

The Determinants of Income Dynamics

Gustavo De Santis and Giambattista Salinari

Abstract Models of income distribution more or less succeed in linking the current level of household (or individual) income to household (or individual) characteristics. However, they typically prove far less satisfactory in explaining income dynamics. In this paper, we use a modified version of Gibrat's model to show that this apparent weakness has a theoretical reason: in the short run (here, 2 to 4 years), randomness prevails and systematic influences are hard to identify. One of these, however, seems to be that small incomes tend to increase more and with more variability than large ones. The traditional version of Gibrat's model cannot explain this peculiarity: a shortcoming that, with a relatively minor modification, can be overcome.

Keywords: Statistical modelling, Measuring poverty and inequality

1 Household Income Dynamics

Let us consider the dynamics of household income and its determinants in Italy during the last decade. By “income dynamics” we mean the probability of a transition from a given income level in year t , to a different income level in year $t+k$, k years later. Can we estimate these probabilities? How? What socio-economic characteristics affect them most?

Already at the individual level, income trajectories are influenced by very many variables: education, age, gender, industry, SES (Socio-Economic Status), changes in the labour market, economic cycles, etc. At the household level, the number of these forces increases exponentially, taking interactions into account, and soon becomes very difficult to handle.

¹

Gustavo De Santis, Dept. of Statistics, Un. of Florence; email: desantis@ds.unifi.it

Giambattista Salinari, Dept. of Statistics, Un. of Florence; email: salinari@ds.unifi.it

Financial support from the Italian MIUR and from the EU is gratefully acknowledged.

In order to simplify the problem, we break it in two: A) we first model income trajectories in a theoretical, perfectly homogeneous population; B) later, we complicate the model, by explicitly considering heterogeneity.

This procedure permits us to disentangle the general features of income dynamics, common to all trajectories (point A), from those that are contingent on households characteristics (point B). Our model proves capable of keeping the two dimensions under control (general process and influence of specific situations), and, both theoretically and empirically, compares favourably with other, more conventional models of income dynamics, as we show below.

2 Gibrat's model: traditional and modified version

Our model of income dynamics in a homogeneous population (see Salinari and De Santis 2010), derives from Gibrat's (1931) seminal paper, where incomes are thought to evolve in a multiplicative form: for every individual, income at time $t+k$ can be obtained by multiplying income at time t by a series of random shocks

$$1) \quad y_{t+k} = y_t \prod_{j=1}^k R_j$$

(y = income, R_i = random variables), or, in logs,

$$2) \quad \log(y_{t+k}) = \log(y_t) + \sum_{i=1}^k R_i$$

The main theoretical consequences of this model are (see e.g. Salinari and De Santis 2010):

- 1) log-growth is independent of log-income,
- 2) the stationary distribution of log-incomes is normal,
- 3) the variance of log-income increases linearly with k ,
- 4) the covariance of log-income increases with t , but decreases with k .

Empirical analyses show that Gibrat's model describes satisfactorily the dynamics of medium and large incomes, but proves somewhat deficient when it comes to the dynamics of small incomes, because their growth-rates are systematically greater and more variable than expected (Fields and others 2006; Salinari and De Santis 2008; 2010). In other words, the first implication of Gibrat's model (independence between log-income and log-increments) does not seem to hold.

This shortcoming, however, may depend on purely statistical reasons. In order to see why, let us transform Gibrat's model into its additive version:

$$3) \quad y_{t+1} = y_t + y_t(R-1)$$

The growth rate $R-1$ can now be thought of as the difference between the gain rate G and the loss rate L , so that eq. (3) can be rewritten as:

$$4) \quad y_{t+1} = y_t + G - L$$

where both G and L are random, non negative variables. Since losses L cannot be greater than $y_t + G$ (otherwise income at time $t+1$ would be negative), L is a truncated random variable with a probability density function ranging between 0 and $y_t + G$. But the impact of truncation is greater for small than for large incomes, so that the growth rate of small incomes tends to be positively affected.

But why is it also more variable? Under our homogeneity assumption, gains and losses can be imagined to occur to individuals as in a Bernoulli experiment, where the probability of getting a "unit" of loss (or gain) is proportional to individual income. In this case, the gains and losses experienced by households distribute as a Poisson, the mean and variance of which increase linearly with income. With this assumption, the variance of the growth-rate $V[R-1] = V[(G-L)/y_t]$ decreases as income increases.

A possible way of improving over Gibrat's model is to use a truncated normal distribution (T) to model income variations ($\Delta y = y_{t+1} - y_t$). Let these variations be normally distributed with mean and variance proportional to y_t . Since variations cannot be smaller than $-y_t$ (otherwise y_{t+1} would be negative), we restrict the domain of the probability distribution to the interval $(-y_t, +\infty)$. The probability density function $f(\cdot)$ of the truncated random variable T is therefore:

$$5) \quad f(\Delta y | y_t, \alpha, \beta) = \frac{\phi\left(\frac{\Delta y - \alpha y_t}{\beta y_t}\right)}{1 - \Phi\left(\frac{-y_t - \alpha y_t}{\beta y_t}\right)}$$

where a and b are two parameters defining the mean (αy_t) and the variance (βy_t) of the normal distribution, $\phi(\cdot)$ is the probability density function of the standard normal variable, and $\Phi(\cdot)$ is its cumulative function.

The "modified Gibrat's model" can now be defined as:

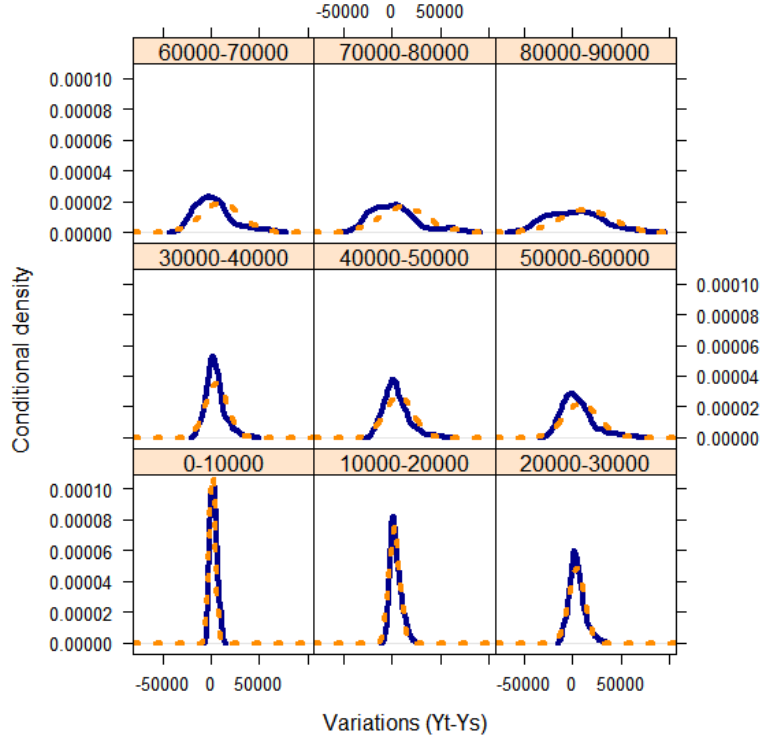
$$6) \quad y_{t+1} = y_t + T$$

and the estimates of the two parameters a and b can be derived numerically, e.g. with a maximum likelihood estimation. Figure 1 shows a comparison between the actual and the expected (eq. 5) income variation registered in 5195 Italian households in 2-year intervals (1998-2000, 2000-2002, ..., 2004-2006), conditional on the income class to which the households belonged at the start of each period. The variability of absolute income variations increases with income, and our model curves approximate reality quite well, especially if one considers that no covariate is taken into consideration here (homogeneity assumption).

The model of eq. 6 can be extended to the case of a heterogeneous population, i.e. where households have different socio-economic characteristics. Assume that $\mathbf{x}_{t,i} = (1, x_{1,i}, \dots, x_{p,i})^T$ is a time-dependent vector of covariates for household i , while $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ are two vectors of parameters describing the effects produced by these covariates respectively on the mean and on the variance of the distribution of the variations. We can now generalize model (5) as follows:

$$7) \quad \begin{cases} a_i = \exp(\boldsymbol{\alpha}^T \mathbf{x}_{t,i}) \\ b_i = \exp(\boldsymbol{\beta}^T \mathbf{x}_{t,i}) \end{cases}$$

Figure 1: Probability density distribution of conditional income variations (Italy 1998-2006). The variations (Δy) observed during the periods 1998-2000, ..., 2004-2006 for 5195 Italian households initially belonging to different income classes (0-10,000, ..., 80,000-90,000 euro). The solid line indicates the empirical densities estimated through Gaussian kernel. The dashed lines indicates the theoretical distribution generated by eq. 5 (with $a=0.16$ and $b=0.32$).



3 A Regression Analysis of Income Dynamics

Before applying "our" modified version of Gibrat's model, let us take a step backward and consider the results produced by an ordinary regression analysis of household income and growth rates, which we will later compare to our own.

We use SHIW data (the Bank of Italy Survey on Household Income and Wealth), for the period 1998 to 2006. This survey is carried out every other year, on about 8 thousand households (about 20 thousand individuals): it is described in detail (and its elementary data are freely accessible) on the website of the Bank¹, and it contains a

¹ <http://www.bancaditalia.it/statistiche/indcamp/bilfait>

panel part that we will exploit in this application. We have five rounds of the survey (years 1998, 2000, 2002, 2004, and 2006), and we study the change in the total net income of all the households whose demographic structure did not change in the period, and whose reference person was, at the beginning of the period, not older than 64 years and in the labour market.

We are interested in modeling income dynamics, using "independent" explanatory variables. Actually, most of the variables that we use are not truly exogenous (as we will briefly discuss below) but our interest here is not so much to "explain" income: merely to show that our approach (modified Gibrat's model) works better than other, more traditional ones. In the traditional version, the dependent variable is the growth rate (table 1, columns 2 and 3).

We, instead, model transition probabilities (table 2). In both cases, we consider two different transitions: over 2 years (e.g. from 1998 to 2000) and over 4 years (e.g. from 1998 to 2002). We drop a few outliers (fewer than 1%, with abnormally high or low growth rates) and we are thus left with about 5100 biannual transitions and 2400 quadrennial transitions, approximately equally distributed among the different time periods considered.

We use a set of "independent" variables that can be classified in two groups:

- a) *Demographic*: number of children (0, 1, 2, 3+), region (North, Center, South), age, and educational attainment of the reference person (1=illiterate or primary school; 2=secondary education; 3=tertiary education; 4=university degree or PhD).
- b) *Labor force*: occupation of the reference person (dependent, self employed, unemployed, other), period.

As for the traditional approach, although we are basically interested in dynamics (growth rates), we also model income levels. As for levels, we use:

$$8) \quad \log(y_t) = \beta^T \mathbf{x}_{t,i} + \varepsilon_{t,i}$$

while for growth rates (both for 2- and 4-year periods) we use:

$$9) \quad \log\left(\frac{y_{t+1,i}}{y_{t,i}}\right) = \beta^T \mathbf{x}_{t,i} + \varepsilon_{t,i}$$

Our results are summarized in Table 1. In all the cases, the baseline is a household living in the north of Italy, with one child and whose reference person is dependent worker, aged 45, with an average level of education. The static regression of income (first column) confirms what is generally known: all the covariates we use affect income in the expected direction: incomes are higher for better educated and older (but younger than 65 years) reference persons, while living in the south, or being a dependent worker (or, worse still, unemployed), depresses income. Notice, also, that income increases with the number of children, but the interpretation must be particularly cautious here: reverse causation, selection, and unobserved heterogeneity surely play a major role. Our simple model explains about 43 per cent of the variance in incomes.

If we now turn to the analysis of dynamics, i.e. of 2-year growth rates (second column) we find that only a few of our covariates exert a measurable effect, and the explained variance is merely 1%. The only variables that seem to matter are those

associated with a low income at the start (which, as we saw above, favors a more rapid increase in the period): being unemployed, or with a low education.

Table 1: Regression analysis of income and income growth rates

<i>Dependent variable:</i>	<i>Income</i>		<i>Growth rate over 2 years</i>		<i>Growth rate over 4 years</i>	
Intercept	10.223	***	0.087	***	0.167	***
Children 0	-0.089	***	-0.025		-0.056	
Children 2	0.077	***	0.010		0.034	
Children 3	0.069	**	0.040	*	0.051	
Age (Stand.)	0.140	***	-0.014	*	-0.011	
Education 1	-0.254	***	0.052	**	0.095	**
Education 3	0.240	***	0.009		0.008	
Education 4	0.562	***	0.023		0.006	
Self Empl.	0.064	***	0.008		0.037	
Unempl.	-0.616	***	0.231	***	0.403	***
Other	-0.160	***	0.036	.	0.084	**
Period 2	0.067	***	-0.011		0.012	
Period 3	0.110	***	0.002		-0.010	
Period 4	0.168	***	-0.023		/	
Center It.	-0.124	***	0.017		0.030	
South It.	-0.476	***	-0.008		-0.028	
R²	0.43		0.01		0.03	

Significance level: ***=1%; **=5%; *=10%.

Periods: 1=from 1998 (ref.); 2=from 2000; 3=from 2002; 4=from 2004.

Education: 1=low; 2=high (ref.); 3=high; 4=very high.

R² for the growth-rates regression have been calculated, for reason of comparability, through the formula: $\text{var}[\exp(\text{fitted.values})]/\text{var}(\text{actual.growth-rates})$

Basically the same happens over a longer time span (4 years, third column), with a very slight improvement in the overall goodness of fit (from 1 to 3%). What can we conclude? How can a few covariates be associated with income levels if they are not associated with growth rates? Notice that:

- income growth is a highly erratic variable;
- growth is a cumulative process: very small systematic differences, even if they go unnoticed in the short run, result in large differences in the long run;
- income trajectories can depend on their starting point;
- what we observe here are net incomes, which are also influenced by the effects of the fiscal policy (progressive taxation, subsidies, etc).

As for the first two points, consider, for example, “Self-employment” in table 1 (2nd and 3rd column): the coefficients are positive (income increases more rapidly for this category), but not significant. However, the mean income of the self-employed is significantly higher than that of the dependent workers. Besides, over a 4 year period, the effect of self employment is slightly more significant than it is over a 2-year period, and this is a pattern that holds for most variables.

But this explanation may not be simply extended to all our covariates. Take “Education 1”, for instance: this has a positive and significant effect on both the biannual and quadrennial growth-rates, but the income of the poorly educated workers is significantly smaller than that of highly educated ones. Our interpretation is that this

apparent contradiction depends on the different starting income of the two groups: highly educated worker are richer at the beginning (they typically come from richer families, and their entry salary is higher), but they tend to lose some of their initial advantage as time goes by.

4 Income Dynamics from Gibrat's Perspective

Let us now apply our modified Gibrat's model to the same data set. Our starting values are the estimates of the coefficients obtained in the previous regression (Table 1, column 2 and 3), and the confidence intervals for the coefficients have been estimated through bootstrapping (500 repetitions). The results of our analysis are summarized in Table 2, where the α coefficients represent the effect produced by the covariates on the mean increment, and, apart from the intercept, are directly comparable with those estimated in the regression of growth-rates. The β coefficients represent the effects of the covariates on the variance of these rates of growth.

Table 2: Analysis of income rates of growth through the modified Gibrat's model

<i>Coeff.</i>	<i>Variation after 2 years</i>		<i>Variation after 4 years</i>	
	<i>Estim. α</i>	<i>Estim. β</i>	<i>Estim. α</i>	<i>Estim. β</i>
Intercept	-1.498 #	-0.599 #	-0.934 #	-0.266 #
Children 0	0.114 #	-0.010	-0.260 #	-0.002
Children 2	-0.105	0.028	-0.067	-0.150 #
Children 3	0.025	0.075	0.081	-0.058
Age (Stand.)	-0.083 #	-0.009	0.059	0.013
Education 1	-0.063	0.233 #	0.150	0.357 #
Education 3	-0.249 #	-0.121 #	-0.155 #	-0.089
Education 4	-0.232 #	-0.057	-1.160 #	-0.597 #
Self Empl.	0.203 #	0.295 #	0.201 #	0.289 #
Unempl.	0.518 #	0.347 #	0.568 #	0.486 #
Other	0.047	0.023	0.294 #	0.080
Period 2	-0.071	-0.306 #	0.045	-0.216 #
Period 3	0.011	-0.111 #	0.004	-0.214 #
Period 4	0.150 #	-0.121	\	\
Center It.	-0.011	-0.077	0.093	0.091 #
South It.	0.118	0.196 #	0.142 #	0.050
R²	0.02		0.05	

significant at 10%.

R² calculated with the formula: $\text{var}(\text{fitted.growth-rates})/\text{var}(\text{actual.growth-rates})$.

The α parameters of Table 2 tell basically the same story we already saw in Table 1: variations in income are, on average, more positive for the unemployed and for the self-employed; the evolution is relatively worse for the well educated, and the effect of age is not so clear. Notice, however, that the cases when the α parameters are significant are now considerably more numerous and that the variance explained by the model, although still low, is twice as high as that explained by the regressions of table 1. For example, self-employment exerts a significant positive effect in all the cases of

Table 2, while this effect is weak or insignificant in the regression analysis of growth rates of Table 1. We advance two tentative explanations for this improvement: 1) our modified Gibrat's model corrects the bias produced by what we referred to as the "differential dynamics of income" (small incomes increase faster and with a greater variability than others); 2) an optimization process based on the likelihood is less influenced by the presence of outliers.

A distinctive feature of our modified Gibrat's model is that it allows us to measure the effects produced by household's characteristics also on the variability of income, and not only on its average. The results are rather clear: low education, unemployment and self-employment all contribute to an increase in the variability of income variations in the subsequent period (of either 2 or 4 years). Note that periods 2 and 3 of both analyses (2-years and 4-years increments) are characterized by a significant lower variability than the baseline period (period 1). This may depend on the macroeconomic situation: the period 2002-2005 is, in fact, for Italy a period of economic imbalance with a low mean annual growth rate of the GDP (around 0.6%). And in times of economic recession, the variability of income variations typically shrinks.

5 Conclusions

Our modified Gibrat's model seems to describe income dynamics better than other models: it eliminates the bias due to the more rapid increase of smaller incomes and it can better take into account the (after all, non-negligible) effects produced by a few selected covariates. The coefficients produced by this model can be used for the estimation of the transition probabilities between income levels in a markovian framework. This information may be used to forecast the long-term evolution of incomes of any particular sub-group: among other things, this permits researchers to model and quantify the risk and the mean duration of poverty spells in each sub-group.

References

1. Fields, G.S., Duval Hernández, R.D., Freije Rodríguez, S.F., Sánchez Puerta, M.L.: Income Mobility in Latin America, mimeo, (2006)
<http://www.cid.harvard.edu/Economia/Mexico06%20Files/fieldsetal%20102306.pdf>
2. Gibrat R.: Les inégalités économiques: applications aux inégalité des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques de familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel, Librairie du Recueil Sirey, Paris (1931)
3. Salinari, G., De Santis, G.: On the Evolution of Household Income, Luxemb. Income Study Work. Pap. Ser., No. 488, July (2008)
4. Salinari, G., De Santis, G.: The Evolution and Distribution of Income in a Life-Cycle Perspective, Riv. Ital. di Econ., Demogr. e Stat., 63(1-2), pp. 257-274, (2010).