Households Surveys

Pauline Grosjean

Field Research Methodology

March 31st, 2008

Outline

- Survey Design and Sampling
  - Overview of issues in sampling
  - Sampling Practice
  - Correction and weights
- Survey Content
  - Reporting period and recall data
  - Measuring welfare
  - Stated preferences
- Some Econometric Issues with Survey Data
  - Group data and clustering
  - Heteroskedasticity
  - Simultaneity and endogeneity
  - Covariates

1. Survey Design and Sampling
1.1. Overview of Issues in Sample Design

- Ideally: list all households in the country, each household has an equal probability of being selected, and each household selected agrees to participate
- Survey "frame": list of households, from which households should be randomly selected
  - Often, a census, or other administrative lists: election registry for example
  - Problems: outdated, incorrect or incomplete coverage

1. Survey Design and Sampling
1.1. Overview of Issues in Sample design

- **Sampling Error**: error inherent in making inference for a whole population from observing only some of its members
- 2 conclusions from sampling theory:
  - Diminishing returns: Sampling error inversely proportional to the square root of the sample size. To reduce error of a sample by half, need to quadruple sample.
  - Sample size for a given level of precision is almost independent of the total population. For ex, a 500 households sample would give ~ same sampling error if extracted from a population of 10000, 100000 or infinite

1. Survey Design and Sampling
1.1. Overview of Issues in Sample design

- **Non Sampling Error**: refusals, respondent fatigue, interviewer errors.
  - Generally likely to increase with sample size
  - Not random

1. Survey Design and Sampling
1.1. Overview of Issues in Sample design

- **Multi-stage sampling**:
  - Usually: 2 or 3 stages
  - Primary Sampling Units: smallest geographic units in census: Census Enumeration Areas: 50 to 200 HH generally. If too many HH, divide into segments
  - Cluster effects: larger error than simple random sample because neighboring HH tend to have similar characteristics.
  - Cluster effect grows with number of HH selected from each PSU

1. Survey Design and Sampling
1.1. Overview of Issues in Sample design

- **Stratification / Analytical domains**:
  - Can lead to more precise estimates: stratification has largest effect on reducing variance when stratum means are different and there is little variation within strata
  - Necessary to weigh sample when calculate estimates for the whole population: “raising” or “inflation” factors = 1/sampling probability

    $$w_k = N_k / n_k$$

  - N_k total number of households of domain k in the population
  - n_k number of households of domain k sampled
  - Multi stage sampling procedure applied independently within each of the domains

1. Survey Design and Sampling
1.1. Overview of Issues in Sample design

- **Sample Size**:
  - Sample size for a given level of precision is almost independent of the total population
  - Living Standard Measurement Study: 2000 to 5000 HH
  - Life in Transition Survey: 1000 HH
- **Dynamics:**
  - Panel data. Pb with panel data: attrition
  - Rotating panel
  - Supplement cross section with administrative data
  - Repeated cross sections
  - Recall data

1. Survey Design and Sampling
1.2. Random Sampling Practice

- Definition of the sample
- Example: LITS:
  - Two stage sampling. 50 PSUs, 20 HH per PSU. 1,000 respondents per country
  - Analytical unit is the Household
  - A household is defined as a person or group of people who normally live at the property who share a living or a sitting room or at least one meal a day
  - 1 or sometimes 2 respondents per household, between 18 and 65 years old. Interviews face to face.

1. Survey Design and Sampling
1.2. Random Sampling Practice

- Non response and household replacement
  - Non responding households are not a random sample of all households (migrants, rich, urban)
  - Leads to a bias
  - Objective: keep non response to a low. Repeat visits (3 times)
  - Choice: non responding households are replaced by other randomly selected households (see later)

1. Survey Design and Sampling
1.2. Random Sampling Practice

- Implementation of the First Sampling Stage
  1) List all PSUs of the country, possibly with implicit stratification (for ex. sorted out by metropolitan, urban and rural areas); and population in each.
  2) Sampling of PSUs with **Probability Proportional to Size**:
  - Add cumulative number of households
  - Scale cumulative distribution: (cumulative population*Number of PSUs to be selected)/whole population
  - Choose a random number: “random start” RS
  - Add random number to scale distribution, and select all PSUs above integer

M/U/R	PSU Name	Population 18 and over	Cumulative population	Scaled cumulative population	Random scaling of cumulative population	Selected PSU
				15	0.29	
m	Strzelce	4,272	4,272	0.14	0.43	
m	Buczek	4,869	9,141	0.29	0.58	
m	Sędziejowice	6,533	15,674	0.50	0.79	
m	Widawa	8,078	23,752	0.76	1.05	1
m	Dłutów	4,090	27,842	0.89	1.18	
m	Wierzchlas	6,640	34,482	1.10	1.39	
m	Bolesławiec	4,147	38,629	1.23	1.52	
m	Czastary	4,054	42,683	1.36	1.65	
u	Bedlno	6,285	48,968	1.56	1.85	
u	Łanięta	2,684	51,652	1.64	1.93	
u	Nowe Ostrowy	3,944	55,596	1.77	2.06	2
u	Oporów	2,804	58,400	1.86	2.15	
u	Wodzierady	3,131	61,531	1.96	2.25	
u	Daszyna	4,202	65,733	2.09	2.38	
u	Góra Świętej	4,677	70,410	2.24	2.53	
u	Grabów	6,555	76,965	2.45	2.74	
u	Brójce	5,301	82,266	2.62	2.91	
u	Nowosolna	3,579	85,845	2.73	3.02	3

1. Survey Design and Sampling
1.2. Random Sampling Practice

- Second sampling stage:
  1) List all dwellings in selected PSU
  2) Select (20+a reserve for replacement) households randomly:
     - select a random starting point and then picking the Kth element in the succession from the sampling frame
     - Sampling interval: K. If N=100 people, n=20, K=100/20=5
     - Start with the random number in the list and take every Kth interval

1. Survey Design and Sampling
1.2. Random Sampling Practice

- Selection of respondent within household
  - Last Birthday method: list birthdates of all members in household roster, select last one
  - Kish grid (L. Kish , A Procedure for Objective Respondent Selection Within the Household, Journal of the American Statistical Association, 1949)

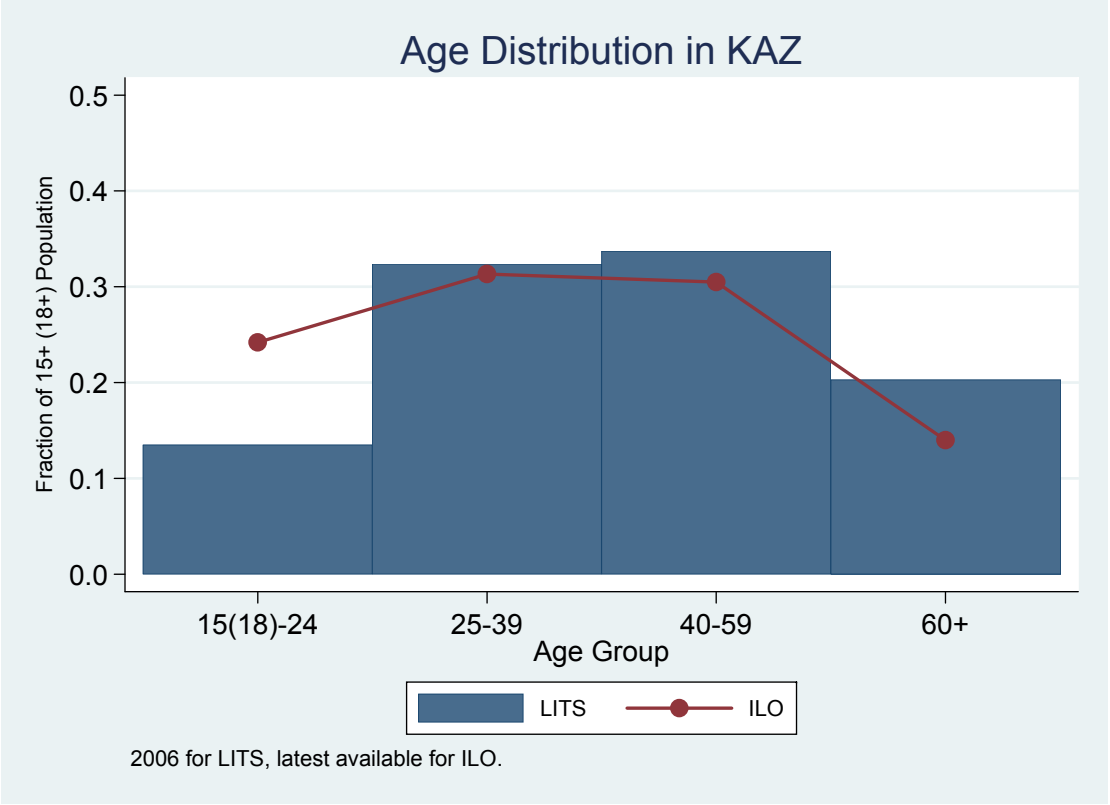
- Objective: minimize free choice of the interviewers
  - Quality control:
    - Sampling: 20% call back visits should be made to verify that households were selected according to the sampling procedure and that the "*principal*" respondent was selected according to the "*Last Birthday*" rule
    - Data quality: 30 % call backs
    - Questionnaire: if more than 80% of answers are DN or NA: discarded and HH replaced

1. Survey Design and Sampling
1.3. Correction and Weights

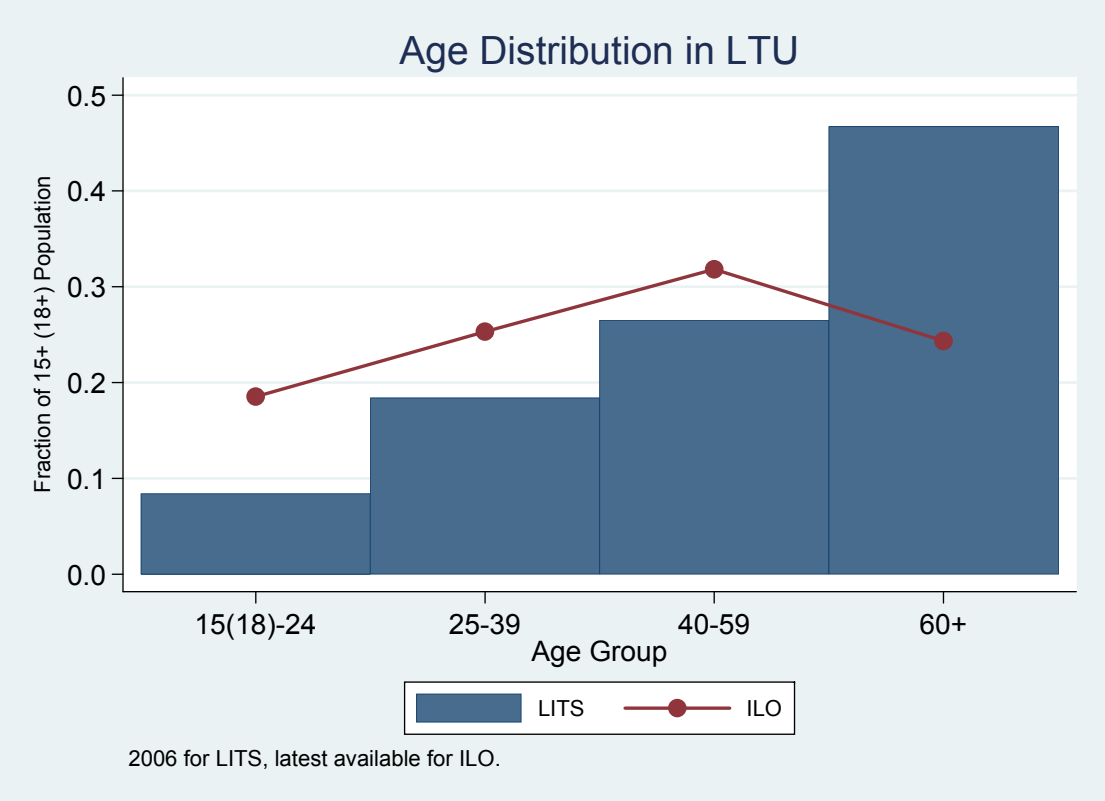
- In a self-weighted sample, the proportions and averages are unbiased estimates for the proportions and averages in the population

- However, if stratified sample, or because of “errors”, this might not be the case
  - Correct “errors” by reweighting to get unbiased estimates and compute sampling errors, CI…

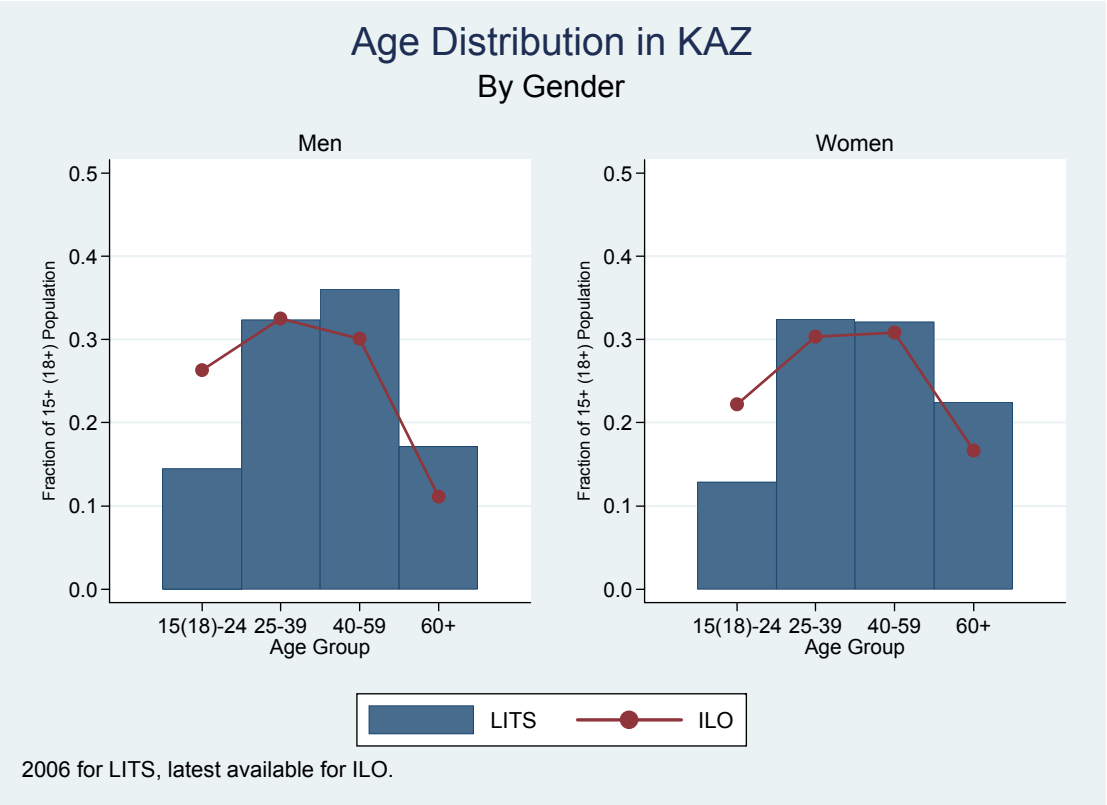
1. Survey Design and Sampling
1.3. Correction and Weights



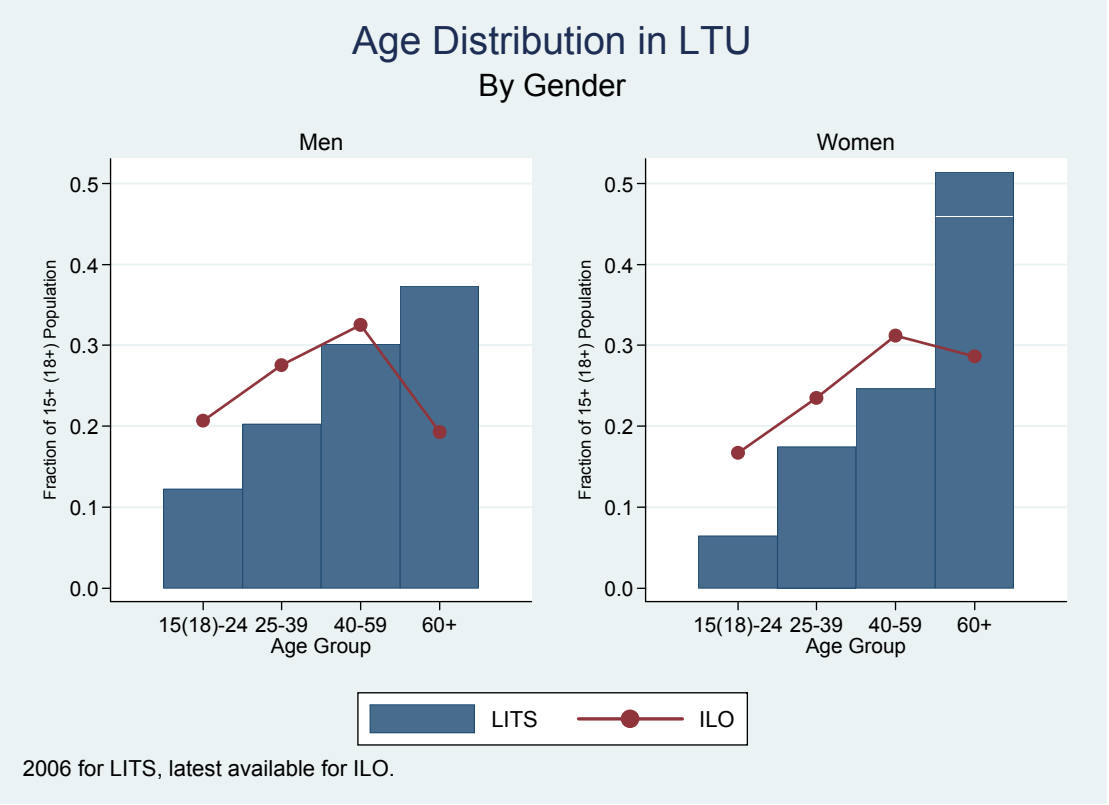
1. Survey Design and Sampling
1.3. Correction and Weights



1. Survey Design and Sampling
1.3. Correction and Weights



1. Survey Design and Sampling
1.3. Correction and Weights



1. Survey Design and Sampling
1.3. Correction and Weights

- Weights = 1 / Selection probability
- Need true probability of selection:

  P = P1 x P2 x P3

  P1 = {probability of selecting the PSU in the country}
  = 50 * {size of the PSU} / {total size of the country}

  P2 = {probability of selecting the household in the segment}
  = 5 / {number of households in the segment}

  P3 = {probability of selecting the respondent in the household}
  1 / {number of persons 18 years and older in the household}

1. Survey Design and Sampling
1.3. Correction and Weights

- If this is impossible (and only if): “refined” weights to make sample match known characteristics of the population

  w=(Population 1865/Sample 1865)*(1/eligible HH member 1865)

- Reestablish representative the sample

2. Survey Content
2.1. Reporting Period and Recall Data

- Trade off potential recall bias from long period reporting against potential variance from short periods
- For consumption:
  - Frequently purchased items (e.g. food), recall period between one week and one month
  - When practical: diaries for a week or two
  - Durable goods: annual recall basis
  - Gives good estimates of means, but overestimate dispersion

- Graphical representation and recallable events might help

- Example 1: Multiple job holding

JOB SERIAL NUMBER	**Q4.02** Please describe all the different jobs you have done for income over the past 12 months	**Q4.03** During which of the past 12 months did you work in this job? **CIRCLE IN APPROPRIATE MONTHS**											
		October, 2005	November, 2005	December, 2005	January, 2006	February, 2006	March, 2006	April, 2006	May, 2006	June, 2006	July, 2006	August, 2006	September, 2006
Job A		01	02	03	04	05	06	07	08	09	10	11	12
Job B		01	02	03	04	05	06	07	08	09	10	11	12
Job C		01	02	03	04	05	06	07	08	09	10	11	12

- Example 2: Work trajectories

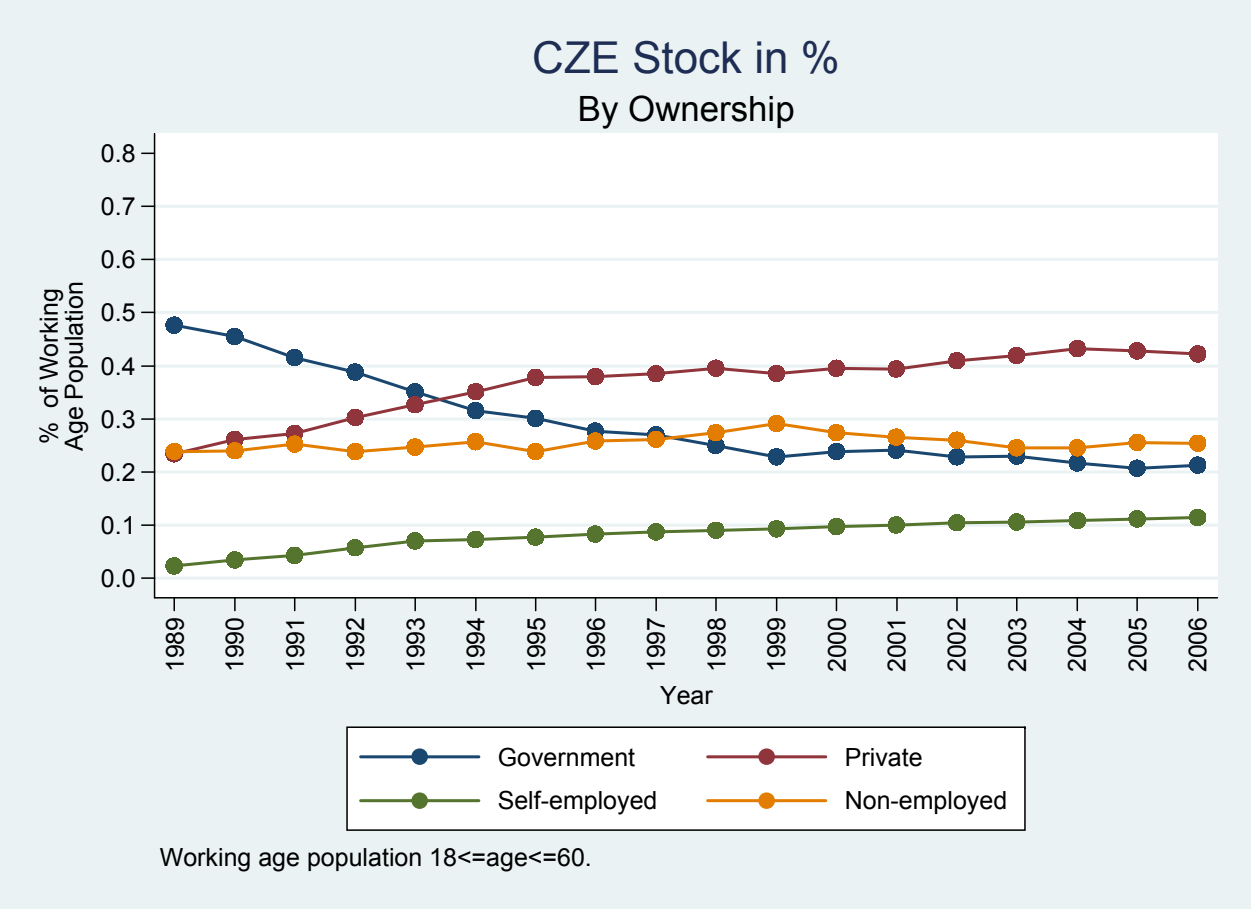
Q6.01a First tell me about the important events in your life. – Starting from 1989 in which your (s) did you….

Year	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6
Do your military service																		
Get married																		
Have a child																		
Divorce																		
Officially retire																		

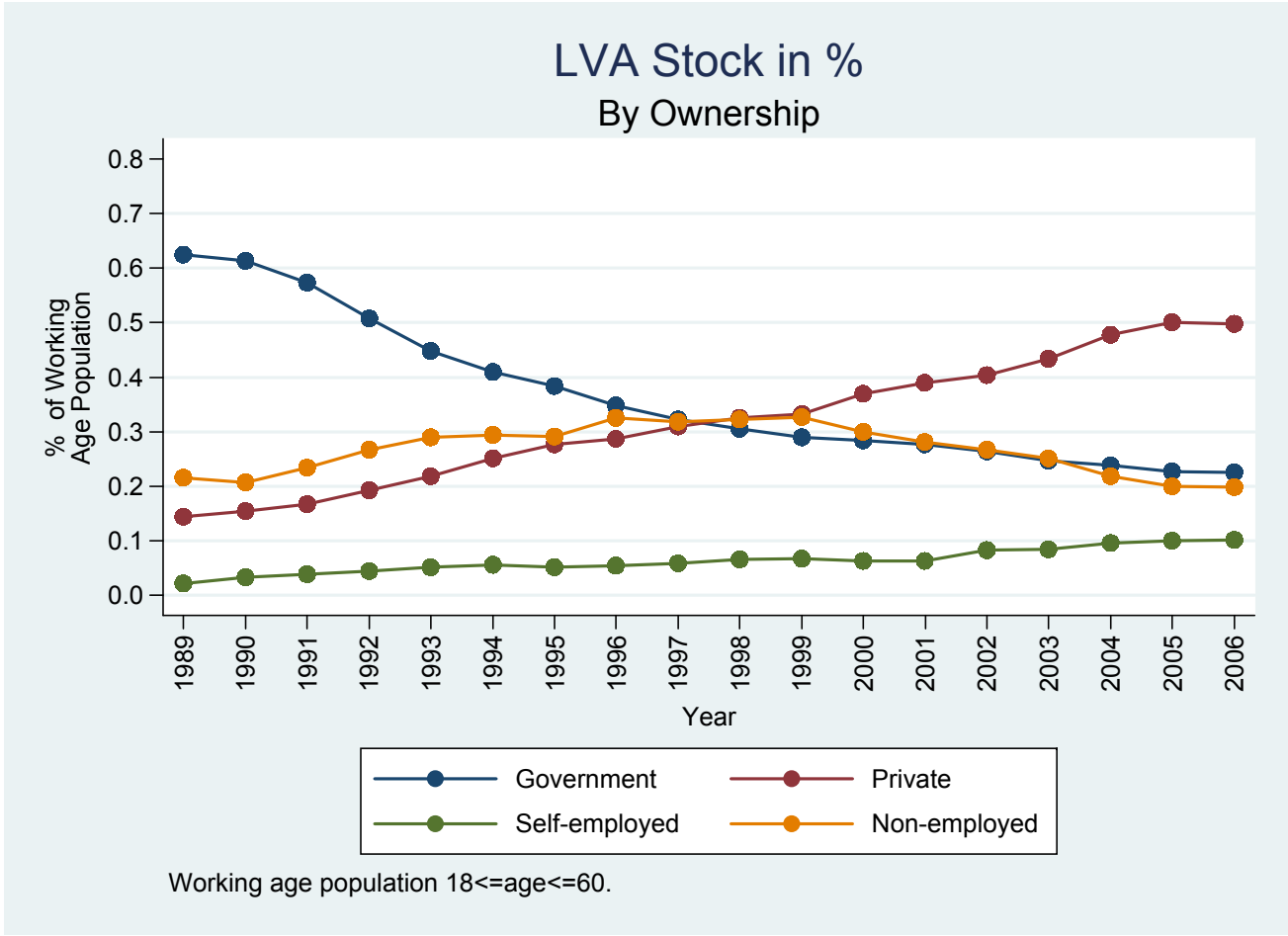
Please tell me in which years you have worked for wages (for an employer) since 1989?

Year	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	Occupation	Industry
Job A: ----------																				
Job B: ----------																				
Job C: ----------																				
Job D: ----------																				

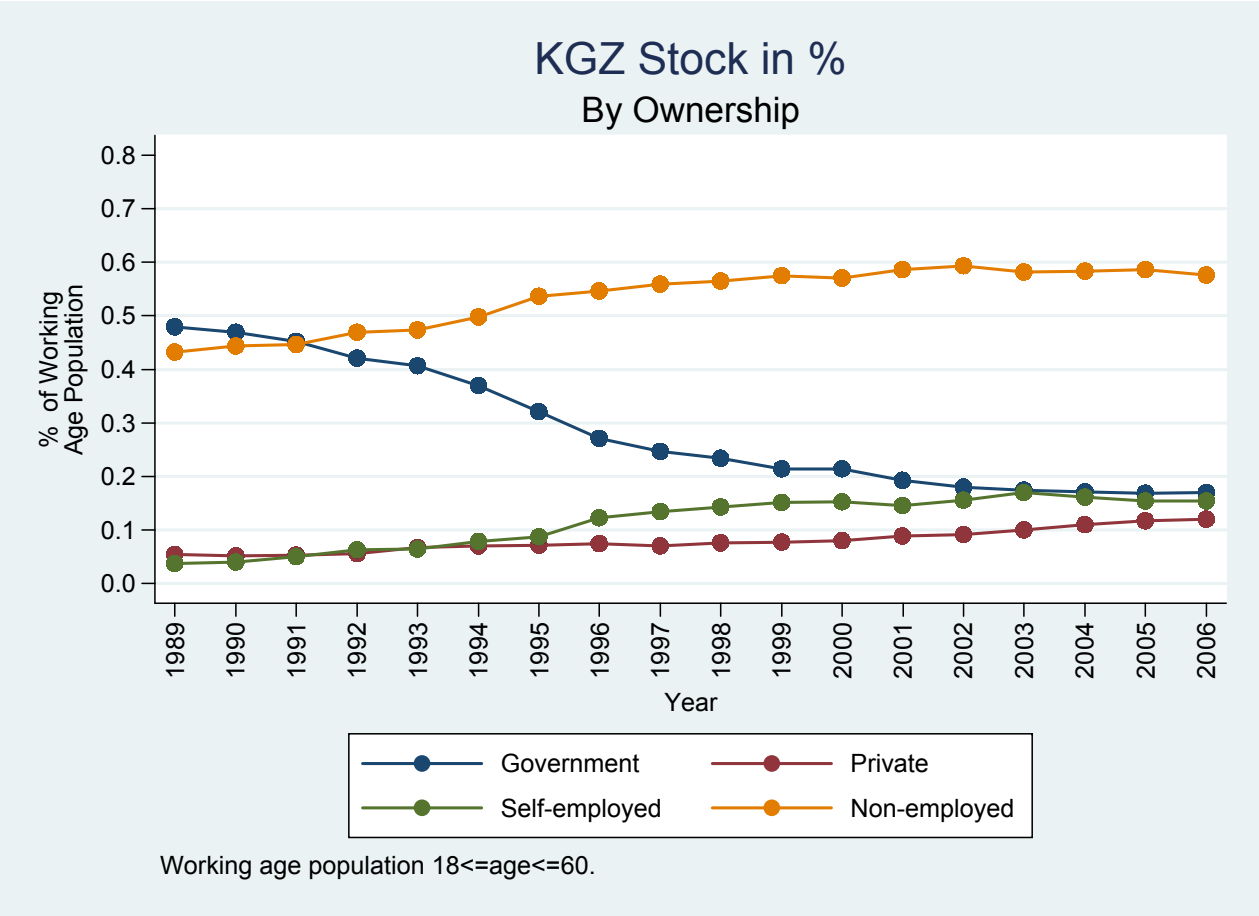
From Work Trajectories: Transition



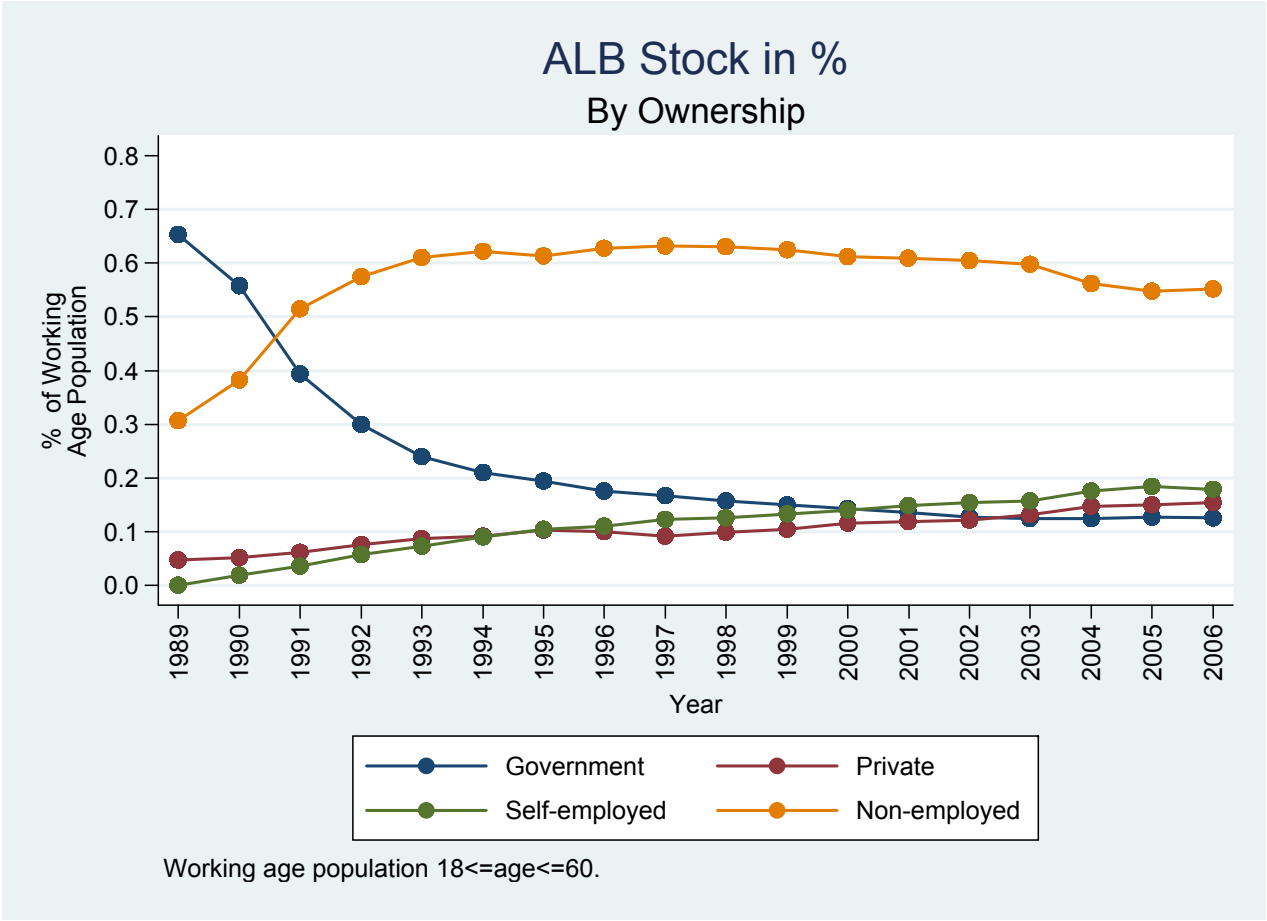
From Work Trajectories: Transition



From Work Trajectories: Transition



From Work Trajectories: Transition



2. Survey Content
2.2. Measuring Welfare

- *How to obtain accurate estimates of total consumption from small number of questions?*
- NSS (India), SURENAS (Indonesia): list of 200 food items…
- Lose accuracy with shorter list, but may be worthwhile
- Principal problems: production and autoconsommation
- **Expenditure vs. income**:
  - Measuring income is even more difficult. Same problems: imputations, seasonality, recall bias; and in addition: misreporting and issue of returns on assets

2. Survey Content
2.2. Measuring Welfare

- In practice: alternative welfare measures:
  - Broad consumption aggregate
  - Short aggregate
  - Subjective income rank
  - Assets

2. Survey Content
2.2. Measuring Welfare

- **Broad consumption aggregate:**

During the <u>past 30 days</u>, approximately how much did your household spend on:

- Food, beverages and tobacco
- Clothing and footwear
- Transport and communication (fixed line phone, mobile phone, internet) expenses
- Recreation, entertainment, meals outside the home, etc

During the <u>past 12 months</u>, approximately how much did your household spend on:

- Education (including tuition, books, kindergarten expenses)
- Health (including health insurance)
- Furnishings (sheets, towels, blankets, linen, etc.)
- Household durable goods (e.g. furniture, household appliances, TV, car, etc.)
- Other expenses (DO NOT PROMPT – RECORD HERE ANY ADDITIONAL EXPENSES THAT THE RESPONDENT WOULD LIKE TO REPORT)

2. Survey Content
2.2. Measuring Welfare

Annualized household expenditures turned into 3 **welfare metrics:**

(1) Equivalized consumption using OECD equivalence scale (weight=1 for first adult, 0.5 for other adults, 0.3 for children)

(2) Consumption aggregate per capita (weight of 1 for all)

(3) Welfare metrics based on share of non-food expenditures

2. Survey Content
2.2. Measuring Welfare

- **Short consumption aggregate:**

  "Living in this dwelling and doing what you do, what would be the minimum amount of money that this household would need to make ends meet at the end of each month?"

- **Subjective income measure:**

  "Please imagine a ten-step ladder where on the bottom, the first step, stand the poorest people and on the highest step, the tenth, stand the richest. On which step of the ten is you household today?"

1	2	3	4	5	6	7	8	9	10

2. Survey Content
2.2. Measuring Welfare

- Comparing welfare measures based on asset ownership:

  "Does anyone in your household have
  - A car
  - A secondary residence
  - A bank account
  - A credit/debit card
  - A mobile phone
  - A computer
  - Access to internet at home"
  - Divide respondents by welfare levels (deciles, or poor-middle-rich) and compare asset ownership rates for each group
  - Equivalized broad consumption aggregate does best job, but other measures are OK too (short aggregate least discriminant)

2. Survey Content
2.3. Stated Preferences

- Stated versus revealed preferences
- Useful to check internal validity of responses
- Discrete choice models: describe decision makers' choices among alternatives
- The decision makers can be people, households, firms, or any other decision-making unit, and the alternatives might represent competing products, courses of action…
- Choice set (set of alternative) must have 3 characteristics:
  - *mutually exclusive* alternatives
  - choice set must be *exhaustive*
  - number of alternatives must be *finite*

2. Survey Content
2.3. Stated Preferences

- Some interesting techniques used by economists:
  - Implicit Assumption Tests
    - The IAT measures implicit attitudes and beliefs that people are either unwilling or unable to report.
    - https://implicit.harvard.edu/implicit/
    - Beaman et al 2008: Powerful Women: does Exposure Reduce Bias http://www.povertyactionlab.com/papers/powerful%20women.pdf
  - Choice experiment

2. Survey Content
2.3. Stated Preferences

- Choice Experiment:
  - Present different options that differ in the levels of different choice attributes
    - Good valuation, policy evaluation
  - Illustrate which attributes are the most important determinant of a policy, or a good
  - Implicit price of each attribute
  - Value trade offs between different attributes of policy or good
  - WTP or WTA a policy

Hensher, Rose & Greene (2005). *Applied choice analysis,* Cambridge University Press
Louviere, Hensher & Swait (2000). *Stated choice methods: analysis and applications*. Cambridge University Press

3. Some Econometric Issues

- With survey data, unlikely to fit textbook assumption that, conditional on independent variables, the dependent variables are *independently, identically and normally distributed*
  - **Group data and clustering**: violation of "independence" assumption
  - **Heteroskedasticity:** violation of "identically distributed" assumption

3. Some Econometric Issues
3.1. Group Data and Clustering

- Causes: neighborhood effects, similar local conditions, similar time of survey
- With group data, each observation contains less information than when observations are independent

$$y_{ic} = x_c'b + a_c + e_{ic} = x_c'b + u_{ic}$$

- If cluster nature ignored, estimated variance is too small and t-values too large
- If errors are correlated, OLS not efficient

3. Some Econometric Issues
3.1. Group Data and Clustering

- If place structure on the covariance: ex: assume that errors are uncorrelated between clusters and errors equicorrelated within clusters with correlation coefficient ρ
  - OLS and FGLS are equivalent and OLS is efficient
  - Scale up (multiply) variance with the **Design Effect**:

  $DE = 1+(m-1)\rho$

  m: number of HH in each cluster (identical)

  ρ: correlation coefficient within clusters (estimated)

3. Some Econometric Issues
3.1. Group Data and Clustering

- If you don't want to assume an error structure: Use Huber-White estimator , but only when number of clusters is large.
  - Also called White's General Estimating Equation
  - It's what STATA *cluster* command does
  - Generates correct error terms, but does not improve efficiency
  - Requires a large number of clusters for validity: Performs poorly when *Number Clusters<50*

3. Some Econometric Issues
3.2. Heteroskedasticity

- As for clustering, consequences of heteroskedasticity in regression analysis is to invalidate usual formulas for the calculation of std. errors, and OLS is inefficient
- Examples:
  - Expenditures: variance higher for the rich
  - Manufacturing output: variance higher in rich countries

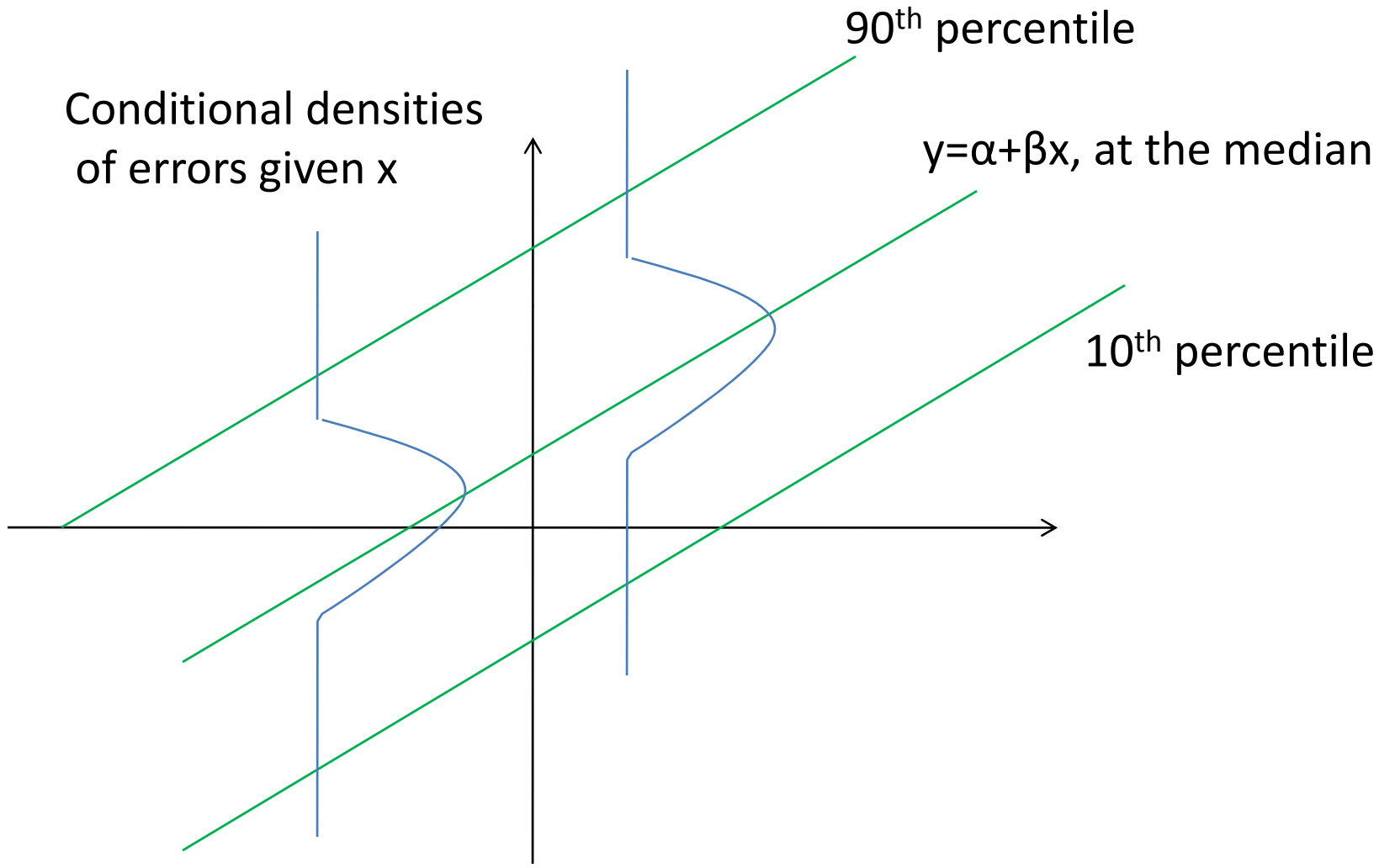
$$V(y_i \mid x_i) = \sigma\text{^}2 + x_i'\Omega x_i$$

- To adjust st. errors: can act as if each households was its own cluster (if there is no cluster effect)
- A useful tool: Quantile regressions

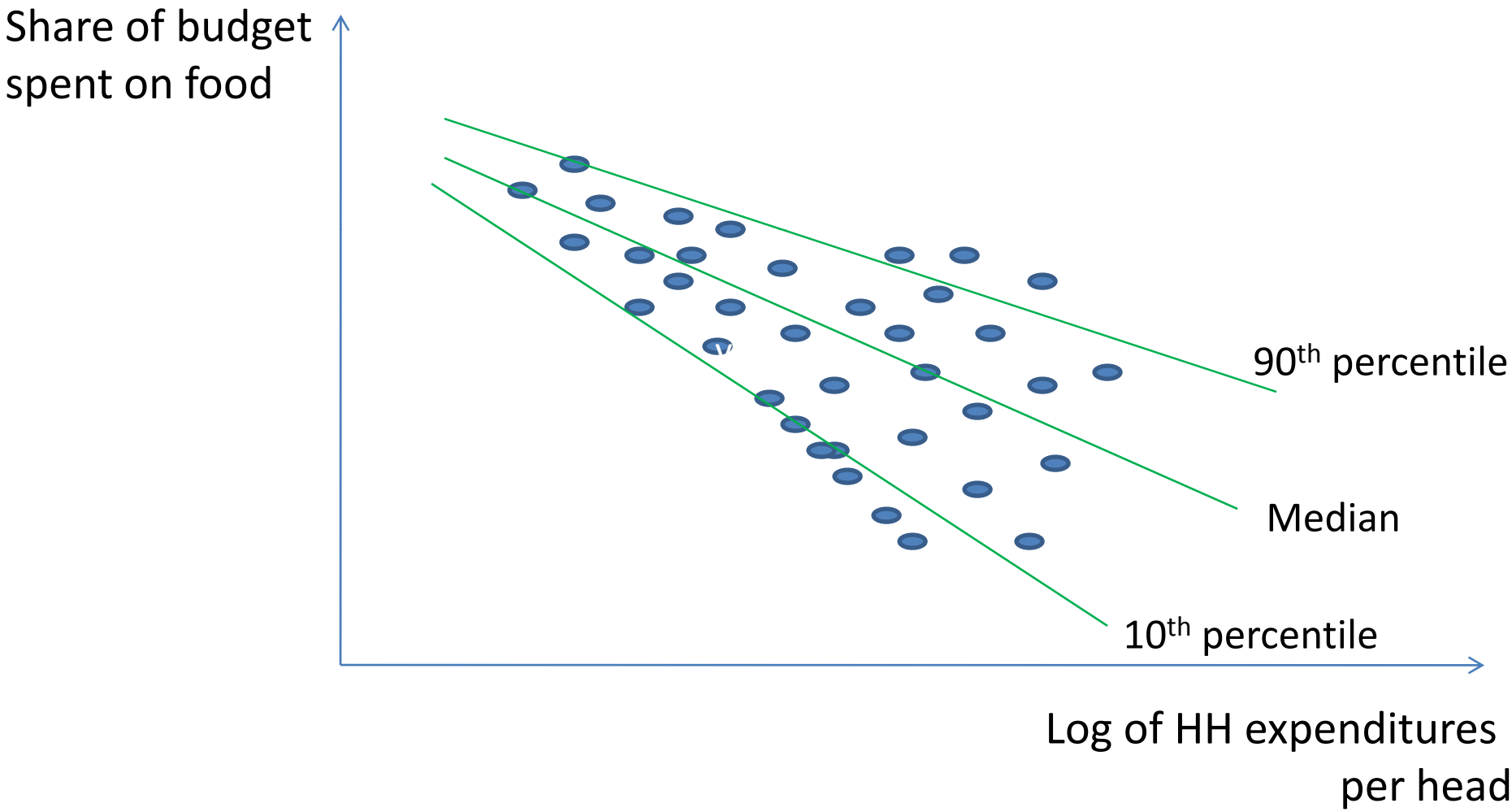
3. Some Econometric Issues
3.2. Heteroskedasticity

- **Quantile regression**:
  - If residuals are heteroskedastic, the distance of each percentile from the regression line will be different at different values of *x*
  - Joining up the percentiles for different values of *x* will not lead to straight lines or simple curves. However, can still fit straight lines to the percentiles: quantile regressions
  - Command “qreg” in stata
  - Tests for heteroskedasticity: Breusch-Pagan (1979) or White’s (1980) tests
  - Can detect heteroskedastcity graphically

A homoskedastic linear regression function



Heteroskedasticity



3. Some Econometric Issues
3.3. Simultaneity and Endogeneity

- Correlation between error terms and explanatory variables
- OLS estimates are biased and inconsistent
- Classic example: interdependence of supply and demand and price in local markets
- One solution: IV approach: ex: rainfall as instrument for price

3. Some Econometric Issues
3.4. Covariates

- Controlling for covariates that are affected by your treatment will bias results
- Controlling for covariates that do not affect outcomes weakens precision
- Results are more believable if you specify intended covariates ex ante

  $\Rightarrow$Always better to base choice of covariates on theory

  $\Rightarrow$Do not add covariates blindly!

References

- Deaton, Angus. The Analysis of Household Surveys, 1997, the John Hopkins University Press
- LSMS (World Bank) http://www.worldbank.org/LSMS/
- UN handbook on household surveys: http://unstats.un.org/unsd/hhsurveys/sectiona_new.htm
- International Social Survey Program http://www.issp.org/
- European Social Survey http://www.europeansocialsurvey.org/