line with Grimm and Klasen (2008) we agree that there is value-added to consider this debate at the micro level within a country as particularly questions of parameter heterogeneity and unobserved heterogeneity are likely to be smaller than between countries. Moreover, at the micro level it is possible to identify more precise transmission mechanisms from rural transport infrastructure to socio-economic development outcomes. This is done empirically by analyzing a UNDP&UNCDF financed rural development project in Zambia's Eastern Province running from 1997-2002.

The secondary datasets consist of respectively a series of repeated cross-sectional living conditions monitoring surveys (LCMSs). The LCMSs were collected in 1998 (baseline) and 2004 (follow-up), that is both prior, during and after the project implementation. Our aim is to assess the ability of the parametric and semi-parametric models as well as using a time- series of cross-sections to provide an adequate description of the logarithm of per adult equivalent consumption of rural household conditional on few covariates, including an infrastructure treatment dummy variable. Although, the mean cotton sales share of household income has more than doubled despite the fact that the mean distance to the input market remained unchanged from 1998 to 2004, the parametric and semi-parametric estimation results are only small and statistically insignificant in terms of gains to mean consumption emerged in the longer-term. The main results are robust to corrections for various sources of selection bias.

Keywords: Parametric and Semi-parametric Regression Models; Time-series Model from Successive cross sections; Cohort Data; Poverty Measures, poor rural area development projects, feeder roads, household surveys, impact evaluation, Zambia.

JEL: C14, C21, D12, I32, O1, Q1.

---

[1] **Acknowledgement:** The research reported here would not have been possible without the support of the Living Conditions Monitoring team of Zambia's Central Statistical Office, which kindly provided us with access to the two Living Conditions Monitoring Surveys used for our analysis. The paper benefited from comments from Michael Grimm and Salvatore di Falco as well as participants at the workshop on *"Poverty and Social Programmes"* held during ISS The Hague 6th Conference "Development Dialogue: Current Trends in Development Studies." We have also benefited from discussions with Colin Thirtle, Bhavani Shankar, Peter Hazell, Jonathan Kydd and Andrew Dorward.
Address for correspondence: Overseas Development Institute, 111 Westminster Bridge Road, London SE1 7JD and Visiting Research Fellow, Development Studies, Graduate Institute of International and Development Studies (IHEID).
E-mail: c.kingombe@odi.org.uk or Christian.kingombe@graduateinstitute.ch

Felix qui potuit rerum cognoscere causas.[2]
Virgil.

1. Introduction

Generally speaking the World Bank(2007a) recently in its' *"World Development Report, 2008"* entitled *"Agriculture for Development"* emphasized that it was because insufficient attention to the managing *the political economy of agricultural policies* to overcome policy biases and underinvestment as well as the governance challenges for the implementation of agricultural policies, that trade liberalization, increased investments in infrastructure and R&D in Africa were not fully implemented as recommended in its' 1982 World Development Report.

In the case of Zambia, despite the relative abundance of data and analysis there is much that needs to be understood about policies that promote rural growth and poverty reduction, as well as the distributional impact of public expenditure on rural roads.

More specifically, trade facilitated and augmented due to rural transport infrastructure improvements introduces new opportunities and new hazards. Households are affected both as consumers and as producers or income earners. As consumers, households are affected when there are changes in the prices of goods consumed by the family. As income earners, households are affected when there are responses in wages and in agricultural income (Winters 2002; Balat and Porto 2005b). Hence, targeting aid in the form of transport infrastructure development to poor areas has been an important vehicle for development assistance (Chen, Mu et al. 2006).

Our paper studies one such poor-area rural development project, namely the UNDP&UNCDF-supported Eastern Province Feeder Road Project (EPFRP). The development objective of the EPFRP was to contribute to the sustainable economic development of Zambia's Eastern Province through the establishment of a comprehensive integrated strategy for rural infrastructure development of around 404 km of deteriorated rural roads in five out of eight districts in Eastern Province. The achievement of this objective relies to the extent possible on locally available private sector resources and the technical and administrative capacity of selected district councils in Chadiza, Chipata, Katete, Lundazi, and

[2] Happy he, who could understand the causes of things. Georgics: Book 2, Line 490.

Petauke districts.[3] This was done using grants to improve access to and within the areas of the project determined in relation to economic potential and social activities of the respective areas of influence through the rehabilitation and sustainable use of rural roads in these selected districts.

Assessing the returns to this rural transport infrastructure project is problematic. Not only is its contribution to local agricultural production and trade difficult to measure and to attribute, it is difficult assessing precisely who benefits from community-level assets such as these feeder roads. Often, any indicators of impact on the poor are indirect (Devereux 2002).[4]

Thus, what happens with the local economic development at the moment inaccessibility no longer is perceived to be a major problem to economic activities due to the rehabilitation and maintenance of the feeder roads in the rural areas. This is the key question, which our paper is going to investigate. In other words, this paper tries to assess whether the EPFRP had any welfare impacts on the district levels both within the disbursement period (12th of June 1996 and 31st of December 2001) and beyond that period.

The attempt to *identify* to what extent the EPFRP contributed or constrained pro-poor rural consumption growth in five out of eight districts constituting Eastern Province is complicated by *confounding factors* such as the MMD Government's reform-based policies, which are difficult to separate from other policies more directly aimed at and typically associated with generating growth and poverty-reduction. In many respects *the limitation* of our paper, is that it can only attempt to explain whether certain households within the catchment districts of these rehabilitated feeder roads coped better than the non-beneficiary households in the counterfactual districts.

Following Chen, Mu et al. (2006) we must deal with the selective geographic placement of the EPFRP, whereby it was targeted to areas with particular agricultural potential and the resulting *selection bias* is unlikely to be time-invariant. There is also a concern about (time-varying) selection bias due to spill-over effects to non-participating areas (control districts). Some local spillover effects

[3] The remaining control districts: Chama, Mambwe and Nyimba were not encompassed by the project.
[4] Rural road networks are needed not only for transporting passengers and commodities, but also for market integration, which reduces price seasonality and enhances food security.

e.g. arising from the spending responses from the ZMK2,065 billion (US$480,233) that was paid in wages within the concerned districts of Eastern Province through the achievement of 870,000 workdays of the entire project are also expected to occur.

It is rare to assess project impacts by repeated panel observations over a relatively long period. A few recent exceptions include e.g. (Dercon and Hoddinott 2005; Chen, Mu et al. 2006; Dercon, Gilligan et al. 2007; Mu and van de Walle 2007). In our case much of the data required is readily available from the Living Conditions Monitoring Surveys (LCMS) conducted by Zambia's Central Statistical Office (CSO) on a regular basis.[5] Unfortunately, the LCMS dataset is only constituted by successive independent cross-sections collected in 1996, 1998, 2003, 2004, and 2006.

Our paper first applies one class of semi-parametric models known as *partially linear models*, which is also available for estimating conditional quantile.[6] There has been relatively little research done on the ability of semi-parametric models to fit datasets that are encountered in applications. Horowitz and Lee (2002); He, Fung et al. (2005); Yatchew (2005); and Zhoua, Zhub et al. (2008) all report the results of some investigations on this issue. On the whole, however, the usefulness of semi-parametric representations of real LCMS datasets remains hitherto largely unexplored. This approach is complemented by *cohort type approach*, where cohorts are defined by date of birth, to facilitate the identification of the gainers and losers from the EPFRP.

The remainder of this paper is organized as follows. Section 2 presents the background and the socio-economic setting as well as describes the LCMS datasets. Section 3 describes the various estimation methods used. Regression model estimation results are presented in Section 4. Section 5 discusses the estimation results, and section 6 concludes the paper.

[5] Caveat: Readers should be particularly cautious concerning the validity of our inferences based upon the analysis of the LCMS IV 2004 dataset, which is constrained by lack of unique identifiers (HID) in the original dataset. This has made it inevitable for us carry out a number of *data manipulation* in order to transform that dataset into a new dataset with all the usual negative consequences associated with these operations.
[6] Other classes are: Nonparametric additive models, and nonparametric additive models with interactions.

2. Background and setting

In this section we will exclusively focus on some of the relevant characteristics of Zambia's Eastern Province as well as the cross-sectional datasets that we will use in our analysis.

2.1. Eastern Province Characteristics

There are several unique features to the Eastern Province. The total number of agricultural holdings in Zambia was 520,520 in 1990, which rapidly increased to a total of 1,305,783 as of October 2000. Eastern Province accounted for 17.7 percent of these in 2000 down from 25% of the total in 1990. However despite experiencing the least percentage growth among Zambia's nine provinces, the agricultural households in Eastern Province still constitutes Zambia's largest population.[7] The 2000 census of population also found that the province had the highest number (231,120) and the second highest percentage of female headed households (19.8%). Along with Central and Southern Provinces both benefiting from the line of rail, the Eastern Province is considered to be among the most agriculturally advanced areas in Zambia. Moreover, Eastern Province has 6,910,000 hectares, of which only 10% was cultivated in the early 1990s (CSO 2001).

Of the 221,703 agricultural households in rural Eastern Province in 2000, the majority (38.6 percent of the total) were engaged in the three major agricultural activities: Crop growing, livestock- and poultry rearing while 28.8% were involved in crop growing and poultry rearing only. Only 20.6% exclusively engaged in crop growing. Moreover, there is potential for high yields of maize, tobacco, sunflower and several other annual crops of which the most relevant cash crop activity is *cotton* (due to soil characteristics) and the region produces a high share of many unmarketed crops (CSO 1994).[8]

Zambia has been classified into *4 broad agro-ecological region zones* on the basis of the average precipitation pattern and the quality of the soils.[9] Eastern Province has two distinct agro-ecological regions—the Eastern Plateau and the Luangwa Valley (see **Map A2** in appendix). Central, Southern and Eastern Plateau known as *agro-ecological region II* covers the Central, Southern and Eastern

[7] At the same time in 2000 Eastern Province had the lowest proportion of urban population at only 9 percent. Moreover, 4.1% of the agricultural households had an urban residence in 2000.
[8] See the specialized CSO report, which provides a statistical profile on agro-ecological zones in Zambia.
[9] Region I The Luangwa - Zambezi River valley zone. Region IIA The Central, Southern and Eastern Plateau. Region IIB the Western, semi-arid plains. Region III the Northern, North western high rainfall zone (Siacinji-Musiwa, 1999).

fertile plateau of Zambia (CSO 1994). It is characterised by: Moderate rainfall ranging from 800 and 1,000 mm of annual rainfall of which approximately 85 per cent falls during the four wettest months, i.e. December to March.[10] In Eastern Province, there is a clear difference between the dry season and the rainy season. The rainfall is concentrated between October/November and April/May, during the other months there is no rainfall at all.

Table 2.1: Rainfall (12-months moving avg.) (mm.), Agricultural season 1994/95 – 2004/2005

Year	1994/1995	1995/1996	1996/1997	1997/1998	1998/1999	1999/2000	2000/2001	2001/2002	2002/2003	2003/2004	2004/2005	Long-term Mean
Eastern	*528,55*	*805,71*	*813,72*	*719,02*	*759,38*	*678,55*	*914,78*	*700,40*	*835,42*	*781,11*	*788,61*	*756,84*
Chadiza (301) (i)	610,83	*868,50*	*945,33*	*708,50*	*901,42*	*584,83*	*1165,17*	*771,92*	*871,33*	*915,92*	*1007,92*	*850,15*
Chama (302) (iii)	494,08	*688,58*	*708,00*	*760,00*	*574,00*	*695,25*	*562,58*	*564,67*	*685,50*	*746,00*	*750,17*	*657,17*
Chipata (303) (i)	610,83	*868,50*	*945,33*	*708,50*	*901,42*	*584,83*	*1165,17*	*771,92*	*871,33*	*915,92*	*1007,92*	*850,15*
Katete (304)	214,17	*936,83*	*826,00*	*874,75*	*791,08*	*711,83*	*1163,50*	*846,17*	*877,47*	*878,01*	*895,40*	*819,56*
Lundazi (305)	693,58	*721,67*	*609,08*	*696,58*	*543,42*	*563,75*	*761,58*	*607,67*	*815,42*	*658,37*	*681,36*	*668,41*
Mambwe (306) (iii)	494,08	*688,58*	*708,00*	*760,00*	*574,00*	*695,25*	*562,58*	*564,67*	*685,50*	*746,00*	*750,17*	*657,17*
Nyimba (307) (ii)	555,42	*836,50*	*884,00*	*621,92*	*894,83*	*796,33*	*968,83*	*738,08*	*938,42*	*694,33*	*608,00*	*776,06*
Petauke (308) (ii)	555,42	*836,50*	*884,00*	*621,92*	*894,83*	*796,33*	*968,83*	*738,08*	*938,42*	*694,33*	*608,00*	*776,06*
Long-term Mean	756,84	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	*756,84*	

Notes: We assume the following coverage for the five weather stations: Chipata covers Chipata and Chadiza districts; Lundazi covers Lundazi district; Petauke covers Petauke and Nyimba districts; Msekere covers Katete district; and Mfuwe covers Chama and Mambwe districts. Lundazi and Katete (Msekere) was closed in respectively 2002/2003-2004/2005 and 2003/2004 - 2004/2005 due to lack of manpower. Hence, the figures in italic have been extrapolated as 5 year moving averages for these same years.
Source: Author's calculations based on Zambia Meteorological Service data.

Zambia Meteorological Service datasets include monthly rainfall from 1993 collected from weather stations in the four districts of Chipata (Chipata and Msekere stations); Lundazi; Petauke; and Mambwe (Mfuwe station). The Eastern Province time-series in **table 2.1** represents the provincial average, given our assumption. The years 1996/1997 and 2003/2004 were periods of above average rainfall levels, whereas the agricultural seasons 1997/98 and 2004/2005 were respectively below and above the average rainfall levels in Eastern Province.

2.2. The Zambian Living Conditions Monitoring Survey data

The CSO carried out the second Living Condition Monitoring Survey (LCMS II) in November-December, 1998, whereas the comparable LCMS IV in terms of the survey designs was conducted between October 2004 and January 2005 covering the whole country on a sample basis.[11] These

[10] In the valley areas, the rainy season tends to begin and end earlier than elsewhere.
[11] In order to have equal precision in the estimates in all the districts and at the same time take into account variation in the sizes of the district, the survey adopted the Square Root sample allocation method, (Lesli Kish, 1987). This approach offers a better compromise between equal and proportional allocation methods in terms of reliability of both combined and separate estimates. The allocation of the sample points (PSUs) to rural and urban strata was almost proportional.

household surveys contain no panel element and are simply repeated independent cross-sections or one-spot (single interview), which make welfare measures imprecise both due to sampling and non-sampling errors (e.g. under - or overestimation of household incomes and expenditures).

Table 2.2: Sample Allocation (Standard Enumeration Areas -Primary Sampling Units)

Surveys	Province	Total (% of total SEA)	Rural	Urban
LCMS IV**	Zambia	1048 (6.28% of 16,683)	n.a.	n.a.
2004	Eastern	n.a.	n.a.	n.a.
LCMS III**	Zambia	520 (3.12% of 16,683)	326	194
2002/2003	Eastern	60	50	10
LCMS II*	Zambia	820 (6.31% of 12,999)	492	328
1998	Eastern	n.a.	n.a.	n.a.
LCMS I*	Zambia	610 (4.69% of 12,999)	348	262
1996	Eastern	68	54	14

Source: Author based upon metadata of the LCMSs.
Notes: * The sampling frame developed from the 1990 census of population and housing.
** The sampling frame developed from the 2000 census of population and housing.

Households had been selected using *a two-stage stratified cluster sample design*. In the *first stage*, a sample of Standard Enumeration Area (SEA) was selected within each stratum (centrality) according to the number allocated to that stratum. Selection had been done systematically with probability proportional to the number of households within each SEA as registered in the 1990 / 2000 Population Census.[12] In the *second stage* in each selected SEA, households were listed and each eligible household was given a unique sampling serial number.[13] Given the stratified surveys the ex-ante probability of being surveyed is not constant across households (i.e. the sample is not purely probabilistic) and some re-weighting must take place before obtaining population's estimates. Each household has to be re-weighted with the inverse of its probability of being sampled.[14]

Economic Activities in Rural Eastern Province

Agriculture is the overwhelming dominant activity in the rural areas in Eastern Province. In 63% of the sampled rural households, the principal activity of the household head is farming in 1998. However, in 2004 the overall share had lowered, while increasing for the upper quintile (**table 2.3**).

[12] Sample allocation was done using the "Probability Proportional to size" (PPS) method. This entailed allocating the total sample (1048 / 820) proportionately to each province according to its population share. Thereafter, allocation of the provincial sample was done proportionately to each district according to the population share from the provincial population. Similarly allocation was done by centrality within a district.

[13] Sample selected from roster of household members was obtained from a responsible adult household member. This may lead to unequal weighting in order to account for household size.

[14] Hence while making statistical inferences about the population we should account for *survey design effects*, i.e. re-weight the sample accordingly to get the correct point estimates and take into account stratification and clustering to get the correct standard deviation (see, Deaton, 1997).

Table 2.3: Principal Economic Activity of Household Head, Rural Areas, Numbers of Household Heads by Quintile of Consumption

	1998			2004		
	All	Poorest 20%	Richest 20%	All	Poorest 20%	Richest 20%
In Wage Employment	3%	0,70%	22,92%	3,07%	1,40%	6,82%
Running a Busines/Self Employed	3,01%	1,63%	6,25%	1,20%	0,47%	1,62%
Farming, Fishing, Forestry	63,09%	67,13%	35,42%	54,80%	58,97%	53,57%
Piecework	n.a.	n.a.	n.a.	0,93%	0,93%	0,65%
Unpaid family workers	n.a.	n.a.	n.a.	0,27%	0,00%	0,32%
Not working but looking for work/means to do business	1,16%	0,93%	0,00%	0,40%	0,00%	0,32%
Not working and not looking for work/means to do business but available or wishing to do so	7,49%	0,58%	0,00%	0,27%	0,23%	0,65%
Full time student	14,98%	13,59%	20,83%	26,20%	25,17%	24,68%
Ful time at home/home duties (homemaker)	7,49%	7,43%	10,42%	2,00%	1,17%	0,97%
Retired	0,00%	0,00%	0,00%	0,07%	0,00%	0,00%
Too old to work	1,47%	1,74%	0,00%	1,47%	1,17%	0,97%
Other	5,10%	6,16%	0,00%	9,33%	10,49%	9,42%
Total	1295	861	48	1500	429	308

Source: Author's calculations.

Material Assets

The large majority of sampled rural households again fall within the bottom quintile. Of the assets listed, only residential building, radios, bicycles and basic farm tools such as a plough and crop sprayer are owned by more than 10 percent of the sampled rural households. Motorized vehicles are practically non-existent; instead 10% of the poorest rural households (those in the bottom quintile) own a scotch cart in 1998.

Table 2.4: Percentage of Households in Rural Areas Owning Particular Assets by Quintile

	1998			2004		
	All	**Poorest 20%**	**Richest 20%**	**All**	**Poorest 20%**	**Richest 20%**
1.1. Plough	18,73%	19,13%	10,20%	25,05%	23,49%	23,74%
1.2. Crop sprayer	12,39%	12,71%	8,16%	17,47%	17,67%	15,73%
1.3. Fishing Boat	0,15%	0,12%	0,00%	0,00%	0,00%	0,00%
1.4. Canoe	0,69%	0,24%	0,00%	1,44%	0,67%	2,97%
Brazier / Mbaula	n.a.	n.a.	n.a.	36,57%	28,41%	37,69%
1.5. Fishing net	3,13%	2,42%	2,04%	1,50%	2,46%	0,59%
1.6. Bicycle	50,84%	51,09%	59,18%	56,42%	51,68%	49,85%
1.7. Motor cycle	1,30%	1,45%	2,04%	0,13%	0,00%	0,00%
1.8. Motor vehicle	1,83%	1,94%	2,04%	1,57%	0,67%	1,48%
1.9. Tractor	0,38%	0,24%	0,00%	0,44%	0,00%	0,00%
1.10. Television	1,61%	1,21%	0,00%	5,32%	2,01%	7,72%
1.11. Video player	0,84%	0,36%	0,00%	3,07%	0,22%	4,15%
1.12. Radio	43,88%	43,58%	48,98%	50,53%	48,10%	49,55%
Grinding/Hammermill (powered)	n.a.	n.a.	n.a.	2,19%	2,46%	2,67%
1.13. Electric (and non-electric) iron	1,15%	0,97%	2,04%	19,54%	11,63%	25,22%
1.14. Refrigerator/Deep freezer	0,61%	0,24%	2,04%	1,88%	0,45%	3,26%
1.15. Telephone (including cellular phone)	0,08%	0,00%	0,00%	1,57%	0,67%	0,30%
1.16. Sewing/knitting machine	7,11%	6,17%	2,04%	4,88%	2,68%	8,01%
1.17. (Electric/gas) Stove/cooker	1,53%	1,09%	2,04%	1,69%	0,67%	2,97%
1.18. Non-residential building	2,60%	2,18%	0,00%	2,44%	2,91%	1,48%
1.19. Residential house/building	85,17%	82,45%	97,96%	87,98%	91,50%	85,76%
1.20. Scotch cart	9,33%	9,93%	6,12%	13,34%	9,84%	12,46%
1.21. Donkeys	0,31%	0,36%	0,00%	1,63%	2,68%	2,67%
Oxens	n.a.	n.a.	n.a.	19,66%	11,63%	22,55%
Total number of households in Eastern Province sample	**1308**	**826**	**49**	**1597**	**447**	**337**

Source: Author's calculations.

If we take a closer look at the district level (cf. **Table A4.a-b**) we find that the catchment districts fare much better than the control districts at both the lowest quintile as well as the two highest quintiles in terms of improvements to the assets base with regards to a much steeper increased ownership of: Ploughs, crop sprayer, bicycle and a scotch cart. However, the rural households in the catchment districts experienced a fall in the ownership of motor cycles and motor vehicles in the same period from 1998 to 2004, with this kind of asset ownership being practically non-existent in the control areas.

Access to Infrastructure, Service and Community Assets: Distance to Markets

In 1998, on average the rural households in Eastern Province had to travel almost 40 km to reach an agricultural input market, which sell fertilizer and seeds. This distance had fallen significantly to around 13 km in 2004, although with no noteworthy differences between rural households belonging to separate consumption quintiles in 2004.

Table 2.5: Mean distance to services and Community Assets, by Household, Rural Areas

		Quintile of Provincial Distriction, 1998						Quintile of Provincial Distriction, 2004				
	All	Poorest 20%	2	3	4	Richest 20%	All	Poorest 20%	2	3	4	Richest 20%
1.1. Food Market	20,4	20,4	52,0	n.a.	8,0	27,5	9,5	10,3	9,2	9,8	9,2	8,8
1.2. Post Office/postal agency	32,1	32,2	52,0	n.a.	10,0	27,5	13,1	12,9	11,8	13,0	12,3	15,5
1.3. Primary School	3,3	3,3	4,0	n.a.	n.a.	1,5	3,8	3,4	2,4	2,9	6,0	4,3
Distance to Low Basic School(1-4)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2,3	1,7	2,1	1,7	2,8	3,1
Distance to Middle Basic School(1-7)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3,0	2,8	2,9	2,5	2,8	3,7
Distance to Upper Basic School(1-9)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	4,3	4,2	4,1	4,6	4,1	4,4
Distance to High School	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	28,8	27,7	26,9	33,1	30,6	27,5
1.4. Secondary School	32,5	32,6	52,0	n.a.	n.a.	29,0	16,4	15,5	16,8	18,7	15,9	16,1
1.5. Health Facility (Health post/Centre/Clinic/Hospital)	15,9	16,0	24,0	n.a.	6,0	13,0	6,1	5,9	6,1	6,2	6,4	6,1
1.6. Hammermill	6,0	6,0	0,0	n.a.	0,0	0,5	4,1	4,3	3,4	3,6	4,0	4,9
1.7. Input market (for seeds, fertilizer, agricultural implements)	40,0	40,3	52,0	n.a.	6,0	29,0	12,6	11,8	12,1	13,5	12,5	13,6
1.8. Police station/post	34,7	34,8	52,0	n.a.	n.a.	29,0	10,9	10,5	10,8	10,9	10,4	12,0
1.9. Bank							21,7	21,4	20,5	23,4	22,1	21,9
1.10. Public transport (road, or rail, or water transport)	8,6	8,6	47,0	n.a.	3,0	29,0	5,7	5,7	5,4	5,4	5,1	6,8

Source: Author's calculations.

3. Theoretical Framework and Estimation methods

In this section we first present the analytical framework. Then we turn to a description of *the range of models* addressed in our paper to estimate the conditional mean function of the logarithm of the total per adult equivalent (p.a.e.) expenditure of rural household in Eastern Province.

3.1. Analytical Framework

An analytical framework has been constructed to *identify* at the national level the various channels through which *price changes* associated with the removal of border trade barriers (analogous to the creation of transport network connectivity) are "passed through" the economic system to influence the welfare of richer and poorer households (McCulloch, Baulch et al. 2001; Winters 2002; UNCTAD 2004; Winters, McCulloch et al. 2004).

We are focusing on households located in the rural areas of Eastern Province, which all were poorly served by transportation infrastructure as well as being plagued by high marketing costs prior to the launch of the EPFRP in 1996 (Chiwele, Muyatwa-Sipula et al. 1998). Within this analytical framework, transport network improvement through the EPFRP is seen as *a price shock*, which has:

- *Expenditure effects* arising because of changes in the prices of the goods that are consumed; and
- *Income and employment effects* arising because of changes in the remuneration of factors of production.

Assessing the impact of rural transport infrastructure improvement on poverty in Eastern Province is not an easy task as emphasized by Winters (2002) *"Tracing the links between trade [facilitated by rural transport infrastructure improvements] and poverty is going to be a detailed and frustrating task, for much of what one wishes to know is just unknown."* Since most of the links are very case specific we narrow the scope by carrying out an in-depth study of the EPFRP focusing on the longer-term impacts in the attempt to make these transmission mechanisms a little less opaque.

The best way of thinking about poor self-employed rural households is in terms of the "farm household," which produces goods or services, sells its labour and consumes (Inderjit Singh, Lyn Squire, and John Strauss 1986 referred to in Winters et al., 2004). An increase in the price of something of which the household is a net seller (labour, goods, services) increases its real income, while a decrease reduces it. Winters et al., (2004) argue that the framework needs to ask how trade liberalization [or termination of geographic isolation through rural roads improvement in our case] affects all of the different sources of income, as well as considering consumption. Winters et al. (2004) further argue that if *price changes* are an important pathway through which liberalization affects the poor, then we must ask how trade liberalization and/or trade facilitation through rural roads improvement affects prices.

More important than price changes according to Winters et al. (2004) is whether markets exist at all: Trade reform and/or the opening of previously inaccessible remote rural areas can both create and destroy markets. Winters et al, (2004) mention that a common worry is that opening up an economy, e.g. through rural roads rehabilitation, will expose it and its component households to increased risk. Certainly, it will expose them to new risks, but the net effect can be to reduce overall risk because world markets (which have many players) are often more stable than domestic ones.

The application of this analytical framework evidently entails a number of shortcomings. In addition, our discussion is conducted entirely at the level of the household. The starting point for the *definition of a household* is the SNA93. In Zambia a household is described as follows:

> *A household is a group of persons who normally live and eat together. These people may or may not be related by blood, but make common provision for food or other essentials for living and they have only one person whom they all regard as the head of the household. A household may also consist of one member (CSO 2003).*[15]

In the absence of an internationally applied definition of a household, the definition of a household to be adopted in this paper is that used in Zambian national household surveys. This will be based on the single-dwelling concept.

3.2. Models and Estimators

We consider three basic approaches to estimate of our welfare measure the logarithm of p.a.e. consumption. The first ***semi-parametric approach*** maintains the functional form assumptions but

[15] Collecting data using households as the unit of analysis obscures intra-household inequalities in income and consumption.

partially relaxes the distribution assumption. The second approach is ***parametric*** and is based on strong assumptions about the conditional data distribution and functional forms. The final approach shows the variance of the logarithm of consumption as measured by tracking randomly selected representatives of semi-aggregated cohorts defined by date of birth through ***a time series of cross sections***.

3.2.1. Semiparametric models

Nonparametric regression estimators are very flexible but their statistical precision decreases greatly if we as in our case include a vector **x** of a dimension exceeding two explanatory variables in the model. The latter caveat has been appropriately termed *the curse of dimensionality*. Consequently, researchers have tried to develop what is usually referred to as *semiparametric* models and estimators, which offer more flexibility than standard parametric regression but overcome the curse of dimensionality by employing some form of *dimension reduction*. Such methods usually combine features of parametric and nonparametric techniques to yield the semi-parametric regression model that could help obtain consistent estimates of the parameters of interest.[16]

Further advantages of semiparametric methods are the possible inclusion of categorical variables (which can often only be included in a parametric way), an easy (economic) interpretation of the results, and the possibility of a part specification of a model. One such example is *the partial linear model*, which takes the pragmatic point of fixing the error distribution but let the index be of non – or semiparametric structure (Härdle, Müller et al. 2004).

According to (Lokshin 2006) the econometric problem of estimating *a partial linear model* arises in a variety of settings. We apply the partial linear regression technique to estimation of the household consumption of poverty attributes. Parametric variables include household size; age and education of head of household; asset ownership; distance to input market area; and cotton share of total household income. The (EPFRP) location effect, which has no natural parametric specification, is incorporated non-parametrically.

[16] That is, when estimating semiparametric models we usually have to use nonparametric techniques.

3.2.2. Partially linear models

Thus, as we extend beyond two dimensions, one compromise is *the partially linear regression model* (PLRM) originally studied by Robinson (1988), which has the form:

(3.1)	y_i	=	$X_i\beta + f(z_i) + \varepsilon_i$
(3.2)	$E[\varepsilon \mid x, z]$	=	0
(3.3)	$E[y \mid x, z]$	=	$X_i\beta + f(z_i)$,

where one part of the model is linear - the **Xs** - and a single variable has potentially nonlinear relationship with **y**. The *p*-dimensional random variable $\mathbf{x} \in \mathbf{R}^{\mathbf{p}}$ and the random variable $\mathbf{z} \in \mathbf{R}^{\mathbf{q}}$ do not have common variables, that is, they are non-overlapping sub-vectors. If they do, then the common variables would be regarded as part of **z** but not **x** and the coefficients that correspond to the common variables would not be identifiable.[17] $\boldsymbol{\varepsilon}_{\boldsymbol{i}}$ is i.i.d. mean-zero error term, such that $Var[y \mid x, z] = \sigma^2_\varepsilon$. The function ***f*** is a smooth, single valued function with a bounded first derivative. Standard errors of $\boldsymbol{\beta}$ are adjusted according to the (Yatchew 1998) method.

If there is no cross terms of **z** among **x**'s, then the model presumes additive separability of non-parametric $\mathbf{f(z_i)}$ and the parametric $\mathbf{X_i\beta}$ parts, which may be too restrictive in some applications. In the PLRM, the convergence rate $\mathbf{n^{-1/2}}$ depends only on the number of continuous regressors among **z** (Hidehiko 2005).[18] Thus, the curse of dimensionality is avoided in estimating $\boldsymbol{\beta}$.

According to (Lokshin 2006) following the methodology suggested by (Yatchew 1998), to estimate the *partial linear model* (3.1) we first rearrange (sort) the data in such a way that $z_1 < z_2 < \ldots < z_T$ where ***T*** is the number of observations in the sample. Then the *first difference* of (3.1) results in:

(3.4) $$(y_{i(n)} - y_{i(n-1)}) = (f(z_{i(n)} - f(z_{i(n-1)}) + \beta(x_{i(n)} - x_{i(n-1)}) + \varepsilon_{i(n)} - \varepsilon_{i(n-1)}, \qquad n = 2,\ldots, T.$$

[17] That is the *identification of β* requires the exclusion restriction that none of the components of **X** are perfectly predictable by components of **z**.
[18] The nonparametric method would be to use y = m(x, z) + ε model. The convergence rate would depend on the number of continuous regressors among (x, z).

When the sample size increases, $f(z_{i(n)} - f(z_{i(n-1)}) \rightarrow 0$ because the derivative of **f** is bounded. Under standard assumptions, equations (3.4) could be estimated by the ordinary least squares (OLS). The vector of estimated parameters $\hat{\beta}_{Diff}$ has the approximated sampling distribution:

(3.5) $$\hat{\beta}_{diff} \rightarrow N\left(\beta, \frac{1}{T}, \frac{1.5\sigma_{\varepsilon}^{2}}{\sigma_{u}^{2}}\right)$$

Where σ^2_u = Var[x | z] is conditional variance of ***x*** given ***z***. The error term in (3.4) has a MA(1) structure, thus reducing efficiency of the OLS estimator. The efficiency could be improved by using higher order differences (Yatchew, 1998). The generalization of (3.4) for the *mth*-order differencing can according to Lokshin (2006) is expressed as:

(3.6) $$\sum_{j=1}^{m} d_j y_{i-j} = \beta\left(\sum_{j=1}^{m} d_j x_{i-t}\right) + \sum_{j=1}^{m} d_j f(z_{i-j}) + \sum_{j=1}^{m} d_j v_{i-j}$$

Where $d_0, \ldots, d_m$ are differencing weights satisfying the conditions:

(3.7) $$\sum_{j=1}^{m} d_j = 0 \qquad \text{and} \qquad \sum_{j=1}^{m} d_j^2 = 1$$

The first condition in (3.7) ensures that the differencing removes the non-parametric component in (3.6) as the sample size increases. The second normalization condition implies that the residuals in (3.6) have variance of σ^2_u. With the optimal choice of weights equation (3.6) could be estimated by OLS. By selecting ***m*** sufficiently large, the estimator approaches asymptotic efficiency.

Define $\Delta\mathbf{y}$ to be the (T–m) x1 vector with elements $[\Delta y]_i = \sum_{j=1}^{m} d_j y_{i-j}$ and $\Delta\mathbf{x}$ to be the (T–m)xp matrix with elements $[\Delta x]_i = \left(\sum_{j=1}^{m} d_j x_{i-j}\right)$, then:

(3.8) $$\hat{\beta}_{Diff} = (\Delta x'\Delta x)^{-1}\Delta x'\Delta y \quad \rightarrow \quad N(\beta, \frac{1}{T}(1+\frac{1}{2m})\sigma_{\varepsilon}^{2}\sum_{(x|z)}^{-1}$$

(3.9) $$s_{diff}^{2} = \frac{1}{T}(\Delta y - \Delta x\,\hat{\beta}_{diff})'(\Delta y - \Delta x\,\hat{\beta}_{diff}) \rightarrow \sigma_{\varepsilon}^{2}$$

(3.10) $$\sum_{(x|z)} = \frac{1}{T}(\Delta x)'(\Delta x) \rightarrow \sum_{(x|z)}$$

This method allows performing inferences on **β** as if there were no non-parametric component ***f*** in the model. Once $\hat{\beta}_{Diff}$ is estimated, a variety of non-parametric techniques could be applied to estimate ***f*** as if **β** were known. Formally, subtracting the estimated parametric part from both sides of (3.1), we get:

$$y_i = x_i \hat{\beta}_{Diff} = x_i (\beta - \hat{\beta}_{Diff}) + f(z_i) + \varepsilon_i \approx f(x_i) + \varepsilon_i \quad (3.11)$$

Because $\hat{\beta}_{Diff}$ converges sufficiently quickly to true **β**, the consistency, optimal rate of convergence, and construction of confidence intervals for ***f*** remain valid and could be estimated by the standard smoothing methods (Lokshin 2006).

Using estimates (3.8) it is possible to perform the differencing test for the parametric specification of ***f***. Suppose $g(z, \pi)$ is the known function of z and some unknown parameter π. We want to test the null hypothesis that $y_i = g(z_i, \pi) + x_i\beta_p$ against the alternative hypothesis that $y_i = f(z_i) + x_i\beta$. Parameters π and β_p and mean square residual could be obtained by estimating the parametric regression of y on x and z. Then:

$$V = \sqrt{mT}\left(\frac{s_{res}^2 - s_{diff}^2}{s_{diff}^2}\right) \rightarrow N(0,1). \quad (3.12)$$

3.2.3. Model using a Times Series of Cross Sections

Finally, we use the same two cross-country data sets (1998 and 2004) to follow cohorts of groups over time from one survey to another. No assumptions are made about the distribution of household consumption in the population; what is being estimated is simply the average consumption in the population in the survey year, not the parameter of a distribution (Deaton 1997).

In **table 3.1** below the 19 rural constituencies in Eastern Province are used to make comparisons over time. These comparisons are often referred to as measurement of net changes in the population (Glewwe and Jacoby 2000).[19] These averages, which relate to households living in the same

[19] However, repeated cross-sectional data reveal nothing about the movements of individuals or households over time, often referred to as gross changes, because different households and individuals are interviewed in each survey. Measurement of

constituency, have many of the properties of panel data. This approach enables us to address the key question about the gainers and losers in the catchment and counterfactual districts from EPFRP by following the same *'Constituency or Age'* cohort over time. In this quest it is necessary to verify that the measured changes are statistically significant and thus unlikely to be caused by chance alone (Glewwe and Jacoby 2000). Hence, in **table 3.1** we compare the case of two independent cross-sectional surveys carried out in the same Constituency. The change in average consumption, say, would be estimated by the difference in average consumptions in 1998 and 2004. The variance of the estimate would be the sum of the variances in the two periods because each cross-sectional sample is drawn independently (Deaton 1997).[20]

Because, our Constituency and Age Cohort data are constructed from two fresh samples, there is no attrition. The way the constituency/age data are used will often be less susceptible to measurement error than in the case with panels. The quantity that is being tracked over time is the mean (average) and the averaging will nearly always reduce the effects of measurement error and enhance the signal-to-noise ratio. In this sense, the constituency/age method can be regarded as *an IV method* (Deaton 1997).

Despite the fact that we record that two catchment constituencies experienced a negative *annual growth rate per capita*, of the 19 constituencies in Eastern Province the four control constituencies are only ranked respectively 12, 15-17. The rankings are almost equivalent when measured in *p.a.e.* terms. The catchment constituencies also largely outperformed the control constituencies with regards to the *change in poverty gap*, where the latter again only are ranked 14, 16, 17 and 19 far behind the catchment constituencies.

---

gross changes over time can only be addressed using panel data. For example, random measurement error in household income or expenditures will exaggerate movements into and out of poverty over time.

[20] The greatest precision will be obtained from a panel, a rotating panel, or independent cross sections depending on the degree of temporal autocorrelation in the quantity being estimated. The higher the autocorrelation, the larger the fraction of households that should be retained from one period to the next.

Table 3.1: Ranking of Consumption Growth and Poverty Change at the Constituency level, 1998 & 2004

Location level	District	301		302		303				304			305			306	307	308		
	Constituency	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
Year	1998	45	75	45	30	149	94	15	15	59	46	44	77	14	78	105	121	75	105	54
	2004	64	38	11	12	66	22	58	53	36	56	47	81	51	77	74	85	81	54	33
1998	Consumption pc	17905,2	13686,3	25703,2	14311,4	12887,3	44972,0	5717,5	11254,4	15539,4	39104,0	8606,9	32889,3	19234,3	34740,4	23470,2	26129,9	9809,9	22116,9	12311,0
2004	Consumption pc*	42230,6	37086,3	26397,4	17856,1	22802,3	30435,6	52567,3	50329,6	21482,0	27012,8	25784,3	55100,3	35997,7	47037,8	33772,9	23092,0	21270,5	35461,7	49565,4
2004-1998	Growth in pc	24325,4	23400,0	694,1	3544,7	9915,0	-14536,3	46849,8	39075,2	5942,6	-12091,2	17177,4	22211,0	16763,4	12297,4	10302,7	-3037,9	11460,6	13344,9	37254,4
2004-1998	Annual growth in pc	15,4%	18,1%	0,4%	3,8%	10,0%	-6,3%	44,7%	28,4%	5,5%	-6,0%	20,1%	9,0%	11,0%	5,2%	6,3%	-2,0%	13,8%	8,2%	26,1%
Rank of Constituency wrt Growth		6	5	16	15	9	19	1	2	13	18	4	10	8	14	12	17	7	11	3
1998	Consumption pae	17988,8	13234,1	25259,7	13194,6	12277,4	44417,0	5533,0	9791,8	14861,8	38709,7	8345,7	32841,2	17424,7	33812,9	22808,2	25745,6	9528,4	21714,1	11521,6
2004	Consumption pae*	59969,9	32425,8	26087,4	17818,7	21771,4	35798,4	48661,6	57187,1	26486,1	26632,0	28764,7	53013,7	30298,2	62419,4	32035,4	29284,8	21577,7	19700,9	52942,2
2004-1998	Growth in pae	41981,1	19191,7	827,7	4624,0	9494,0	-8618,6	43128,6	47395,3	11624,3	-12077,7	20419,0	20172,5	12873,5	28606,5	9227,1	3539,2	12049,3	-2013,3	41420,7
2004-1998	Annual growth in pae	22,2%	16,1%	0,5%	5,1%	10,0%	-3,5%	43,7%	34,2%	10,1%	-6,0%	22,9%	8,3%	9,7%	10,8%	5,8%	2,2%	14,6%	-1,6%	28,9%
Rank of Constituency wrt Growth		5	6	16	14	10	18	1	2	9	19	4	12	11	8	13	15	7	17	3
1998	Below Food Poverty Line pae	6176,1	10930,8	-1094,8	10970,3	11887,5	-20252,1	18631,9	14373,2	9303,1	-14544,8	15819,3	-8676,3	6740,2	-9647,9	1356,7	-1580,7	14636,5	2450,8	12643,4
2004	Below Food Poverty Line pae	-35805,0	-8260,8	-1922,5	6346,3	2393,5	-11633,5	-24496,7	-33022,2	-2321,2	-2467,1	-4599,7	-28848,8	-6133,3	-38254,5	-7870,4	-5119,8	2587,2	4464,1	-28777,3
1998-2004	Change in Poverty Gap	-41981,1	-19191,7	-827,7	-4624,0	-9494,0	8618,6	-43128,6	-47395,4	-11624,3	12077,7	-20419,0	-20172,5	-12873,5	-28606,5	-9227,1	-3539,2	-12049,3	2013,2	-41420,7
Rank of Constituency wrt Poverty Change		3	8	19	16	13	15	2	1	12	10	6	7	9	5	14	17	11	18	4
1998	Below Total Poverty Line pae	15573,6	20328,3	8302,7	20367,8	21285,0	-10854,6	28029,3	23770,6	18700,6	-5147,3	25216,7	721,1	16137,7	-250,5	10754,1	7816,8	24034,0	11848,3	22040,8
2004	Below Total Poverty Line pae	-26407,5	1136,6	7475,0	15743,7	11791,0	-2236,0	-15099,2	-23624,7	7076,3	6930,3	4797,7	-19451,3	3264,2	-28857,0	1527,0	4277,6	11984,7	13861,5	-19379,9
1998-2004	Change in Poverty Gap	-41981,1	-19191,7	-827,7	-4624,0	-9494,0	8618,6	-43128,6	-47395,3	-11624,3	12077,7	-20419,0	-20172,5	-12873,5	-28606,5	-9227,1	-3539,2	-12049,3	2013,3	-41420,7
Rank of Constituency wrt Poverty Change		3	8	19	16	13	15	2	1	12	10	6	7	9	5	14	17	11	18	4

Notes: * Constant values in 1998 prices.
Source: Authors' computations.

Fig.3.1: Household Expenditure by Age, 1998 & 2004.

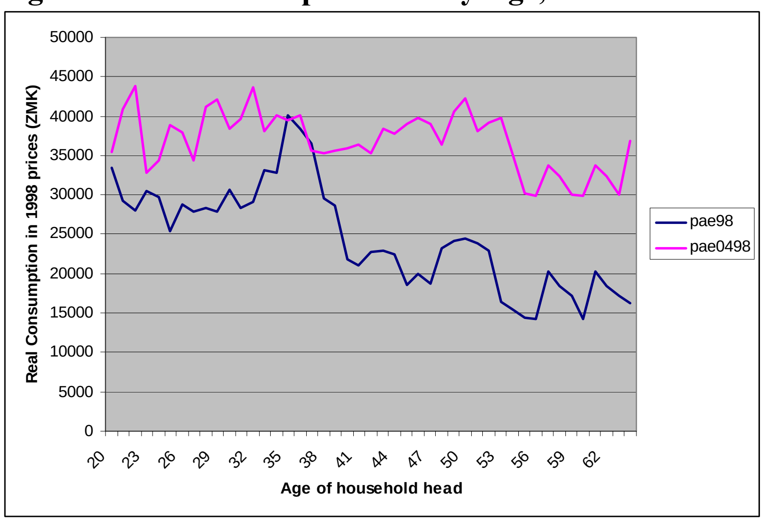
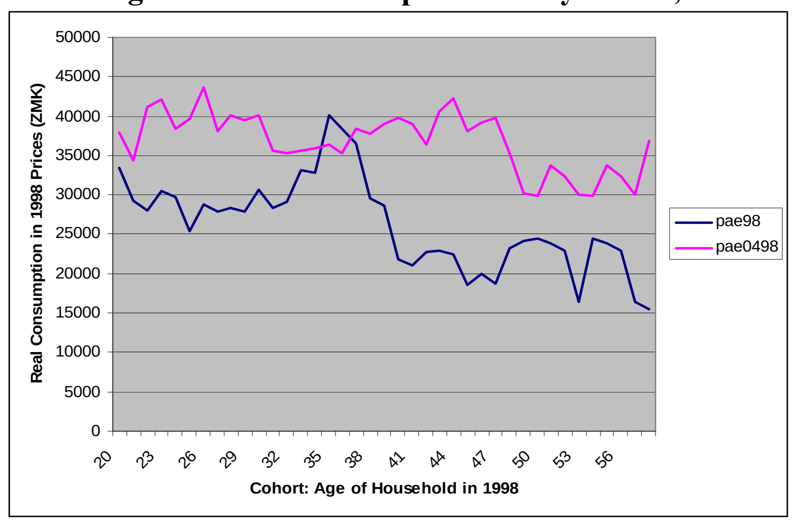


Fig.3.2: Household Expenditure by cohort, 1998 & 2004



Notes: Five-year moving averages are shown in the figure.
Source: Authors' computations.

In **table 3.2** we show that we can also use the survey data to follow cohorts of groups over time, where cohorts are defined by *date of birth*.[21] We use the successive LCMSs to follow each cohort over the 6 year period from 1998 to 2004 by looking at the members of the cohort who are randomly selected into each survey. For example, we look at *the mean p.a.e. consumption of 20-year-olds* in the 1998 survey, *of 26-year-olds* in the 2004 survey, etc. These averages, because they relate to the same group of people, have many of the properties of panel data.[22] Because there are many cohorts alive at one time, cohort data are more diverse and richer than the aggregate constituency data, but the semi-aggregated structure provides a link between the micro-level household level and the macro-economic data from national accounts.

Table 3.2 shows, that there were 1287 members of the cohort in the 1998 survey and 907 in the 2004 survey (in which the sample size was increased for reasons explained above). The table also illustrates the same process for ten cohorts, born in 1978, 1974 and 1973, 1969, and so backward at five-year intervals until the oldest, which was born in 1933, 1929, and the members which were from 65 to 69 years old in 1998. Tracking different cohorts through successive surveys allows us to disentangle the generational from life-cycle components in consumption profiles as shown in **figure 3.1** and **figure 3.2** (Deaton 1997).

Table 3.2: Number of Persons in Selected Cohorts by Survey Year, 1998 & 2004

	Cohort: Age in 1998	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	>84	Total
Whole Group	1998	6	114	191	168	159	126	116	103	78	75	73	47	22	9	6	1287
	2004	90	159	142	132	102	96	59	64	46	54	28	17	4	4	0	907
Catchment Group	1998	6	75	142	120	126	92	83	77	58	59	59	29	15	7	4	946
	2004	66	136	114	104	83	80	48	45	34	40	26	14	4	4	0	732
Counterfactual Group	1998	0	39	49	48	33	34	33	26	20	16	14	18	7	2	2	341
	2004	24	23	28	28	19	16	11	19	12	14	4	1	0	0	0	175

Notes: The Age groups matches the cohorts in 1998, so that the age in Year 2004 is exactly 6 years older than in 1998, corresponding to the difference in year between 1998 and 2004.

The **figures 3.1 and 3.2** above shows the average real household expenditure p.a.e. (in 1998 prices) of all households with heads of that age against various cross-sectional age profile cohorts in

[21] We have little choice but to classify households by *the age of the head*, even if it is far from clear that the large households that exist in many rural areas can be adequately described in terms of the simple life cycle of the nuclear family.
[22] *Cohorts* are frequently interesting in their own right, and questions about the gainers and losers from economic development are often conveniently addressed by following such groups over time (Deaton, 1997:117).

Zambia's Eastern Province observed from 1998 through to 2004.[23] For example, the data in **figure 3.1** compares the consumption of a 20 year-old in 1998 with the consumption of 20 year-old in 2004, whereas in **figure 3.2** show e.g. the cohort born in 1970, who were 28 years old in 1998 are compared to those aged 34 in 2004. The 1998 result is plotted in *the blue pae98 curve* in the **figure 3.2**. The average consumption of 34-year olds in the 2004 LCSM IV survey is calculated and is depicted in the *red pae0498 curve* on the same segment as the pae98. The rest of the two lines depict the relationship between age and consumption as captured by these two surveys.[24] The graphs, especially the one depicting the 1998 survey, do not look much like *the stylized life-cycle profiles*, which sees individuals smoothing their consumption over their life-time.

Moreover, **figure 3.1** shows a clear age effects on consumption. In 33 out of 45 ages (i.e. 73%), when the cohorts are observed as *averages for a single year of age* in each year, and not smoothed by combining adjacent years into *moving averages* as in figures 3.1-2 (e.g. 30 year-old in 1998 vs. 30 year-old in 2004), was the 2004 cohort above the 1998 cohort. This is an indication of *the growth of real consumption* raising the profiles through this distinct time period. This age effect was even more pronounced when the line for the younger 2004 cohort with a very few exceptions (13%) are always above the line for the older 1998 cohorts (e.g. 41 year-olds in 1998 vs. 47 year-olds in 2004). This could be a reflection of *the macroeconomic growth* in Zambia from 1999 to 2006 (**chapter 2**), which was making younger generations better-off. In other words, the average real household consumption in Eastern Province increases by ZMK11,248 per month, equivalent to a 42.4% increase from 1998 to 2004 (i.e. an annual increase of 6.1%) for the 'same age cohort' of different households tracked over time (**figure 3.2**).[25]

There are a number of other difficulties with age profiles of consumption as those in **figure 3.1-2** above. First, these profiles are simply the cross sections for the households in the surveys, and there is

[23] That is the *age profiles of consumption* for 1200 and 900 rural households in the LCMS II and IV for 1998 and 2004. To plot consumption against age is perhaps the most obvious way to examine *the life-cycle behaviour of consumption* from survey data (Deaton, 1997:339).
[24] In order to keep the age profiles of consumption relatively smooth, I have used the averages for each age to calculate the *five-year moving averages* shown in the two figures. Five-year smoothing also eliminates problems of "age-heaping" when some people report their age rounded to the nearest five years (ibid).
[25] For an enumeration of a number of the advantages, which cohort data have over most panels see Deaton (1997:120f).

no reason to suppose that the profiles represent the typical or expected experience for any individual household or group of households. We are not looking across ages for the same household or same cohort of households, but at the experience of different ages of different groups of households, whose members were born at different dates and have had quite different lifetime experiences of education, wealth, earnings, and accessibility e.g. through the EPFRP. Without controlling for these other variables, many of which are likely to affect the level and shape of the age profiles, we cannot isolate the pure effect of age on consumption or of EPFRP on consumption.

Panel Data from Successive Cross Sections

Following Deaton (1997:121) we briefly consider the issues that arise when using cohort data as if they were repeated observations on individual units. We first consider the simplest *univariate model with fixed effects*, so that at the level of the individual household, we have a parametric the linear regression model with a single variable:

(3.13) $$y_{it} = \alpha + \beta' x_{it} + \theta_i + \mu_t + u_{it},$$

where the μ_t are dummies and θ_i is an individual-specific fixed effect.[26] If we average (3.13) over the members of the age group who appear in the survey, and who will be different in 1998 and 2004, the *"fixed effect"* will not be fixed, because it is the average of the fixed effects of different households in each year. Because of this *sampling effect*, we cannot remove the *age group fixed effects* by differencing or using within-estimators.

(Deaton 1997) considers an alternative approach based on the unobservable population means for each cohort. Starting from the cohort version of (3.13), and denoting population means in cohorts by the subscripts **c**, so that, simply changing the subscript **i** to **c**, we have:

(3.14) $$y_{ct} = \alpha + \beta' x_{ct} + \theta_c + \mu_t + u_{ct},$$

And take *first differences* – to eliminate the fixed effects so that:

[26] For an exposition of *the 'Fixed Effects Model in the two-period case'* see section 12.6 in Johnston&Dinardo(1996:395ff).

(3.15) $$\Delta y_{ct} = y_{c04} - y_{c98} = \beta \Delta x_{ct} + \Delta \mu_t + \Delta u_{ct},$$

where the first time is a constant in any given year. This procedure has eliminated the fixed effects, but we are left with the unobservable changes in the population cohort means in place of the sample cohort means, which is what we observe. If we replace **Δy** and **Δx** in (3.15) by observed changes in the sample means, we generate an *error-in-variables problems*, and the estimates will be attenuated.[27]

Decompositions by age, cohort, and year

In the case of lifetime consumption profile, if *the growth in living standards* acts so as to move up the consumption-age profiles proportionately, it makes sense to work in logarithms, and to write the *logarithm of consumption* as:

(3.16) $$\ln c_{ct} = \beta + \alpha_a + \gamma_c + \psi_t + u_{ct},$$

Where the superscripts **c** and **t** (as usual) refer to cohort and time (year), and **a** refers to age, defined here as the age of the cohort **c** in year **t** (1998). A convenient way to label cohorts is according to Deaton (1997) to choose **c** as the age in year t=0 (i.e. 1998). By this, **c** is just a number like **a** and **t**. Given the way cohorts have been defined, with bigger values of **c** corresponding to older cohorts, we would expect $\boldsymbol{\gamma_c}$ to be declining with **c** (i.e. the age of the cohort in year 0):

(3.17) $$a_{ct} = c + t,$$

That is in year 2004 (t=6) the cohort who in 1998 (t=0) was 30 years-old (c=30) would have the age 36 ($a_{30,6} = 36$).

Using LCMS data to extract information on growth in per adult equivalent consumption

In order to impose structure, we follow (Dercon and Hoddinott 2005) who borrow from the conceptual frameworks used to understand *growth at the national or cross-country level* such as that

[27] There are at least two ways of dealing with *the error-in-variables problem*, see Deaton(1997:122f).

found in (Mankiw, Romer et al. 1992) (cf. Kingombe, 2010a). In the context of *cohort panel data* on p.a.e. consumption, $\mathbf{y_{ct}}$, of **N** cohorts **c** (c=1, ... , N) across periods **t**, a version of this empirical model can be written as in (Islam 1995):

$$\ln y_{ct} - \ln y_{ct-1} = \alpha + \beta \ln y_{ct-1} - \gamma k_{ct-1} + \theta \Delta Z_{ct} + \delta X_c + u_{ct}. \tag{3.18}$$

$\mathbf{\Delta Z_{ct}}$ are changes in *time-varying characteristics* of cohorts and communities that help to explain growth and $\mathbf{X_c}$ are *fixed characteristics* of the cohorts and the community. Examples of $\mathbf{\Delta Z_{ct}}$ could according to (Dercon and Hoddinott 2005) be changing levels of different (exogenous) assets (i.e. not due to investment decisions, but *exogenously changing endowments* at the cohort and community). They also include exogenous shocks in the specification, for example, rainfall shocks. The presence of $\mathbf{X_c}$ would suggest that different types of cohorts may have a particular growth path, linked to fixed characteristics (such as distance to commercial bank in town, etc.). Following, (Dercon and Hoddinott 2005) we add further *"initial" conditions* to the specification, i.e., variables related to assets whose presence may have growth effects ($\mathbf{k_{ct-1}}$). Examples are levels of landholdings or infrastructure.

Thus, we use equation **(3.18)** for our test to see whether the EPFRP infrastructure and accessibility matter for understanding *growth in consumption outcomes* in the period from 1998 to 2004. Because we want to focus on age cohort variables, it makes sense to run our regression using the most complete controls for household-level variables that do not change over time. This is accomplished by estimating *a fixed-effect regression* – essentially including a dummy variable for each household in the sample – that controls for all household characteristics that might affect the growth of consumption but do not change over time. Consequently, all our covariates are identified using *changes over time*. The attraction of such an approach is according to Dercon and Hoddinott (2005:16) that "we avoid some standard issues, such as *placement effects* due to fixed factors and other sources of endogeneity affecting accessibility. However, while desirable in terms of ensuring that we can confidently identify the impact of variables that change over time, this approach comes with a cost: that we *cannot identify factors that do not change over time*."

4. Estimation Results

Our analysis use the data collected through respectively the 1998 LCMS (II) and the 2004 LCMS (IV) to evaluate *the effects of the EPFRP* on the well being of rural households in the catchment and control districts of Zambia's Eastern Province as well as the changes in living standards from 1998 to 2004, using LCMS 1998 as a useful benchmark of living standards, coming as it does at the end of the most intensive part of the economic reform period from 1991 to 1996.

Table 4.1 Descriptive Statistics of the Data covering Rural Eastern Province

	LCMS II 1998						LCMS IV 2004					
	Full Sample (i)		Poor Households (ii)		Extremely Poor Households (iii)		Full Sample (iv)		Poor Households (v)		Extremely Poor Households (vi)	
Variables	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Household Size	5,06	2,931	5,16	2,93	5,19	2,93	5,43	3,01	5,70	2,96	5,80	2,96
Age of household head	43,06	15,772	43,43	15,87	43,64	15,84	42,42	15,23	43,44	15,45	43,80	15,66
Age 0-14	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.
Age 15-24	120*	n.a.	112*	n.a.	104*	n.a.	120*	n.a.	20*	n.a.	9*	n.a.
Age 25-65	1037*	n.a.	970*	n.a.	927*	n.a.	1286*	n.a.	187*	n.a.	122*	n.a.
Age 66 and over	149*	n.a.	146*	n.a.	141*	n.a.	177*	n.a.	52*	n.a.	37*	n.a.
Education level of household head	5,85	2,843	5,78	4,54	5,68	4,58	5,03	3,88	4,50	3,61	4,32	3,54
No education	408*	n.a.	398*	n.a.	386*	n.a.	344*	n.a.	76*	n.a.	57*	n.a.
Grade 1-4	188*	n.a.	183*	n.a.	175*	n.a.	389*	n.a.	82*	n.a.	57*	n.a.
Grade 5-7	255*	n.a.	237*	n.a.	228*	n.a.	532*	n.a.	70*	n.a.	40*	n.a.
Grade 8-9	67*	n.a.	54*	n.a.	48*	n.a.	171*	n.a.	24*	n.a.	12*	n.a.
Grade 10-12	37*	n.a.	27*	n.a.	23*	n.a.	112*	n.a.	5*	n.a.	2*	n.a.
Grade 12 GCE (A)	11*	n.a.	8*	n.a.	4*	n.a.	3*	n.a.	0*	n.a.	0*	
College/undergraduate/Bachelor's degree and above	2*	n.a.	1*	n.a.	1*	n.a.	22*	n.a.	1*	n.a.	0*	
Maritual status of head	3,98	2,864	4,04	2,89	4,06	2,89	2,43	1,03	2,82	1,24	3,02	1,30
Never married	240*	n.a.	223*	n.a.	215*	n.a.	36*	n.a.	3*	n.a.	1*	n.a.
Married	497*	n.a.	458*	n.a.	433*	n.a.	1643*	n.a.	225*	n.a.	114*	n.a.
Separated	22*	n.a.	20*	n.a.	19*	n.a.	41*	n.a.	17*	n.a.	15*	n.a.
Divorced	51*	n.a.	43*	n.a.	40*	n.a.	88*	n.a.	32*	n.a.	25*	n.a.
Widower	95*	n.a.	92*	n.a.	88*	n.a.	235*	n.a.	68*	n.a.	49*	n.a.
Not stated	401*	n.a.	392*	n.a.	377*	n.a.						
Employment Status?	4,65	3,44	4,74	3,45	4,77	3,45	3,85	4,10	4,24	4,33	4,10	4,29
SELF EMPLOYED	403*	n.a.	377*	n.a.	357*	n.a.	869*	n.a.	155*	n.a.	99*	n.a.
CENTRAL GOVT EMPLOYEE	20*	n.a.	13*	n.a.	8*	n.a.	27*	n.a.	2*	n.a.	0*	n.a.
LOCAL GOVT EMPLOYEE	2*	n.a.	1*	n.a.	1*	n.a.	3*	n.a.	0*	n.a.	0*	n.a.
PARASTATAL EMPLOYEE	0*	n.a.	0*	n.a.	0*	n.a.	7*	n.a.	0*	n.a.	0*	n.a.
PRIVATE SECTOR EMPLOYEE	16*	n.a.	13*	n.a.	12*	n.a.	34*	n.a.	0*	n.a.	0*	n.a.
NGO Employee	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0*	n.a.	0*	n.a.	0*	n.a.
Intern.Org. & Embassy Employee	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.
EMPLOYER/PARTNER	5*	n.a.	5*	n.a.	5*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.
HOUSEHOLD EMPLOYEE	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.						
UNPAID FAMILY WORKER	452*	n.a.	439*	n.a.	419*	n.a.	0*	n.a.	0*	n.a.	0*	n.a.
PIECE WORKER	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.						
OTHER	1*	n.a.	1*	n.a.	1*		377*	n.a.	81*	n.a.	48*	n.a.
Food exp / total exp	0,41	0,26	0,41	0,27	0,41	0,27	0,79	0,20	0,78	0,19	0,78	0,20
Total HH Income (ZMK / month)	345402,5	8112332,0	346626,0	8346002,0	351442,0	8542870,0						
Total HH expenditure (ZMK / month)	67543,43	110622,30	52668,73	62883,78	46711,38	52116,29	483340,70	567464,10	390953,70	468119,50	369773,40	458534,30
Food exp (ZMK / month)	32338,62	40705,97	28239,41	32459,35	25740,02	27650,76	381419,10	514630,20	305285,10	411886,20	289068,50	406648,00
Total area under cultivation	1,97	3,637	1,83	2,41	1,81	2,40						
Sample size	1300		1228		1172		1583		1072		889	

Note: * Number of heads of households.
Source: Author's calculations.

The set of transmission mechanisms suggested in the theoretical framework of (Winters 2002), namely prices, wages and employment, are associated with the consumption outcomes varying by personal characteristics such as age and education. The basic rural household characteristics are summarized in **table 4.1**. The two LCMSs only show small differences in the characteristics of rural households amongst the whole sample versus the moderately and extremely poor rural households. The extremely poor households in rural Eastern Province are on average larger and have more children per working age adult than better off households. Although the average year of schooling has increased since 1998, the poorest heads of households still have less schooling than the heads of better off households. Moreover, they are also older.

4.1. Model Diagnostics

We share the belief that better models and a better understanding of one's data result from focused data analysis as well as the inclusion and exclusion of respectively exogenous and endogenous regressors for the structural model guided by substantive theory.[28]

The descriptive statistics for the data of the covariates are presented in **table 4.2**. *The sample size* of the reduced model's dependent variable is now 1287 in 1998 and 999 in 2004, which is significantly larger than the full model.[29] The sample size is distributed between the catchment and the control districts. That is, 948 and 339 rural households were sampled in 1998 within the catchment and control districts respectively. Likewise 817 and 182 rural households were sampled in 2004 within the same catchment and control districts (**Table A7**).

Surprisingly, *the mean of cotton sales as a share of household income* increased faster between 1998 and 2004 in the control districts (294%) than in the catchment districts (157%). Even more unexpectedly is the finding that the average distance to input markets in Eastern Province decreased faster in the in the control districts (-76%) compared to the catchment districts (-42%). Yet the

[28] *Economic theory* often provides some guidance in model specification but may not explicitly indicate how a specific variable should enter the model, identify the functional form, or spell out how the stochastic elements (e_i) enter the model. Comparative static results that provide expected signs for derivatives do not indicate which *functional specification* to use for the model (Baum, 2006).

[29] Our *initial specification* reported all theoretically possible variables found in the existing CSO datasets. However, only 9 out of 70 coefficients are statistically significant at the 10% level in the 1998 dataset, using a 2-tailed p-value testing the null hypothesis that the coefficient is null. This is probably because our sample size in 1998 dataset (n=198) is so reduced by missing data on the large number of explanatory variables (70) covering the rural areas in Eastern Province.

percentage change in *the logarithm of p.a.e. expenditures* increased at a higher rate in the catchment districts (+16%) than in the control districts (+10%) (**Table A7**).

Table 4.2: Descriptive Statistics of covariates, 1998 and 2004

			1998					2004					Percentage
Type	**Variable Name**	**Variable**	**Obs**	**Mean**	**Std.Dev.**	**Min.**	**Max.**	**Obs**	**Mean**	**Std.Dev.**	**Min.**	**Max.**	**change**
CV	Log pae monthly household expenditure	LNPAE98*	1287	8,996	1,234	3,912	14	999	10,293	1,160	6,563	13,771	14%
CV	Cotton Sales share of household income	Cotincshare	1274	0,104	0,202	0,010	1	999	0,293	0,338	0	1	183%
DV	Stratum, excl. Large AHH	Stratum124**	1304	1,510	1,028	1	4	999	1,193	0,417	1	4	n.a.
CV	Distance to Inputmarket	Distiput	1311	22,692	22,957	0,000	99	999	12,139	17,566	0	99	-47%
DV	EPFRP Treatment	Infrastructure***	1311	0	0	0	0	999	0,818	0,386	0	1	n.a.
DV	Plough Ownership	Plough	1311	0,799	0	0	1	999	0,683	0,466	0	1	n.a.
DV	Bicycle Ownership	Bicycle	1311	0,483	0	0	1	999	0,337	0,473	0	1	n.a.
DV	Scotchcart Ownership	Scotchcart	1311	0,901	0	0	1	999	0,822	0,383	0	1	n.a.
DV	Motorvehicle Ownership	Motorvehicle	1311	0,976	0,152	0	1	999	0,979	0,144	0	1	n.a.
CV	Age of Head of Household	Age	1306	43,054	15,745	15,000	99	999	42,888	15,024	20	90	0%
CV	Age Squared	Agesq	1306	2101,322	1522,004	225	9801	999	2064,876	1464,556	400	8100	-2%
DV	Head of HH ever attended School	s4q5	968	0,421	0,494	0	1	997	0,217	0,412	0	1	n.a.

Notes: * Logarithm of the Per Adult Equivalent total household expenditure; ** Small - and medium scale farmers, and non-agricultural rural households. *** Dummy / indicator variable takes the values 0 or 1 to indicate the absence or presence of some categorical (EPFRP) effect that may be expected to shift the outcome.
Source: Author's calculations.

The dependent variable is still this natural logarithm of p.a.e. monthly expenditure in both 1998 and 2004. Our aim is to measure the differences between the estimate of the 1998 mean of *Y=LNpae98* conditional on five covariates and the estimate of the 2004 mean of *Y=LNpae04* conditional on the same covariates.

These covariates described in **table 4.2** are rural stratum excluding the small number of large scale farmers, the distance to the nearest input market, the head of household's ownership of a motor vehicle and the EPFRP Treatment dummy variable (denoted *infrastructure*) and finally the share of income derived from cotton sales to total rural household income (denoted *cotincshare*).[30] In other words we explore the link between the rural households response to the price changes ensuing from improved rural road transport infrastructure investment by focusing on the single most important cash crop namely cotton. That is, *the critical factor* in our case is the share of household income generated by the sales of the cotton production (see Govereh, Jayne et al., 2000; Zambia Food Security Research Project, 2000; Zulu and Tschirley 2002; Balat and Porto 2005a; Balat and Porto 2005b; Brambilla and Porto 2007; Kabwe and Tschirley 2007; Tschirley and Kabwe 2007; Kabwe 2009).

[30] Alternatively we could use a dummy variable of the cotton variable to show whether the household derived any income from cotton cultivation (=1) or not (=0).

4.2. Estimation results of Parametric and Semiparametric Models

This section reports the results of estimating the mean of *Y* conditional on the five covariates of a linear parametric model (4.1) and of a semiparametric model (4.2). The parametric model is more parsimonious, and thereby more interpretable than models proposed through a selection estimation procedure. The models that we estimate are as follows for respectively the 1998 and 2004 datasets:

(4.1) E(LNpae | Stratum124, Cotton sale / Household income, Distance Input market, Motor vehicle, Infrastructure) = β_0 + β_1*Stratum124 + β_2*(Cotton Sale / HH Income) + β_3*Distance Input market + β_4*Motorvehicle ownership + β_5*Infrastructure + e_{ht},[31]

(4.2) E(LNpae | Stratum124, Cotton sale / Household income, Distance Input market, Motor vehicle, Infrastructure) = m(X) = β1*Stratum124 + β2*Cotincshare + β3*Distiput + β4*s10q8 + G(β5*Infrastructure)

Where **G** is an unknown function and the **βs** are unknown scalar parameters. The error term $\mathbf{e_{ht}}$ comprises the rural household-level error term of household **h** at time **t** (1998 or 2004). Model **(4.2)** is *a semiparametric partially linear model **(3.1)***, which is a further example of semi-parametric **GLM** that is able to handle (additional) nonparametric components as discussed in **section 3.1**.[32]

In order to estimate the coefficients of **(4.1)** we had to choose a correct OLS analysis for the LCMS II and LCMS IV survey design, in other words we do OLS regression with clusters.[33] Rather than the standard linear predictor **(4.1)** the partially linear model allows **(4.2)** one predictor - infrastructure - to be nonlinear.

In **table 4.3** *the OLS model* is compared to its flexible generalization *the standard Generalized Linear Model (GLM)* with the unknown parameters, **β**, fitted using Newton-Raphson (maximum likelihood) optimization.[34] The dependent variable is assumed to be generated by the distribution function, **f**, from the Gaussian(normal) probability distribution family (Hardin and Hilbe 2007). The

---

[31] To avoid that our statistics becomes misleading from outliers associated with the different population of the seven large scale farmers in rural Eastern Province in 1998, which is different from the rest of the sample set. Hence these seven data points will all be censored in the remaining part of this paper.

[32] *A partially linear model*, that is a semiparametric model of the type: Y = Xtâ + ö(Z) + U, where X and Z are multivariate explanatory variables. The apparent asymmetry between the effect of the variables X and Z in the partially linear model brings the analyst to include all dummy or categorical variables in the parametric component of the model.
The GPLM can be viewed as a compromise between the GLM and a fully nonparametric model (He et al., 2005).

[33] In **Appendix A6.a-b** the tables provides a summary of the various parametric models shown that we tried. Some of the models only adjust the standard error by computing a cluster robust standard error for the coefficient. Other procedures do more complex modeling of the multilevel structure. And there are some procedures that do various combinations of the two.

[34] The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a **link function** and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

results show that *cotton's income share of total income* went from negative and insignificant in 1998 to positive coefficients and statistical significance at the 0.05 level.

Table 4.3: Comparison of Generalized linear models, 1998 and 2004

		1998				2004			
Variable Names	**Covariates**	**OLS**	**OLS**	**GLM**	**GLM**	**OLS**	**OLS**	**GLM**	**GLM**
Cotton Sales share of	cotincshare	-0.1237	-0.1650	-0.1237	-0.1650	-0.0673	-0.1241	-0.0673	-0.1241
household income		(0.1888)	(0.1699)	(0.2360)	(0.2674)	(0.1073)	(0.1068)	(0.0776)	(0.0852)
Stratum, excl. Large	stratum124	0.2432***	0.2672***	0.2432***	0.2672***	-0.2783***	-0.3609***	-0.2783	-0.3609**
AHH		(0.0383)	(0.0336)	(0.0315)	(0.0248)	(0.0964)	(0.0885)	(0.1827)	(0.1415)
Distance to	Distiput	-0.0033*	-0.0044***	-0.0033	-0.0044***	-0.0018	-0.0027	-0.0018	-0.0027
Inputmarket		(0.0017)	(0.0015)	(0.0030)	(0.0016)	(0.0021)	(0.0021)	(0.0018)	(0.0021)
	Infrastructure					0.3228***	0.3035***	0.3228***	0.3035***
EPFRP Treatment						(0.0941)	(0.0949)	(0.1150)	(0.1135)
Motorvehicle	Motorvehicle	-1.5649***	-1.5959***	-1.5649***	-1.5959***	0.9974***	1.0724***	0.9974***	1.0724***
Ownership		(0.2942)	(0.2519)	(0.4180)	(0.1968)	(0.2532)	(0.2562)	(0.1490)	(0.1387)
	Plough	0.0266		0.0266		0.1083		0.1083	
Plough Ownership		(0.1193)		(0.0902)		(0.1011)		(0.0772)	
	Bicycle	-0.3326***		-0.3326***		0.2172***		0.2172***	
Bicycle Ownership		(0.0799)		(0.0663)		(0.0803)		(0.0792)	
	Scotchcart	-0.1089		-0.1089		-0.0431		-0.0431	
Scotchcart Ownership		(0.1672)		(0.1133)		(0.1209)		(0.0825)	
Age of Head of	Age	-0.0297**		-0.0297*		-0.0656***		-0.0656**	
Household		(0.0137)		(0.0164)		(0.0145)		(0.0297)	
	Agesq	0.0002*		0.0002		0.0007***		0.0007**	
Age Squared		(0.0001)		(0.0002)		(0.0001)		(0.0003)	
Head of HH ever	s4q5	-0.4115***		-0.4115***		-0.1055		-0.1055	
attended School		(0.0787)		(0.0454)		(0.0883)		(0.1356)	
	_cons	11.4715***	10.2921***	11.4715***	10.2921***	10.6893***	9.4946***	10.6893***	9.4946***
Constant		(0.4447)	(0.2605)	(0.5691)	(0.2157)	(0.4659)	(0.3038)	(0.4732)	(0.2407)
	N	920	1246	920	1246	997	999	997	999
	r2	0.1554	0.0888			0.0873	0.0521		
	F	16.7288	30.2329			8.5654	10.9101		
	ll	-1420.2349	-1960.7012	-1420.2349	-1960.7012	-1517.3167	-1538.6864	-1517.3167	-1538.6864

Notes: Distribution of depvar (LNpae) is family (Gaussian). The Generalized Linear Models (GLM) are fitted using Newton-Raphson (maximum likelihood) optimization with robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.
Source: Author's calculations.

Alternatively in order to calculate the *linear partial regression* we follow Michael Lokshin, who estimates the following semiparametric regression model by the method of differencing:

$$y_i = X_i\beta + f(z_i) + \varepsilon_i. \quad (4.3)$$

The weighting matrix (**W**) accounts for the correlations among the set of l-moments when the errors are not i.i.d. The weighting matrix only plays a role in the presence of over-identifying restrictions.[35] *The significance test* of the Treatment Infrastructure variable that enters the *Yatchew*

[35] If the equation to be estimated is overidentified, l > k, we have more equations than we do unknowns.

specification non-linearly indicates that the *infrastructure dummy variable* is highly significant (P-value of 0.000) in 2004 (**column 6 table 4.4**). The same is the case if the form of the vector of differencing weights $d_1, ..., d_m$ are specified using *Hall et. al. (1990) weights* for differencing (**column 8 table 4.4**, cf. **figures A8.1-2**).

Compared with the estimation of the fully parametric model (**table 4.3**) one finds that while the signs of the coefficients are almost the same in 1998 between the two specifications, the magnitudes of some coefficients are different. For example, the effect of *'distance to the input market'* on the logarithm of p.a.e. consumption of rural household change from -0.0044 in the fully parametric model to -0.0031 in the partial linear model estimation with *Yatchew's weighing matrix* in 1998. In the 2004 dataset the coefficient of the same covariate change from -0.0027 to -0.0007, whereas the *'cotton sales share'* changes from wrong sign – 0.1241 to the correct positive coefficient 1.3910 in 2004 (**table 4.4**).

Table 4.4: Partial Linear regression models, 1998 and 2004

The optimal weighting matrix is that which produces the most efficient estimate (Baum, 2006).

		1998				2004			
		Weighting Matrix				Weighting Matrix			
Variable Names	**Covariates**	**Yatchew**	**Yatchew**	**Hall**	**Hall**	**Yatchew**	**Yatchew**	**Hall**	**Hall**
Cotton Sales share of	cotincshare	-11.5638	-15.1956	-0.3689	-1.7352	-0.3784	2.5665	-0.2792	1.3910*
household income		(17.8541)	(16.7153)	(4.4799)	(4.9766)	(1.8683)	(1.7332)	(0.7920)	(0.7657)
Stratum, excl. Large	stratum124	0.1725	0.2327**	0.3116**	0.2865***	-0.2070	-0.3197***	-0.3481***	-0.4463***
AHH		(0.1697)	(0.1132)	(0.1297)	(0.1076)	(0.1362)	(0.1188)	(0.1136)	(0.1025)
Distance to	Distiput	0.0024	0.0031	0.0021	-0.0004	-0.0000	-0.0004	0.0007	-0.0007
Inputmarket		(0.0037)	(0.0025)	(0.0029)	(0.0025)	(0.0031)	(0.0028)	(0.0026)	(0.0025)
Motorvehicle	Motorvehicle	-1.7509***	-1.4151***	-1.5850***	-1.7253***	0.8686***	0.8220***	0.8870***	0.9726***
Ownership		(0.3543)	(0.2524)	(0.2969)	(0.2466)	(0.3278)	(0.3045)	(0.2634)	(0.2565)
	Plough	0.2388		0.0676		0.0438		0.1061	
Plough Ownership		(0.1631)		(0.1230)		(0.1328)		(0.1060)	
	Bicycle	-0.2198*		-0.3005***		0.2127**		0.1939**	
Bicycle Ownership		(0.1186)		(0.0839)		(0.1071)		(0.0881)	
	Scotchcart	-0.1415		-0.0269		-0.0141		-0.0390	
Scotchcart Ownership		(0.2090)		(0.1728)		(0.1540)		(0.1288)	
Age of Head of	Age	-0.0124		-0.0247*		-0.0720***		-0.0727***	
Household		(0.0174)		(0.0139)		(0.0190)		(0.0151)	
	Agesq	0.0001		0.0002		0.0007***		0.0008***	
Age Squared		(0.0002)		(0.0001)		(0.0002)		(0.0002)	
Head of HH ever	s4q5	-0.3860***		-0.3557***		-0.0304		-0.0685	
attended School		(0.0937)		(0.0796)		(0.1051)		(0.0922)	
	N	919	1245	913	1239	996	998	990	992
	r2	0.0923	0.0473	0.1024	0.0499	0.0526	0.0284	0.0745	0.0440
	F	9.2443	15.4142	10.3005	16.2273	5.4765	7.2561	7.8838	11.3670
	ll	-1385.5857	-1625.6319	-1384.8174	-1863.7249	-1494.5794	-1428.8287	-1505.5017	-1486.7673
	V(i)	2.266	25.016	4.445	13.901	1.768	8.067	1.197	8.168
	P>|V|	0.012	0.000	0.000	0.000	0.039	0.000	0.116	0.000

Notes: (i) Significant test on Infrastructure. Standard errors in parentheses * p<0.10, ** p<0.05, *** p<0.01

Source: Author's estimations.

4.3. Estimation results of Panel Data from Successive Cross Sections

In this section we use cohort panel data, which we likewise draw from the LCMS II and LCMS IV. There are **45** cohort-year pair observations. The data contain one cohort(age) identifier, two years of data on monthly p.a.e. expenditure, two years of cotton income shares, two years of landownership per household member, two years of rainfall data and of lagged rainfall data, two years of highest grades attained, and two years of distances to public or private services (input market and formal banks). The data are for 45 ages. By converting the wide form to long form we expand the dataset from 45 observations (45 individual ages) to 90 observations (90 cohort *age-year pairs*). A year-identifier variable, year (**t-1**=1998 and **t** = 2004), has been created.

By providing separate time-series plot for the 45 LNpae individual units in the sample, we see a clear upward going tendency between 1998 and 2004. If we begin our graphical analysis by looking at a scatter plot of the dependent variable (LNpae) on the following key regressors (distance to the input market; highest grade attained by head of household; rain; and cotton income share), using data from all panel observations we find a clear downward trend between log expenditure p.a.e. and distance to input market, whereas log expenditure steeply increases until around attainment of six grade after which the log expenditure gradually declines. Rain lagged seems to have a higher upward going effect on log expenditure than rainfall in the actual agricultural season. Finally, we see a clear incremental upward going trend in log expenditure against cotton income share.

According to the panel summary statistics on within and between variation, we find that for the variables (rain; household size; distance to food market; distance to input market), there is more variation across individuals (*between variation*) than over time (*within variation*), so within estimation may lead to considerable efficiency loss.

Pooled OLS regression

We start our pseudo-panel analysis with the pooled OLS regression for log p.a.e. expenditure using data for all cohorts in both years. We include as regressors: Education (highest grade attained by head of household); experience (Age) and quadratic in experience (Age squared); rain; rain lagged; landownership per household member; cotton income share of total income; cotton income share squared; distance to input market; and distance to nearest formal bank.[36] Experience is *time-varying* in a deterministic way as the sample comprises people who work full-time in all years, so experience increases by one year as **t** increments by one.[37]

Regressing $\mathbf{y_{ct}}$ on $\mathbf{x_{ct}}$ yields consistent estimates of $\boldsymbol{\beta}$, if the composite $\mathbf{u_{ct}}$ are uncorrelated with the regressors in the pooled model or population-averaged model, which assume regressors are exogenous:

$$y_{ct} = \alpha + x'_{ct}\beta + u_{ct} = \alpha + x'_{ct}\beta + \alpha_c + \varepsilon_{ct}, \qquad (4.4)$$

[36] The few commercial formal banks are all exclusively situated in the district centres. Therefore the distance to the formal banks could be considered as a proxy for the distance to the district centre. Moreover this is the only possible variable, which the limited dataset allows us to consider as an instrument variable. See the discussion in section 5 below.

[37] Education is a *time-invariant* regressor, taking the same value each year for a given individual. However, since the education variable is the mean of all the head of households in the same cohort, it could be justified that education is *varying*, if we by education also mean life-long learning, which includes technical vocational education and training.

The error $\mathbf{u_{ct}}$ is likely to be correlated over time for a given cohort, so we use cluster-robust standard errors that cluster on the cohort.

Table 4.5: Pooled OLS with and without cluster-robust standard errors

	Whole Sample				Catchment Sub-sample				Counterfactual Sub-sample			
		Robust		Default		Robust		Default		Robust		Default
Lnpae	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Highgrade	0,078	0,051	,0783*	0,043	0,042	0,055	0,042	0,034	0,048	0,040	0,048	0,035
Age	,0640019**	0,026	,064**	0,026	-0,001	0,020	-0,001	0,022	-0,030	0,033	-0,030	0,042
Agesquare	-,00086***	0,000	-,00086***	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Rain	-0,025	0,016	-0,025	0,016	-0,0378***	0,011	-0,0378***	0,012	-0,002	0,008	-0,002	0,011
Rain Lagged	,0881***	0,020	,0881***	0,021	0,028	0,018	0,0282*	0,016	0,055**	0,023	0,055**	0,023
Landownership pc	,4568***	0,098	,4568***	0,081	0,193***	0,061	0,193***	0,062	0,195	0,154	0,195	0,135
Cotton Income share	2,975	2,481	2,975	2,223	0,284	1,228	0,284	1,792	-0,691	1,968	-0,691	1,547
Cotton Inc Share squared	-2,880	3,126	-2,880	2,614	-0,003	1,912	-0,003	2,028	0,838	2,199	0,838	1,746
Distance to Input market	-0,006	0,008	-0,006	0,008	-0,001	0,006	-0,001	0,007	-0,004	0,009	-0,004	0,008
Distance to Bank	-,01898***	0,007	-,01898***	0,006	-0,005	0,004	-0,005	0,005	-0,005	0,006	-0,005	0,006
Constant	4,138***	1,142	4,138***	1,270	10,43***	0,895	10,434***	1,050	5,847***	1,911	5,847***	1,782
R2	0,895		0,895		0,672		0,672		0,447		0,447	
Adj R2			0,882				0,631				0,364	
N	90		90		90		90		78		78	

Source: Authors' estimations.

The standard errors are small except for the two cotton share regressors. Moreover, *the cluster-robust standard errors* are smaller than the default standard errors for the following regressors: Age; Age square; Rain; Rain lagged (excl. Catchment sample); and distance to input market (excl. Control sample), and they are larger for the remaining regressors. Given the very high **R2** it is almost certain that log p.a.e. expenditure is *overpredicted* in both 1998 and 2004. This is probably associated with how the cohort dataset was constructed. The failure to control for this error correlation leads to underestimation of standard errors. On the other hand, the difference between default and cluster-robust standard errors for pooled OLS is quite small given the small number of time periods (T=2).

Table 4.6: Between Estimator with Default Standard Errors

Lnpae	Whole Sample Between Estimator Coef.	Default Std. Err.	Catchment Sub-sample Between Estimator Coef.	Default Std. Err.	Counterfactual Sub-sample Between Estimator Coef.	Default Std. Err.
Highgrade	-0,032	0,045	-0,017	0,048	0,070	0,044
Age	0,015	0,022	0,018	0,023	0,020	0,044
Agesquare	0,000	0,000	0,000	0,000	0,000	0,001
Rain	0,024	0,015	-0,020	0,014	-0,012	0,012
Rain Lagged	0,014	0,023	0,002	0,026	0,063**	0,024
Landownership pc	-0,054	0,117	0,095	0,095	0,406*	0,183
Cotton Income share	0,223	2,747	-1,191	2,538	-0,217	2,212
Cotton Inc Share squared	-0,712	3,094	0,297	2,942	0,088	2,461
Distance to Input market	-0,012	0,008	-0,011	0,012	-0,009	0,012
Distance to Bank	-,0124*	0,006	-0,012	0,007	-0,007	0,009
Constant	8,161***	1,452	11,814***	1,551	4,722	1,917
R2: within	0,431		0,131		0,426	
R2: between	0,492		0,544		0,592	
R2: overall	0,269		0,225		0,426	
N	90		90		78	
Number of groups	45		45		44	
Obs per group	2		2		2	
sd(u_i + avg(e_i.))	0,189		0,223		0,456	
F(10, 34)	3,290		4,060		4,780	
Prob > F	0,005		0,001		0,000	

Source: Authors' estimations.

We next calculate the first-order autocorrelation coefficient for LNpae to be -0.1230 (whole sample). 45 observations are used to compute the autocorrelation at lag 1, -0.151, which provides a rough estimate of the intraclass correlation coefficient of the residuals that is far from indicating perfect anti-correlation (-1).

The *between estimator* uses only between or cross-section variation in the data and is the OLS estimator from the regression of $\bar{y}_i$ on $\bar{x}_i$. The Between Estimator estimates and standard errors are closer to those obtained from Pooled OLS (cf. **table 4.5**) than those that could have been obtained from within estimation.[38]

Table 4.6 shows that the distance to the formal bank (district centre) correctly is negatively associated with the dependent variable, but only significant for the whole sample. The signs of the coefficients for the Lagged rain and Landownership regressors correct for both the catchment and control areas, and they are significant at respectively the 0.05 and 0.1 level, but only for the control sample.

[38] Some groups have fewer than 3 observations therefore it was not possible to estimate correlations for those groups. 45 groups omitted from estimation.

Table 4.7 presents the results of our comparison of three cross-sectional time-series regression models. And in **table 4.8**, several features emerge when we compare some of the panel estimators and associated standard errors, variance components estimates, and R^2. The estimated coefficients vary considerably across estimators, especially for the time-varying regressors. This reflects quite different results according to whether within variation or between variation is used. Cluster-robust standard errors for the FE and RE models exceeds the default standard errors except for distance to input market and distance to bank in the former case and rain; rain lagged and distance to bank in the latter case. The various R^2 measures and variance-components estimates also vary considerably across models.

First-Difference Estimator

Consistent estimation of $\boldsymbol{\beta}$ in the fixed-effects (FE) model requires eliminating the $\boldsymbol{\alpha_i}$. One way to do so is mean-difference, yielding the *within estimator*.[39] An alternative way is to first-difference, leading to *the first-difference estimator*. This alternative has the advantage of relying on weaker exogeneity assumptions (Cameron and Trivedi 2009). The first-difference (**FD**) estimator is obtained by performing OLS on the first-differenced variables:

$$(y_{it} - y_{i,t-1}) = (x_{it} - x_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1}) \tag{4.5}$$

[39] The within estimator is traditionally favoured as it is the more efficient estimator if the ε_{it} are i.i.d.

Table 4.7: RE Estimator: Comparison of cross-sectional time-series regression models

	Catchment Sub-sample						Counterfactual Sub-sample					
	GLS random-effects estimator		Maximum-likelihood random-effects estimator		GEE Population-averaged estimator(i)		GLS random-effects estimator		Maximum-likelihood random-effects estimator		GEE Population-averaged estimator(i)	
		Robust		Default		Default		Robust		Default		Default
Lnpae	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Highgrade	0,045	0,055	0,042	0,031	0,039	0,031	0,048	0,040	0,048	0,032	0,049	0,032
Age	-0,005	0,020	-0,001	0,021	0,002	0,020	-0,030	0,033	-0,030	0,039	-0,019	0,037
Agesquare	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Rain	-0.039***	0,012	-0.038***	0,011	-0.037***	0,011	-0,002	0,008	-0,002	0,010	-0,002	0,010
Rain Lagged	0,029	0,018	0.028**	0,015	0.028*	0,015	0.055**	0,023	0.055**	0,022	0.055**	0,022
Landownership pc	0.188***	0,058	0.193***	0,058	0.198***	0,058	0,195	0,154	0,195	0,125	0.213*	0,125
Cotton Income share	0,340	1,256	0,284	1,679	0,213	1,688	-0,691	1,968	-0,691	1,434	-0,486	1,432
Cotton Inc Share squared	0,027	1,951	-0,003	1,900	-0,026	1,910	0,838	2,199	0,838	1,618	0,495	1,603
Distance to Input market	-0,001	0,005	-0,001	0,007	-0,001	0,007	-0,004	0,009	-0,004	0,008	-0,004	0,008
Distance to Bank	-0,004	0,005	-0,005	0,005	-0,006	0,005	-0,005	0,006	-0,005	0,006	-0,004	0,006
Constant	10,487***	0,901	10.435***	0,983	10.39***	0,975	5,847***	1,911	5,847***	1,652	5,592***	1,627
R2: within	0,776						0,476					
R2: between	0,354						0,544					
R2: overall	0,672						0,447					
N	90		90		90		78		78		78	
Number of groups	45		45		45		44		44		44	
Obs per group	2		2		2		1,8		1,8		1,8	
theta	0,103											
Wald chi2(11) / chi2(10)	93553,2				170,11		44,25				57,75	
Prob > chi2	0				0		0				0	
sigma_u	0,099		0	0,125			0		0	0,175		
sigma_e	0,283		0,340	0,025			0,747		0,611	0,049		
rho	0,108		0	.			0		0			
Log likelihood			-30,668						-72,265			
LR chi2(10)			100,35						46,2			
Prob > chi2			0						0			
Likelihood-ratio test of sigma_u=0:			0						0			
Prob > chibar2			1						1			

Notes: (i) Correlation exchangeable.
Source: Authors' estimations.

Table 4.8a Panel Estimator Comparison, Catchment

Variable	OLS_rob	Between-effects (BE) model	Fixed-effects (FE) model	FE_rob (i)	GLS random-effects (RE) model	RE_rob (i)
highgrade	0.0419	-0.0174	0.0502	0.0502	0.0448	0.0448
	0.0546	0.0484	0.0378	0.0424	0.0336	0.0549
Age	-0.001	0.0176	0.0944	0.0944	-0.0047	-0.0047
	0.0202	0.0229	0.0551	0.0508	0.0227	0.0226
Agesq	-0.0001	-0.0004	0.0012	0.0012	-0.0001	-0.0001
	0.0002	0.0003	0.0004	0.0003	0.0003	0.0003
rain	-0.0378	-0.0201	-0.0009	-0.0009	-0.0386	-0.0386
	0.0113	0.0141	0.0181	0.0129	0.0124	0.0123
rainlagged	0.0282	0.0017	-0.0206	-0.0206	0.0285	0.0285
	0.0183	0.0259	0.0201	0.0159	0.0162	0.0187
Landowners~c	0.1931	0.0945	-0.1426	-0.1426	0.1877	0.1877
	0.061	0.0951	0.0889	0.0696	0.0621	0.0634
cotincshare	0.2839	-1.1913	-1.7587	-1.7587	0.3402	0.3402
	1.2282	2.5376	1.929	2.0335	1.7816	1.4581
cotincshar~q	-0.0029	0.2974	2.4718	2.4718	0.0267	0.0267
	1.9123	2.942	2.147	2.1065	2.0157	2.0074
distiput	-0.0013	-0.011	-0.0039	-0.0039	-0.0014	-0.0014
	0.0055	0.0123	0.0066	0.0069	0.0069	0.006
distbank	-0.0047	-0.0121	-0.0039	-0.0039	-0.0036	-0.0036
	0.0043	0.0075	0.006	0.0068	0.0054	0.0046
_cons	10.4348	11.8145	4.4949	4.4949	10.4869	10.4869
	0.8947	1.5507	2.122	1.9112	1.0586	0.9965
N	90	90	90	90	90	90
r2	0.6721	0.544	0.8844	0.8844		
r2_o		0.2245	0.0172	0.0172	0.6715	0.6715
r2_b		0.544	0.3606	0.3606	0.3541	0.3541
r2_w		0.1307	0.8844	0.8844	0.7763	0.7763
sigma_u			2.774	2.774	0.0988	0.0988
sigma_e			0.2835	0.2835	0.2835	0.2835
rho			0.9897	0.9897	0.1084	0.1084

Table 4.8b Panel Estimator Comparison, Control

Variable	OLS_rob	Between-effects (BE) model	Fixed-effects (FE) model	FE_rob (i)	GLS random-effects (RE) model	RE_rob (i)
highgrade	0.0478	0.0701	0.0198	0.0198	0.0478	0.0478
	0.0402	0.0438	0.0631	0.0565	0.035	0.0445
Age	-0.0304	0.0196	-0.1497	-0.1497	-0.0304	-0.0304
	0.0328	0.044	0.2338	0.2421	0.0421	0.0432
Agesq	0.0002	-0.0003	0.0021	0.0021	0.0002	0.0002
	0.0004	0.0005	0.0013	0.0014	0.0005	0.0005
rain	-0.0024	-0.0121	0.0037	0.0037	-0.0024	-0.0024
	0.0085	0.0117	0.0255	0.0231	0.0106	0.0089
rainlagged	0.055	0.0635	0.0347	0.0347	0.055	0.055
	0.023	0.0239	0.0474	0.0457	0.0234	0.0214
Landowners~c	0.1954	0.4058	-0.0024	-0.0024	0.1954	0.1954
	0.1541	0.1832	0.2674	0.2438	0.1354	0.1556
cotincshare	-0.6905	-0.2174	-1.4411	-1.4411	-0.6905	-0.6905
	1.9683	2.2119	2.5926	2.7286	1.547	1.8223
cotincshar~q	0.838	0.0881	2.3725	2.3725	0.838	0.838
	2.1987	2.4608	3.056	3.0741	1.7458	2.0804
distiput	-0.0036	-0.0085	0.0014	0.0014	-0.0036	-0.0036
	0.0094	0.0125	0.0149	0.0122	0.0084	0.0098
distbank	-0.0045	-0.0068	-0.0055	-0.0055	-0.0045	-0.0045
	0.0065	0.009	0.0163	0.0159	0.006	0.0065
_cons	5.8466	4.722	8.5311	8.5311	5.8466	5.8466
	1.9113	1.9174	8.7127	8.3592	1.7821	1.6822
N	78	78	78	78	78	78
r2	0.4469	0.5917	0.5468	0.5468		
r2_o		0.4256	0.0814	0.0814	0.4469	0.4469
r2_b		0.5917	0.0009	0.0009	0.5441	0.5441
r2_w		0.4257	0.5468	0.5468	0.4763	0.4763
sigma_u			0.8738	0.8738	0	0
sigma_e			0.7473	0.7473	0.7473	0.7473
rho			0.5776	0.5776	0	0

Notes: (i) vce(robust) uses the robust or sandwich estimator of variance.
Source: Authors' estimations.

First-differencing has eliminated $\boldsymbol{\alpha_i}$ in (**4.4**), so OLS estimation of this model leads to consistent estimates of $\boldsymbol{\beta}$ in the FE model. The coefficients of time-invariant regressors are not identified, because then $x_{it} - x_{i,t-1} = 0$, as was the case for the within estimator (ibid.).

As expected, the coefficient for education is not identified because *Age* here is time-invariant in the whole sample only. The signs of the coefficients for the other regressors do not change compared with the other estimators (cf. **table 4.8**). And again it is the regressors: Lagged rain; landownership and distance to bank (cf. **table 4.7**) that are statistically significant in the whole sample, whereas it is the regressors: Age; age squared and rain in the catchment sub-sample.

The FD estimator like the within estimator, provides consistent estimators when the individual effects are fixed. For our panel with **T=2**, the *FD* and *within estimators* are equivalent. Thus, **table 4.9** gives the results of a simplified version of **model (3.18**) above that explain the growth in p.a.e. consumption between 1998 ("**t-1**") and 2004 ("**t**").

Table 4.9: First-Differences Estimator with Cluster-Robust Standard Errors

	Whole				Catchment				Counter Factual			
D.Lnpae	Coef.	Robust Std.Err.	Coef.	Robust Std.Err.	Coef.	Robust Std.Err.	Coef.	Robust Std.Err.	Coef.	Robust Std.Err.	Coef.	Robust Std.Err.
highgrade												
D1.	0,066	0,082	,132*	0,070	0,050	0,045	0,051	0,049	0,020	0,062	0,000	0,063
Age												
D1.	(dropped)				0,094*	0,054			-0,150	0,266		
Agesq												
D1.	(dropped)				0,001***	0,000			0,002	0,002		
rain												
D1.	-0,049	0,032			-0,001	0,014	-0,050**	0,021	0,004	0,025	0,010	0,023
rain_lag												
D1.	0,117**	0,047	,071**	0,029	-0,021	0,017	0,032	0,022	0,035	0,050	0,040	0,054
Landowners~c												
D1.	0,477***	0,128	,544***	0,118	-0,143*	0,074	0,121**	0,053	-0,002	0,268	-0,054	0,239
cotincshare												
D1.	3,174	3,408	3,888	3,636	-1,759	2,148	0,315	2,105	-1,441	2,995	-1,918	3,088
cotincshar~q												
D1.	-4,204	4,151	-5,377	4,813	2,472	2,225	0,646	2,741	2,373	3,374	2,500	3,278
distiput												
D1.	-0,005	0,014			-0,004	0,007	-0,005	0,007	0,001	0,013	0,001	0,013
distbank												
D1.	-0,0205**	0,009	-,024**	0,010	-0,004	0,007	0,005	0,007	-0,006	0,017	-0,008	0,013
R2	0,940		0,935		0,884		0,803		0,547		0,498	
N	45		45		45		45		34		34	

Source: Authors' estimations.

5. Discussion of Estimation Results

5.1. Specification tests of the functional form

A key assumption maintained in section 4.2 is that *the functional form* was correctly specified for the estimated relationship of the list of included regressors. In this section we will check the validity of this assumption. *Formal specification tests* can distinguish between systematic lack of fit and random sampling errors. If the zero-conditional-mean assumption:

$$E[e \mid x_1, x_2, \dots , x_5] = 0 \tag{5.1}$$

is violated, the coefficient estimates are inconsistent. The three main problems that cause the zero-conditional-mean assumption to fail in a regression model are: Improper specification of the model; endogeneity of one or more regressors; or measurement error of one or more regressors (Baum 2006).

This section reports the results of *formal specification tests* of models (4.1)–(4.2), whereas section 5.2 addresses endogeneity and measurement errors.

First consider testing the specification of the parametric model (4.1). The consistency of the linear regression estimator requires that the sample regression function corresponds to the underlying regression function or true model for the response variable y = LNpae98 (or LNpae04). The cost of *omitting relevant variables* is high. A variable mistakenly excluded from our model is unlikely to be uncorrelated in the population or in the sample with the regressors (**Table A8.1-2**).

Graphically analyzing our regression data

Within our specification analysis, we want to examine the simple *bivariate relationships* between y (LNpae98) and the five regressors in x underlying our parametric linear regression model (4.1). From the scatter plots there doesn't appear to be any collinerarity problems among the regressors neither at the provincial (**Figures A3-4**) nor at the district levels, which is confirmed by the correlation matrices (**Tables A8.1-2**).

In each of the **six panes of figures A5.1-5.2**,[40] we see that several of the 1245 observations in the 1998 dataset are far from the straight line linking the response variable (LNpae98/LNpae04) and that regressor. The same applies to an even greater number of the 649 observations in the 2004 dataset. The *t statistics* shown in each of the six panels test *the hypothesis* that the least-squares regression line has a slope significantly different from zero.[41] Based on the 1998 dataset the outlying values are particularly evident in the graphs for *s10q8* (household ownership of motor vehicle), especially in the catchment districts (**figure A5.1.2**), and *age2* (age of head of household squared) where low values of E[s10q8|x] and E[age2|x] are associated with LNpae much higher than those predicated by the model. On the other hand, the slope of the *assetown09* (household ownership of motor vehicle) is positive in the 2004 dataset associated with the positive slope in the catchment districts [550 observations], given the few (99) observations in the control districts, where the slope is negative.

Moreover, including *irrelevant regressors* does not violate the zero-conditional mean assumption (Baum 2006).[42] Fortunately, in all six cases for both the 1998 and 2004 dataset most of the points are not clustered around the horizontal line at ordinate zero. Nor is it possible to say that the slope in several of the six cases is significantly different from zero, which means that some $\mathbf{x_g}$ (e.g. stratum, cotton income share, distance and infrastructure in both 1998 and 2004 and age2 in 2004) can't be considered to be making an important contribution to the model beyond that of the other regressors in 1998. Finally, the points in each of the six panes in both 1998 and 2004 are closely enough grouped around the straight line in the plot, not to cast in doubt the linear specification of $\mathbf{x_g}$ in the model.

Misspecification of the functional form

Our model **(5.1)** that includes the appropriate five regressors may be misspecified because of the model may not reflect *the algebraic form of the relationship* between the response variable and those

---

[40] *The added-variable plot* identifies the important variables in a relationship by decomposing the multivariate relationship into a set of two-dimensional plots (see Cook and Weisberg, 1994:191-194).
[41] The Student's *t distribution* tells us that, for $v = n\text{-}1 = 5$ degrees of freedom, the probability that $t > 2.571$ is 0.025 (see table of selected t values). Also, the probability that $t < -2.571$ is 0.025. Using the formula for t with $t = \pm\ 2.571$ a 95% confidence interval for the population's mean may be found.
[42] The long model delivers unbiased and consistent estimates of all its parameters, including those of the irrelevant regressors, which tend to zero (Baum, 2006). Despite *the risk of collinearity*, Baum(2006), in line with the works by David Hendry, recommends *a model selection strategy* that starts with a general specification and seeks to refine it by imposing appropriate restrictions.

regressors. In the words of (Baum 2006), this problem may be easier to deal with than the omission of relevant variables. In a misspecification of the functional form, we have all the appropriate variables at hand and only have to choose the appropriate form in which they enter the regression function.

Ramsey's RESET

We see (**table A9**) that both the parsimonious *Ramsey's (1969) omitted-variable regression specification error test (RESET)*, [43] which augments the regression using the second, third, and fourth powers of the fitted values $\hat{y}$ series (of LNpae) as well as the Ramsey RESET test using the powers of the individual independent regressors themselves *reject* RESET's null hypothesis of no omitted variables for the model, albeit at the 10% significant level in the first test.[44] Since the hypothesis is rejected, then these powers cannot be excluded from the regression without compromising the level of explication of the dependent variable. That indicates that the original regression was not specified correctly.

We re-specify equation **(4.1)** to include: The square of age (age2), whether household own bicycle (s10q6/assetownd07) and whether head of household ever attended school (s4q5).[45] We see that the respecified and extended model's values *no longer reject the RESET* at the 1% significance level and so we may conclude that the regression was specified correctly in 1998 only (**table A9**). The relationship between squared age (age2), school attendance (s4q5) and bicycle ownership (s10q6) appears to be nonlinear (although with wrong pattern of signs on their coefficients).

Specification plots

Next, we evaluate the specifications of the model and the extended model by use of two types of plots. First we graph the residuals on the y-axis versus the predicted (i.e. fitted) values on the x-axis.[46] Then we plot the residuals against a specific regressor. The residuals "bounce randomly" around the 0

[43] The **Wald test** is a statistical test, typically used to test whether an effect exists or not. In other words, it tests whether an independent variable has a statistically significant relationship with a dependent variable.
[44] The idea that *a polynomial* constructed from these estimated values can be seen as a "reduced form" for many different combinations of powers and cross-products involving the independent variables.
[45] Unfortunately, there are only 24 observations in rural Eastern Province in 1998 if we use the variable "highest grade ever attended by head of household," which would have given us an opportunity to measure the association between the level of education and level of poverty.
[46] This helps to identify non-linearity, outliers, and non-constant variance.

line (i.e. linear is reasonable). No one residual "stands out" from the basic random pattern of residuals (i.e. no outliers).[47] The residuals roughly form a "horizontal band" around 0 line (i.e. constant variance). This in turn indicates that there doesn't seem to be a problem with either models.

With regards to the residuals for *the continuous variables* – age2, distance to input market, and cotton income share – the assumption of homoskedasticity doesn't seem to be challenged by the residuals on the y axis -versus- the values of a predictor on the x axis plots, which through the random scatter indicated that the models are probably good. The only exception might be the case of *'household ownership of motor vehicles'* in the original model, where the spread in the estimator (i.e. the observed residuals) $e_i = y_i - \hat{y}$ are larger when the household expressed no ownership, leading to the suspicion that the errors are *heteroscedastic* in both 1998 and in 2004.

Outlier statistics and measures of leverage

To evaluate the adequacy of the specification of the fitted models, we will consider evidence relating to the models' *robustness to influential data*.[48] We calculate a measure of each data point's *leverage*, calculated from the diagonal elements of the "hat matrix", $h_j = x_j(X'X)^{-1}x'_j$, where $\mathbf{x_j}$ is the jth row of the regressor matrix.

From **table 4.2** displaying summary statistics of the dependent variable LNpae98 (LNpae04), we see that it ranges from minimum 4,605 (6,629) to maximum 12,662 (13.226). The five largest values of the leverage measure are listed in **tables A10.1-2**. The five largest squared residuals are listed in **tables A11.1-2**. These leverage values versus the (normalized) squared residuals are displayed in figures **A6.1-2**, which show that there in 1998 are around a dozen of points with very high leverage or very large squared residuals.[49] Several of the largest values of leverage or the squared residuals correspond to the extreme values of the pae98 recorded in the dataset.

[47] We have deliberately from the outset chosen to disregards the *commercial large-scale farmers stratum*.
[48] *An outlier* in a regression relationship is a data point with an unusual value and a high degree of leverage on the estimates. An outlier may be an observation associated with a large residual (in absolute terms), a datapoint that the model fits poorly. On the other hand, an unusual data point that is far from the center of mass of the $\mathbf{x_j}$ distribution may also be an outlier, although the residual associated with that data point will often be small because the least-squares process attaches a squared penalty to the residual in forming the least-squares criterion (Baum, 2006).
[49] A large value of leverage does not imply a large squared residual (Baum, 2006).

DFITS statistics

Every data point affects the estimates of the slope, the intercept, the predicted values, and the error variance of the regression line to some degree; an influential observation does so more than other points. Hence, we will next compute the *DFITS statistics* of Welsch and Kuh (1977) to provide a summary of the leverage values and magnitudes of residuals.[50] The DFITS measure is a scaled difference between the in-sample and out-of-sample predicted values for the *j*th observation. DFITS evaluates the result of fitting the regression model including and excluding that observation.

Tables A.12.1-2 displays the 55 large values of DFITS for the 1998 dataset and 27 large values of DFITS for the 2004 dataset for which cut-off = 1. That is, about 6% of the 1998 observations and 4% of the 2004 observations are flagged by the DFITS cut-off criterion. Many of those observations associated with large positive DFITS have the top-coded values for pae98, however the magnitude of the positive and negative DFITS are more or less the same in both the 1998 and 2004 datasets. The identification of *top-coded values* that represent an arbitrary maximum recorded pae98 suggests that we consider a *different estimation technique* for this model that can properly account for the *censored* nature of the pae98.

DFBETA statistics

Finally, we compute DFBETAs for the discrete Treatment dummy variable *"infrastructure"* (ℓ) regress or in our regression model, which measures the distance that this regression coefficient would shift when the jth observation is included or excluded from the regression, scaled by the estimated standard error of the coefficient.[51]

Compared to the DFITS measure, in **tables A.13.1-2** we see a similar pattern for the DFBETA for *"infrastructure"* with 7.6% of the 1998 sample of 919 observations and 7.1% of the 2004 sample of

[50] Since RStudent is the first component of DFFits, *an outlier* is more likely than another observation to be identified as strongly influencing the predicted value. Even though outliers are more distant from the regression line than others points are, they still tend to tilt the regression line towards them. The second component of DFFits is the ratio, which increases as the observation's leverage increases. This means that observations with independent variable values far from the mean tend to tilt the regression line more strongly than do the central observations.

[51] One rule of thumb suggests that *a DFBETA value greater than unity* in absolute value might be reason for concern since this observation might shift the estimated coefficient by more than one standard error (Baum, 2006).

647 observations exhibiting large values of this measure. As with DFITS, the large positive values are more or less the same size in magnitude as their negative counterparts in both 1998 and 2004. Around 45% of the positive values are associated with the top-coded pae98 values above ZMK50,000 in 1998. These presumably better off rural households have values well in excess of its minimum or mean.

Thus, there is evidence of many data points with *a high degree of leverage*. Whether these pae98 / pae04 data have been improperly measured, we can only speculate about this. But these observations in particular have been identified by the *DFITS and DFBETA measures*. Removing the bottom-coded and top-coded observations from the sample would remove rural households from the sample non-randomly, affecting the wealthiest and poorest rural households.[52]

Mathematically, *measurement error* (commonly termed errors-in-variables) has the same effect on an OLS regression model as *endogeneity* of one or more regressors. This measurement error is of concern, because the economic behaviour we want to model - that of the rural households in Eastern Province - presumably is driven by the actual measures, not our mis-measured approximations of those factors. So if we fail to capture the actual measure, we may misinterpret the behavioural response (Baum 2006).

5.2. Instrumental-variable estimators

The zero-conditional mean assumption must hold for us to use linear OLS regression. There are three common instances where this assumption may be violated: *Endogeneity* (simultaneous determination of response variable and regressors), *omitted-variable bias*, and *errors in variables* (measurement error in the regressors) since households are unlikely to be able to recall household expenditure. The solution to each is the same econometric tool: the *instrumental-variables (IV) estimator* (Baum 2006).

To derive consistent estimators of **(4.1)**, we must find an **IV** that satisfies two properties: The instrument $\mathbf{z}$ must be uncorrelated with $\mathbf{e}$ (that is, the *orthogonality assumption*) but must be highly correlated with $\mathbf{x_j}$. A variable that meets those two conditions is an IV or instrument for $\mathbf{x_j}$ that deals

[52] A version of the tobit model, two-limit tobit, can handle censoring of both lower and upper limits (Baum, 2006).

with the correlation of $\mathbf{x_j}$ and the error term (Johnston and DiNardo 1997; Wooldridge 2002; Baum 2006).

In order to capture the transitory component in household expenditure, we use *rainfall as an instrument*.[53] As most of the rural households in our sample rely on crop yields as the main source of income, rainfall can explain a non-trivial share of the *intertemporal variation* in total household expenditure (LNpae). We argue that *a rainfall-induced variation* in household income will be less tainted by measurement error. Rainfall fluctuations arguably captures a transitory and exogenous component in household's income and expenditure, *uncorrelated* to life-cycle decisions, knowledge or other variables that may enter the households preferences over choice of cash crops. Thus, our ***identification strategy*** rests on the assumption that rainfall affects consumption outcomes only via the total income variable and not via some omitted variable.

A potential scenario is that households are able to completely smooth output in the event of a *weather shock* by adjusting their consumption of leisure. This type of income smoothing suggests that a realized weather shock can affect the consumption of leisure without affecting the observed level of rural production. As noted by (Rosenzweig and Wolpin 2000), the credence of using *weather variation* as an instrument for rural income rests on how the market structure is defined, and on how expenditure and income is observed. Rainfall has a decisive impact on expenditure in our sample, implying that rural households in Eastern Province are unable to borrow and save across transitory income shocks. The rural households are completely liquidity constrained, because they are unable to save or borrow across aggregated income shocks to achieve at least perfect intertemporal consumption smoothing.

[53] We do not have more than one potential *candidate instrument*.
A valid instrument variable $\mathbf{z}$ requires $E[u|z]= 0$ (exogeneity, i.e. $\mathbf{z}$ is validly excluded from the outcome equation of interest. In other words, $\mathbf{z}$ has no direct effect on $\mathbf{y}$ but only indirectly through its impact on $\mathbf{x}$) and $Cov(z,x) \neq 0$ (relevance of the instrument: $F>0$). If the instrument is weak, 2SLS is no longer reliable since the estimator will not only be badly biased but the estimator will also have a nonnormal sampling distribution making statistical inference meaningless.

Empirical implementation[54]

Our extended baseline regression equation is (4.1), where we model the logarithm of total household expenditure p.a.e. (LNpae) as a function of a couple of continuous variables: distiput and age2 (distance to input market and age of head of household squared); and a set of indicator variables: stratum124, s10q8, s10q6, and s4q5 (rural household stratum, motor vehicle ownership, bicycle ownership, and school attendance), and infrastructure, and indicator for residency in one of the 5 districts affected by the EPFRP. The endogenous variable is *cotincshare*, cotton sales as a share of total household income. Here we do not consider LNpae and cotincshare are simultaneously determined, but rather that cotincshare cannot be assumed independent of the error term: the same correlation that arises in the context of an endogenous regressor in a structural equation (Baum 2006). The cotincshare is instrumented with two factors excluded from the equation (4.1): the average yearly rainfall in each district in Eastern Province for the agricultural season of the year of the survey (1998/1999 and 2004/2005) and the average yearly rainfall lagged with one year (1997/1998 and 2003/2004).[55]

Tables A5.1-2 presents the descriptive statistics. Then we fit the IV model using first-stage regression to evaluate the degree of correlation between these two factors and the endogenous regressor. From **tables A14.a-b** the *first-stage regression results* suggest that one of the two excluded instruments, namely the one year lagged rainfall variable is highly correlated with the endogenous variable (cotincshare) in the 1998/99 agricultural season and that this is the case for both the lagged (rain04) and rainfall during the 2004/05 agricultural season (rain05). The exception is the rainfall variable rain99 (rainfall during the agricultural season where the survey was carried out).

Tables A15.a-b show that the endogenous regressor *cotincshare* has a distinguishable negative IV coefficient in 1998 but a positive coefficient in 2004 as expected in the latter case. Thus, conditioning on the other factors included in the equation, *cotincshare* does seem to play a role in determining the LNpae98 and especially in the case of LNpae04, although the sign of the coefficient estimate doesn't agree with the predictions of theory and empirical findings in 1998. Furthermore, in

---

[54] Since we do *not have two candidate instruments*, we won't use the alternative approach, **2SLS**, which combines multiple instruments into one optimal instrument, which can be used in the simple **IV** estimator. The order condition is often stated as requiring that there be at least as many instruments as endogenous variables. The order condition is necessary, but not sufficient, for the rank condition to hold (Baum, 2006).

[55] Instruments that satisfy the rank condition but are not sufficiently correlated with the endogenous variables for the large-sample approximations to be useful are known as *weak instruments* (Baum, 2006).

1998 only the distance to input market coefficient estimate seem to agree with the predictions of theory.

One cannot directly *test for the exogeneity assumption* but there are indirect ways of testing it. **Table A16.1** shows the Anderson-Rubin-Sargan-Basmann test results for the validity of overidentifying structural restrictions, which signals a strong rejection of the null hypothesis that the instruments are uncorrelated with the error term (i.e. the disturbance process) and thereby suggests that we should not be satisfied with this specification of the equation according to the 1998 dataset.

In the 2004 dataset the instrument is weak given the F-statistics in our specification is less than 2. This leads to the *weak instruments problem* as discussed by Bound, Jaeger, and Baker (1995),[56] that is, instruments that are only weakly correlated with the included endogenous variables. Unfortunately, weak instruments pose considerable challenges to inference using GMM and IV methods (Stock et al., 2002).

From **table A17.1** we see that the endogenous regressor *cotincshare* still does play a role in the equation. *The Hansen J statistics* is the GMM equivalent to the Sargan test.[57] The independence of the instruments and the disturbance process is called into question by the strong rejection of the J test null hypothesis.

In **tables A.18.1-2** we test whether the subset of the excluded rainfall instruments is appropriately exogenous. The equation estimated without suspect instruments, free of one additional orthogonality condition on rain99, doesn't have a *Hansen J statistics*, whereas the *C statistic* for the instrument, rain99, tested is highly significant. Hence rain99 does not appear to be valid in this context.
To evaluate whether we have found a more appropriate specification, we reestimate the equation with the remaining instrument rain98. From the results shown in **tables A.19.1-2** we see that *cotincshare* no longer appears a significant regressor and the equation's J statistics is zero again. The following

[56] It is not useful to think of weak instruments as a “small sample” problem. Bound et al.,(1995) provided an empirical example of weak instruments despite having 329,000 observations (Stock et al., 2002).
[57] Hansen's J is the most common diagnostic used in the GMM estimation to evaluate the suitablity of the model. a rejection of the null hypothesis implies that the instruments do not satisfy the required orthogonality conditions (Baum, 2006).

regressors: stratum124, s10q8, s10q6, and s4q5 all seem to be playing a role in this from of the estimated equation.

Detection of heteroskedasticity (unequal variance of the errors) can be achieved by many different tests under the assumption of a linear statistical model of the form (4.1). In **table 5.1** below we compute several of the tests for heteroskedasticity appropriate in the IV context from the last regression reported in **table A.19.1**. All of the tests using the 1998 dataset signal a problem of heteroskedasticity in the estimated equation's disturbance process.

Table 5.1: Testing for heteroskedasticity in the IV context in 1998 and 2004

IV heteroskedasticity test(s) using levels of IVs only										
Ho: Disturbance is homoskedastic			1998					2004		
Pagan-Hall general test statistic	19,4	Chi-sq(9)	P-value	=	0,0220	5.802	Chi-sq(9)	P-value	=	0.7596
Pagan-Hall test w/assumed normality	22,574	Chi-sq(9)	P-value	=	0,0072	4.786	Chi-sq(9)	P-value	=	0.8525
White/Koenker nR2 test statistic	21,554	Chi-sq(9)	P-value	=	0,0104	12.062	Chi-sq(9)	P-value	=	0.2099
Breusch-Pagan/Godfrey/Cook-Weisberg	24,855	Chi-sq(9)	P-value	=	0,0031	12.177	Chi-sq(9)	P-value	=	0.2035
IV heteroskedasticity test(s) using fitted value (X-hat*beta-hat) & its square										
Ho: Disturbance is homoskedastic			1998					2004		
Pagan-Hall general test statistic	7,711	Chi-sq(2)	P-value	=	0,0212	2.767	Chi-sq(2)	P-value	=	0.2507
Pagan-Hall test w/assumed normality	8,961	Chi-sq(2)	P-value	=	0,0113	2.267	Chi-sq(2)	P-value	=	0.3219
White/Koenker nR2 test statistic	8,381	Chi-sq(2)	P-value	=	0,0151	4.851	Chi-sq(2)	P-value	=	0.0884
Breusch-Pagan/Godfrey/Cook-Weisberg	9,665	Chi-sq(2)	P-value	=	0,008	4.897	Chi-sq(2)	P-value	=	0.0864

Notes: Pagan-Hall statistics is robust to the presence of heteroskedasticity elsewhere in a system of simultaneous equations and to non-normally distributed disturbances. White's general test (White, 1980), or its generalization by Koenker (1981), also relaxes the assumption of normality underlying the Breusch-Pagan test (see Deaton, 1997:79).[58] The Breusch-Pagan (1979) and White tests for heteroskedasticity can be applied in 2SLS models, but Pagan and Hall(1983) point out that they will be valid only if heteroskedasticity is present in that equation and nowhere less in the system. The other structural equations in the system corresponding to the endogenous regressors must also be homoskedastic even though they are not being explicitly estimated.
Source: Author's calculations.

Weighted Least Squares

Weighted least squares provides one method for dealing with heteroscedasticity. For our data, we obtain the results shown in **table 5.2**. In the 1998 data the sign of the coefficient of cotton income share change from negative in the OLS estimation to positive in the WLS estimation, plus with a smaller magnitude. The other noteworthy change is the R2 falls from 19.22% in the OLS model to 15.9% in the WLS model in 1998. In the 2004 data the coefficient sign change for three regressors: Distance to input

[58] *The Breusch-Pagan test* for heteroskedasticity uses a test-equation: the squared residuals divided by the residual variance are explained by all exogenous variables. The test statistic is computed as half the difference between the Total Sum of Squares and the Sum of Squared Residuals, which has a Chi-square distribution. Warning: this test should only be used if the endogenous variable is NOT used as lagged exogenous variable and if the number of observations is VERY LARGE. All OLS assumptions should be satisfied, including normality of the error term.

market; Infrastructure dummy; and School attendance. Moreover, stratum becomes significant and motor vehicle significant at the 0.01 level from the 0.1 level. Finally, the R2 contrary to the 1998 data goes slightly up in the 2004 dataset.

Table 5.2: Weighted Least-Squares Estimator vs OLS estimator, 1998 & 2004

	1998				2004			
	WLS		OLS		WLS		OLS	
LNpae98	Coef.	Robust Std. Err.	Coef.	Std.Err.	Coef.	Robust Std. Err.	Coef.	Std.Err.
Cotton Income Share	0,083	0,219	-0,104	0,188	0,158	0,158	0,132	0,139
Stratum, excl. Large	0,226***	0,048	0,247***	0,038	-0,258**	0,118	-0,058	0,118
Distance Inputmkt	-0,002	0,002	-0,003*	0,002	0,001	0,002	-0,001	0,002
Motorvehicle Ownership	-1,885***	0,314	-1,591***	0,286	0,726***	0,257	0,54*	0,314
Infrastructure	-0,093	0,107	-0,006	0,085	0,064	0,151	-0,037	0,126
Age	-0,041**	0,020	-0,015	0,014	-0,029	0,021	-0,024	0,019
Age Squared	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Bike Ownership	-0,427***	0,100	-0,368***	0,076	-0,049	0,110	-0,005	0,100
School Attendance	-0,353***	0,109	-0,420***	0,078	0,058	0,123	-0,012	0,109
Constant	11,974***	0,548	11,220***	0,433	11,380***	0,577	11,363	0,570
N	873		919		647		647	
R2	0,159		0,192		0,019		0,017	

Source: Author's estimations.

Testing the relevance of instruments

We will now test whether the instrument variable are highly correlated with the included endogenous variable - *cotincshare* - by examining the fit of the first-stage regressions.[59] The relevant test statistics here relate to the explanatory power of the excluded instruments in these regressions. The statistics proposed by Bound, Jaeger, and Baker (1995) can diagnose instrument relevance only in the presence of one endogenous regressor (Baum 2006).[60]

The **tables A20.1-2** illustrate the weak-instrument problem with a variation on the LNpae equation using rain98 and rain99 as instruments.[61] In order to ensure that we have found a strong instrumental variable we provide *Shea's (1997) partial R2 statistic* and its associated F-statistic. In the first-stage regression results, *Shea's partial R2 statistic* is very small for this equation indicating that the instrument is insufficiently relevant to explain the endogenous regressor in both 1998 and 2004, and

[59] The first-stage regressions are reduced-form regressions of the endogenous regressors, x_1, on the full set of instruments, z.
[60] When multiple endogenous regressors are used, other statistics are required.
[61] There is the familiar difficulty of finding convincing instruments, and it is usually easier to justify the role of instruments such as assets and lagged income as predictors of e.g. permanent income than it is to define their absence from the direct determination of consumption in the equation of interest (Deaton, 1997:352). Although average rainfall is predictably difference from place to place, the deviation of each year's rainfall from its local mean is serially uncorrelated and thus unpredictable (op.cit., p.353).

the *Cragg-Donald statistics* rejects its null hypothesis of under identification. *The Anderson canonical correlation statistic* rejects its null hypothesis, suggesting that the instrument may be adequate to identify the equation.[62] Finally, the redundant (rain98) option indicates that rain98 does provide useful information to identify the equation. For one endogenous regressor - *cotincshare -,* an F statistics less than 10 is not the case for the 1998 dataset,[63] contrary to the 2004 dataset, which is cause for concern (Staiger and Stock, 1997:557)

Durbin-Wu-Hausman tests for endogeneity in IV estimation

There are three equivalent ways of obtaining the Durbin component of the Durbin-Wu-Hausmann (DWH) statistics. The different commands implement distinct versions of the tests, which although asymptotically equivalent can lead to different inference from finite samples (Baum 2006).

The first method fit the less efficient but consistent model using IV. Then fit the fully efficient model. The comparison in **table 5.3** is restricted to the point estimate and estimated standard error of the endogenous regressor, *cotincshare*; the Haussmann test statistics accept the exogeneity of the rain98 variable. The command also warns of difficulties computing a positive-definite covariance matrix. The small chi2 value indicates that estimation of the equation with regress yields consistent results.

Table 5.3: Durbin-Wu-Haussmann Tests for Endogeneity in IV estimation

	---- Coefficients ----					---- Coefficients ----			
1998	**(b) iv**	**(B) .**	**(b-B) Difference**	**sqrt(diag(V_b-V_B)) S.E.**	**2004**	**(b) iv**	**(B) .**	**(b-B) Difference**	**sqrt(diag(V_b-V_B)) S.E.**
stratum124	.2031568	.2510582	-.0479013	.0215904	stratum124	-.4984822	-.0484214	-.4500609	.0763337
distiput	-.0039727	-.0032202	-.0007525	.0003487	distance11	.0019052	-.0012524	.0031576	.0005356
s10q8	-1.690421	-1.583606	-.1068143	.0481315	assetownd09	-.8796261	.57119	-1.450816	.2460694
infrastruc~e	-.0330978	-.0025029	-.0305949	.0138882	infrastruc~e	-.0354586	-.0369459	.0014874	.0002523
s1q3b	-.0121234	-.0151452	.0030217	.0013891	s1q3b	-.0585523	-.023063	-.0354893	.0060193
age2	-.0000538	-6.21e-06	-.0000476	.0000215	age2	.0007242	.0001812	.0005429	.0000921
s10q6	-.4330964	-.3649632	-.0681332	.0316846	assetownd07	.3288292	-.0119875	.3408167	.0578051
s4q5	-.4429078	-.4174164	-.0254913	.0114684	s4q5	.2676141	-.018386	.2860001	.0485078
_cons	11.5889	11.19786	.3910402	.1785454	_cons	11.82469	11.3528	.471887	.0800356

b = consistent under Ho and Ha; obtained from ivreg2			b = consistent under Ho and Ha; obtained from ivreg2		
B = inconsistent under Ha, efficient under Ho; obtained from regress			B = inconsistent under Ha, efficient under Ho; obtained from regress		
Test: Ho: difference in coefficients not systematic			Test: Ho: difference in coefficients not systematic		
chi2(2)	=	= (b-B)'[(V_b-V_B)^(-1)](b-B)	chi2(1)	=	(b-B)'[(V_b-V_B)^(-1)](b-B)
	=	5.02		=	34.76
Prob>chi2	=	0.0812	Prob>chi2	=	0.0000
(V_b-V_B is not positive definite)			(V_b-V_B is not positive definite)		

[62] Canonical correlation Measure of the strength of the overall relationships between *canonical variates* (Also referred to as linear composites, linear compounds, and linear combinations) for the independent and dependent variables. In effect, it represents the bivariate correlation between the two canonical variates.
[63] Given that, one recommendation when faced with a weak-instrument problem is to be parsimonius in the choice of instruments.

Source: Author's calculations.

Tables A20.3-4 in appendix illustrates the second method, which fits the fully efficient model and specifies the regresssors to be tested. The second method's C test statistic agrees qualitatively with that from Haussmann by accepting the exogeneity/orthogonality of the rain98 variable.

Finally **table A20.5** in appendix illustrates the method. The test statistic is identical to that provided by the C Statistic in **table A20.1**. All forms of the test agree that the estimation of this equation with linear regression yields consistent results. The regressor *cotincshare* must be considered exogenous in the fitted model.

Quantile Regression

We now turn to the "identically distributed" assumption, and consider the consequences of heteroskedasticity. Just as lack of independence appears to be the rule rather than the exception, so does heteroskedasticity, which seems to be almost always present in survey data (Deaton 1997).[64] We follow Deaton(1997) who suggests that the computation of quantile regressions is useful, both in its own right, because *quantile regression estimates* will often have better properties than OLS, as a way of assessing the heteroskedasticity in the conditional distribution of the logarithmic transformation of pae, and as a stepping stone to the nonparametric methods.

The **tables 5.4.1-4** and **tables A.21.1-2** in appendix shows the quantile regression outputs corresponding to the 20th, 40th, 60th and 80th percentile in the distribution of LNpae98 calculated using 942 rural household observations and the distribution of LNpae04. The slopes that are the response coefficients β of the covariates of these four regression functions differ. These differences and the different spread between the regression functions show the increase in the conditional variance of the regression among better-off rural households.[65]

[64] Even when individual behaviour generates homoskedastic regression functions within strata or villages, but there is heterogeneity between villages, there will be heteroskedasticity in the overall regression function. In the presence of heteroskedasticity, OLS is inefficient and the usual formulas for standard errors are incorrect (Deaton, 1997).
[65] These regressions do not tell us anything about the causal processes that generate the differences, but they present the data in an interesting way that can be suggestive of ideas for a deeper investigation (Deaton, 1997).

Table 5.4.1 Quantile 1: 20%, 1998

	stratum124	cotincshare	distiput	s10q8	infrastructure	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Quantile Regression										
LNpae98	.2716251**	-.0782083	.001294	-.4652535	.1687908	.0015369	-.0001599	-.6087141**	-.1944722	7.488913
	(.1373107)	(.5336642)	(.0065642)	(.3969689)	(.2379402)	(.034054)	(.000338)	(.2355047)	.2209168	1.050488
observations	186	186	186	186	186	186	186	186	186	186
Panel B: OLS Regression										
LNpae98	.168185**	-.0343721	-.0014761	-.1687206	.0944224	.0054081	-.0001351	-.450483***	-.1376368	7.509946
	(.0684282)	(.2667184)	(.0029171)	(.7416053)	(.1163264)	(.0186177)	(.0001829)	(.113021)	(.1075116)	(.9047165)
observations	186	186	186	186	186	186	186	186	186	186
Panel C: IV Regression										
LNpae98	.1700681**	.1138567	-.001169	-.0933139	.0933699	.0056862	-.0001356	-.4348589***	-.1381655	7.388749
	(.069249)	(.8488862)	(.0033632)	(.8479329)	(.1165689)	(.0186953)	(.000183)	(.1414588)	(.1076443)	(1.119847)
observations	186	186	186	186	186	186	186	186	186	186

Standard errors in brackets, * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5.4.2 Quantile 1: 20%, 2004

	stratum124	cotincshare	distiput	s10q8	infrastructure	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Quantile Regression										
LNpae98	.1877515	-.4743539	-.0004347	.3145396	.1307892	-.034932	.0004011	-.4790948**	-.1251921	9.632436
	(.2221348)	(.3468837)	(.0052569)	(.4758589)	(.3070168)	(.059044)	(.0006046)	(.2295752)	(.2738291)	(1.569067)
observations	130	130	130	130	130	130	130	130	130	130
Panel B: OLS Regression										
LNpae98	.0086425	-.072652	.0029035	.0541163	.0571551	-.0430928**	.0004307**	-.1509163	-.1656449	10.52856
	(.1093686)	(.1359383)	(.0022261)	(.2024322)	(.1214499)	(.0204027)	(.0002074)	(.0936526)	(.1023013)	(.5196166)
observations	130	130	130	130	130	130	130	130	130	130
Panel C: IV Regression										
LNpae98	.1117769	-.7570402	.0014271	.1494557	-.0130489	-.0370546	.0003344	-.1774808	-.1686605	10.52329
	(.246006)	(1.431557)	(.0039288)	(.2982788)	(.1979758)	(.0257289)	(.0003037)	(.1169498)	(.1127629)	(.5719717)
observations	130	130	130	130	130	130	130	130	130	130

Standard errors in brackets, * significant at 10%, ** significant at 5%, *** significant at 1%.
Source: Author's calculations.

Table 5.4.3 Quantile 4: 80%, 1998

	stratum124	cotincshare	distiput	s10q8	infrastructure	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Quantile Regression										
LNpae98	.0656825	-.1010467	-.004214	-1.378079***	.5049315**	-.0442934	.0004334	.112295	-.1021425	13.07247
	(.0703631)	(.4745047)	(.0041808)	(.3869489)	(.2162816)	(.0402143)	(.0004523)	(.1939267)	(.2243541)	(.9556869)
observations	188	188	188	188	188	188	188	188	188	188
Panel B: OLS Regression										
LNpae98	.0366734	-.1632588	-.0034037**	-.7335591***	.3297382***	-.0097287	.0000602	.0220449	-.0932895	11.47478
	(.0312563)	(.2288021)	(.0017186)	(.1617527)	(.090622)	(.0167093)	(.0001824)	(.0809641)	(.0956322)	(.4021212)
observations	188	188	188	188	188	188	188	188	188	188
Panel C: IV Regression										
LNpae98	.0434536	.1065753	-.0033837**	-.7149364***	.3453743***	-.0111327	.0000793	.0248216	-.1114647	11.43999
	.0334405	.5142713	.0017256	.1654593	.0948021	.0169444	.000186	.0814176	.1008837	.4080262
observations	188	188	188	188	188	188	188	188	188	188

Standard errors in brackets, * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5.4.4 Quantile 4: 80%, 2004

	stratum124	cotincshare	distiput	s10q8	infrastructure	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Quantile Regression										
LNpae98	-.0969562	.3841689	.0006647	.0823007	-.0983912	.0149554	-.0001947	.1317238	.1019841	12.79436
	(.4026205)	(.4181356)	(.0080136)	(.63335)	(.349107)	(.0444247)	(.0004248)	(.3000069)	(.2936413)	(1.443234)
observations	142	142	142	142	142	142	142	142	142	142
Panel B: OLS Regression										
LNpae98	.0168404	.1963199	.0002025	-.0637579	-.0974462	.0000327	-.0000102	.114323	.0201334	12.76118
	(.1321313)	(.1407112)	(.0024202)	(.3917926)	(.1155226)	(.0169465)	(.0001717)	(.0975078)	(.1065241)	(.5901007)
observations	142	142	142	142	142	142	142	142	142	142
Panel C: IV Regression										
LNpae98	.1104283	-.6054509	.0002117	.2312543	-.020249	.0109664	-.0001538	.1024231	.0208296	12.3533
	(.2428376)	(1.660179)	(.0027015)	(.7490474)	(.2048184)	(.0294243)	(.0003527)	(.1115709)	(.1189138)	(1.068066)
observations	142	142	142	142	142	142	142	142	142	142

Standard errors in brackets, * significant at 10%, ** significant at 5%, *** significant at 1%.
Source: Author's calculations.

5.3. Testing Linear Panel-Data Models

The essential distinction in microeconometric analysis of panel data is that between FE and RE models. If effects are fixed, then the pooled OLS and RE estimators are inconsistent, and instead the within (or FE) estimator needs to be used. The within estimator is otherwise less desirable, because using only within variation leads to less efficient estimation and inability to estimate coefficients of time-invariant regressors (Cameron & Trivedi, 2009).

Haussmann test for fixed effects

Under the null hypothesis that individual effects are random, these estimators should be similar because both are consistent. Under the alternative, these estimators diverge. This juxtaposition is a natural setting for **a Haussmann test**, comparing FE and RE estimators (Cameron & Trivedi, 2009).[66]

In **tables 5.5a-b** we compare the estimable coefficients of time-varying regressors. *Sigmamore* specifies that both covariance matrices are based on the (same) estimated disturbance variance from the efficient estimator. We obtain that for all the coefficients of the regressors, that the test of RE against FE yields t-values that are higher in the catchment case than in the control areas for both the default version of the Haussmann as well as for sigmamore and sigmaless. Moreover, the t-values for Age; Age square; rain; rainlagged; landownership; and cotton income share indicate a highly significant difference for the catchment areas only, except for age square, which is significant at the 10% level in the control areas. Finally, the overall statistics for the catchment areas only has $p<=0.001$ for all three Haussmann tests, which leads to a strong rejection of the null hypothesis that the RE provides consistent estimates.

[66] Or can be applied to a key subset of these (often one key regressor).

Table 5.5a: Haussmann Test for Fixed Effects, catchment areas

	Coefficients			Hausmann FE RE	Hausman Sigmamore	Hausman Sigmaless
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B)	sqrt(diag(V_b-V_B)	sqrt(diag(V_b-V_B)
	FE	RE	Difference	S.E.	S.E.	S.E.
highgrade	0,050	0,045	0,005	0,017	0,032	0,026
Age	0,094	-0,005	0,099	0,050**	0,063*	0,052**
Agesq	0,001	0,000	0,001	0,0003***	0,0005**	0,0004***
rain	-0,001	-0,039	0,038	0,013***	0,0184**	0,015**
rainlagged	-0,021	0,029	-0,049	0,012***	0,019**	0,0152***
Landowners~c	-0,143	0,188	-0,330	0,064***	0,089***	0,073***
cotincshare	-1,759	0,340	-2,099	0,74***	1,541	1,262*
cotincshar~q	2,472	0,027	2,445	0,7393***	1,676*	1,373*
distiput	-0,004	-0,001	-0,003	.	0,004	0,003
distbank	-0,004	-0,004	0,000	0,003	0,005	0,004
chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)				27,210	26,820	39,990
Prob>chi2				0,001	0,002	0,000

Notes: b = consistent under Ho and Ha; obtained from xtreg. B = inconsistent under Ha, efficient under Ho; obtained from xtreg. Test: Ho: difference in coefficients not systematic.
sigmamore and sigmaless specify that the two covariance matrices used in the test be based on a common estimate of disturbance variance (sigma2). **sigmamore** specifies that the covariance matrices be based on the estimated disturbance variance from the efficient estimator. This option provides a proper estimate of the contrast variance for so-called tests of exogeneity and overidentification in instrumental variables regression. **sigmaless** specifies that the covariance matrices be based on the estimated disturbance variance from the consistent estimator (Stata Manual).[67]
Source: Author's estimations.

Table 5.5b: Haussmann Test for Fixed Effects, Control Areas

	Coefficients			Hausmann FE RE	Hausman Sigmamore	Hausman Sigmaless
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B)	sqrt(diag(V_b-V_B)	sqrt(diag(V_b-V_B)
	FE	RE	Difference	S.E.	S.E.	S.E.
highgrade	0,020	0,048	-0,028	0,052	0,043	0,049
Age	-0,150	-0,030	-0,119	0,230	0,202	0,229
Agesq	0,002	0,000	0,002	0,001*	0,001*	0,001*
rain	0,004	-0,002	0,006	0,023	0,020	0,023
rainlagged	0,035	0,055	-0,020	0,041	0,035	0,039
Landowners~c	-0,002	0,195	-0,198	0,231	0,193	0,219
cotincshare	-1,441	-0,691	-0,751	2,080	1,685	1,910
cotincshar~q	2,373	0,838	1,535	2,508	2,055	2,329
distiput	0,001	-0,004	0,005	0,012	0,010	0,011
distbank	-0,006	-0,005	-0,001	0,015	0,013	0,015
chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)				5,070	7,310	5,690
Prob>chi2				0,828	0,605	0,770

Source: Author's estimations.

[67] *Sigmamore or sigmaless* are recommended when comparing fixed-effects and random-effects linear regression because they are much less likely to produce a nonpositive-definite differenced covariance matrix (although the tests are asymptotically equivalent whether or not one of the options is specified) (Stata on-line Manuel, 2009).

6. Conclusions

In this paper we have presented evidence through the analysis of existing Zambian rural household survey data to help us understand the linkage between consumption growth and rural feeder road improvements in rural areas in Zambia's Eastern Province. Our results from both cross-section and pseudo-panel data analysis have proven that this is indeed a difficult undertaking. One important reason for this is certainly the measurement of expenditure and the other key regressors, especially with regards to the imperfect 2004 dataset the quality of which has been affected by our resort to using the collapsing technique. This approach seriously affected the variation, and hence explains the low R2s in 2004. This evidently, makes it difficult to base policy prescriptions on our unclear results from a number of different models with different policy implications also.

Agriculture is the overwhelming dominant activity in the rural areas in Zambia's agriculture-based Eastern Province and therefore the main source of pro-poor economic growth. Despite the fact that farming as the principal activity among the sampled rural households had fallen from 63 % in 1998 to 55% in 2004 (**table 3.3**) as a testimony of a diversification of income sources towards the labour market and the rural nonfarm economy and successful migration out of rural areas, the cotton production still generated the largest share of these *households' income* in both 1998 and 2004.

This happened notwithstanding the fact that the world cotton market in the late 1990s was severely marked by the collapse in the world prices as illustrated by *price indices of cotton products* in **figure A9** and likewise observed by (Hertel & Winters, 2006; Balat & Porto, 2005b). Hence, it is somewhat surprising that cotton's share in income among the rural households in Eastern province rose by 234% for the total sample, and 220% and 310% for respectively the catchment and control districts over the same period.

Moreover, on average, the sampled rural households *farm just less than one hectare* for food crops and 1.28 hectare for non-food crops hadn't change noteworthy between 1998 and 2004. On the other hand, *the mean distance to services and community assets* for the rural households had diminished significantly in the same period, mainly due to the improvement in rural transport infrastructure not only associated with the EPFRP catchment districts but probably also due to similar projects being implemented in the province (Kingombe, 2009b).

One weakness of our approach is that although we stratify the rural households into small-, medium and rural-nonfarm the district level approach doesn't allow us to properly capture the heterogeneity both within among the households e.g. subsistence small scale farmers versus commercial smallholder farming. The same applies to the eight districts that together constitutes Eastern Province, e.g. according to their agricultural potential and access to markets.

We try to address this shortcoming through infrastructure dummy variable in the case of the whole sample, with the catchment districts considered as proxies for agricultural potential and access to markets but with the drawback of not capturing the diverse local conditions within each district according to distinct agro-ecologies, which produce a wide range of farming systems and crops.

Following Horowitz and Lee (2002), we also believe that it might be useful to develop additional semi-parametric models that achieve both good estimation precision and a high degree of flexibility. Although the usefulness of semiparametric models in econometrics applied on rural development and statistics is not fully understood, virtually any new application of these models could provide useful additional information for the ex-post policy evaluation of poor area public infrastructure projects.[68]

Finally, *the cohort data approach* seem has many advantages over both the parametric and the semi-parametric approach based on independent cross-sectional household surveys, which could be explored further in future research.

[68] The GLM framework is the standard nonlinear model framework in many areas of applied statistics. Cameron&Trivedi(2009:321) mention that it is little used in econometrics.

References

Balat, J. F. and G. G. Porto (2005a). "The WTO Doha Round, Cotton Sector Dynamics and Poverty Trends in Zambia." Policy Research Working Paper Series **3697**.

Balat, J. F. and G. G. Porto (2005b). "Globalization and complementary policies: poverty impacts in rural Zambia." Working Paper 11175: 42.

Baum, C. F. (2006). An Introduction to Modern Econometrics Using Stata, Stata Press.

Brambilla, I. and G. G. Porto (2007). "Market Structure, Outgrower Contracts and Farm Output. Evidence From Cotton Reforms in Zambia."

Cameron, A. C. and P. K. Trivedi (2009). Microeonometrics Using Stata, Stata Press.

Chen, S., R. Mu, et al. (2006). "'Are There Lasting Impacts of a Poor-Area Development Project?' " World Bank Policy Research Working Paper 4084: 48.

Chiwele, D. K., P. Muyatwa-Sipula, et al. (1998). Private Sector Response to Agricultural Marketing Liberalisation in Zambia. A Case Study of Eastern Province Maize Markets. Research report. A. Olukoshi. Uppsala, Nordiska Afrikainstitutet. Research Report No.107: 90.

CSO (1994). National Census of Agriculture 1990/92. Census Report (Part I). Lusaka, Central Statistical Office.

CSO (2001). 2000 Census of Population and Housing. Preliminary Report. R. o. Z. Central Statistical Office**:** 50.

CSO (2003). Zambia 2000 Census of Population and Housing. Agriculture Analytical Report. C. S. Office. Lusaka, Republic of Zambia Central Statistical Office**:** 68.

Deaton, A. (1997). The analysis of household surveys: a microeconometric approach to development policy. Washington, D.C., The Johns Hopkins University Press.

Dercon, S., D. O. Gilligan, et al. (2007). "The impact of roads and agricultural extension on consumption growth and poverty in fifteen Ethiopian villages." CSAE WPS/2007-01.

Dercon, S. and J. Hoddinott (2005). "Livelihoods, growth, and links to market towns in 15 Ethiopian villages " FCND Discussion Paper 194.

Devereux, S. (2002). "FROM WORKFARE TO FAIR WORK. The Contribution of Public Works and other Labour–Based Infrastructure Programmes to Poverty Alleviation." Issues in Employment and Poverty (5): 45.

Glewwe, P. and H. Jacoby (2000). Recommendations for Collecting Panel Data. Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 years of the Living Standards Measurement Study. M. Grosh and P. Glewwe. Washington, D.C., World Bank. **2:** 275-314.

Govereh, J., T. S. Jayne, et al. (2000). "Improving Smallholder and Agribusiness Opportunities in Zambia's Cotton Sector: Key Challenges and Options." FSRP Working Paper No. 1.

Hardin, J. and J. Hilbe (2007). "Generalized Linear Models and Extensions."

Härdle, W., M. Müller, et al. (2004). Nonparametric and Semiparametric Models: An Introduction. Berlin, Heidelberg, New York, Springer.

He, X., W. K. Fung, et al. (2005). "Robust estimation in generalized partial linear models for clustered data." J. Amer. Statist. Assoc. **100**(1176-1184).

Hidehiko, I. (2005). Semiparametric Data Analysis, UCL**:** 14.

Horowitz, J. L. and S. Lee (2002). "Semiparametric methods in applied econometrics: do the models fit the data?" Statistical Modelling **2**: 3–22.

Islam, N. (1995). "Growth Empirics: A panel data approach." Quarterly Journal of Economics **110**(4): 1127-1170.
Johnston, J. and J. DiNardo (1997). Econometric Methods, McGraw-Hill International Editions.
Kabwe, S. (2009). "The Evolution of the Cotton and Textile Industry in Zambia. A presentation at ICAC Research Associate Program, Washington DC, 10 April, 2009.".
Kabwe, S. and D. Tschirley (2007). "An Effective Public-Private Coordination in Zambia's Cotton Sector: Deliberation on the Cotton Act. Presented at the Agricultural Consultative Forum. May, 2007. ."
Lokshin, M. (2006). "Semi-parametric difference-based estimation of partial linear regression models.".
Mankiw, G. N., D. Romer, et al. (1992). "A Contribution to the Empirics of Economic Growth." The Quarterly Journal of Economics **107**(2): 407-437
McCulloch, N., B. Baulch, et al. (2001). "Poverty, Inequality and Growth in Zambia during the 1990s." Discussion Paper No. 2001/123: 47.
Mu, R. and D. van de Walle (2007). "Rural roads and poor area development in Vietnam." Policy Research Working Paper Series **4340**.
Robinson, P. M. (1988). "Root-N-Consistent Semiparametric Regression." Econometrica **56**(4): 931-954.
Rosenzweig, M. R. and K. I. Wolpin (2000). "Natural .Natural Experiments in Economics." Journal of Economic Literature **38(4)**: 827-874.
Tschirley, D. and S. Kabwe (2007). "Cotton in Zambia: 2007 Assessment of its Organization, Performance, Current Policy Initiatives, and Challenges for the Future. Working Paper No. 26. ."
UNCTAD (2004). The Least Developed Countries Report 2004. Linking International Trade with Poverty Reduction, United Nations.
Winters, L. A. (2002). "Trade Liberalisation and Poverty: What are the Links?" The World Economy **25**(9): 1339-1367.
Winters, L. A. (2002). "Trade Liberalisation and Poverty: What are the Links? ." World Economy **25**: 1339-1367.
Winters, L. A., N. McCulloch, et al. (2004). "Trade Liberalization and Poverty: The Evidence So Far." Journal of Economic Literature, American Economic Association **42**(1): 72-115.
Wooldridge, J. (2002). Econometric Analysis of Cross-Section and Panel Data,. Cambridge, Mass., MIT Press.
World Bank (2007a). World Development Report, 2008. Agriculture for Development. Washington, DC., The World Bank,.
Yatchew, A. (1998). "Nonparametric regression techniques in economics." Journal of Economic Literature **36**: 669-721.
Yatchew, A. (1998). "Nonparametric regression techniques in economics. ." Journal of Economic Literature **36**: 669-721.
Yatchew, A. (2005). Semiparametric Regression for the Applied Econometrician, Cambridge University Press.

Zambia Food Security Research Project (2000). "Improving smallholder and agri-business opportunities in Zambia's cotton sector: Key challenges and options. Lusaka, Zambia."

Zhoua, J., Z. Zhub, et al. (2008). "Robust testing with generalized partial linear models for longitudinal data." Journal of Statistical Planning and Inference **138**(6): 1871-1883.

Zulu, B. and D. Tschirley (2002). "An Overview of the Cotton Sub-Sector in Zambia."

Web Links

Partnership **in** **S**tatistics for Development in the 21st Century (PARIS21)
http://www.paris21.org/

The **I**nternational **H**ousehold **S**urvey **N**etwork (IHSN)
http://www.internationalsurveynetwork.org/home/

Country questionnaires (database)[69]
http://www.internationalsurveynetwork.org/home/?lvl1=tools&lvl2=questionnaire&lvl3=country#

Other survey databanks, incl. IFPRI
http://www.internationalsurveynetwork.org/home/?lvl1=activities&lvl2=catalog&lvl3=databank

Central Statistical Office. 2008. Poverty in Zambia -1991 – 2006.
http://www.zamstats.gov.zm/lcm.php

World Bank. 2008. Africa Household Survey Databank.
http://www4.worldbank.org/afr/poverty/databank/default.cfm

World Bank. 2008. ZAMBIA, 1996 & 1998. Living Conditions Monitoring Survey (LCMS). This CD-ROM contains 2 surveys.
http://www4.worldbank.org/afr/poverty/databank/cdroms/in_stock_zmb_96_98.cfm?cd=zmb_96_98&CFID=2260647&CFTOKEN=99110ce0ffa34585-A8C54B66-06F7-FAC6-18EBE63C159DC9B4&jsessionid=98302753a9902a643156

UNCTAD electronic portal on commodities http://www.unctad.org/infocomm/

Information on cotton http://www.unctad.org/infocomm/anglais/cotton/sitemap.htm

International Cotton Advisory Committee (ICAC) http://www.icac.org/

UNCTAD Commodity Price Statistics on-line
http://www.unctad.org/Templates/Page.asp?intItemID=1889&lang=1
http://stats.unctad.org/handbook/ReportFolders/ReportFolders.aspx?CS_referer=&CS_ChosenLang=en

STATA http://www.stata.com

[69] Zambia, 1950-2008: Found 32 questionnaire(s) in 15 survey(s).

Annexes

Table A1: Eastern Province Districts codes

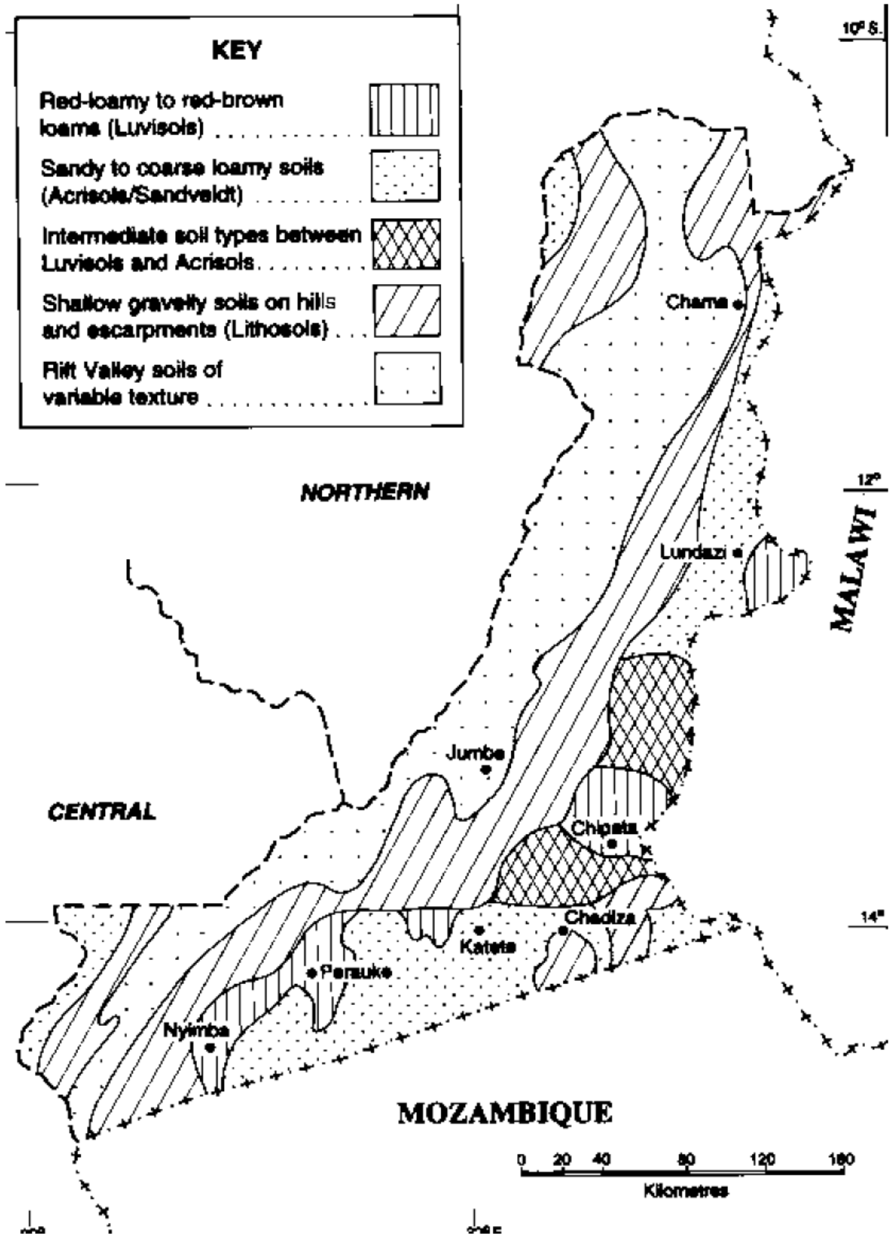
District Name	District Code	Catchment	Control	Sampled Households 1998	Sampled Households 2004
Chadiza	301	Yes	No	120	102
Chama	302	No	Yes	90	23
Chipata	303	Yes	No	286	199
Katete	304	Yes	No	150	139
Lundazi	305	Yes	No	170	209
Mambwe	306	No	Yes	135	74
Nyimba	307	No	Yes	120	85
Petauke	308	Yes	No	235	168
Total	1998	961	345	1306	
	2004	817	182		999

Source: CSO. Living Conditions Monitoring Survey II 1998 data user's guide.

Household's unique identification (1998): This variable is a concatenation of the first variables of the Survey minus the panel number.
hid = province+district+const+ward+csa+sea+rururb+stratum+centrlty+hhn

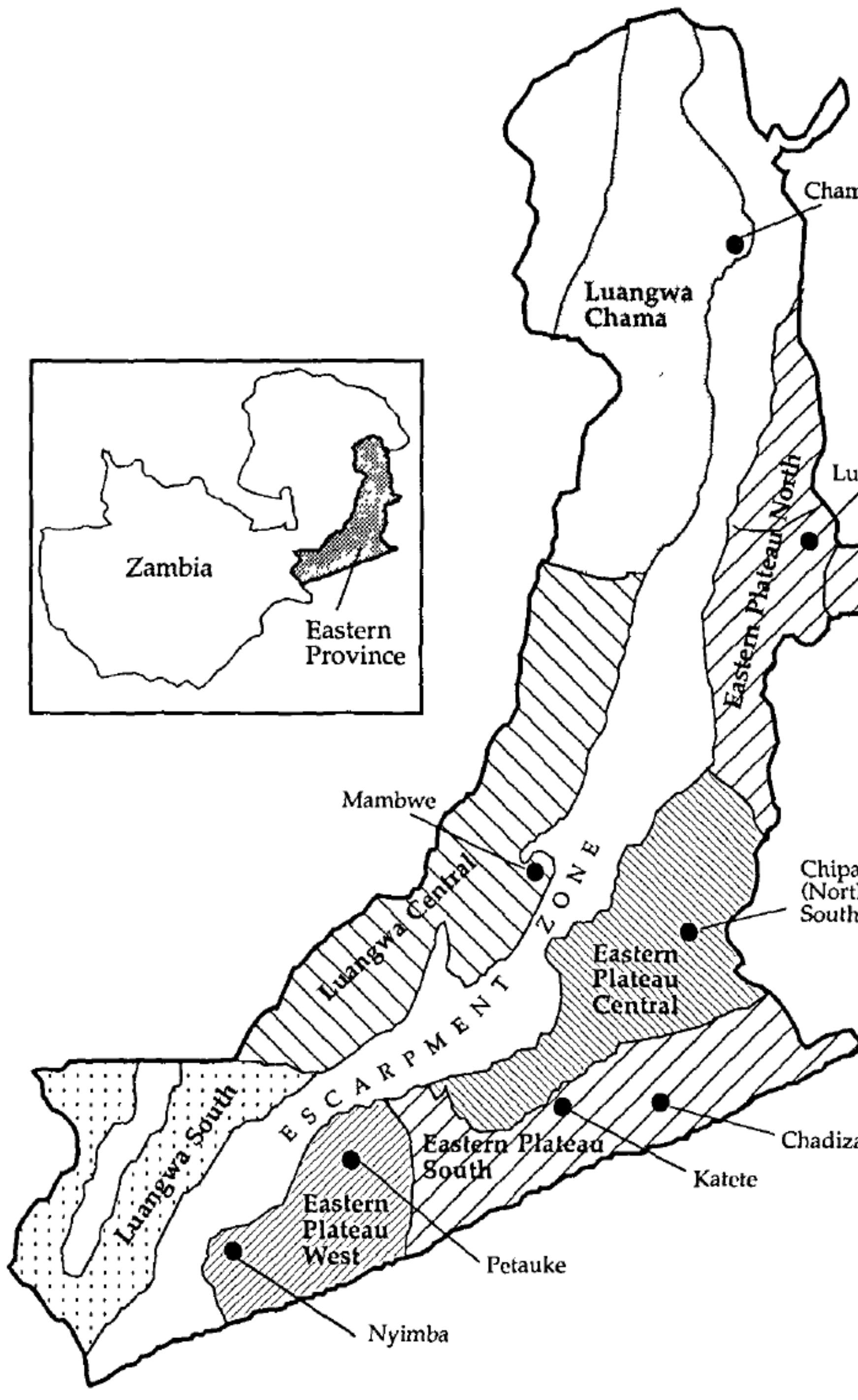
Household's unique identification (2004): This variable is a concatenation of the first variables of the Survey minus the centrality and the panel number.
hid = province+district+const+ward+csa+sea+rururb+stratum+hhn

Map A1: The major soil types of Eastern Zambia



Source: Samuel Simute, C. L. Phiri, and Bo Tengnäs, 1998: Agroforestry Extension Manual for Eastern Zambia.

Map A2: Agroecological zones and agricultural districts, Eastern Province, Zambia



Source: ARPT (Adaptive Research Planning Team), Eastern Province Agricultural Development Project, "Annual Report, 1985-86" (Chipata, Zambia, 1986, mimeographed).

Table A.1.1: Poverty Lines, Current Prices, Zambian Kwacha (ZMK)

	OER	UPL	LPL	CPI	CPI Ch
1996	1275	28979	20181	175	43,1
1998	2195	47187	32861	100	24,5
2003	4737	92185	64530	87,3	21,4
2004	4848	111747	78223	73,5	18

Notes: Official Exchange Rate (OER=USD/ZMK); Upper Poverty Line (UPL) and Lower Poverty Line (LPL) per month; Consumer Price Index (CPI) (1998=100); CPI Annual percentage change (Ch).
Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A.1.2: Poverty Lines, Constant Prices, Zambian Kwacha (ZMK)

	OER	UPL	LPL	Implicit GDP Deflator 2000=100*	Implicit GDP Deflator 1998=100
1996	1919,18	43620,38	30377,27	42,10	66,43
1998	2195,00	47187	32861	63,37	100,00
2003	1681,57	32724,45	22907,29	178,51	281,70
2004	1428,82	32934,38	23054,1	215,01	339,30

Sources: Author's calculations. * IMF, World Economic Outlook (WEO) Database.

Table A.2. Rural Summary statistics on Outcome Indicators

	1998		2004		Percentage change	
	Treatment	**Control**	**Treatment**	**Control**	**Treatment**	**Control**
*Mean Income***	64853,44	29071,57	59651,33	46688,28	-8%	61%
Standard Deviation	(1184530)	(47889.78)	(250051.9)	(65130.34)		
Observations	962	346	817	182		
Mean Consumption	15076,05	18845,14	37570,73	27289,34	149%	45%
Standard Deviation*	(34193.13)	(29965.03)	(53322.62)	(40485.26)		
Observations	960	343	817	182		
*Mean Consumption***	15076,05	18845,14	58276,76	42329,07	287%	125%
Standard Deviation	(34193.13)	(29965.03)	(82709.85)	(62797.55)		
Observations	960	343	817	182		
*Mean P.A.E. Consumption***	15596,61	19603,87	59891,14	43848,03	284%	124%
Standard Deviation	(34647.19)	(31105)	(86171.3)	(64447.27)		
Observations	958	342	817	182		
Income Poverty Rate						
Upper Poverty Line = ZMK47187	0,918	0,917	0,550	0,655	-40%	-29%
Lower (Food) Poverty Line = ZMK32861	0,855	0,853	0,404	0,489	-53%	-43%
Consumption Poverty Rate						
Upper Poverty Line = ZMK47187	0,952	0,955	0,882	0,769	-7%	-19%
Lower (Food) Poverty Line = ZMK32861	0,917	0,917	0,647	0,777	-29%	-15%

Notes: * Using the CPI04 = 335,5 Deflator; ** Using the Foodbasket = 216,3 Deflator.
What we are measuring when we do not take into account survey design is the moments (mean and standard deviation) of the sample distribution whist when survey design is been applied we get the moments of the population.
Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A3.a Principal Economic Activity of Household Head, Rural Areas, Numbers of Household Heads by Quintile of Consumption, Catchment Districts

		1998			2004		
	Main economic activities status	All	Lowest 20%	Highest 40%	All	Lowest 20%	Highest 40%
1	In Wage Employment	2,83%	0,00%	6,64%	3,83%	3,24%	4,70%
2	Running Business / Self-Employed	1,47%	1,29%	4,15%	1,36%	0,54%	1,25%
3	Farming, Fishing, Forestry	65,45%	68,24%	57,26%	94,68%	95,68%	94,04%
4	Not working but looking for works / Means to do business	1,47%	0,00%	2,49%	0,00%	0,00%	0,00%
5	Not working not looking for works / Means to do business, but available to do so	0,52%	0,86%	0,00%	0,00%	0,00%	0,00%
6	Full Time student	14,03%	10,73%	18,67%	0,12%	0,54%	0,00%
7	Full Time at home / home duties	7,23%	8,58%	7,05%	0,00%	0,00%	0,00%
8	Retired	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	Too old to work	1,68%	4,29%	0,83%	0,00%	0,00%	0,00%
10	Other	5,34%	6,01%	2,90%	0,00%	0,00%	0,00%
	Total	956	233	241	809	185	319

Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A3.b Principal Economic Activity of Household Head, Rural Areas, Numbers of Household Heads by Quintile of Consumption, Control Districts

		1998			2004		
	Main economic activities status	All	Lowest 20%	Highest 40%	All	Lowest 20%	Highest 40%
1	In Wage Employment	4,35%	0,00%	0,00%	8,33%	14,81%	12,24%
2	Running Business / Self-Employed	7,25%	9,33%	6,36%	8,33%	5,56%	2,04%
3	Farming, Fishing, Forestry	56,81%	60,00%	40,91%	27,22%	303,70%	93,88%
4	Not working but looking for works / Means to do business	0,29%	0,00%	0,00%	0,00%	7,41%	2,04%
5	Not working not looking for works / Means to do business, but available to do so	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	Full Time student	17,39%	16,00%	10,91%	11,11%	1,85%	0,00%
7	Full Time at home / home duties	8,41%	8,00%	5,45%	4,44%	0,00%	0,00%
8	Retired	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	Too old to work	0,87%	0,00%	0,00%	0,00%	0,00%	0,00%
10	Other	4,64%	6,67%	4,55%	1,67%	0,00%	0,00%
	Total	345	75	110	180	54	49

Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A4.a: Pct. of Households in Rural Catchment Areas Owning Particular Assets by Quintile

		1998			2004		
	Asset Ownership	All	Lowest 20%	Highest 40%	All	Lowest 20%	Highest 40%
1	plough	79,36%	81,86%	80,57%	66,71%	56,99%	73,37%
2	crop sprayer	85,71%	92,83%	79,35%	74,42%	70,43%	77,09%
3	bicycle	48,67%	62,03%	43,52%	41,25%	26,88%	39,94%
4	motorcycle	98,19%	100,00%	96,56%	99,76%	100,00%	99,69%
5	motorvehicle	96,15%	99,58%	90,89%	97,67%	94,62%	99,69%
6	tractor	99,14%	100,00%	97,77%			
7	radio	50,94%	75,95%	29,15%			
8	telephone	98,59%	100,00%	96,56%			
9	scotch cart	89,40%	91,56%	88,66%	81,76%	78,49%	87,62%
10	donkey	99,37%	100,00%	99,19%	98,65%	97,31%	99,07%
	Total	1274	237	494	817	186	323

Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A4.b: Pct. of Households in Rural Control Areas Owning Particular Assets by Quintile

		1998			2004		
	Asset Ownership	All	Lowest 20%	Highest 40%	All	Lowest 20%	Highest 40%
1	plough	93,22%	94,94%	90,55%	75,27%	77,78%	81,63%
2	crop sprayer	90,04%	92,41%	88,56%	80,22%	79,63%	79,59%
3	bicycle	52,12%	69,62%	45,27%	36,81%	29,63%	46,94%
4	motorcycle	97,88%	100,00%	96,02%	98,90%	96,30%	100,00%
5	motorvehicle	97,03%	100,00%	94,03%	98,90%	96,30%	100,00%
6	tractor	98,94%	100,00%	97,51%			
7	radio	49,15%	74,68%	30,85%			
8	telephone	97,88%	100,00%	95,02%			
9	scotch cart	96,40%	97,47%	94,53%	84,07%	83,33%	91,84%
10	donkey	99,36%	100,00%	98,51%	97,80%	77,78%	100,00%
	Total	472	79	201	182	54	49

Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A7.b: Descriptive Statistics of covariates, 1998 and 2004, Catchment Districts

Type	Variable Name	Variable	1998 Obs	1998 Mean	1998 Std.Dev.	1998 Min.	1998 Max.	2004 Obs	2004 Mean	2004 Std.Dev.	2004 Min.	2004 Max.	Percentage change
CV	Log pae monthly household expenditure	LNPAE98	948	8,942	1,215	3,912	13,534	817	10,338	1,169	6,563	13,771	16%
CV	Cotton Sales share of household income	cotincshare	929	0,110	0,207	0,010	1	817	0,283	0,332	0	1	157%
DV	Stratum, excl. Large AHH	straturnl24	958	1,553	1,049	1	4	817	1,206	0,431	1	4	n.a.
CV	Distance to Inputmarket	Distiput	964	23,498	22,825	0	99	817	13,742	18,493	0	99	-42%
DV	EPFRP Treatment	infrastructure	964	0	0	0	0	817	1	0	1	1	n.a.
DV	Plough Ownership	Plough	964	0,756	0,430	0	1	817	0,667	0,472	0	1	n.a.
DV	Bicycle Ownership	Bicycle	964	0,470	0,499	0	1	817	0,330	0,471	0	1	n.a.
DV	Scotchcart Ownership	Scotchcart	964	0,880	0,326	0	1	817	0,818	0,386	0	1	n.a.
DV	Motorvehicle Ownership	Motorvehicle	964	0,975	0,156	0	1	817	0,977	0,151	0	1	n.a.
CV	Age of Head of Household	Age	961	43,751	15,843	15	99	817	42,765	14,984	20	90	-2%
CV	Age Squared	Agesq	961	2164,902	1545,234	225	9801	817	2053,103	1459,178	400	8100	-5%
DV	Head of HH ever attended School	s4q5	720	0,461	0,499	0	1	815	0,237	0,425	0	1	n.a.

Catchment Districts: Chadiza, Chipata, Katete, Lundazi, and Petauke districts.
Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A7.c: Descriptive Statistics of covariates, 1998 and 2004, Control districts

Type	Variable Name	Variable	1998 Obs	1998 Mean	1998 Std.Dev.	1998 Min.	1998 Max.	2004 Obs	2004 Mean	2004 Std.Dev.	2004 Min.	2004 Max.	Percentage change
CV	Log pae monthly household expenditure	LNPAE98	339	9,146	1,276	4,605	12,662	182	10,090	1,101	6,629	13,226	10%
CV	Cotton Sales share of household income	Cotincshare	345	0,085	0,185	0,010	1	182	0,336	0,362	0	1	294%
DV	Stratum, excl, Large AHH	Stratuml24	346	1,390	0,958	1	4	182	1,137	0,345	1	2	n.a.
CV	Distance to Inputmarket	Distiput	347	20,452	23,208	0	99	182	4,945	9,793	0	43	-76%
DV	EPFRP Treatment	Infrastructure	347	0	0	0	0	182	0	0	0	0	n.a.
DV	Plough Ownership	Plough	347	0,916	0,277	0	1	182	0,753	0,433	0	1	n.a.
DV	Bicycle Ownership	Bicycle	347	0,519	0,500	0	1	182	0,368	0,484	0	1	n.a.
DV	Scotchcart Ownership	Scotchcart	347	0,960	0,197	0	1	182	0,841	0,367	0	1	n.a.
DV	Motorvehicle Ownership	Motorvehicle	347	0,980	0,141	0	1	182	0,989	0,105	0	1	n.a.
CV	Age of Head of Household	Age	345	41,110	15,325	20	86	182	43,440	15,232	21	80	6%
CV	Age Squared	Agesq	345	1924,217	1442,787	400	7396	182	2117,725	1491,381	441	6400	10%
DV	Head of HH ever attended School	s4q5	248	0,306	0,462	0	1	182	0,126	0,333	0	1	n.a.

Control districts: Chama, Nyimba, and Mambwe.
Source: Author's calculations based upon the LCMS II and LCMS IV datasets.

Table A9: Ramsey's null hypothesis of no omitted variables for the model

	1998	2004
	F(3, 1236)*	F(3, 640)
Using powers of the fitted values of Lnpae	2.48	1.99
	F(7, 1232)***	F(7, 636)**
Using powers of the independent variables	7.21	2.60
	F(3, 907)	F(3, 635) *
Using powers of the fitted values of Lnpae	0.78	2.22
	F(10, 900)***	F(10, 628) **
Using powers of the independent variables	3.99	2.18

Source: Author's calculations based upon the LCMS II and LCMS IV.

Table A20.3: OLS estimation: Estimates efficient for homoskedasticity only. Statistics consistent for homoskedasticity only, 1998

			Number of obs	=	919
			F(8, 910)	=	26,88
			Prob > F	=	0
Total (centered) SS	=	1439,834	Centered R2	=	0,1912
Total (uncentered) SS	=	74624,53	Uncentered R2	=	0,9844
Residual SS	=	1164,607	Root MSE	=	1,131

LNpae98	Coef.	Std.Err.	t	P>t	[95% Conf.	Interval]
stratum124	0,256558	0,037226	6,89	0,000	0,183499	0,329617
distiput	-0,00328	0,001706	-1,93	0,055	-0,00663	6,41E-05
s10q8	-1,58715	0,286279	-5,54	0,000	-2,149	-1,02531
infrastruc~e	-0,00322	0,085248	-0,04	0,970	-0,17053	0,164083
age2	-0,00016	2,41E-05	-6,68	0,000	-0,00021	-0,00011
s10q6	-0,36427	0,076342	-4,77	0,000	-0,51409	-0,21444
s4q5	-0,42729	0,077976	-5,48	0,000	-0,58032	-0,27426
cotincshare	-0,1091	0,187597	-0,58	0,561	-0,47728	0,259069
_cons	10,88423	0,301723	36,07	0,000	10,29208	11,47639

Sargan statistic (Lagrange multiplier test of excluded instruments): 1.197
Chi-sq(1) P-val = 0.2739

-orthog- option:

Sargan statistic (eqn. excluding suspect orthogonality conditions): 0.000
Chi-sq(0) P-val = .

C statistic (exogeneity/orthogonality of suspect instruments): 1.197
Chi-sq(1) P-val = 0.2739

Instruments tested: cotincshare

Included instruments: stratum124 distiput s10q8 infrastructure age2 s10q6 s4q5 cotincshare

Excluded instruments: rain98

Table A20.4: OLS estimation: Estimates efficient for homoskedasticity only. Statistics consistent for homoskedasticity only, 2004

Table A20.5 Tests of endogeneity of cottonshare of total income, 1998 and 2004

Tests of endogeneity of: cotincshare					
H0: Regressor is exogenous	1998				
Wu-Hausman F test:	1,18563	F(1,909)	P-value	=	0,2765
Durbin-Wu-Hausman chi-sq test	1,19711	Chi-sq(1)	P-value	=	0,2739
Tests of endogeneity of: cotincshare					
H0: Regressor is exogenous	2004				
Wu-Hausman F test:					
Durbin-Wu-Hausman chi-sq test					

Source: Author's estimations.

Table A21.1: Quantile 2 (40%), 1998

	stratum124	cotincshare	distiput	s10q8	infrastruc~e	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Quantile Regression									
LNpae98	.0306036	.0838091	-.0020793*	.1379662**	.0440118	.0217552**	-.0002333**	.0230573	-.0544589	7.756616
	(.0291806)	(.1214851)	(.0010685)	(.0656821)	(.0514329)	(.0089519)	(.0000917)	(.0490552)	(.0487617)	(.2266492)
observations	183	183	183	183	183	183	183	183	183	183
	Panel B: OLS Regression									
LNpae98	-.0046445	.1052395	-.0012672*	.1649384	.0215499	.0165729***	-.0001865***	.033443	-.0011833	7.893203
	(.0192763)	(.0847555)	(.0007452)	(.2112359)	(.034651)	(.0061266)	(.000063)	(.0330267)	(.0333019)	(.2580066)
observations	183	183	183	183	183	183	183	183	183	183
	Panel C: IV Regression									
LNpae98	.0542128	2.930644	-.0017576	-.0475819	.0696887	.0073809	-.0000588	.2413556	.141309	7.700467
	(.1051758)	(4.380422)	(.0021677)	(.6629515)	(.1202833)	(.0219358)	(.0002619)	(.3342344)	(.2385394)	(.7636825)
observations	183	183	183	183	183	183	183	183	183	183

Source: Author's calculations based upon the LCMS II.

Table A21.2: Quantile 2 (40%), 2004

	stratum124	cotincshare	distance11	assetownd09	infrastruc~e	s1q3b	age2	assetownd07	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Quantile Regression									
LNpae04	.0111791	-.0643832	.000694	-.0299201	.0283091	.0036939	-.0000392	-.098744	-.1200001	10.45
	(.0710609)	(.0854265)	(.0017299)	(.1451867)	(.0782728)	(.0118197)	(.0001212)	(.0659405)	(.0712755)	(.3413457)
observations	122	122	122	122	122	122	122	122	122	122
	Panel B: OLS Regression									
LNpae04	.0035191	-.0071594	.0003387	-.0581507	.0200911	-.0003744	.0000113	-.0635094*	-.1091965***	10.5874
	(.0380326)	(.0451215)	(.0009131)	(.1184886)	(.0418787)	(.0061095)	(.0000628)	(.0356337)	(.0369827)	(.1987075)
observations	121	121	121	121	121	121	121	121	121	121
	Panel C: IV Regression									
LNpae04	.0013715	.0157283	.000291	-.0664228	.0212123	-.0003076	.0000113	-.061623	-.1050834**	10.58732
	(.0421088)	(.1968866)	(.0009977)	(.1373647)	(.0429653)	(.0061421)	(.0000629)	(.0390149)	(.050566)	(.1989388)
observations	121	121	121	121	121	121	121	121	121	121

Source: Author's calculations based upon the LCMS IV.

Table A.21.3: Quantile 3 (60%), 1998

	stratum124	cotincshare	distiput	s10q8	infrastruc~e	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Quantile Regression									
LNpae98	.0131241	.0182534	-.0004084	(i)	.0773357**	.0111856**	-.000118**	.0118405	-.0285177	8.756667
	(.0170845)	(.0645665)	(.0007415)		(.0328427)	(.0051563)	(.0000518)	(.0303626)	(.0316825)	(.130224)
observations	179	179	179	179	179	179	179	179	179	179
	Panel B: OLS Regression									
LNpae98	.018657	-.0039638	.0002326	(dropped)	.0733414***	.0080515*	-.0000778*	-.0158918	-.0400004	8.760756
	(.0146426)	(.0544468)	(.0005954)		(.0282064)	(.0045343)	(.000046)	(.0256738)	(.0259596)	(.1119348)
observations	179	179	179	179	179	179	179	179	179	179
	Panel C: IV Regression									
LNpae98	.013529	-.1442804	.0002699	(dropped)	.0760096***	.0084445*	-.0000842*	-.0191792	-.0460521	8.786202
	.0172762	.2444354	.0006102		.0291063	.0046699	.0000481	.0267582	.0283837	.1219941
observations	179	179	179	179	179	179	179	179	179	179

Note: (i) s10q8 dropped due to collinearity.

Table A21.4: Quantile 3 (60%), 2004

	stratum124	cotincshare	distiput	s10q8	infrastruc~e	s1q3b	age2	s10q6	s4q5	_cons
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Quantile Regression									
LNpae04	-.1210904**	.0650898	-.0007462	-.0753203	-.1920056***	.0036829	-.0000166	.0100093	-.0253618	11.4021
	(.0578442)	(.0704968)	(.0009632)	(.1274591)	(.0671596)	(.0098971)	(.0001027)	(.0483314)	(.0496774)	(.2700357)
observations	120	120	120	120	120	120	120	120	120	120
	Panel B: OLS Regression									
LNpae04	-.0771451*	.0281332	.0001323	-.112156	-.1140004**	-.0061481	.0000622	.0055806	.0063169	11.50263
	(.0434194)	(.0533395)	(.0007735)	(.1044471)	(.0520285)	(.0078154)	(.0000825)	(.0369743)	(.0397268)	(.2137675)
observations	120	120	120	120	120	120	120	120	120	120
	Panel C: IV Regression									
LNpae04	-.0274513	.3761425	.0010553	-.1658994	-.1568088	-.0070213	.0000763	.0601966	-.0007482	11.42693
	(.1070676)	(.6617499)	(.0019704)	(.1596262)	(.1015923)	(.0093515)	(.0001008)	(.1121809)	(.0486602)	(.2896831)
observations	120	120	120	120	120	120	120	120	120	120

Source: Author's calculations based upon the LCMS IV.

The response of rural household pae to changes in covariates, using one year lagged rainfall as an instrument for household total income,

Table A23.1: 1998

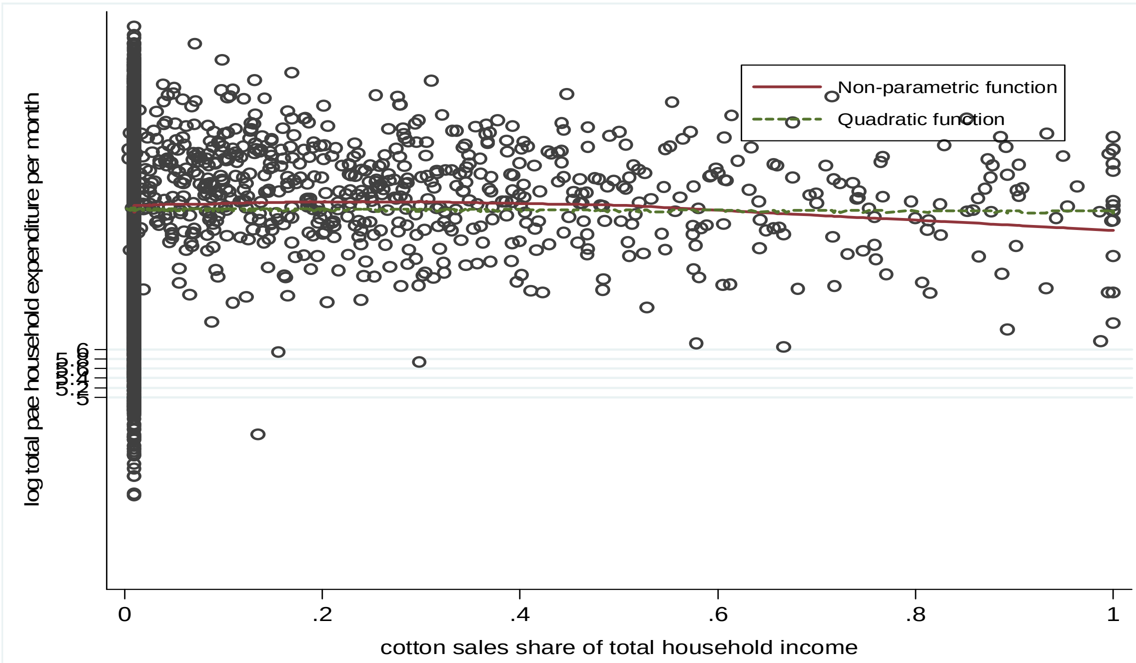
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
cotincshare	-1.155389	-.8419328	-.9221791	-.7696741	-.8154724	-.774537	-1.134514	-.9443361
	(.9183744)	(.871716)	(.8619486)	(.7984894)	(.7807687)	(.7721495)	(.7808524)	(.7749301)
stratum124	.2498329***	.2539297***	.2458618***	.2495341***	.2255256***	.2352431***	.2448666***	.2234603***
	(.0430891)	(.0422414)	(.0418052)	(.0408097)	(.0403459)	(.0399179)	(.0389216)	(.0442302)
distiput		-.0047652***	-.0047825***	-.0046938***	-.0053502***	-.005375***	-.0057705***	-.0036463**
		(.0016435)	(.0016205)	(.0016029)	(.0015733)	(.0015736)	(.0015543)	(.0017596)
s10q8			-1.665408***	-1.661231***	-1.664813***	-1.652807***	-1.50967***	-1.645161***
			(.2647215)	(.2635249)	(.2572347)	(.2566613)	(.2512013)	(.2934461)
infrastruc~e				-.0559229	-.0633752	-.0584501	-.0239623	-.0208506
				(.0778764)	(.0760542)	(.0759079)	(.0743726)	(.0872116)
s1q3b					-.0172192***	-.0013572	-.006715	-.0133198
					(.0022048)	(.01268)	(.0124171)	(.0138254)
age2						-.0001667	-.0001123	-.0000347
						(.0001327)	(.0001298)	(.0001426)
s10q6							-.5411348***	-.403476***
							(.075096)	(.0835142)
s4q5								-.4325963***
								(.0798648)
Observations	1246	1245	1245	1245	1245	1245	1245	919
R^2	0.0178	0.0400	0.0677	0.0743	0.1185	0.1209	0.1531	0.1744

Table A23.2: 2004

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
cotincshare	48.082	19.07374	18.74465	18.90642	20.14417	20.06111	56.85585	25.23182
	(602.5159)	(48.60581)	(46.7002)	(46.14526)	(52.30744)	(51.80792)	(410.6017)	(78.44945)
stratum124	-2.526446	-1.223312	-1.477908	-1.490572	-1.99464	-1.882525	-4.445558	-1.894739
	(31.0412)	(2.828579)	(3.496587)	(3.463053)	(5.08833)	(4.79789)	(31.83854)	(5.802258)
distiput		.0089651	.0091082	.0091843	.0120271	.012552	.0349126	.0117013
		(.0286511)	(.0284554)	(.0281838)	(.0371967)	(.0381299)	(.2637287)	(.0438474)
s10q8			-3.685252	-3.721869	-3.98182	-4.082824	-12.71405	-5.3806
			(10.7666)	(10.64287)	(11.98193)	(12.1471)	(96.06082)	(18.64155)
infrastruc~e				.0078374	.1514407	.0398085	-.1452633	-.0308442
				(.6860589)	(.8639049)	(.7511902)	(2.171083)	(.9074778)
s1q3b					.055252	-.1423722	-.3531945	-.1686533
					(.1591745)	(.3262393)	(2.40122)	(.4730236)
age2						.0020507	.0052984	.0024084
						(.0049495)	(.0370629)	(.0070666)
s10q6							3.241787	1.38617
							(23.53907)	(4.40603)
s4q5								1.154893
								(3.731657)
Observations	999	649	649	649	649	649	649	647
R^2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

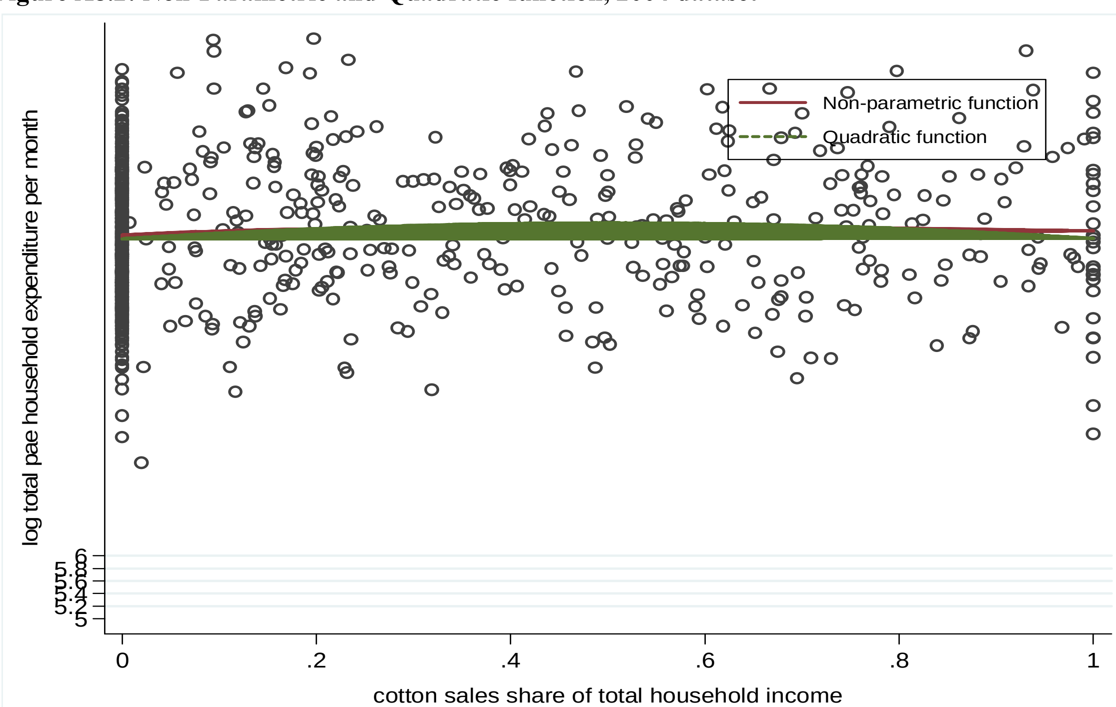
Source: Author's calculations based upon the LCMS II and LCMS IV.

Figure A8.1: Non-Parametric and Quadratic function, 1998 baseline dataset



Source: Author's calculations based upon the LCMS II.

Figure A8.2: Non-Parametric and Quadratic function, 2004 dataset



Source: Author's calculations based upon the LCMS IV.

Table A24: Fixed-effects (within) Panel IV regression with Cotton Income instrumented by external instrument (distance to input market)

	Catchment		Control	
	Coef.	**Std.Err.**	**Coef.**	**Std.Err.**
cotincshare	10.06956	20.95206	-0.9114254	25.86078
highgrade	0.0721656	0.07569	-0.0020926	0.0889257
rain	-0.0250642	0.057645	0.0078505	0.0504679
rainlagged	0.0029041	0.065073	0.0429133	0.0995224
Landowners~c	0.1493332	0.105086	-0.0609962	0.332604
cotincshar~q	-9.979608	22.84006	1.427766	27.42889
distbank	0.0068756	0.010582	-0.0077815	0.0147865
_cons	9.706605	2.433925	5.738843	5.690627
R2 within	0.7095		0.4951	
R2 between	0.0145		0.2798	
R2 overall	0.4991		0.3574	
corr(u_i, Xb)	-0.2025		0.0215	
N of obs	90		78	
N of groups	45		44	
Obs per group	2		1.8	
Wald chi2(7)	43245.28		12338.95	
Prob>chi2	0		0	
Sigma_u	0.33814856		0.53700608	
Sigma_e	0.43122517		0.74367172	
rho	0.38076791		0.34272354	
F tests that all u_i=0: F(44,38)	0.9		0.65	
Prob>F	0.6405		0.8993	

Notes: Instrumented: cotincshare. Instruments: highgrade rain rainlagged Landownershippc cotincsharesq distbank distiput.
Source: Authors' estimations.