

# LINDA – Longitudinal INdividual DAta for Sweden\*

by

Per-Anders Edin and Peter Fredriksson†

November 20, 2000

## Abstract

This paper presents LINDA – a register-based longitudinal data set for Sweden. LINDA consists of a large panel of individuals, and their household members, which is representative for the population during the period 1960 to 1998. As future years become available, this information will be added to the data set. LINDA also includes a specific sample of immigrants. This sample has the same design and covers the same time period as the population sample. We provide a description of the sources of data, the sampling frame as well as the sampling procedure. Moreover, to illustrate the usefulness and particular features of LINDA, we give the development of some of the key variables in the data set.

Keywords: Longitudinal data, Population sample, Immigrant sample, Sweden.

JEL classification: D10, D30, H31, J15, J60.

---

\* We thank Tommy Blomquist, Pebbe Selander, and Ed Palmer for helpful comments on earlier drafts. Fredriksson acknowledges the financial support of the Swedish Council for Work Life Research (RALF).

† Department of Economics, Uppsala University, P.O. Box 513, SE-751 20 Uppsala, Sweden.  
Email: per-anders.edin@nek.uu.se; peter.fredriksson@nek.uu.se.

## 1. Introduction

Suitable data sets for conducting longitudinal analyses for research and policy purposes have long been lacking in most countries. LINDA – a register-based longitudinal data set – is an attempt to bridge this gap for Sweden. The purpose of this paper is to present LINDA. We describe the sources of data, the sampling frame as well as the sampling procedure. Moreover, we give the development of some of the key variables in the data set.

LINDA consists of a large panel of individuals, and their household members, which is representative for the population from 1960 through 1998. As future years become available, new information will be added to the data set. The sampling procedure ensures that each new year is representative for that year.

Within LINDA's realm there is also a specific sample of immigrants. This particular sample has the same design and covers the same time period as the sample which is representative for the population.

The data base is intended to be a general research base – a complement to other Swedish surveys such as The Swedish Level of Livings Survey (*LNU*) and Household Market and Non-market Activities (*HUS*).<sup>1</sup> The core registers are the Income Registers and Population Censuses.

Access to the data will be granted for all researchers through Statistics Sweden subject to the normal confidentiality provisions. Statistics Sweden will charge researchers on a marginal cost basis.

LINDA is a joint endeavor between the Department of Economics at Uppsala University, The National Social Insurance Board (RFV), Statistics Sweden, and the Ministry of Finance. The collection of historical data (1960-1995) was funded by the Swedish Council for Planning and Coordination of Research (FRN), Swedish Council for Social Research (SFR), and RFV. Future updating of the data will be funded mainly by the Ministry of Finance.

---

<sup>1</sup> LNU is described in Eriksson and Åberg (1987) and HUS in Klevmarken and Olovsson (1993).

## 1.1 General characteristics of LINDA and an outline of this presentation

The principal data sources are the Income Registers – available annually from 1968 to 1998 – and the Population Censuses – available every fifth year from 1960 to 1990; no census has been taken after 1990. All variables in these registers are included in the data base. In Section 2, we give a brief account of these two registers and other registers included in the data set.

LINDA contains two distinct samples: a *population sample*, representative for the entire population, and an *immigrant sample*, covering almost 20 percent of immigrants to Sweden.<sup>2</sup> Both samples are random. The population sample covers 3.35 percent of the population annually.<sup>3</sup> In 1994, this amounts to 300,000 individuals. There is no overlap between the population and immigrant samples so, in practice, the coverage of immigrants exceeds 20 percent.<sup>4</sup>

Apart from being a panel which is representative for the population, the sampling procedure ensures that the data are representative for each year. Starting with a representative sample a particular year, we sample from the inflow to replace the outflow to obtain the next year's sample; thus, the data are also cross-sectionally representative. Section 3 provides more details about the sampling procedure and the sampling frames for the two samples.

For each year, information on all family members of the sampled individuals are added to the data set. Family members are only included in the sample as long as they stay in the family. The definition of a “family” differs between the Population and Housing Census and the Income Register; the family concept of the Census is based on whether individuals actually reside together, while the Income Registers define a family as they are regarded for tax purposes. For Census years both family definitions are available. Section 4, among other things, provides more substance to the different definitions of a family. This section also presents the development of some key variables over time.

For a panel covering such a long time span, documentation is obviously crucial. At the time of writing, however, the variables in the data are only rudimentarily described in

---

<sup>2</sup> The coverage is generally around 19.5 %.

<sup>3</sup> There is some variation in the coverage. In the early 1960's 3.31 % of the population is covered. From then on fraction of the population included in the sample increases steadily to reach 3.35 % around 1990.

<sup>4</sup> At present, the immigrant sample is funded until 1998.

Swedish at each point in time. The process of documenting changes in variable definitions has just begun. The documentation will be continuously updated at the website: <http://www.nek.uu.se/linda>. In Section 5, we give some additional information on the documentation of the data set, how the data are accessed, and the organization of the data.

A general feature of the data is that the information becomes richer over time. For the period 1960-1967, there is only Census data for 1960 and 1965 along with yearly (rudimentary) information on individual income giving pension rights from RFV's Pension Register. From 1968 and onwards, there is yearly information on income and some background characteristics from Statistics Sweden's Income Registers. As time passes, the Income Registers become more detailed, including more components of income beginning with transfers in the mid 1970's. In the 1990's, the data base has been expanded in a significant way since other register information – such as information on unemployment duration – is included. In the Appendix, we give a more detailed account of how the contents of LINDA have changed over time.

## **2. The sources of data**

In this section we give a summary account of the sources of the data in LINDA. Since the bulk of the data come from the Income Register and the Population and Housing Censuses, we focus on these two registers; see Sections 2.1 and 2.2. Section 2.3 turns to a brief description of the other registers included in the data set.

### 2.1 The Income Registers

The Income Registers (*Inkomst- och Förmögenhetsstatistiken*) are based on filed tax reports. Therefore, the information contained in these registers generally refers to the entire (calendar) year. For example, the different income concepts are annual and the individual industry of employment refers to the main industry of employment during the particular year (to the extent that it is reported).

The fact that filed tax reports constitute the basis for the Income Registers has important implications. For instance, the content of labor income may vary over time

because of changes in the tax base; in general, the data are contingent on the tax legislation in a particular year.

An appealing feature of the Income Registers is that supplementary information in the form of statements of income (*kontrolluppgifter*) is increasingly available (employers file statements of labor income, commercial banks file statements of different sources of capital income, the Social Insurance Board files statements on social security, etc). This fact should increase the quality of the data and provides information on sources of income that had not been available previously.

The information in the Income Registers becomes richer over time. The tax base has been broadened to include various transfers, beginning in the mid 1970's. In addition, data on untaxed transfers have been gradually added to the register and, hence, data on transfer receipts are more or less complete today. Over the years, the Income Registers have also come to include more data on individual background characteristics. These include sex, age, country of birth, education, marital status etc. In the Appendix we provide an overview of the variables in LINDA.

## 2.2 The Population and Housing Censuses

The Population and Housing Census (*Folk- och Bostadsräkningen*) has been conducted every fifth year from 1960 to 1990.<sup>5</sup> The population of the Census is registered residents at the 1<sup>st</sup> of November each year and the information generally pertains to this point in time.<sup>6</sup> The Censuses are based on questionnaires that all Swedish residents over 15 years-of-age are required to fill in. The response rate has been extraordinary high, but has declined somewhat over time; in, e.g., 1975 the response rate was 99.1 percent and 1990 it was 97.5 percent.

The Censuses broadly consist of three types of information: individual, household, and housing characteristics. The information included from these three categories varies somewhat across Census years. They always include the demographic characteristics, the employment status, the occupation, the industry, the place of residence, and the

---

<sup>5</sup> The census in 1960 is referred to as a *Population Census*, while the later censuses are referred to as *Population and Housing Censuses*. The Population Censuses started in 1860. The Censuses were conducted every tenth year during 1860-1930. During 1930 to 1990 they were conducted every fifth year with the exception of 1955.

<sup>6</sup> The only exception to this rule is the Census of 1980, when the date of reference was the 15<sup>th</sup> of September.

location of the workplace of the individual. Household characteristics consistently include family size and the number of children. The housing characteristics comprise the type of housing and the size of the dwelling. Detailed information on the design of the Censuses and their contents are available in *Folkräkningen* (1960) and *Folk- och Bostadsräkningen* (1965, 1970, 1975, 1980, 1985, 1990).

### 2.3 Other registers

Primarily during the 1990's, there is a significant increase in the number of registers included in LINDA. Below we give a summary account of the contents of the registers that have been added to the data set. For each register, we indicate the year when the data become available.

There are two registers related to retiree income. The Pension Register (*Pensionsregistret*, available from 1994) provides detailed information on different forms of old-age pensions and transfers accruing to retirees. The Pensionable Income Register (*Pensionspoängsregistret*) contains information on annual income forming the basis for old-age pensions; this register has data for this type of income since 1960.<sup>7</sup>

The Sick-leave Register (*Sjukfallsregistret*, available from 1997) contains information on sick-leave spells paid by the Social Insurance Offices (*Försäkringskassorna*) and compensation during these spells. Notice that employers are currently required to pay sick-leave compensation for employees during the first two weeks of the spell.<sup>8</sup> Consequently, the initial weeks are not included in the data. The Parental-leave Register (*Föräldrapenningsregistret*, available from 1997) is another register that is added during the 1990's. It holds information on the number of days on parental leave for the mother as well as the father.

There are two registers providing information on unemployment spells and unemployment compensation. The first of these, the Unemployment Register (*HÄNDEL*, available from 1991), contains information on all individuals registered at the employment offices. For each individual, it gives, e.g., the duration of the

---

<sup>7</sup> The Pensionable Income Register is available from 1995. Notice, however, that a separate file with pensionable income histories for all individuals that have ever been in LINDA can be obtained from the Social Insurance Board. This means that one has this kind of income information for representative samples for 1960, 1965, and 1968 and onwards.

<sup>8</sup> Prior to 1992, the Social Insurance Offices paid all spells. From 1992 to 1997, employers paid for the initial two weeks of sick-leave. In 1998, employers were required to pay for spells lasting up to four weeks; and in 1999 matters have been changed yet again and employers are obliged to pay for the initial two weeks.

unemployment spell, the reason for ending the spell, the education and the labor market experience related to the type of work the individual is applying for, and placements in labor market programs. The Unemployment Compensation Register (*AKSTAT*, available from 1994) contains information on, e.g., the number of days on unemployment insurance and/or cash assistance, the amount received, and the previous wage.

From 1998 and onwards there is an interesting addition to LINDA when wage data have been collected for the LINDA sample.<sup>9</sup> The “wage register” includes data on monthly salaries (in full time equivalents), occupation, and information on whether the individual worked part time or full time. In general, the wage measure includes all taxable components of the wage.<sup>10</sup>

*Table 1: Included registers in LINDA*

<b>Register</b>	<b>Swedish name</b>	<b>Time period</b>
Income Register	Inkomst- och Förmögenhetsstatistiken	1968--
Population and Housing Census	Folk- och Bostadsräkningen	1960, -65, -70, -75, -80, -85, -90
Pensionable Income Register	Pensionspoängsregistret	1995--
Pension Income Register	Pensionsregistret	1994--
Unemployment Register	HÄNDEL	1991--
Unemployment Compensation Register	AKSTAT	1994--
Higher Education Register	Högskoleregistret	1977-95
Local Adult Education Register	Komvuxregistret	1988-95
Sick-leave Register	Sjukfallsregistret	1997--
Parental-leave Register	Föräldrapenningsregistret	1997--
Wage Register	Löneregistret	1998--

Finally, there are two registers dealing with enrollment in education. First, there are data on enrollment and exams from specific programs within higher (tertiary)

<sup>9</sup> To be precise, wages in 1998 exist for the LINDA sample of 1997. This will change in 1999 such that wages in 1999 will pertain to the LINDA sample of 1999.

<sup>10</sup> Notice that from 1992 to 1997, wages also exist for around 70 percent of the LINDA sample. The reason for the incomplete coverage is that wage data were sampled in the private sector. Public sector wage data are complete.

education. This information is available from 1977 to 1995. Second, there are data on enrollment in local adult education. These data are available from 1988 to 1995. Both data sources have very detailed information on the actual program of enrollment. These two registers will not be automatically updated in LINDA.

Table 1 summarizes this section by presenting the registers that are included in LINDA and the time period for which data in the particular register are available.

### **3. The sampling frame and sampling procedure<sup>11</sup>**

In this section we give a description of the sampling frame and the sampling procedure. Moreover, we provide the formulas for computing population aggregates. These population aggregates are useful for validation of the data.

Apart from the fact that an individual is defined as an immigrant if he or she was born abroad, the features of the sampling frame are analogous for the population and immigrant samples. However, there are two essential differences. First, since there is no information on country of birth prior to 1970, the immigrant sample has the sampling frame of 1970 during the 1960's. As a consequence, the immigrant sample is not cross-sectionally representative during the 1960's. Second, to be included in the immigrant sampling frame individuals must have an assigned country of birth. The final assignment of the country of birth indicator can take as long as two years, however. Therefore, the immigrant sample is only representative for the population of immigrants that stayed in Sweden long enough to get an assigned country of birth. Apart from these discrepancies the fundamentals of the sampling frames and sampling procedures are identical for the two samples and therefore we only describe them for the population sample. The extension to the immigrant sample should be straightforward.

---

<sup>11</sup> This section builds on Lindström (1999).



### 3.1 The sampling frame

From 1991 to 1996, the sampling frame consists of all individuals who have lived in Sweden during a particular year. Thus, those individuals who lived in Sweden continuously, those who were born, those who died, those who immigrated, and those who emigrated during a particular year, collectively comprise the sampling frame. From 1997 and onwards, the sampling frame also includes the estates of deceased persons. One can exclude these estates using the information in *BOBJTYP*; *BOBJTYP* = 03 if the observation pertains to the estate of a deceased individual. From 1987 and onwards one can also restrict the population to individuals who resided in Sweden at the end of the year by using the variables *RTBTRIND* (1987-92) and *BRTBTR* (1993-98); *RTBTRIND* = *BRTBTR* = 1 if the individual lived in Sweden, *RTBTRIND* = *BRTBTR* = 0 otherwise.

Prior to 1991, the sampling frame is slightly different. Since there is no information in the Income Registers about individuals aged 0-15, Statistics Sweden has used information from the Population Register (*Registret för totalbefolkningen*, *RTB*) at the end of the year for these youths. This means that individuals aged 0-15 who have emigrated or died during the particular year are not included in the sampling frame. Although this change of the sampling frame can impair demographic analyses, it is likely to be of minor importance for most applications.

### 3.2 The sampling procedure

A Poisson sampling procedure with permanent random numbers was employed. The basic idea is that all individuals in the sampling frame a particular year are assigned a random number. The individuals keep their random number as long as they stay in the sampling frame. Newcomers in the sampling frame are assigned a “permanent” random number in an analogous fashion.

The samples have been constructed by assigning all individuals in the sampling frame of a specific year,  $t$ , a uniform random number,  $x$ , distributed on the  $[0,1]$  interval. Individuals with a random number less than a pre-specified value,  $X$ , are included in the sample. In the following year,  $t+1$ , the sampling frames from periods  $t$  and  $t+1$  are matched with one another. Individuals who were in the sampling frame in

year  $t$  are assigned their previous random number, while individuals who are in the sampling frame in  $t+1$  *only* are assigned a uniform random number – again distributed on the  $[0,1]$  interval. The individuals in year  $t+1$  with a random number less than  $X$  are included in the sample in  $t+1$ . Thus, individuals who disappeared from the sample between periods  $t$  and  $t+1$  are replaced by a random sample of the inflow between  $t$  and  $t+1$ , i.e. a random sample of newborns and immigrants. This sampling procedure ensures that each cross-section of LINDA is representative for the population in that particular year.

To construct the sample in  $t+2$ , Statistics Sweden started by creating a “merged” sampling frame with members of the sampling frame for one ( $t$  or  $t+1$ ) or both years ( $t$  and  $t+1$ ). This merged sampling frame is matched with the sampling frame of  $t+2$  and the above sampling procedure is applied to create the sample in  $t+2$ . An analogous sampling procedure is applied to the remaining years.

For each year, information on all household members of the sampled individuals is added to the data set. Households are defined for tax purposes, implying that two adult individuals belong to the same family if they are married or if they are cohabiting and have children in common. Cohabiting individuals without common children are separate households for tax purposes. Family members are only included in the sample as long as they stay in the family. Children who have been sampled as family members stay in the sample until they turn 18. The Census does not apply this definition of a household; household members are defined on the basis of whether individuals live together.

The variable *BURVKODP* can be used to restrict the population to individuals who have actually been sampled: *BURVKODP*=1 if an individual has been sampled; *BURVKODP* = 2 if an individual has been included as a family member.

### 3.3 Constructing population aggregates<sup>12</sup>

In principle, there are two different ways of constructing an unbiased estimate of a population aggregate (and the variance of that estimate): one can use sampled individuals only; the alternative is to use all individuals, i.e. the sampled individuals and their household members. To illustrate these different procedures, let us introduce

---

<sup>12</sup> The formulas in this section are based on Rosén (2000).

some notation; the names within parentheses are variable names in LINDA. Let  $i$  index individuals and  $h$  households. Moreover, let  $N$  denote the size of the sampling frame (*BSTORANP*);  $n$  the number of sampled individuals (*BLILLANP*);  $M_h$  the size of household  $h$  (*BANT*); and  $m_h$  the number of sampled individuals in a particular household,  $h$  (*BANTP*).

To be concrete, suppose we want an estimate of aggregate income,  $\omega$ . The simplest approach is to base the estimate only on individual income,  $w_{ih}$ , for individuals who were actually sampled (*BURVKODP=1*).

$$\hat{\omega} = \frac{N}{n} \sum_{i \in D} w_{ih} \quad (1)$$

where  $\hat{\omega}$  denotes the estimator and  $D$  the set of individuals for which *BURVKODP=1*. An unbiased estimate of the variance of  $\hat{\omega}$  is given by

$$\text{v\hat{a}r}(\hat{\omega}) = \frac{N(N-n)}{n(n-1)} \left\{ \sum_{i \in D} w_{ih}^2 - \frac{n\hat{\omega}^2}{N} \right\} \quad (2)$$

where  $\text{v\hat{a}r}(\cdot)$  denotes the estimated variance.

Another approach is to use all observations, i.e. the sampled individuals and their family members. An unbiased estimate of  $\omega$  is then given by

$$\hat{\omega} = \frac{N}{n} \sum_h \frac{m_h}{M_h} \sum_i w_{ih} \quad (3)$$

The weight,  $m_h/M_h$ , corrects for the fact that there are two routes for an individual to end up in the full sample: either he/she is sampled (*BURVKODP=1*) or another member of the household is sampled. An unbiased estimate of the variance of  $\hat{\omega}$  is obtained using

$$\text{v\hat{a}r}(\hat{\omega}) = \frac{N(N-n)}{n(n-1)} \left\{ \left[ \sum_h \frac{m_h}{M_h^2} \left( \sum_i w_{ih} \right)^2 \right] - \frac{n\hat{\omega}^2}{N} \right\} \quad (4)$$

In the case where observations on household income,  $w_h = \sum_i w_{ih}$ , are immediately available, then instead of (3) one can use

$$\hat{\omega} = \frac{N}{n} \sum_h \frac{m_h}{M_h} w_h \quad (3')$$

and instead of (4)

$$\text{vâr}(\hat{\omega}) = \frac{N(N-n)}{n(n-1)} \left\{ \sum_h \frac{m_h}{M_h^2} w_h^2 - \frac{n\hat{\omega}^2}{N^2} \right\} \quad (4')$$

## 4. The Development of some Key Variables

In this section we give the development of some of the variables included in the data set. The reasons for doing this are basically twofold: (i) we want to illustrate the usefulness and drawbacks of LINDA; (ii) we want to give future users of LINDA the means for judging whether the data they have received are reasonable. Throughout we use the population sample, restrict attention to individuals who were actually sampled, and mostly report statistics for the population of 20-64 year-olds. In the Appendix we give the total size of the population sample and the number of sampled individuals aged 20-64, which should provide quick and rudimentary checks on whether the data received are correct.

### 4.1 Earnings

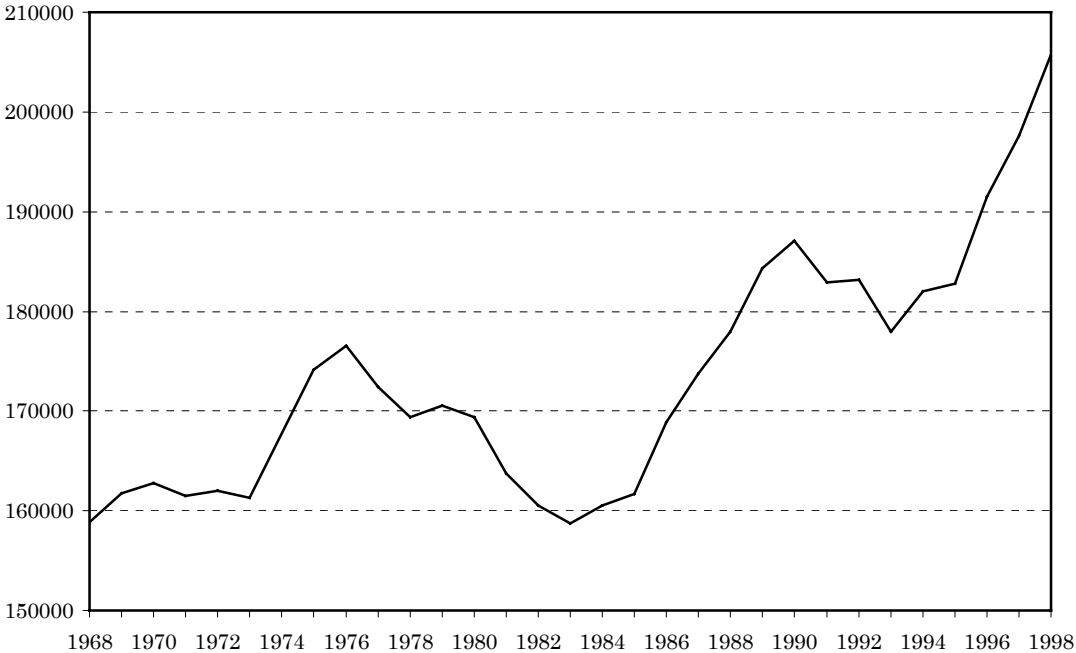
We begin by taking a look at the development of annual earnings between 1968 and 1998.<sup>13</sup> *Figures 1 and 2* present two aspects of the earnings development for individuals aged 20-64: *Figure 1* gives the development of mean earnings, conditional on having

---

<sup>13</sup> From 1978 and onwards, the measure of earnings is directly available in the data. During 1968-73 we have calculated earnings by adding income from employment (*A-inkomst av tjänst + sjöinkomst*) and income from business (*A-inkomst av jordbruk + A-inkomst av rörelse*). During 1974-77 we subtracted the sum of pensions (*pension*), unemployment compensation (*dagpenning vid arbetslöshet + KAS*), and compensation during labor market training (*utbildningsbidrag*) from income from employment and business. Unemployment compensation and compensation during labor market training became taxable income, and hence part of income from employment, in 1974. The earnings measure that we are able to construct during 1968-73 is not entirely consistent with the measure during later years, since we have no information on sickness benefits (which should be added to income from employment) and pensions. The sickness benefit became taxable income in 1974; hence it is part of income from employment from then on.

earnings greater than the basic amount;<sup>14</sup> and *Figure 2* the fraction of individuals with earnings less than the basic amount. We separate the development in the latter figure by gender since the evolution of “zero earnings” – i.e. earnings less than the basic amount – is markedly different for males and females. The development of mean earnings is very similar apart from a trend increase in female earnings relative to male earnings.<sup>15</sup>

*Figure 1: Average earnings in 1997 prices (SEK), 20-64 year-olds, 1968-98.*



*Notes:* Calculations based on the Income Registers using all individuals aged 20-64 in the sample. Deflated by the CPI. Only individuals with earnings greater than the basic amount are included in the calculation of the average.

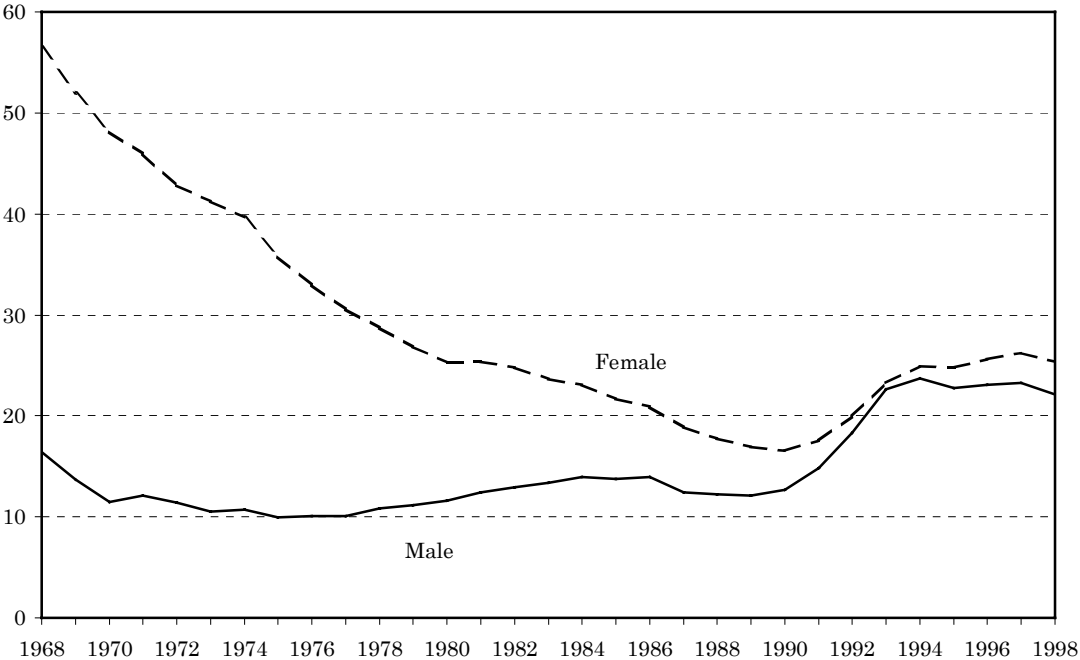
It comes as no surprise that there is a marked cycle in the development of average earnings, since labor income is the product of wages and hours. There are two significant drops in mean earnings. The drop in the beginning of the 1980’s was the consequence of successive devaluations and an economic downturn. The unemployment shock of the 1990’s – when unemployment rose from less than 2 percent in 1989 to more than 9 percent in late 1993 – is also clearly visible in the data. We noted earlier that the

<sup>14</sup> The basic amount is the minimum amount of earnings that qualifies to the earnings related part of the public pension system. In 1998, this amount was 36,400 SEK.  
<sup>15</sup> In 1968, female earnings were 60.8 percent of male earnings; in 1998, the gender earnings gap stood at 71.3 percent.

information on, e.g., income is conditional on the state of the tax code. At first glance, there seem to be no major “jumps” in the evolution of earnings. However, users of LINDA should notice that income post 1990 is not comparable to income for previous years, because of the “tax reform of the century”; see Agell *et al.* (1996). This tax reform implied a significant broadening of the tax base and, hence, an increase in average earnings as calculated from filed tax reports.

The extent of the increase in female labor force participation is vividly illustrated in *Figure 2*. In 1968, more than 57 percent of women had earnings of less than the basic amount. At the same point in time, 16 percent of males earned less than this threshold. By 1993, the gender difference had almost disappeared; in 1993, the difference between males and females was only 0.7 percentage points.

*Figure 2:* The proportion earning no more than the basic amount, percent, 1968-98.



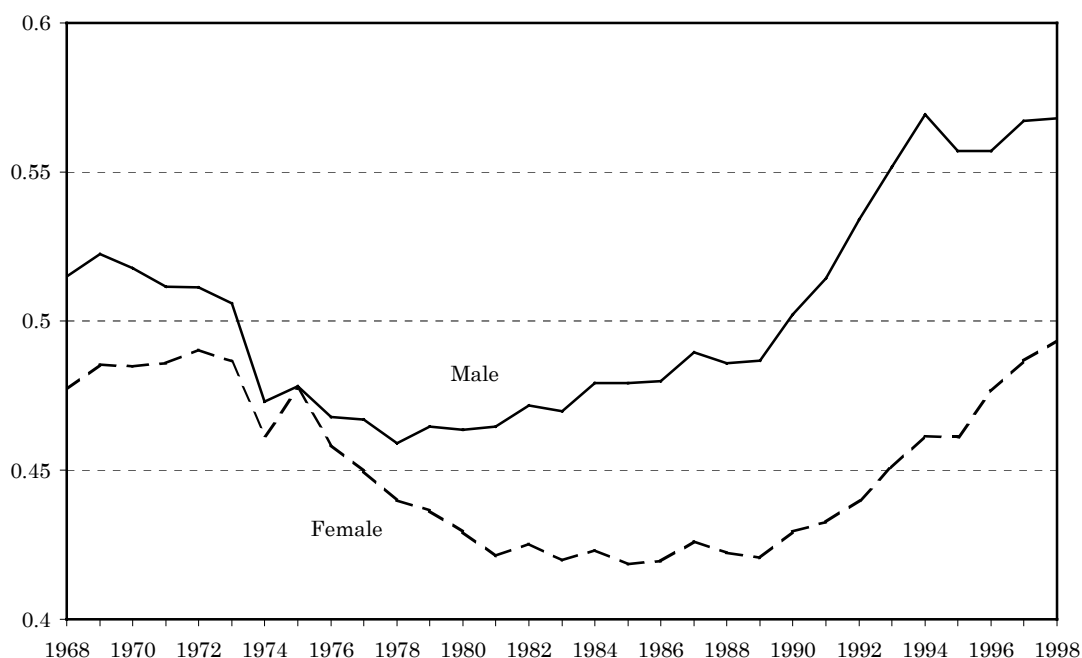
*Notes:* Calculations based on the Income Registers using all individuals aged 20-64 in the sample.

*Figure 3* shows a third aspect of the earnings development during the past thirty years, namely the dispersion of earnings. The overall development of earnings dispersion corresponds well to the evolution of wage dispersion; see Edin and Holmlund

(1995). The spread of the (log) wage distribution decreased precipitously from the late 1960's to the early 1980's to increase somewhat in the second half of the 1980's.

A similar pattern is evident in *Figure 3*, although the spread of the male earnings distribution starts to increase around 1980 and the spread of the female earnings distribution around 1990. In the wake of the unemployment hike of the 1990's, earnings dispersion has increased to reach the level of the late 1960's. Of course, this is probably an overestimate of the increase in wage dispersion, since earnings dispersion is a function of the distribution of hours as well.<sup>16</sup>

*Figure 3*: The standard deviation of log earnings, 1968-98.



*Notes:* Calculations based on the Income Registers using all individuals aged 20-64 in the sample. Only individuals with earnings greater than the basic amount are included in the calculation of the standard deviation.

#### 4.2 The family concept

We noted earlier that households are defined for tax purposes in the Income registers while the Census uses the conventional definition. To illustrate the consequences of this

<sup>16</sup> We have no explanation for the downturn of earnings dispersion in 1974. The difference between the 90<sup>th</sup> and 10<sup>th</sup> percentile of the log earnings distribution exhibits a similar (and unmotivated) drop in 1974. The “anomaly” may partly be the result of changes in the definition of taxable income in 1974; see footnote 13.

difference *Table 2*, *inter alia*, reports the percentage of married and cohabiting individuals according to the Income registers and the Censuses in 1980, 1985, and 1990.

The proportion of married individuals in the Income Register corresponds well to the Census. However, the fraction of cohabiting individuals deviates significantly. The difference, which stems from the fact that the Income Register generally codes cohabitants with no common children as singles, is around 10 percentage points.

The development during the 1990's seemingly suggests that the proportion of cohabiting individuals has decreased. However, this is at least partly the result of a change in the procedures for filing tax returns. A new and simplified tax return form was introduced in 1990. A side-effect of this introduction is that the quality of the information pertaining to cohabitants who have children in common may have declined.<sup>17</sup>

*Table 2: Married and cohabiting individuals in the Income Registers and the Censuses, 20-64 year-olds, percent.*

Year	Income Register		Census	
	Married	Cohabiting	Married	Cohabiting
1980	58.9	1.5	58.4	12.0
1985	54.3	4.0	53.9	15.2
1990	53.8	4.0	53.1	14.6
1993	49.8	4.5	n.a.	n.a.
1997	46.4	3.3	n.a.	n.a.

*Notes:* In the Income Register, an individual was classified as cohabiting if he/she was jointly taxed with someone else and not married (to do this we used the variables *TAXKOD* and *CIV*). The Census information is distilled from a direct question. The samples in 1980, 1985, and 1990 refer to individuals who responded to the Population and Housing Census. The samples in 1993 and 1997 are restricted (using the information in *BOBJTYP* and *BRTBTR*) to refer to individuals who were alive and resided in Sweden on December 31 each year.

In sum, the household information of the Income Registers in principle generalizes to married individuals and cohabitants with children in common. However, the latter group may be underreported – in particular during the 1990's. If the researcher wants to

<sup>17</sup> Work is currently underway to improve the family concept of the Income Register during the 1990's.

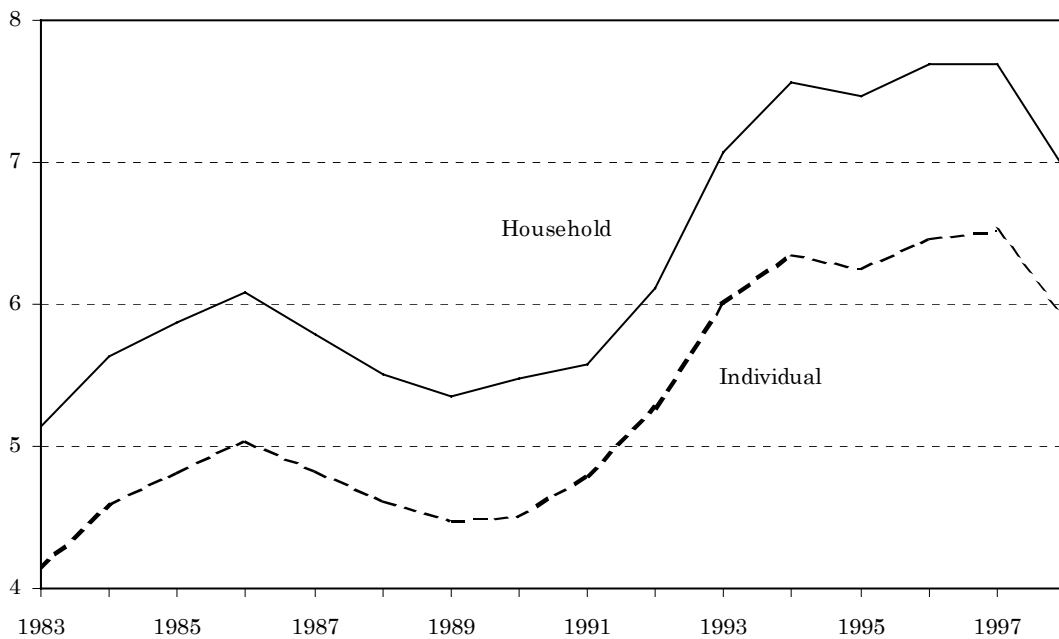


generalize to all individuals residing together the Censuses should be used instead.<sup>18</sup>

### 4.3 Social assistance

The information on transfers in LINDA is more or less complete from 1983 and onwards. In *Figure 4*, we give an example pertaining to the development of the proportion of social assistance recipients from 1983 to 1998.

*Figure 4*: The incidence of social assistance receipt, percent, 1983-98.



*Notes*: Calculations based on the Income Registers using all individuals aged 20-64 in the sample.

Social assistance is a means-tested benefit program. Family income and family composition determine the benefit entitlement. The Income Register contains the amount of social assistance received at the individual level. This information, however, has the special feature that the household benefit may be allocated in full to a household head designated by the social worker. To illustrate the consequence of this feature, *Figure 4* presents two ways of calculating the incidence of social assistance receipts.

<sup>18</sup> An alternative way to calculate the number of cohabiting individuals would be to sum the number of adult (over age 17) members of a household using the household identifier available in LINDA. Using this alternative, we get the following numbers for the fraction of cohabiting individuals: -0.4 % (1980); 4.7 % (1985); 4.8 % (1990); 4.3 % (1993); 3.3 % (1997). Clearly, this household definition and the information on marital status are not consistent in 1980, which, again, is an indication that the household information is most reliable for married individuals.

Firstly, we classify an individual as a social assistance recipient if he or she is a member of a social assistance recipient household; see the solid curve labeled “Household”. Secondly, we classify an individual as a recipient if he or she has received a positive amount of assistance; see the curve labeled “Individual”.

The definition of a recipient is clearly important as *Figure 4* illustrates. There is a difference of around one percentage point between the two definitions. The definition using only information at the individual level underestimates the number of individuals on “welfare” by 15-20 percent in comparison to the appropriate definition.

Looking at the evolution over time, there is a slight trend and a marked cycle in social assistance receipt. The consequence of the economic downturn of the 1990’s is clearly illustrated in the data. The incidence of welfare receipt increased by 44 percent (2.3 percentage points) from 1989 to 1997.

#### 4.4 Education

In any empirical analysis using individual data one would probably like to control for schooling. The Population and Housing Censuses of 1960, 1970, and 1990 contain information on education and the Income Register contains the Educational Register from 1990 and onwards. Moreover, there is the Socio-economic classification during all Census years with the exception of 1965. The Socio-economic classification is based on the *normal* education requirement to hold a specific position. This information is potentially useful in instances when the researcher is not interested in the educational premium *per se*, but wants a control for skill differences across individuals.

The Census of 1960 holds information on all university degrees and a sub-set of degrees from upper-secondary school. The 1970 Census contain information on all degrees for individuals aged 16-59 and the 1990 Census holds the same kind of information but for 16-64 year olds.

In *Table 3* we have restricted the population to 25-59 year olds. The upper limit was chosen in order to be able to compare the figures across time. The increase in the educational level over time is striking. As a consequence of the expansion of university education during the 1960’s, the fraction of individuals with a university degree

quadrupled between 1960 and 1970. Furthermore, the fraction of the population with the lowest level of schooling declined by almost two thirds between 1970 and 1997.

*Table 3* is also meant to illustrate the improving quality of the information in the Educational Register. In 1990 there are two observations on education – one from the Census and the other from the Income Register. There is a fair amount of missing values in the register and it underestimates the fraction of individuals with a university degree. In choosing between which data source to use in 1990 one should use the Census. In 1991, however, the Educational Register was updated using the information of the 1990 Census. Moreover, Statistics Sweden put some effort into getting information on the educational level of immigrants in 1995. As a consequence, the Educational Register has become more complete over time and there are relatively few missing values in 1997.

*Table 3: The population of 25-59 year-olds by highest education, percent.*

	1960	1970	1990		1997
	FoB	FoB	FoB	IoF	IoF
Less than upper-secondary degree ( $\leq 10$ years)	n.a.	62.2	34.1	29.4	22.4
Upper-secondary degree	n.a.	26.1	37.6	42.2	48.0
Tertiary degree	1.7	8.0	22.9	21.1	28.3
Missing	n.a.	3.7	5.5	7.3	1.2

*Notes:* “FoB” refers to the Population and Housing Census; “IoF” to the Income Register. The sample in 1960, 1970, and 1990 refers to individuals who responded to the Population Census. The sample in 1997 is restricted (using the information in *BOBJTYP* and *BRTBTR*) to refer to individuals who were alive and resided in Sweden on December 31.

#### 4.5 Immigration

Another frequently used individual characteristic is immigrant status. *Table 4* reports the proportion of individuals aged 20-64 born abroad. Moreover, it gives some information pertaining to the structure of immigrant stock at each point in time, by reporting region of origin for the stock of foreign-born individuals. The data until 1990 come from the Population Census, while post-1990 data are from the Income Register.

Table 4. The proportion of foreign-born individuals and immigrants by region of origin, 20-64 year-olds, percent.

	1960	1970	1975	1980	1985	1990	1995	1997
Immigrants	5.5	8.7	9.1	10.2	10.7	12.1	13.4	13.4
Thereof								
...Nordic	57.7	59.3	59.0	57.2	51.7	43.8	34.1	31.6
...OECD (excl. Nordic)	23.2	13.7	18.4	15.2	16.5	12.9	12.1	11.9
...Non-OECD	19.2	27.0	22.6	27.7	31.9	43.2	53.8	56.5

*Notes:* The samples in 1960 to 1990 refer to individuals who responded to the Population and Housing Census. The samples in 1993 and 1997 are restricted (using the information in *BOBJTYP* and *BRTBTR*) to refer to individuals who were alive and resided in Sweden on December 31 each year. "OECD" pertains to the membership as of 1985.

There are two developments that we think are noteworthy. First, the number of foreign-born individuals has risen rather drastically from 1960-97. In 1997, first-generation immigrants in Sweden constitute 13.4 percent of the population. Relative to the overall population, their number exceeds the number of first-generation immigrants in the United States – a country sometimes referred to as the “nation of immigrants”. Second, there has been a drastic shift in the structure of immigration. Beginning in the mid 1970’s with an inflow of political refugees from Chile, the relative share of immigration for labor market reasons has decreased while the immigration for asylum reasons has increased. This shift is evident in the data at least from 1985 and onwards. In 1980, immigrants of Nordic descent constituted 57 percent of the immigrant stock, while immigrants from non-OECD countries constituted 28 percent; by 1997 the share of immigrants from Nordic countries had decreased to 32 percent; concomitantly, the share of immigrants of non-OECD origin had risen to 56 percent.

## 5. Accessibility, organization and documentation

Access to the data will be granted for all researchers through Statistics Sweden subject to the normal confidentiality provisions. An application for usage of the data should be sent to the head of the Department for Welfare statistics (*Välfärdsstatistiken*) at Statistics Sweden who decides on whether access is granted. In the event of a successful application, the head of the research project will have to sign an agreement where he or she agrees to adhere to the normal disclosure provisions.

Statistics Sweden will charge researchers and policy analysts on a marginal cost basis. The maximum cost will be in the order of SEK 1,500 per year, provided that data are delivered in SAS-format. When orders are for several years of LINDA, the cost per year will be lower. Statistics Sweden can also supply tailored runs based on LINDA. In such an instance, there are no pre-specified prices available – the price will have to be determined case by case.

At present, the data from the Income Registers and Population Censuses are organized by cross-section and origin register. For the remaining registers, the file pertaining to a particular register contains all individuals who have ever appeared in LINDA; moreover, years prior to 1996 are lumped together and the updates (i.e. 1996-1998) are in separate files. The data are de-identified and delivered on CD's.

At the time of writing, documentation is only rudimentary. For each year, variable definitions are available in Swedish and there is no “longitudinal” documentation of the contents of LINDA. However, the members of “the Consortium” are putting considerable effort into providing a longitudinal documentation for a sub-set of the variables. The results of these, and future, efforts will be made available at the web-site: <http://www.nek.uu.se/linda>.

## References

- Agell, J., P. Englund and J. Södersten (1996), “Tax Reform of the Century – The Swedish Experiment”, *National Tax Journal* 49, 643-64.
- Edin P-A. and B. Holmlund (1995), “The Swedish Wage Structure: The Rise and Fall of Solidarity Wage Policy?”, in R.B. Freeman and L.F. Katz (Eds.), *Differences and Changes in Wage Structures*, National Bureau of Economic Research Comparative Labor Markets Series, University of Chicago Press, Chicago, 307-43.
- Eriksson, R. and R. Åberg (1987), *Welfare in Transition*, Clarendon Press, Oxford.
- Folkräkningen* (1960), Statistics Sweden.
- Folk- och Bostadsräkningen* (1965, 1970, 1975, 1980, 1985, 1990), Statistics Sweden.
- Klevmarken, N. and P. Olovsson (1993), *Household Market and Nonmarket Activities*, Industriens Utredningsinstitut.
- Lindström, K. (1999), “LINDA: Urvalsförfarande och urvalsramar”, mimeo, Statistics Sweden.
- Rosén, B. (2000), “Teori för urvalsundersökningar”, mimeo, Statistics Sweden.

## Appendix

### A.1 The number of sampled individuals

To give users of LINDA the means to quickly check whether the data they have received are correct, Table A.1 reports the number of observations for the total sample as well as for individuals aged 20-64.

Table A.1: Number of individuals in the Income Register, population sample.

Year	No. of observations, full sample	No. of observations, 20-64 year-olds
1968	264,754	155,279
1969	268,562	157,938
1970	271,351	159,358
1971	274,022	160,243
1972	274,671	159,910
1973	275,417	159,654
1974	276,344	159,537
1975	277,791	160,103
1976	279,045	160,572
1977	279,689	160,389
1978	280,185	160,452
1979	281,271	160,998
1980	282,070	161,272
1981	282,400	161,370
1982	282,377	161,155
1983	280,961	161,403
1984	282,892	162,134
1985	283,621	162,538
1986	284,571	163,291
1987	285,593	163,920
1988	287,205	164,923
1989	289,448	166,171
1990	292,548	167,891
1991	293,550	168,268
1992	295,361	169,483
1993	297,576	170,879
1994	300,000	172,629
1995	300,978	173,335
1996	306,404	173,723
1997	306,931	174,341
1998	307,390	174,821

*Notes:* The number of observations refers to the number of *sampled* individuals, i.e. individuals which have *BURVKODP* = 1. Prior to 1978 a single individual may appear in the data twice; we have deleted one of these observations. The number of individuals with duplicate observations equals: 26 (1968); 20 (1969); 29 (1970); 23 (1971); 16 (1972); 0 (1973); 5 (1974); 6 (1975); 5 (1976); 4 (1977). STATA-code that eliminates duplicate observations is available at the website: <http://www.nek.uu.se/linda>.

## A.2 An overview of the variables in LINDA

Table A.2 describes the availability for a selection of variables in the Population and Housing Censuses, the Income Registers, and the Pensionable Income Register. Notice that the list of variables is far from exhaustive. The reader is referred to the website: <http://www.nek.uu.se/linda> for a full description.





Year

<b>Variable</b>	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	
Industry of employment (5)	x					x					x					x				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Occupation (6)	x					x			x	x	x	x	x	x	x	x	x	x			x									x										
Sector of employment																					x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Location of workplace	x					x					x					x					x																			
<i>Housing</i>																																								
Type of housing	x					x					x					x					x																			
Size of housing	x					x					x					x					x																			
Tax assessment values																											x	x	x	x	x	x	x	x	x	x	x	x	x	x
Housing coordinates																																								x

*Notes*

- (1) Pre-1968 immigrants have missing values except for the Population Census in 1970, but one can always use the Census of 1970 to get the true immigration year.
- (2) Households are defined for tax purposes during non-Census years.
- (3) All income and wealth measures are defined for tax purposes.
- (4) These data are not representative for the population of 1961-64 and 1966-67
- (5) Industry-coding according to ISIC-69 1970-92 and according to ISIC-92 1993-98. 4-digit ISIC in 1970 and 1975; 3-digit in 1978; 5-digit 1980-98.
- (6) Occupational-coding differs slightly between Census years. Rough classification during 1968-69, 1971-74, and 1976-77.