Fiscal federalism, patient mobility and the soft budget constraint: a theoretical approach.

Rosella Levaggi* and Francesco Menoncin
Dipartimento di Scienze Economiche
Università degli Studi di Brescia
Via S. Faustino 74b
25122 BRESCIA

e-mail of corresponding author: levaggi@eco.unibs.it

October 3, 2007

Abstract

In some countries the reform of public health care provision has been accompanied by a parallel process of devolution that has also entailed the organisation of health care becoming a regional competence. However, the application of fiscal federalism in the context of the provision of health care is not so straightforward due to the nature of the services involved. In this paper we will concentrate on the related phenomenon of the soft budget constraint phenomenon. This framework can be traced back to that of a game where less efficient local authorities prefer to send their citizens to receive services outside their region instead of becoming more efficient. In order to improve the probability of being bailed out, the users are sent to other local authorities where there is excess capacity. The lack of coordination between local objectives and total welfare means that this policy is optimal at local level, but inefficient at Central Government level.

Kewwords: Soft budget constraint, health care provision, patients mobility

J.E.L. I18,H77

1 Introduction

Public health care provision has undergone important reforms throughout the past few decades. In some countries the creation of an internal market for health care has been accompanied by a parallel process of devolution that has also entailed the organisation of health care becoming a regional competence. However, the application of fiscal federalism in the context of the provision of health care is not so straightforward as might be expected.

The traditional literature on fiscal federalism is centered on the production of goods and services that have mainly the characteristic of local public goods whose cost may be mainly financed through benefit taxes. Health care [1] is something quite different; from an economic point of view it might be classified as an impure public good or a merit good, i.e. it is rival in consumption, it can be supplied to local residents by providers located outside the boundaries of the local authority, and it may be used as an indirect tool to equalise income distribution. The provision of these goods opens a very interesting debate on fiscal federalism as regards local taxes and grants in aid.[2] In this paper we will concentrate on the related phenomenon of Local Governments systematically running into a deficit (with local expenditure higher than local revenues) which is often related to user mobility between local authorities.

If Central Government tolerates this behaviour and, sooner or later, bails out the local authority, we have a soft budget constraint phenomenon. In this way, fiscal federalism generates perverse effects on economic growth.

This topic has not yet received due attention from the literature which mainly deals with bailing out firms that run into deficit. Levaggi and Zanola (2004) and Bordignon and Turati (2006) show empirical evidence for the soft-budget constraint. In both cases the focus is on expenditure rather than explanation of the soft budget constraint itself. Wildasin (2004) proposes one of the few theoretical models of bailing out at local government level. Wildasin assumes the existence of two neighbour local authorities that produce local public goods. However, the good produced by one local authority (B) spills its beneficial effect also to residents of the other local authority (A). If the willingness to pay for the good is different in the two regions, a possible equilibrium is one where the local authority A provides the local public good in B and finances such provision. In this way the residents of B free ride on use of the local public goods. In other words we can say that local authority A bails out B as regards the provision of that specific local public good.

The model presented in this paper shares some assumptions with Wildasin's but has some differences:

1. The presence of a two-tier government structure. In our model we assume the presence of a Central Government with the function of regulator and

[1] Health care, especially elective care, can be provided outside the area of residence of the patient since the patient can travel, and often does, a long distance to receive the service; the same argument applies to higher education.

[2] See Levaggi (2006).

grants-in-aid distributor. The production of local public goods takes place at local level and it is financed locally and through grants-in-aid. Both tiers have a budget constraint, but while the Central Government's one is binding, Local Governments are allowed to run into a deficit (so- called soft budget constraint). Accordingly, a local deficit implies an increase in central taxes.

2. Local authorities supply health care and differ in their level of productive efficiency. The less efficient region (B) may decide not to locally produce the health care and send its patients to the more efficient region (A) in order to receive the service. In such a way a part of the cost is paid by residents in A through the soft budget constraint.

This framework can be traced back to that of a game where less efficient local authorities prefer to send their citizens to receive services outside their region instead of becoming more efficient. This policy is supported by the more efficient region which, due to the shape of its utility function, prefers to produce more goods than are locally needed. The lack of coordination between local objectives and total welfare means that this policy is optimal at local level, but inefficient at Central Government level. The outcome of such game is a welfare loss. There are in fact two clear losers:

a) the whole community, which would be better-off if hard budget constraint rules were imposed;

b) the users of the services in the regions where soft budget constraint is widespread who have to travel and incur private costs.

The paper is organised as follows: in Section 2 we present the peculiar characteristics of the soft budget constraint in Italy, in Section 3 the model is presented, Section 4 discusses the results and Section 5 draws the conclusions.

2 Soft budget constraint in Italy

Health care in Italy represents a very good example of fiscal federalism applied in a problematic context. Income is in fact unevenly distributed across regions, hence the need to finance a great proportion of expenditure through grants-in-aid. In Italy the regionalisation of health care expenditure is characterized by two important facts which are presented in table one:

- some regions persistently run into deficit;
- patient mobility across regions is fairly well-developed.

Table 1 about here

The data proposed in table 1 refer to 2005 and show that most of the regions with deficits also have a negative balance as regards health care; the only

important exception is Lazio, but the behaviour of this region has probably been dominated by other factors. Levaggi and Ruocco (2007) show some other interesting facts: the regions that run into deficit heavily rely on grants from Central Government to finance health care (about 75% against a national average of 55%) because their tax base is fairly low compared with the other regions. The mobility of patients across regions follows quite interesting patterns: in some cases (Trento, Bolzano, Valle d'Aosta, Abruzzo, Molise, Umbria) it depends on economies of scale. Small regions are not able to produce all the services locally and prefer to specialise in a few of them; through mobility they are then able to supply health care to all their population. The mobility pattern is concentrated among adjacent regions and it is usually two-sided. This behaviour can be consistently explained by the theory of fiscal federalism: it corresponds to spillovers in the production of health services. However, only a limited part of the mobility flow can be explained by this mechanism: most of it is represented by a one-sided flow of patients from southern regions (running into a deficit) to northern regions. In the next sections we propose a model that explains this evidence.

3 The model

In this article we present a very simple model. A county is divided into two local authorities, A and B, which should provide a fixed quantity S of health care to their population. In Italy we can think of this level as the bundle of services that make up LEA (Livelli Essenziali di Assistenza[3] - Essential Assistence Levels). Such provision is financed through a linear income tax which is partly levied at national level (τ_c) and partly at local level at rate τ_i, $i \in \{A, B\}$. To simplify matters we assume that the two authorities are identical but?? in their income and the cost to produce health care. Local authority A is richer and more efficient so that the same level of fiscal effort produces more services in A than in B. When resources are unevenly distributed and Central Government wishes to pursue horizontal equity, a grant should be used to balance resources distribution. The literature (King, 1994) has long debated whether resources or expenditure should be equalised. Here, we use a mixed model where a social planner maximises the trade-off between utility of local residents and public costs, by taking account of the fact that health care should not necessarily be fully supplied in both local authorities. This allows a trade-off to be developed between general taxation and local welfare which will then be used to explain why decentralised decisions can be different from those of a benevolent social planner. The devolution of the decision of how much health care to provide locally, in fact, means that each region might not choose the provision which would be optimal at central level. In this context we show that the combined

[3] LEA were made operative in Italy after the enhancement of fiscal federalism for health care. They define a package of minimum services that each citizen in Italy is entitled to receive independently of where *he/she lives*. Central Gover*n*ment assures that each Region has the financial resources to provide such level of services through a system of lump-sum grants.

interests of the two local authorities might produce a soft budget constraint equilibrium.

4 Central Government decision

In this section we present the optimal solution for an economy where Central Government is the only actor and may decide how to allocate the production of health care between the two regions. In a context of symmetric information on local preferences and resources this solution represents the first best for that economy. The theory for fiscal federalism (King, 1994, Tresh, 2002 and Levaggi, 1991) has long argued that such solution cannot be replicated in the actual world because the Central Government is not able to observe all the relevant parameters. In this model we assume symmetrical information in order to isolate the effect of decentralised decisions and soft budget constraint. This solution also represents the first step of the decision process for a decentralised economy. At this stage, in fact, grants and optimal tax levels are set.

To explain the working of fiscal federalism, resources equalisation and local autonomy we divide the process into two separate stages: Central Government sets the level of production in A and B, and the total tax rate $(\tau_c + \tau_i)$ in order to maximise total welfare. In the second stage, Central Government sets all the tax rates and grants in order to equalise resources. If local authorities choose not to change any of these parameters, the fiscal federalism solution would be equivalent to the centralised one. However, this is seldom the case because of the asymmetry in the objective functions of Central and Local Governments. While the former maximise the sum of all local authorities welfare, the latter maximise their own welfares and these objectives do not necesarily coincide.

The two regions are equal and this means that the need for health care is equal to $\frac{S}{2}$. Income is fixed and it is equal to: Y_A and Y_B with $Y_A > Y_B$. The marginal costs to produce health are equal to v_A and v_B with $v_A < v_B$, i.e. local authority A is more efficient than B in producing health care. Individuals derive utility from their net income and the production of health care in their region. The utility is linear in yield (Y_i) and quadratic in the local health care production (S_i) and, for the region i, is

$$U_i = Y_i(1 - \tau_c - \tau_i) + z_i S_i - \frac{1}{2}{S_i}^2,$$

where τ_c and τ_i are the central and the local taxes, respectively, and z_i is a preference parameter coinciding with the level of S_i maximising the utility. We will assume that $z_B = \frac{S_B}{2}$ and $z_A > \frac{S_A}{2}$ i.e. the less efficient region prefers to set its production in such a way to supply health care to all its residents whereas the more efficient region (A) prefers to attract patients from outside. This assumption can be justified on several grounds: it may depend on a reputational effect which attracts better physicians, it may depend on the existing structure of production which may be set to provide care for a number of people greater than the local needs.

Central Government is assumed to be a benevolent regulator which should set the national tax level (τ_c) and the grant (G) so that total welfare is maximised and resources are equalised. Equalisation of resources means that for a specific equilibrium (S_A, S_B) the total tax rate should be uniform and set to $t = \tau_c + \tau_i$ which implies $\tau_A = \tau_B = \tau_l$.

The problem faced by the Central Government can be written as

$$\max_{S_A,S_B,t} \left\{ Y_A(1-t) + z_A S_A - \frac{1}{2}{S_A}^2 + Y_B(1-t) + z_B S_B - \frac{1}{2}{S_B}^2 \right\}, \quad (1)$$

$$\text{s.t.}$$

$$v_A S_A + v_B S_B = t\left(Y_A + Y_B\right),$$

$$S_A + S_B = S.$$

The two constraints imply that the total supply $(S_A + S_B)$ must equate the total demand (S) for health care. Furthermore, the Central Government revenue $(tY_A + tY_B)$ must equate the total expenditures for producing the health care $(v_A S_A + v_B S_B)$.

The only solution to Problem (1) is

$$S_A^C = \frac{S}{2} + \frac{z_A - z_B}{2} + \frac{v_B - v_A}{2}, \quad (2)$$

$$S_B^C = \frac{S}{2} - \left(\frac{z_A - z_B}{2} + \frac{v_B - v_A}{2} \right), \quad (3)$$

$$t = \frac{1}{2} \frac{S\left(v_A + v_B\right) + \left(v_A - v_B\right)\left(z_A - z_B + v_B - v_A\right)}{Y_A + Y_B}, \quad (4)$$

where the upperscript C stands for «centralized» solution. Since $z_A > z_B$ and $v_A < v_B$ then $S_A^C > S_B^C$ and, in particular, $S_A^C > \frac{S}{2}$ and $S_B^C < \frac{S}{2}$. Under the hypothesis that the health care demand is equally split between the two regions, this implies that there is a flow of patients going from region B to region A. This flow is measured by

$$S_I^C \equiv S_A^C - S_B^C = z_A - z_B + v_B - v_A. \quad (5)$$

The solution presented above shows that in allocating production, the Central Government takes into account local preferences and costs.

4.1 Grants central government tax and preferred local additional taxes

If a Central Government wishes to delegate some activities to local authorities, then two main changes must be made in the previous problem (1):

1. the total tax rate t must be split into two taxes: a central one (τ_c) and a local one (τ_l) which is imposed in order to achieve equalisation of the tax bases; accordingly, we must substitute t by $\tau_c + \tau_l$;

2. the first constraint in problem (1) becomes

$$v_A S_A^C + v_B S_B^C + G_A + G_B = t\left(Y_A + Y_B\right),$$

where G_A and G_B are grants allocated to the local authorities by the central government. This means that the tax rate t will be given by (4) plus $\frac{G_A+G_B}{Y_A+Y_B}$.

Now, we assume the Central Government will set τ_c, τ_l, G_A, and G_B in order to achieve the following goals:

1. the local tax levied in A together with the grant must finance the production of $\frac{S}{2}$:

$$\tau_l Y_A + G_A = v_A \frac{S}{2},$$

2. the local tax levied in B together with the grant must finance all the production in B and the production in A not already financed by the local taxes and grant in A:

$$\tau_l Y_B + G_B = v_B S_B^C + v_A \left(S_A^C - \frac{S}{2}\right),$$

3. the central tax must finance all the grants

$$\tau_c \left(Y_A + Y_B\right) = G_B + G_A.$$

If we put these three constraints together, we obviously have infinite solutions for τ_c, τ_l, G_A, and G_B. The choice will then mainly depend on political reasons. In what follows we choose the solution that allows us to minimize the algebraic complexity of the formulas and set $G_A = 0$. Grants and optimal tax rates can be summarized as follows:

$$\begin{aligned}
\tau_l &= \frac{1}{Y_A}\frac{S}{2}v_A, \\
\tau_c &= \frac{v_B S_B^C + v_A S_A^C}{Y_A + Y_B} - \frac{1}{Y_A}\frac{S}{2}v_A, \\
G_B &= v_B S_B^C + v_A S_A^C - \frac{Y_B + Y_A}{Y_A}\frac{S}{2}v_A.
\end{aligned}$$

5 Decentralised decision

Once Central Government has set national taxes and the grant, each local authority decides the level of taxation and local production. The local authority can perceive its budget as hard or as soft, i.e. it can have different expectations about Central Government bail out decisions. We will examine both settings by starting from a decentralised model where both local authorities have to be budget-balanced and we will then move to a setting where only one of them has to strictly respect its budget constraint. In this environment we will study the opportunity for the two local authorities to bargain on the price for the fraction of health care provided using mobility.

5.1 Hard budget constraint

The optimal sharing of production that has been outlined in Section 4 may not be the outcome of a process of decentralisation at local government level, even in a setting where there is symmetric information between Central Government and local authorities. This usually happens because the local authority does not fully take into account the consequences of its actions on welfare (Petretto, 1999). This behaviour usually leads to a sub-optimal solution; in this section we show the classical problem arising from lack of coordination: each local authority maximises its own utility function whereas in the centralized model the total welfare is maximized. Each local authority takes the national tax τ_c and the grant G_i as given. The decision is somehow asymmetric for the two local authorities given that each of them has a different variable cost to produce health care.

5.1.1 Local authority B

The authority with higher marginal cost (B) does not like ($z_B = S/2$) to attract patients from the other local authority. In fact, providing health care to a number of patients higher than $S/2$ would be very expensive because of B's high marginal cost. Its decision will then relate to the split between local production (S_B^D) and mobility ($\frac{S}{2} - S_B^D$). Such a decision depends on B's preference and on the price p_A charged by A for health care. Some health care systems (Italy among them) have set p_A at national level, but it might become a strategic variable in the game between local authorities. For this reason, in our model we use it as a strategic variable.

In this way the game between the two local authorities is as follows: B sets the number of cases to be sent to A as a function of p_A and A accordingly sets p_A in order to maximise its utility. In this game A plays the role of a Stackelberg leader given its advantage in terms of income and productivity.

The problem faced by the local authority B can be written as

$$\max_{S_B, \tau_B} \left\{ Y_B(1 - \tau_c - \tau_B) + \frac{S}{2} S_B - \frac{1}{2} {S_B}^2 \right\},$$

$$\text{s.t.}$$

$$v_B S_B + p_A \left(\frac{S}{2} - S_B \right) = \tau_B Y_B + G_B.$$

where τ_B is the tax on local income which can be used to finance a different split between local provision and mobility than the one foreseen by Central Government.

The optimal local tax is

$$\tau_B = \frac{v_B \frac{S}{2} - (v_B - p_A)^2 - G_B}{Y_B}.$$

The optimal amount of patients to treat (let's call it S_B^D) is given by (see Appendix 2)

$$S_B^D = \frac{S}{2} - (v_B - p_A).$$

which in general is different from the production Central Government optimally set (3). The difference arises from the fact that when B maximises its utility, it does not take into account the positive effect that mobility has on its residents; the second important difference is that S_B^D depends on the price charged by A. In a decentralised system the latter may not necessarily set a price equal to the marginal cost.

5.1.2 Local authority A

The local authority A would like to attract patients ($z_A > S/2$) from the nearby region; to do so it may set a price (p_A) it makes attractive for B to send its patients to A (which is more efficient in producing health care). This competitive advantage puts A in a Stackelberg leader position since by setting p_A it can make B choose its preferred level of mobility.

Afer solving the decentralized problem for B, we have obtained S_B^D which is lower than $S/2$. This means that A must satisfy the health care demand for a number of patients given by its residents ($S/2$) plus the patients not treated in B (i.e. $S/2 - S_B^D$). Since we have

$$\frac{S}{2} - S_B^D = v_B - p_A,$$

the optimization problem for A can be written as

$$\max_{p_A,\tau_A} \left\{ Y_A (1 - \tau_c - \tau_A) + z_A \left(\frac{S}{2} + v_B - p_A \right) - \frac{1}{2} \left(\frac{S}{2} + v_B - p_A \right)^2 \right\},$$

s.t.

$$v_A \frac{S}{2} + (v_A - p_A)(v_B - p_A) = \tau_A Y_A.$$

which is maximised with respect to the price charged to residents in B and the local level of taxation τ_A.

The optimal price is given by (see Appendix 3)

$$\begin{aligned} p_A^* &= \frac{1}{3} \left(v_A + \frac{S}{2} - z_A + 2v_B \right), \\ \tau_A &= \tau_l + \frac{(v_A - p_A^*)(v_B - p_A^*)}{Y_A} \end{aligned}$$

which, in principle, can be higher or lower than v_A. In fact,

$$p_A^* < v_A \Leftrightarrow v_B - v_A < \frac{1}{2} \left(z_A - \frac{S}{2} \right)$$

i.e. the price will be set lower than the marginal cost if the difference in productivity (measured by $v_B - v_A$) is lower than half the marginal utility of increasing the production of health care in A.

The amount of patients who travel from B to A in order to receive treatment is given by

$$\begin{aligned} S_I^D &\equiv S_A^D - S_B^D = \left(\frac{S}{2} + v_B - p_A\right) - \left(\frac{S}{2} - (v_B - p_A)\right) = 2\,(v_B - p_A) \\ &= \frac{2}{3}\left(v_B - v_A + z_A - \frac{S}{2}\right) \end{aligned} \tag{6}$$

The comparison between the number of patients moving in the centralized and in the decentralized models ((5) and (6) respectively) allow us to conclude that (recall that $v_B = \frac{S}{2}$)

$$S_I^D \equiv \frac{2}{3}\left(v_B - v_A + z_A - \frac{S}{2}\right) < S_I^C \equiv v_B - v_A + z_A - \frac{S}{2},$$

i.e. the number of patients treated outside B is always lower than in the first best equilibrium. This produces an increase in the utility function of residents in B, but it certainly reduces that of residents in A who are the net losers of this allocation since they suffer an increase in their local tax rate and a reduction in the number of admissions to their hospitals.

5.2 Soft budget constraint

In the previous paragraph we have shown why the parameters chosen by Central Government may become sub-optimal in the context of decentralisation. Another and even more perverse effect may derive from the interpretation of the budget constraint by local authorities as soft.

In Italy there are two main methods to bail out regions that run into deficit. The first one, which could be defined as a direct method, consists of offering the Regions that have a persistent budget deficit to wave half of it in exchange for the agreement to cover the other half with local resources. Since this process is repeated through time, the Central Government pays for the whole deficit. The second method is even more subtle: the allocation of the grant to local authorities for a certain year is provisional; it may increase in the future if the Region has a deficit. This means that ex post the deficit of a Region may appear to be lower than it actually is. The rationale for this policy is that Central Government recognizes that the process of allocating resources may be flawed since preferences and local needs are difficult to estimate, but it creates a ratchet effect mechanism that is very dangerous.

The model assumes a bargaining process between the two local authorities. Each of them maximises its own utility function separately in a sort of Nash equilibrium framework. The basic idea behind the model we are going to present is that an agreement exists between the two local authorities. A, the richest and most efficient, would like to increase its production beyond its local need

$S/2$. To do so, it tries to attract patients from B in exchange for a reduction in the cost to provide health care. A possible solution is the one presented in the previous paragraph, but such equilibrium is unsatisfactory for several reasons: Central Government may impose a price on interregional mobility, and the local authority A may find it difficult to increase τ_l from a political point of view. The second possibility is to induce Central Government to supply B with extra resources provided they are used to finance mobility from B to A. The mechanism of the soft budget constraint in our model responds to this objective. The rules of the game are such that if both local authorities agree that the deficit should be repaid, Central Government does so by increasing the national tax level t. A agrees to pay for the debt of B only if it is supported by mobility of patients from B to A.

In this way, while the price for mobility from B to A is set to its marginal cost v_A, B finances only a fraction p_A and runs into a deficit. The actual level of the deficit which B incurs is determined by A by setting the number of patients that A is willing to receive from B (S_I^D).

Both local authorities are apparently better-off than in the previous decentralised solutions:

- local authority A can share the cost of offering health care at a price lower than its marginal cost with the residents in B. In the presence of a soft budget constraint, in fact, the mobility from B to A is reimbursed at its marginal cost so that A does not need to increase its local tax rate. The deficit will be paid through an increase of the national tax rate which is paid by both local authority residents. In this way A also has a political advantage: the burden in terms of popularity of increasing the fiscal pressure is borne by Central Government;
- local authority B can decrease local taxation and can bargain with A how much of its health care cost to shift. This policy allows the local authority to offer a sort of compensation to the residents that need to incur an extra cost to receive health care outside B.

In this context, although local authority A has more income and is more productive, it cannot act as a Stackelberg leader because both local authorities need to agree on a common strategy. For this reason in this case the first local authority decides.

5.2.1 Local authority B

The local authority B decides how much of the total cost of health care provided by A it will finance with local taxes and to some extent it will also fix the deficit it is prepared to incur. If we define r as the fraction of the price p_A B is prepared to finance with local taxes, the budget constraint for local authority B can be written as:

$$v_B\left(\frac{S}{2}-S_I^D\right)+S_I^D v_A = Y_B\tau_B + G_B + S_I^D r v_A.$$

The local authority foresees that its deficit $S_I^D(1-r)p_A$ will have to be met by an increase in the national tax rate, i.e. Central Government will have to set

$$\tau_c = \frac{S_I^D(1-r)v_A + G_B}{Y_A + Y_B}.$$

The problem for the local authority can be written as:

$$\max_{r,\tau_B} \left\{ Y_B \left(1 - \frac{S_I^D(1-r)p_A + G_B}{Y_A + Y_B} - \tau_B \right) + z_B \left(\frac{S}{2} - S_I^D \right) - \frac{1}{2} \left(\frac{S}{2} - S_I^D \right)^2 \right\}$$

s.t.

$$v_B \left(\frac{S}{2} - S_I^D \right) + S_I^D r v_A - G_B = Y_B \tau_B$$

and the above constraint. The solution, presented in appendix four, can be written as:

$$\begin{aligned} r &= 0 \\ \tau_B &= \frac{v_B \left(\frac{S}{2} - S_I^D \right) - G_B}{Y_B}. \end{aligned}$$

5.2.2 Local authority A

By the agreement we have described above, A knows that the patients that it receives will be financed using the soft budget constraint and that the tax rate set by Central Government will increase. Given that only the deficit arising from mobility will be repaid, it can decide the amount of deficit B can incur by setting S_I^D. As before, it knows that the policy of soft budget constraint will change the tax rate, i.e.

$$\tau_c = \frac{S_I^D(1-r)v_A + G_B}{Y_A + Y_B}.$$

In this context we also assume that the price A can charge to B is equal to the marginal cost of production, v_A. This level allows A to set its tax rate to the level τ_l.[4]

The problem for local authority A can be written as:

$$\max_{S_I^D} \left\{ Y_A \left(1 - \frac{S_I^D(1-r)v_A + G_B}{Y_A + Y_B} - \frac{v_A \frac{S}{2}}{Y_A} \right) + z_A \left(\frac{S}{2} + S_I^D \right) - \frac{1}{2} \left(\frac{S}{2} + S_I^D \right)^2 \right\}$$

whose solution is

$$S_I^D = \left(z_A - \frac{S}{2} \right) - \frac{Y_A}{Y_A + Y_B} v_A (1-r)$$

[4] This policy can be implemented in several ways. As noted before, Central Government may set a price lists for health care delivered through interregional mobility. An indirect way to obtain the same result would be to fix a minimum tax level τ. The reason for this second policy will be clear in what follows.

5.2.3 Equilibrium

The equilibrium with the soft budget constraint is identified by finding r and S_I^*.This can be done by substituting in the equation for r and S_I^D the optimal values of the reaction functions of the other local authority. In this case, the solution is recursive since r does not depend on S_I^D.

The solution can be written as:

$$\begin{gathered} r = 0 \\ S_I^* = \left(z_A - \frac{1}{2}S\right) - \frac{Y_A}{Y_A+Y_B}v_A \gtrless S_I^C = \frac{1}{2}\left[\left(z_A - \frac{S}{2}\right) + (v_B - v_A)\right] \\ \tau_B^* = \frac{v_B\left(\frac{S}{2} - S_I^D\right) - G_B}{Y_B} \\ \tau_A = \tau_l \\ \tau_c^* = \tau_c + \frac{v_A}{Y_A+Y_B}\left(z_A - \frac{1}{2}S - Y_A\frac{v_A}{Y_A+Y_B}\right) \end{gathered}$$

In the final solution, local authority A is budget balanced and is allowed to produce more than under a decentralised solution with a soft budget constraint. This solution is then preferred by this local authority as it increases production and reduces the tax rate. Local authority B reduces its level of taxation and its internal production, but it is not necessarily better off from a strictly welfare point of view. The actual comparison is rather cumbersome from an algebraic point of view; but what is important for our discussion is that the soft budget constraint policy does not necessarily imply a benefit for the region that incurs it.

6 Discussion

The model presented shows that fiscal federalism may not always be beneficial as the traditional literature shows. The main problem is that at local government level a coordination problem among policies implemented at this level may exist. Each local authority maximises its utility, but does not take account of the effects on the welfare of the other Regions and, in the end, the solution is welfare decreasing. What is even more interesting is that sharing of the benefits of the policy may not be as expected; in particular it is not necessarily the Region that incurs a deficit that benefits most from the soft budget constraint policy.

To show this, we ran a simulation based on our model. We assumed that income in region $A$is 50% higher than in Region B while cost is 30% higher in the latter. Expenditure for health care is about 7% of total GDP. The inital parameters are as follows:

Parameters	Y_A	Y_B	v_A	v_B	S	z_A	z_B
Value	150	100	1	1.3	20	13	10

and the results are summarised in Table 1.

The centralised solution may be used as benchmark and shows that a coordinated policy would maximise total welfare, but it does not maximise the

Table 1: Simulation results

Parameters	**Centralised**	**HBC-1**	**HBC-2**	**SBC**
p_A	-	1	0.2	1
r	-	-	-	0
S_A^C	11.65	10.3	11.1	12.4
S_B^C	8.35	9.7	8.9	7.6
t	0.090	-	-	-
τ_c	0.066	0.023	0.023	0.032
τ_A	0.023	0.066	0.072	0.066
τ_B	0.023	0.070	0.059	0.040
G_B	5.838	5.838	5.838	5.838
U_A	220.08	217.35	218.31	219.37
U_B	139.63	140.54	141.10	139.73
U_T	359.72	357.90	359.42	359.16

utility of people living in B. This opens a space for three types of decentralised solutions depending on whether the two local authorities have a form of bargain in reaching a better solution. When the budget constraint is taken as hard, the HBC-1 solution represents the case in which each local authority maximises its utility and takes the behaviour of the other local authority as given. In this case, A suffers a loss of utility because it receives a small number of patients from B. This solution can be improved upon by allowing A to make B pay a price lower than the marginal cost to produce health care (HBC-2). In the presence of a soft budget constraint (SBC), when local authorities can run into a deficit, A is better off than in the previous simulation, but it is not B that suffers a reduction in its utility. This result is quite interesting. B prefers the soft budget constraint to the centralised solution, but if it could make its policy to respect its budget constraint credible, it would be able to attain a higher level of welfare.

This conclusion is in line with the literature on decentralisation in health care (Petretto, 2000). In our model we add another dimension represented by the presence of a soft budget constraint. Such a policy is usually associated in the political discussion with the deviating behaviour of some local authorities that autonomously decide to spend more than what they should. Given the nature of health care, Central Government has to bail them out and has to make the virtuous Region pay for it.

In our model we have shown that the process may be rather different. Soft budget constraint behaviour arises from a bargaining solution between the Regions in which they anticipate that the deficit will have to be covered at central level.

The presence of a soft budget constraint is welfare decreasing for the community as a whole, but it shares the benefits between the two local authorities in ways that have not been explored so far. The real winner is in fact A, the

local authority that respects its budget and that appears to be the virtuous one.

7 Conclusions

The traditional literature on fiscal federalism suggests that the allocation of functions to local governments should follow efficiency principles. The choice of the quantity to be produced should be left to the lower tier which knows local preferences better than the centre. Nevertheless, in order to finance the local provision, grants might be used for equity and efficiency reasons, and the presence of grants balancing for uneven distribution in both resources and needs is one of the main arguments in favour of the application of strict budget balance rules. This means that any expenditure in excess of that financed by Central Government should be paid by local taxes. However, the traditional literature on fiscal federalism is centered on the production of goods and services that have mainly the characteristic of local public goods. In the recent past the process of fiscal federalism has extended the category of goods and services to be provided at local level to include also services that are both impure public goods and merit goods, i.e. they are rival in consumption and can be supplied to local residents also by providers located outside the boundaries of the local authority.

The use of fiscal federalism in this case is quite problematic for the regulator which is left with very few options for making local government replicate the first best solution. This is a well-known result; what our model adds to the previous literature is the effect of soft budget constraint on welfare. In this article we have shown that soft budget constraint policies are quite welfare worsening in this context and that the true loser is not necessarily the local authority that respects its budget. The use of a soft budget constraint along with passive mobility in fact reduces total welfare of the population that has to move and usually this aspect is not sufficiently taken into account by the decision makers. The solution to this problem is not easy, but it seem that this policy should be avoided as far as possible.

References

[1] Bordignon, M. and G. Turati (2006) Bailing Out Expectations and Health Expenditure in Italy: an empirical approach, http://www4.unicatt.it/Docenti/Bordignon/download/Bailing%20out%20Expectations.pdf

[2] Desai, R. M. and A. Olofsgård (2005) "A Political Model of the Soft Budget Constraint,"European Journal of Political Economy, forthcoming.

[3] Levaggi, R. (2006) Trasferimenti ai governi locali, beni pubblici impuri e strategie di rielezione, in Fossati (ed) Analisi di politiche economiche e fiscali per lo sviluppo, Franco Angeli

[4] Grazzini, L. and A. Petretto (2004) Coordinamento e concorrenza fiscale nei processi di federalismo. Una rassegna teorica e alcune considerazioni sul caso italianoFirenze, http://maceprha.unipv.it/websiep/default.asp

[5] Grazzini, L. and A. Petretto (2005) Regiona fiscal effort with revenue sharing and equalisation, http://maceprha.unipv.it/websiep/default.asp

[6] King, D., (1984), Fiscal tiers: the economics of multi-level Government, Allen & Unwin, London

[7] Maskin, E and C. Xu (2001) "Soft Budget Constraint Theories,"Economics of Transition, 9, 1-27.

[8] Kornai J., (1986), The Soft Budget Constraint, Kyklos, 39(1), 3-30.

[9] Kornai, J., E. Maskin and G. Roland (2003) "Understanding the Soft Budget Constraint,"Journal of Economic Literature, XLI, 1095-1136.

[10] Levaggi, R. (1991)Fiscal Federalism, and grants-in-aid: The problem of asymmetrical information, Avebury, Gower, Aldershot

[11] Levaggi, R. and R. Zanola (2003) Flypaper Effect and Sluggishness: Evidence from Regional Health Expenditure in Italy, International Tax and Public Finance,

[12] Levaggi, R. and A. Ruocco (2007) Regole e concorrenza nella sanità, *mimeo*

[13] Oates, W. (1972) Fiscal Federalism, Harcourt Brace Janovich, New York

[14] Oates, W. (2005) Towards a second generation Theory of Fiscal Federalism, International Tax and Public Finance, 12, 349-374

[15] Rodden, J. (2000), Breaking the Golden Rule: Fiscal Behaviour with Rational Bailout Expectation in German States, Prepared for the Workshop: European Fiscal Federalism in Comparative Perspective Center for European Studies, Harvard University

[16] Petretto, A. (2002) The impact of vertical fiscal competition on the tax structure of a federation with equalising grant, http://maceprha.unipv.it/websiep/default.asp

[17] Petretto, A. (2000) On the cost and benefit of the regionalisation of the NHS, Economics of Governance, 1, 213-232

[18] Robinson, J. and R. Torvic (2006) A political economy theory of the soft budget constraint, NBERWorking Paper 12133, http://www.nber.org/papers/w12133

[19] Tresh, R (2002) Public Finance: A normative view, 2nd Edition, Academic Press, San Diego

[20] Wildasin , D.E. (2004) The institutions of federalism: towards an analytical framework, National Tax Journal, LVII, 247-272

A Appendix one: Solution of the Central Government problem

The Lagrangean can be written as:
$L = Y_A(1-t) + z_A S_A - \frac{1}{2}{S_A}^2 + Y_B(1-t) + z_B S_B - \frac{1}{2}{S_B}^2$
$-\lambda_1 \left(v_A S_A + v_B S_B - t(Y_A + Y_B)\right)$
$-\lambda_2 \left(S_A + S_B - S\right)$

The F.O.C.can be written as:

$\frac{\partial}{\partial t} = -Y_A - Y_B + \lambda_1 Y_A + \lambda_1 Y_B = 0 \quad \lambda_1 = 1$

$\frac{\partial}{\partial S_A} = z_A - S_A - \lambda_2 - v_A \lambda_1 = 0$

$\frac{\partial}{\partial S_B} = z_B - S_B - \lambda_2 - v_B \lambda_1 = 0$
which can be solved to give the solutions presented in the text.

B Appendix two: Hard budget constraint - B's decision

The problem is given by:

$$\max_{S_B, \tau_B} \left\{ Y_B(1 - \tau_c - \tau_B) + \frac{S}{2} S_B - \frac{1}{2}{S_B}^2 \right\},$$

s.t.

$$v_B S_B + p_A \left(\frac{S}{2} - S_B \right) = \tau_B Y_B + G_B.$$

The Lagrangean can be written as:
$L = Y_B(1-t-\tau_B) + z_B S_B - \frac{1}{2}(S_B)^2 - \lambda \left(v_B S_B^D + v_A(\frac{S}{2} - S_B) - \tau_B Y_B + G_B\right)$
$\frac{\partial}{\partial \tau_B} = \lambda Y_B - Y_B = 0 \quad \lambda_1 = 1$

$\frac{\partial}{\partial S_B} = z_B - S_B + \lambda v_A - \lambda v_B = 0$
Substituting the first constraint in the second it is possible to write

$$S_B^D = \frac{S}{2} - \left(v_B - p_A\right).$$

and to obtain τ_B from the budget constraint.

C Appendix three: Hard budget constraint - A's decision

$$\max_{p_A,\tau_A}\left\{Y_A\left(1-\tau_c-\tau_A\right)+z_A\left(\frac{S}{2}+v_B-p_A\right)-\frac{1}{2}\left(\frac{S}{2}+v_B-p_A\right)^2\right\},$$

s.t.

$$v_A\frac{S}{2}+\left(v_A-p_A\right)\left(v_B-p_A\right)=\tau_A Y_A.$$

The Lagrangean can be written as:
$L = Y_A(1-t-\tau_A)+z_A(\frac{S}{2}+v_B-p_A)-\frac{1}{2}(\frac{S}{2}+v_B-p_A)^2$
$\quad - \ \lambda\left(v_A\frac{S}{2}+\left(v_A-p_A\right)\left(v_B-p_A\right)-\tau_A Y_A\right)$
The F.O.C can be written as:
$\frac{\partial}{\partial p_A} = -z_A+\frac{1}{2}S+v_B-p_A+\lambda v_B-2\lambda p_A+\lambda v_A=0$
$\frac{\partial}{\partial \tau_A} = -Y_A+\lambda Y_A \qquad \lambda=1$
Substituting the second equation in the first one we can write:
$p_A^* = \frac{1}{3}(v_A+\frac{S}{2}-z_A+2v_B)$
From the budget constraint we can then write
$\tau_A = \tau_l+\frac{(v_A-p_A^*)(v_B-p_A^*)}{Y_A}$

D Appendix four: Soft budget constraint. Decision by local authority B.

$$\max_{r,\tau_B}\left\{Y_B\left(1-\frac{S_I^D(1-r)v_A+G_B}{Y_A+Y_B}-\tau_B\right)+z_B\left(\frac{S}{2}-S_I^D\right)-\frac{1}{2}\left(\frac{S}{2}-S_I^D\right)^2\right\}$$

s.t

$$v_B\left(\frac{S}{2}-S_I^D\right)+S_I^D r v_A-G_B=Y_B\tau_B$$

$$r \geq 0; r\leq 1$$

The Lagrangean can be written as:
$L = Y_B\left(1-\frac{S_I^D(1-r)v_A+G_B}{Y_A+Y_B}-\tau_B\right)+z_B\left(\frac{S}{2}-S_I^D\right)-\frac{1}{2}\left(\frac{S}{2}-S_I^D\right)^2$
$\quad -\lambda\left(v_B\left(\frac{S}{2}-S_I^D\right)+S_I^D r v_A-G_B-Y_B\tau_B\right)$
The FOC can be written as:
$\frac{\partial L}{\partial r} = Y_B v_A\frac{S_I^D}{Y_A+Y_B}-\lambda v_A S_I^D$
$\frac{\partial L}{\partial \tau_B} = \lambda Y_B - Y_B$
The second derivative is always negative, hence $r=0$ and by substitution $\tau_B = \frac{v_B\left(\frac{S}{2}-S_I^D\right)-G_B}{Y_B}$.

Table 1: Some indicators for health care expenditure and mobility

	Tax revenue (%total evenue)	admission from other regions	Of which nearby	admission to other regions	Of which nearby	Mobility balance	Surplus/Deficit (per capita)
Piedmont	0,403	5,6	0,55	5,6	0,55	-19,003	-0
Aosta Valley	0,403	5,7	0,48	5,7	0,48	-16,282	-111
Lombardy	0,608	8,2	0,44	8,2	0,44	438,503	0
Bolzano	0,411	11,8	0,64	11,8	0,64	6,600	49
Trento	0,416	8,8	0,66	8,8	0,66	-15,381	-4
Veneto	0,491	7,6	0,54	7,6	0,54	116,280	-41
Friuli V.g.	0,409	11,5	0,64	11,5	0,64	15,520	12
Ligury	0,336	9,9	0,57	9,9	0,57	-19,052	-159
Emilia R.	0,486	10,2	0,46	10,2	0,46	270,712	14
Tuscany	0,399	10,7	0,38	10,7	0,38	103,664	5
Umbria	0,308	17,2	0,71	17,2	0,71	27,252	8
Marche	0,390	8,6	0,52	8,6	0,52	-44,959	-11
Lazio	0,524	8,0	0,49	8,0	0,49	42,903	-265
Abruzzo	0,275	9,0	0,50	9,0	0,50	17,377	-152
Molise	0,120	16,0	0,92	16,0	0,92	0,261	-347
Campania	0,211	2,4	0,51	2,4	0,51	-260,570	-248
Puglia	0,237	2,7	0,60	2,7	0,00	-153,548	-16
Basilicata	0,112	12,9	0,89	12,9	0,89	-53,928	-53
Calabcia	0,100	2,5	0,24	2,5	0,24	-210,573	-27
Sicily	0,239	1,5	0,00	1,5	0,00	-195,353	-103
Sardinia	0,287	0,8	0,00	0,8	0,00	-50,023	-155
ITALY	0,395					0,000	-4.387,116