The efficiency case for transit subsidies in the presence of a 'Soft' budget constraint

Andrés Gómez-Lobo
Department of Economics
University of Chile
agomezlo@econ.uchile.cl

October 5, 2013

Abstract

The main contribution of this paper is to discuss the implications of a 'soft' budget constraint on optimal transit fares and subsidies. We find that the effect of productive inefficiencies on optimal fares and subsidy levels depends critically on the way cost reducing effort enters the cost function and on the institutional environment (as measured by the 'tightness' of the budget constraint faced by operators). In particular, recognizing that subsidies may have an adverse effect on productive efficiency does not necessarily imply that transit subsidies should be eliminated. Unsurprisingly, there will be a trade-off between the negative cost efficiency effects of transit subsidies and the welfare enhancing allocative efficiencies related to these subsidies. Under certain circumstances optimal subsidies may be higher when operators face an intermediate budget constraint than when they face a 'tight' budget constraint. We illustrate this last result using a simple numerical model.

Keywords: transit fares, subsidies, cost efficiency

1 Introduction

Winston (2000) cites evidence that 75% of mass transit subsidies in the US benefit transit workers (in the form of above market wages) or suppliers of capital equipment. Only 25% of this spending benefits users in the form of lower fares and improved quality of service.

Furthermore, in a comprehensive review of bus-transit operator performance, De Berger and Kersten (2008) state that "Cost inflation is to some extent related to...transit firms' weak budget constraints due to subsidies" (page 10). They add that "there appears to be sufficient evidence to conclude that subsidies do increase operating costs"(page 12). More recent studies seem to confirm this result.[1]

In spite of the numerous empirical studies analyzing the effects of subsidies on operators' technical and cost efficiency levels there is scant work on the welfare implications of this phenomenon on optimal transit fares. The policy implications emphasized in the literature relate to the privatization of transit operators, the competitive tendering of routes and services or the introduction of more powerful contractual incentives for cost reduction.[2] However, there also seems to be the presumption that the mere existence of inefficiencies and cost inflation is sufficient to justify significant fare increases and sharp reductions in transit subsidies. Winston and Shirley (1998) argue that increasing transit fares would improve net social welfare; the benefits of lower fiscal deficits would more than compensate for the negative effects of higher fares on users' well-being.[3]

To date no systematic study has been undertaken analyzing the welfare implications of a 'soft-budget' constraint on the optimal transit fare or subsidy. The main purpose of this paper is to undertake such an analysis using a simple theoretical model.

[1] See for example Nieswand and Walter (2012) for a recent study of German local bus companies.

[2] The review by De Borger and Kersten (2008) seems to confirm that high powered incentive schemes do reduce bus transit operating costs. This includes risk sharing contracts as in France (Gagnepain and Ivaldi, 2002) and yardstick competition as in Norway (Dalen and Gómez-Lobo, 2003). According to their review, the efficiency effects of ownership structure are mixed.

[3] Parry and Small (2009), in their detailed study of transit fares in Los Angeles, Washington D.C. and London, argue the opposite. Namely, that increasing transit subsidies would improve social welfare. However, they recognize that their model does not account for possible cost inflation effects related to transit subsidies.

In order to analyze this issue we assume that operating costs depend on unobservable cost reducing effort. The incentives to undertake such effort will depend on the budget constraint faced by managers. We model the 'softness' of the budget constraint by a cost-sharing rule. When subsidies cover any operating deficit (an extreme type of 'soft' budget constraint), managers will have low incentives to reduce costs. When subsidies are insensitive to cost over-runs, then operators face a 'tight' budget constraint and cost reducing effort will be socially optimal.

We find that the effect of cost inefficiencies on optimal fares and subsidy levels depends critically on the way cost reducing effort enters the cost function and on the institutional environment (as measured by the 'tightness' of the budget constraint faced by operators). In particular, recognizing that subsidies may have an adverse effect on productive efficiency does not necessarily imply that transit subsidies should be eliminated. Unsurprisingly, there will be a trade-off between the negative cost efficiency effects of transit subsidies and the welfare enhancing allocative efficiencies related to these subsidies.

Interestingly, under certain circumstances optimal subsidies may be higher when operators face an intermediate budget constraint than when they face a 'tight' budget constraint. The intuition for this result is that when cost reducing effort is related to patronage, the second-best transit fare should be reduced in order to increase demand and spur cost reducing effort. If this effect dominates the direct effect of cost inefficiency on fares then the optimal subsidy per passenger will be higher.

In the next section we develop the model where we introduce cost reducing effort in the cost function and specify the managers' objective function. We then explore the consequences of productive inefficiencies on fares and subsidies. Finally, we use a very simple numerical model to illustrate some of the theoretical results.

In this paper we do not consider distributive issues as a motivation for transit subsidies.[4]

[4]See Estupinan, Gómez-Lobo, Munoz-Raskin and Serebrisky (2009) for a discussion of distributive issues in public transit.

2 The model

Except for the inclusion of a cost of public funds and cost inflation, the model presented here follows Small and Verhoef (2007, chapter 4). We assume that gross consumer surplus (CS) can be represented by the following benefit function: $CS = B(q_a, q_p)$ where q_a are the number of trips in private (*automobile*) transport and q_p are the number of trips in public transport.[5]

2.1 Users

The inverse demand function for each type of travel is the derivative of the consumer surplus function:

$$p_a(q_a, q_p,) = \frac{\partial CS}{\partial q_a} = B_a(q_a, q_p) \tag{1}$$

$$p_p(q_a, q_p) = \frac{\partial CS}{\partial q_p} = B_p(q_a, q_p) \tag{2}$$

We define $CT_a(q_a)$, $CT_p^{op}(q_p, e)$, and $CT_p^u(q_p)$ as the total cost of private automobile travel, the total operating costs of the public transit system and the total user costs of public transport, respectively. User costs are related to the access time, waiting time and in-vehicle times that users have to invest when traveling in public transport.[6] Cost reducing effort (e) affects operating costs and we will discuss how we model this effect further below.

There are several restriction that must be taken into account when determining the social optimal quantity of trips in either mode of transport.

First, users of private transport only perceive the average cost of using this mode, $ac_a(q_a) = \frac{CT_a}{q_a}$. As is well know, users will not take into account the additional costs borne by others when making a decision to use private transport. This is the classical

[5]We are implicitly assuming that there are no income effects on travel demand.

[6]We are also assuming with the above definitions that there is no interaction between private and public transport. That is, the cost of private transport is not affected by the level of public transport and vice versa. This would be the case when public transport is rail or bus services offered in specialized segregated corridors (as in BRT systems). Introducing cost interaction between these two modes would not change anything substantial in the results below although the notation would be much more cumbersome. See Ahn (2009) for a model that incorporates such interactions.

congestion externality, although it applies also to accidents or pollution generated by private transport. To see this, note that social marginal cost of using private transport is:

$$mc_a(q_a) = ac_a(q_a) + q_a \cdot ac_a^{'} \tag{3}$$

where $mc_a(q_a) = \frac{\partial CT_a}{\partial q_a}$ is the marginal cost of using private transport and $ac_a^{'} = \frac{\partial ac_a}{\partial q_a}$ is the change in the average cost of private transport as the number of trips in this mode increases.

When deciding to undertake a trip in private transport individuals will only consider the first term on the right hand side of the above expression and will not consider the cost they impose on all other users (q_a) if their additional trip increases average costs (due for example, to higher congestion). Thus, there is an externality that is not internalized by individuals, which is equal to the difference between the marginal social cost of an additional trip in private transport and its average cost:

$$Ext = mc_a(q_a) - ac_a(q_a) = q_a \cdot ac_a^{'} \, . \tag{4}$$

We can also assume that there are other policies that may help to internalize externalities generated by private transport. For example, petrol prices are relatively high in Europe and this may help to reduce the magnitude of externalities.[7] Therefore, we assume that private transport users also have to pay a cost of τ_a per trip.

In the above formulation, τ_a can represent fuel taxes or any other policy that affect private transport users (for example, a congestion tax). However, in this paper we do not optimize with respect to this variable, but rather assume it is parametric to the problem. Thus, we assume that political obstacles or the time-frame of our analysis preclude the introduction of congestion charges to solve externalities in private transport. Further below we will make some more comments regarding this point.

[7]However, as noted by Parry and Small (2005), in so far as externalities are related to the number of kilometers traveled (as in the case of congestion and accidents), fuel taxes may not be as effective as they might appear in reducing these externalities. This is because part of their effect is to change behavior related to vehicle choice (more fuel efficient cars for example) rather than the number of trips or kilometers traveled. In addition, fuel taxes may be a very blunt instrument to control externalities that vary by local area, time of day and other dimensions.

In sum, welfare maximization must consider that the number of private trips taken is endogenous and determined by the equality between the marginal benefit of an additional private trip to its 'perceived' marginal cost for users. That is:

$$B_a\left(q_a, q_p\right) = ac_a(q_a) + \tau_a \ . \tag{5}$$

Second, the number of public transport trips is determined by the equality between the marginal benefit of an additional trip with the "perceived" marginal cost by users:

$$B_p\left(q_a, q_p\right) = ac_p^u(q_p) + \tau_p \ , \tag{6}$$

where τ_p is the public transport fare and $ac_p^u(q_p)$ is the average user cost of using public transport. As in the case of private transport, users only perceive the average cost of using public transit, which is the fare plus the average cost of user time in this mode, that in turn is composed of the cost of access time, waiting time and in-vehicle time.[8]

2.2 Managers and operating costs

As for operating cost and cost reducing effort, we assume that $e \in [0, \infty)$ and $CT_p^{op}(q_p, e) > 0 \ \forall q_p, e > 0$. In addition, we make some standard assumptions:

$$\frac{\partial CT_p^{op}(q_p, e)}{\partial e} < 0 \tag{7}$$

$$\frac{\partial^2 CT_p^{op}(q_p, e)}{\partial e^2} > 0 \tag{8}$$

These last conditions imply that effort reduces costs but at at a decreasing rate.

[8]Private transport also has an in-vehicle time cost. However, in this case we assume it is constant per trip and can be subsumed in $ac_a(q_a)$. However, as will be seen further below, in the case of public transport, operational variables will affect time costs and therefore it makes sense to specify them separately.

Managers' face a disutility from exerting effort, equal to $-\Psi(e)$. This disutility is assumed to be strictly convex in effort and in order to guarantee an interior solution we assume that for zero effort the marginal utility cost is lower than the cost savings:

$$\Psi'(0) < -\frac{\partial CT_p^{op}(q_p, 0)}{\partial e} < -\infty \tag{9}$$

$$\Psi'(e) > 0 \tag{10}$$

$$\Psi''(e) > 0 \tag{11}$$

The manager's utility function is assumed to be:

$$U(q_p, e) = T + \theta \cdot \left(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e)\right) - \Psi(e) \tag{12}$$

where θ is a cost sharing parameter that measures to what extent managers get to keep the savings from cost reducing effort or bare the costs from not exerting effort. In the regulatory literature this parameter determines the 'power' of the contract and summarizes the incentives provided to agents in a contractual relationship. If the operator is private this parameter determines the degree to which the firm is residual claimant to cost savings or cost over-runs. But the above specification of manager's utility function can also accommodate the case of a publicly owned operator run by hired management.

When θ is equal to one managers bare the full effects of cost increases or benefit from all cost reductions. This case reflects a tight budget constraint scenario. The other extreme is when θ equals zero, in which case managers are fully compensated for cost overruns but do not benefit from cost savings. This case reflects a scenario with a very 'soft' budget constraint.

T is a lump-sum transfer that the authorities must make to managers in order to guarantee that utility is positive. In other words, we impose an Individual Rationality (IR) constraint on the problem solved further below.

Before continuing it is important to note that we are using the cost sharing rule as a simple way to describe the positive incentives faced by managers. We are not attempting to answer the normative question as to which cost sharing rule would be optimal. The current perfect information set-up is too simple to answer that question

since a regulator could impose the first-best effort outcome by specifying a regulatory contract that implies a non-negative utility level for managers when first-best effort is exerted and an infinite penalty otherwise. In order to answer the normative question some asymmetry of information must be introduced in the model. This is the approached used in the analysis of optimal linear (cost-sharing) schemes by Schmalensee (1989) or in the more sophisticated optimal menu contracts approach of Baron and Myerson (1982) and Laffont and Tirole (1993).[9]

The first-best effort level, e^s, is given by the following first order condition:

$$-\frac{\partial CT_p^{op}}{\partial e}(q_p, e^s) = \Psi'(e^s) \tag{13}$$

which may depend on q_p.

However, the effort actually expended by managers, e^*, will be the solution to the following first order condition:

$$-\theta \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e^*) = \Psi'(e^*). \tag{14}$$

Therefore, actual effort will be lower than optimal effort when managers face a soft budget constraint ($\theta < 1$). When $\theta = 0$ managers exert no cost reducing effort at all and costs are $CT_p^{op}(q_p, 0)$.

2.3 Transit authority

We make the accounting convention that the authorities receive all revenues and must pay all costs, including payments to managers. The net financial costs to the authorities, S, is the shortfall between operating costs plus compensation to managers minus revenues:

$$S = CT_p^{op}(q_p, e) + T + \theta \cdot \left(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e)\right) - \tau_p \cdot q_p \ . \tag{15}$$

[9] In the transport economics literature, asymmetric information models have been used by Dalen and Gómez-Lobo (1996) and Gagnepain and Ivaldi (2002).

These resources have an opportunity cost that is represented by an exogenous parameter λ. This is the cost of public funds and implies that in order to raise \$1 of funding through distortionary taxation, there is a deadweight loss of $\$\lambda$ in the economy. We do not give a general equilibrium grounding to the cost of public funds parameter under the assumption that transit subsides do not represent a large fraction of public expenditure and thus can be considered exogenous to this sector.[10] This allows us to obtain a simpler formula for the optimal transit fare and subsidy and is also consistent with the modern regulatory economics literature (Laffont and Tirole, 1993).

2.4 Social Welfare and optimal fare

With this set-up social welfare is given by the sum of net consumer surplus, manager's utility and the net financial costs to the authorities given by (15). In this last case, these resources must include the cost of public funds since these transfers have an opportunity cost as discussed above. Social welfare is thus:

$$\begin{aligned} W &= B\left(q_a, q_p\right) - CT_a(q_a) - CT_p^u(q_p) - \tau_p \cdot q_p + \\ & T + \theta \cdot \left(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e)\right) - \Psi(e) + \\ & (1+\lambda) \cdot \Big(\tau_p \cdot q_p - CT_p^{op}(q_p, e) - T \\ & -\theta \cdot \left(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e)\right)\Big) \end{aligned} \tag{16}$$

Social welfare must be maximized with respect to q_a, q_p, τ_p, T and e, taking into account restrictions (5), (6), (12) and (14).[11]

[10] Dodgson and Topham (1987), to cite one example, also use a cost of public funds in their analysis while authors such as Parry and Small (2009) assume a non-distortionary lump-sum tax to fund transit subsidies. Jara-Díaz and Gschwender (2009) also introduce a multiplier but instead of a cost of public funds it is an endogenous Lagrange multiplier related to the transit system's financial constraint.

[11] It must be noted that optimal fares and subsidies are closely linked to frequency, bus size and the network structure. However, in this paper we analyze optimal public transport fares without going into much detail regarding the particular specification for operational variables and user time costs in order to describe in a straight forward manner what the efficiency justifications are for subsidizing public transport. Implicitly we are assuming that frequency, bus size and network structure are also optimized in the solution.

In the Appendix it is shown that the solution to the above optimization problem leads to the following condition:

$$\begin{aligned}\tau_p &= ac_p^{op}(q_p, e^*) + \Big(mc_p^{op}(q_p, e^*) - ac_p^{op}(q_p, e^*)\Big) \\ &+ \frac{(mc_p^u - ac_p^u)}{(1+\lambda)} + \frac{Ext \cdot D_{ap}}{(1+\lambda)} + \frac{\lambda}{(1+\lambda)} \cdot \frac{\tau_p}{|\psi_p|} \\ &+ (1-\theta) \cdot \frac{\partial CT_p^{op}}{\partial e} \cdot \frac{de}{dq_p}. \end{aligned} \tag{17}$$

where Ext is the externality caused by private transport and not internalized by τ_a: $Ext = mc_a(q_a) - \tau_a - ac_a(q_a)$, $|\psi_p|$ is the absolute value of the demand elasticity of public transport with respect to its fare and D_{ap} is the diversion ratio between car use and public transport. This last parameter measures how many of the lost (increased) ridership in public transport due an increase (decrease) in public transport fares go to (come from) the private transport mode. As private and public transport are substitutes this diversion ratio should be negative and less than one in absolute value. Finally, $\frac{de}{dq_p}$ is the change in effort exerted as transport demand increases and is related to the effect that effort has on marginal costs:

$$\frac{de}{dq_p} = \frac{-\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e')}{\Big(-\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial^2 e}(q_p, e') - \Psi''(e')\Big)} \tag{18}$$

The denominator in this last expression is negative so if effort reduces marginal costs then effort will be increasing in output for a given cost-sharing parameter.

3 Determinants of the optimal transit fare

There are several cases worth analyzing. If the cost-sharing rule implies a very 'tight' budget constraint ($\theta = 1$) and there is no cost of public funds ($\lambda = 0$)then it is easy

to verify that the optimal fare formula collapses to:

$$\begin{aligned}\tau_p &= ac_p^{op}(q_p, e^*) + \Big(mc_p^{op}(q_p, e^*) - ac_p^{op}(q_p, e^*)\Big) \\ &\quad +(mc_p^u - ac_p^u) + Ext \cdot D_{ap} \end{aligned} \tag{19}$$

In this case, effort will be first-best and the last two terms of (45) disappear. This last expression shows clearly the main efficiency justifications for subsidizing public transport. Condition (19) has four terms on the RHS. We discuss each of these in turn.

The first term is average operating costs. If the optimal fare is equal to average operating costs then there is no subsidy. Therefore, the last three terms account for adjustments to this break-even fare.

The second term is a correction for scale economies in operating costs, as in the case of a natural monopoly. It is usually considered that there are no economies of scale in bus service provision but they usually exist for rail transport. If operators' cost function do exhibit economies of scale then this term will be negative (since marginal costs will be lower than average costs) and transit fares should be adjusted below average operating costs in order to set fares at the first-best level.

The third term is a correction for economies of scale in user costs. This adjustment is usually referred to as the 'Mohring effect' (Mohring, 1972). The argument runs as follows: as demand increases transit planner will provide a denser route structure, higher frequencies, or both, thus reducing access and waiting times in this transport mode. Through this mechanism additional transit users will reduce user costs for all existing users, implying that the social marginal time cost is lower than the private marginal time cost (which is equal to the average user cost). This positive consumption externality justifies a Piguovian subsidy and the optimal fare is thus reduced. We will not go into more details regarding this term but refer interested readers to the classical references on this topic: Mohring (1972), Turvey and Mohring (1975), Jansson (1979), and Jansson (1993).

The fourth term is a correction for the negative externalities caused by private transport that are not internalized by users.[12] If public and private transport are

[12]Note that if transit is complementary to labor supply then there will be an additional correction

substitutes then the diversion ration will be negative and this last term is also negative. In this case public transport should be optimally subsidized in a second-best world in order to diminish externalities caused by private transport. If there is an optimal congestion tax and fuel taxes are such that accidents and pollution externalities are fully borne by users then this term disappears. Note also that D_{ap} is crucial in this fourth term. If this diversion ratio is zero –implying that the cross-elasticity of demand for private transport is not responsive to transit fares– then no subsidy can be justified based on this second-best argument, irrespective of how high are the externalities generated by private transport.

The second to fourth terms of condition (19) will usually be negative, implying a positive subsidy in the optimal solution. These adjustments to fare levels are the usual justifications for transit subsidies (e.g. Parry and Small, 2009).

If $\lambda > 0$ while still abstracting from incentive effects (i.e. $\theta = 1$) then a new term appears in the optimal fare equation (the fifth term of (45)). This term is the typical 'Ramsey' inverse elasticity rule for natural monopoly pricing when there is a cost of public funds. As λ increases it reduces the importance of the adjustment for economies of scale in user costs and the adjustment for externalities generated by private car use. If the authorities impose the restriction that operating costs must be funded entirely from fares (a self-funding or zero subsidy restriction), then the parameter λ is endogenous and will be equal to whichever value sets the fare level equal to average costs.[13]

This 'Ramsey' adjustment implies that the efficiency of the tax system will matter for transit subsidies. In countries where this system is very inefficient or subsidies are expected to be funded by cross-subsidies, then the optimal transit subsidy will be lower than in countries with more efficient funding sources.

The final case is when $\theta < 1$. That is, when the cost-sharing rule implies a soft budget constraint and cost reducing effort is not optimal. In this case fares will be

term analogous to the one discussed here due to the distortion generated by income taxes. In that case, transit fares should be lowered, particularly during peak-periods, in order to induce higher work hours and reduce the welfare loss in the economy due to income taxation. We have ignored this additional argument for transit subsidies in this paper. For a model that incorporates this issue see Parry and Bento (2001).

[13]This is the approach taken by Jara-Díaz and Gschwender (2009) where they show that a self-funding restriction implies a transit system with lower frequencies and bigger buses compared to the social optimal level (assuming no cost of public funds).

higher since average cost will be higher as effort is reduced. However, there is an additional term in the optimal fare equation (last term of equation (45)). It states that if effort is increasing in output then there is an additional factor affecting optimal fares that was not present before. All else constant, it may be efficient to reduce fares in order to increase demand and through this mechanism induce more cost reducing effort on the part of operators.

This last effect implies that under certain circumstances (that will be explored in a numerical model further below), subsidies may be higher for some intermediate values of the cost-sharing parameter compared to the case of a tight budget constraint. This is an unexpected and counterintuitive result implying that cost inflation may actually increase optimal subsidies in some cases.

However, it should be noted that if effort does not have any effect on marginal costs and only affects fixed costs, the last term of (45) disappears since the second derivative of the cost function with respect to output and effort is zero implying that $\frac{de}{dq_p} = 0$ from equation (18). In addition, since effort does not change marginal costs, the optimal fare is unaffected by the cost-sharing parameter in this case.[14]

It might seem puzzling that in this last case no allowance should be made to fares to accommodate the higher average costs due to low effort. The explanation is somewhat subtle. For a fixed cost-sharing rule, and assuming patronage has no effect on effort, fares will have no effect on the cost inefficiency. Therefore, the question is whether this inefficiency should be funded by users through higher fares or by authority through transfers that have a cost of public funds. However, this is exactly the trade-off implied by the cost of public funds adjustment and is therefore already considered in the analysis. Higher fixed costs should be funded through fares until the deadweight loss associated with these higher fares is equal to the cost of public funds. Since the cost of public funds is exogenous and fixed, starting from an optimal fare level an increase in fixed costs due to inefficiency should be funded through transfers since this is the cheaper option from an economic perspective.

This last result does not imply that no effort should be made to increase efficiency through a change in the cost-sharing rule, a point we will discuss in the conclusions. Rather, for a constant cost-sharing rule nothing is gained by increasing fares beyond the case when there is no cost inflation.

[14] In the regulatory literature this is often called the incentive-pricing dichotomy. See Laffont and Tirole (1993) for more on this issue.

4 A simple numerical illustration

In order to illustrate the above ideas in this section we use a very simple model to solve for the optimal fare and examine how this fare changes as the cost sharing parameter changes. In particular, we want to show that for some parameter configurations optimal subsidies actually increase for intermediate value of the cost sharing parameter.

We assume that the cost function is given by:

$$CT(q_p, e) = k \cdot q^{\alpha} \cdot exp^{-\beta \cdot e} \tag{20}$$

where $k > 0$ is a scale parameter, $\alpha > 0$ determines economies of scale in production and $\beta > 0$ determines the cost reducing effects of managerial effort.

The utility of cost of effort is assumed to be:

$$\psi(e) = exp^{\gamma \cdot e} \tag{21}$$

where γ is a positive parameter.

Demand is assumed to be iso-elastic:

$$q_p(\tau_p) = K \cdot \tau_p^{-\eta} \tag{22}$$

where K is a scale parameter and η is the absolute value of the price elasticity of demand.

As for the 'Mohring' effect, we assume a very simple specification:

$$(mc_p^u - ac_p^u) = M \cdot q_p^{-0.5} \tag{23}$$

where M is a positive parameter. This specification implies that user scale economies decrease at a rate proportional to the square root of demand. This can be justified by the 'square root law' that states that in simple transit models, frequency should increase at rate proportional to the square root of demand. Since waiting times will be inversely proportional to frequency, then user costs will also decrease proportional to this rate.

The externality effect is set to a fixed value of $Ext \cdot D_a p = D$ independent of demand. Finally, the cost of public funds is a fixed parameter λ.[15]

With the above specification it is possible to numerically solve for the optimal fare given different cost-sharing parameters. Table 1 shows the parameter values for the first model solved. Figure 1 shows the average cost and optimal fare for this parameter configuration. It can be seen that as the cost-sharing parameter decreases, average costs and the optimal fare increase. However, from Figure 2 it can be seen that the subsidy per trip increases for intermediate values of the cost sharing parameter compared to the case of optimal effort ($\theta = 1$), but the total subsidy decreases as operators become more inefficient (Figure 3).

Table 1: Parameter values for numerical model 1

Parameter	Value
k	100
α	1
β	0.1
γ	1.5
K	$10,000,000$
η	0.7
M	$-10,000$
D	-5
λ	0.3

Table 2 shows the parameter values of a very similar model to the first one, except that the externality effect was reduced from -5 to -1. With this slight change it can be seen from Figure 4 and 5 that the optimal subsidy per trip is reduced as the cost sharing parameter is lower. As the cost sharing rule becomes very small and approaches zero, subsidies per trip become negative, implying that fares should be taxed.

However, the interesting aspect of this parameter configuration for this second model is that the total subsidy has an inverted U shape as shown in Figure 6. Total subsidies increase for a cost sharing parameter below one, reaching a maximum

[15]Reasonable values for λ are between 0.2-0.4 depending on the fiscal structure of a country.

Table 2: Parameter values for numerical model 2

Parameter	Value
k	100
α	1
β	0.1
γ	1.5
K	$10,000,000$
η	0.7
M	$-10,000$
D	-1
λ	0.3

when θ is equal to 0.74. This is a direct consequence of the last term of equation (45) whereby optimal fares should be adjusted in order to increase demand and through this mechanism indirectly increase cost reducing effort. If this last term is set to zero in this model then the optimal subsidy per trip as well as the total subsidy monotonically decrease as θ decreases.

5 Conclusions

What we have shown in this paper is that the mere existence of cost inflation and a soft-budget constraint does not imply that transit subsidies should be eliminated. Recognizing that there may be productive inefficiencies is not sufficient to eliminate these subsidies and increase fares. If the cost-sharing rule is fixed then no efficiency gains will be made through this policy (the inefficiency will now be funded through fares rather than government transfers) and may actually increase inefficiency if lower output implies less cost reducing effort on the part of operators.

The main policy consequence of our results is that in order to tackle cost inflation, reforms that increase θ must be introduced. Another way to see this is that if θ is not changed the inefficiencies are uncahnged. In this case, the dilemma is whether these inefficiencies should be funded from fares or through subsidies. We have seen that this will depend, on the one hand, on the cost of public funds and, on the other hand, on the efficiency arguments that call for subsidies.

An example may help to illustrate this idea. Lets assume a transit operator is publicly owned. Imposing a tight budget constraint on these types of companies is not easy since employees' and managers' payment structure may be based on public sector regulations that do not allow for profit motives or to link salaries to performance, they may have political power that precludes imposing a tight budget constraint, or for other reasons this option may not be feasible. In this scenario, recognizing that the operator is inefficient does not imply that the solution is to raise fares and lower subsidies. This change will only affect who is paying for these inefficiencies but does not do much to reduce operating costs. In fact, inefficiencies may increase if higher fares reduce demand and thereby also reduce cost reducing effort linked to demand levels. In this case, privatization, competitive tendering or yardstick competition may be the correct policy options but eliminating subsidies by itself may make matters worse.

In this paper we have assumed that the cost-sharing rule and therefore cost inflation is exogenous. Another possibility –and perhaps what many analysts have in mind– is that the potential to receive subsidies affects the cost-sharing rule. That is, the θ parameter in our model could be endogenous and depend on the possibility of making transfers to operators. Analyzing this case would require making a welfare comparison between a situation in which transfers are prohibited and operators must break-even and a situation in which the authorities can make transfers to operators. There are many subtle issues that must be addressed with this approach as discussed in Chapter 15 of Laffont and Tirole (1993). We leave for further research the application of this idea to transit services.

References

[1] Ahn, K. (2007), 'Road pricing and bus service policies', *Journal of Transport Economics and Policy*, 43(1), pp. 25-53.

[2] Baron, D. and R. Myerson (1982), 'Regulating a monopolist with unknown costs', *Econometrica*, 50, pp. 911-930.

[3] Dalen, D.M. and A. Gómez-Lobo (1996), 'Regulation and incentive contracts: An empirical investigation of the Norwegian bus transport industry', IFS Working Papers W96/08, Institute for Fiscal Studies, London.

[4] Dalen, D.M. and A. Gómez-Lobo (2003), 'Yardsticks on the road: Regulatory contracts and cost efficiency in the Norwegian bus industry', *Transportation*, 30:371-386.

[5] De Borger, B. and K. Kerstens (2000), 'The Performance of Bus-transit Operators', in: Hensher, D.A. and Button, K. J. (eds) *Handbook of Transport Modelling*, 2nd Ed., Amsterdam, Elsevier, 693-714.

[6] Dodgson, J.S. and N. Topham (1987), 'Benefit-Cost Rules for Urban Transit Subsidies', *Journal of Transport Economics and Policy*, 21, pp. 57-71.

[7] Estupinan, N., A. Gómez-Lobo, R. Munoz-Raskin and T. Serebrisky (2009), 'Affordability of Public Transport: what do we mean, what can be done?', *Transport Reviews*, 29(6), pp. 715-39.

[8] Gagnepain, P. and M. Ivaldi (2002), 'Incentive Regulatory Policies: The Case of Public Transit Systems in France', *RAND Journal of Economics*, 33(4), pp. 605-629.

[9] Jansson, J.O. (1979), 'Marginal Cost Pricing of Scheduled Transport Services', *Journal of Transport Economics and Policy*, 13, 268-94.

[10] Jansson, K. (1993), 'Optimal Public Transport Price and Service Frequency', *Journal of Transport Economics and Policy*, 27, 33-50.

[11] Jara-Díaz, S.R. and A. Gschwender (2009), 'The Effect of Financial Constraints on the Optimal Design of Public Transport Services', *Transportation*, 36(1), 65-75.

[12] Laffont, J.J. and J. Tirole (1993), *A Theory of Incentives in Procurement and Regulation*, MIT Press.

[13] Mohring, H. (1972), 'Optimization and Scale Economies in Urban Bus Transportation', *American Economic Review*, 62, pp. 591-604.

[14] Nieswand, M. and M. Walter (2012), 'Cost Efficiency and Subsidization in German Local Public Bus Transit', GRASP Working Paper 30, European Commission, October.

[15] Parry, I.W.H. and A. Bento (2001), 'Revenue Recycling and the Welfare Effects of Road Pricing', *The Scandinavian Journal of Economics*, 103(4), pp. 645-671.

[16] Parry, I.W.H. and K.A. Small (2005), 'Does Britain or the United States Have the Right Gasoline Tax?',*American Economic Review*, 95, pp. 1276-89.

[17] Parry, I.W.H. and K.A. Small (2009), 'Should transit subsidies be reduced?', emphAmerican Economic Review, 99, pp. 700-24.

[18] Schmalensee, R. (1989), 'Good Regulatory Regimes', *Rand Journal of Economics*, 20(3), pp. 417-436.

[19] Small, K.A. and E. Verhoef (2007), *The Economics of Urban Transportion*, Second Edition, Routledge.

[20] Winston, C. (2000), 'Government failure in Urban Transportation', *Fiscal Studies*, 21(4), pp. 403-425.

[21] Winston, C. and C. Shirley (1998), *Alternative Route: Toward Efficient Urban Transportation*, Brookings Institution, Washington, DC.

A Derivation of the optimal subsidy formula without cost inflation

The Lagrange function for this problem is:

$$\begin{aligned} L &= B\left(q_a, q_p\right) - CT_a(q_a) - CT_p^u(q_p) - CT_p^{op}(q_p, e) - \Psi(e) \\ &+\omega \cdot \Big(ac_a(q_a) + \tau_a - B_a\left(q_a, q_p\right) \Big) + \gamma \cdot \Big(ac_p^u(q_p) \\ &+\tau_p - B_p\left(q_a, q_p\right) \Big) + \eta \cdot \Big(-\theta \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e) - \Psi'(e) \Big) \\ &+\lambda \cdot \Big(\tau_p \cdot q_p - CT_p^{op}(q_p, e) - \theta \cdot \big(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e) \big) - T \Big) \\ &+\mu \cdot \Big(T + \theta \cdot \big(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e) \big) - \Psi(e) \Big) \end{aligned} \tag{24}$$

where ω, γ, η and μ are the Lagrange multipliers associated with the four restrictions and as before λ is the (exogenous) cost of public funds.

The first order conditions for this problem are:

$$\begin{aligned}\frac{\partial L}{\partial q_a} &= B_a\left(q_a, q_p\right) - mc_a(q_a) + \omega \cdot \Big(ac_a^{'} - B_{aa}(q_a, q_p)\Big) \\ &\quad - \gamma \cdot B_{pa}(q_a, q_p) = 0 \end{aligned} \tag{25}$$

$$\begin{aligned}\frac{\partial L}{\partial q_p} &= B_p\left(q_a, q_p\right) - mc_p^{op} - mc_p^{u} + \omega \cdot \Big(- B_{ap}(q_a, q_p)\Big) \\ &\quad + \gamma \cdot \Big(ac_p^{u'} - B_{pp}(q_a, q_p)\Big) + \eta \cdot \Big(- \theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e)\Big) \\ &\quad + \lambda \cdot \Big(\tau_p - mc_p^{op} - \theta \cdot \Big(mc_p^{op}(q_p, 0) - mc_p^{op}(q_p, e)\Big)\Big) \\ &\quad + \mu \cdot \theta \cdot \Big(mc_p^{op}(q_p, 0) - mc_p^{op}(q_p, e)\Big) = 0 \end{aligned} \tag{26}$$

$$\begin{aligned}\frac{\partial L}{\partial e} &= -\frac{\partial CT_p^{op}}{\partial e}(q_p, e) - \Psi'(e) + \eta \cdot \Big(- \theta \cdot \frac{\partial^2 CT_p^{op}}{\partial^2 e}(q_p, e) - \Psi''(e)\Big) \\ &\quad + \lambda \cdot \Big(- \frac{\partial CT_p^{op}}{\partial e}(q_p, e) + \theta \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e)\Big) \\ &\quad + \mu \cdot \Big(- \theta \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e) - \Psi'(e)\Big) = 0 \end{aligned} \tag{27}$$

$$\frac{\partial L}{\partial \tau_p} = \gamma + \lambda \cdot q_p = 0 \tag{28}$$

$$\frac{\partial L}{\partial T} = -\lambda + \mu = 0 \tag{29}$$

$$\frac{\partial L}{\partial \omega} = ac_a(q_a) + \tau_a - B_a\left(q_a, q_p\right) = 0 \tag{30}$$

$$\frac{\partial L}{\partial \gamma} = ac_p^u(q_p) + \tau_p - B_p\left(q_a, q_p\right) = 0 \tag{31}$$

$$\frac{\partial L}{\partial \eta} = -\theta \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e) - \Psi'(e) = 0 \tag{32}$$

$$\frac{\partial L}{\partial \mu} = T + \theta \cdot \left(CT_p^{op}(q_p, 0) - CT_p^{op}(q_p, e)\right) - \Psi(e) = 0 \tag{33}$$

where in the above expressions $ac_p^{u'}$ is the derivative of the average user cost of public transit:

$$ac_p^{u'} = \frac{\partial ac_p^u(q_p, e)}{\partial q_p} \tag{34}$$

Using (25), (28) and (30) we obtain that:

$$\omega = \frac{Ext - \lambda \cdot q_p \cdot B_{pa}}{\left(ac_a' - B_{aa}\right)} \tag{35}$$

where Ext is the externality caused by private transport and not internalized by τ_a:

$$Ext = mc_a(q_a) - \tau_a - ac_a(q_a) \tag{36}$$

Using (29), (31) and (32) and inserting them into (26) and rearranging we obtain:

$$\begin{aligned}(1+\lambda) \cdot \tau_p &= (1+\lambda) \cdot mc_p^{op}(q_p, e) + (mc_p^u - ac_p^u) \\ &\quad + \omega \cdot \left(B_{ap}(q_a, q_p)\right) - \gamma \cdot \left(ac_p^{u'} - B_{pp}(q_a, q_p)\right)\end{aligned}$$

$$+\eta \cdot \left(\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e)\right) \tag{37}$$

Using the expression for ω (equation (35)) and condition (28) and dividing through by $(1+\lambda)$ we obtain:

$$\begin{aligned}
\tau_p &= mc_p^{op} + \frac{1}{(1+\lambda)} \cdot (mc_p^u - ac_p^u) + \frac{1}{(1+\lambda)} \cdot \frac{Ext \cdot B_{ap}}{\left(ac_a' - B_{aa}\right)} \\
&\quad - \frac{\lambda}{(1+\lambda)} \cdot \frac{q_p \cdot B_{pa} \cdot B_{ap}}{\left(ac_a' - B_{aa}\right)} + \frac{\lambda}{(1+\lambda)} \cdot q_p \cdot \left(ac_p^{u'} - B_{pp}\right) \\
&\quad + \frac{1}{(1+\lambda)} \cdot \eta \cdot \left(\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e)\right)
\end{aligned} \tag{38}$$

The above condition can be expressed in a more concise way if we introduce the diversion ratio D_{ap}, which is defined as the ratio of the change in private car use due to an increase in the public transport fare over the change in public transport users due to an increase of this fare. Totally differentiating condition (30) we can see that:

$$D_{ap} = \frac{\frac{dq_a}{d\tau_p}}{\frac{dq_p}{d\tau_p}} = \frac{B_{ap}}{\left(ac_a' - B_{aa}\right)} \tag{39}$$

Thus, we arrive at a simpler expression of the optimal public transport fare:

$$\begin{aligned}
\tau_p &= ac_p^{op} + (mc_p^{op} - ac_p^{op}) + \frac{(mc_p^u - ac_p^u)}{(1+\lambda)} + \frac{Ext \cdot D_{ap}}{(1+\lambda)} \\
&\quad + \frac{\lambda}{(1+\lambda)} \cdot \frac{\tau_p}{|\psi_p|} + \frac{1}{(1+\lambda)} \cdot \eta \cdot \left(\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e)\right)
\end{aligned} \tag{40}$$

In the above expression $|\psi_p|$ is the total elasticity of public transport with respect to its fare. To see this, totally differentiate condition (6) with respect to all its

arguments:

$$B_{pa} \cdot dq_a + B_{pp} \cdot dq_p = ac_p^{u'} \cdot dq_p + d\tau_p \tag{41}$$

Dividing this last expression by dq_p we obtain:

$$\frac{d\tau_p}{dq_p} = B_{pa} D_{ap} + \left(B_{pp} - ac_p^{u'} \right) \tag{42}$$

Thus, the last two terms of condition (38) are equal to:

$$\frac{\lambda}{(1+\lambda)} \cdot q_p \cdot \frac{-d\tau_p}{dq_p} \tag{43}$$

from which equation (40) follows.

Finally, from condition (27) and using (32) and (29) it is possible to obtain the expression for η:

$$\eta = \frac{(1+\lambda) \cdot (1-\theta) \cdot \frac{\partial CT_p^{op}}{\partial e}(q_p, e)}{\left(-\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial^2 e}(q_p, e) - \Psi''(e) \right)} \tag{44}$$

Inserting this last expression into (40) we obtain:

$$\begin{aligned} \tau_p &= ac_p^{op}(q_p, e) + \left(mc_p^{op}(q_p, e) - ac_p^{op}(q_p, e) \right) \\ &+ \frac{(mc_p^u - ac_p^u)}{(1+\lambda)} + \frac{Ext \cdot D_{ap}}{(1+\lambda)} + \frac{\lambda}{(1+\lambda)} \cdot \frac{\tau_p}{|\psi_p|} \\ &+ (1-\theta) \cdot \frac{\partial CT_p^{op}}{\partial e} \cdot \frac{de}{dq_p}. \end{aligned} \tag{45}$$

where $\frac{de}{dq_p}$ is how effort changes when transit demand increases and is obtained by

totally differentiating (32):

$$\frac{de}{dq_p} = \frac{\theta \cdot \frac{\partial^2 CT_p^{op}}{\partial e \partial q_p}(q_p, e)}{\left(- \theta \cdot \frac{\partial^2 CT_p^{op}}{\partial^2 e}(q_p, e) - \Psi''(e) \right)} \tag{46}$$

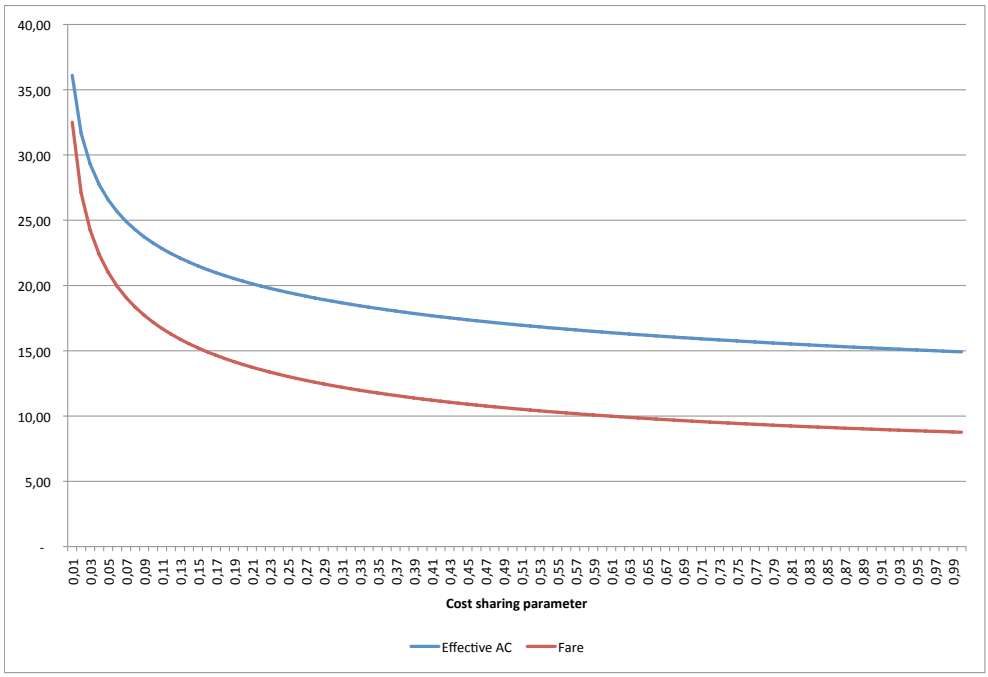

Figure 1: Average cost and optimal fare: model 1

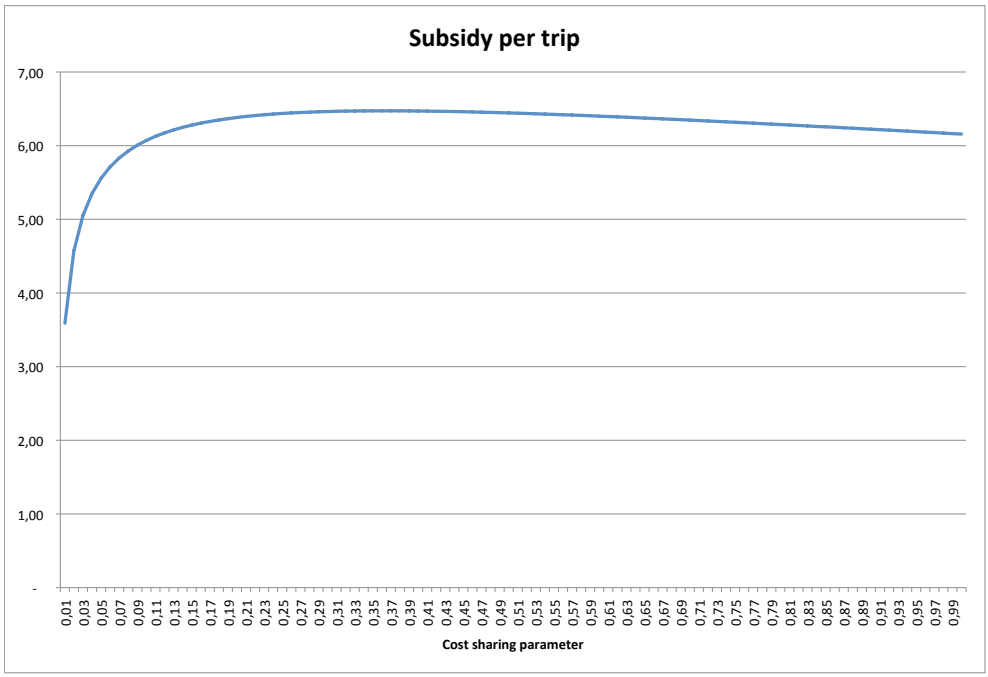

Figure 2: Subsidy per trip: model 1

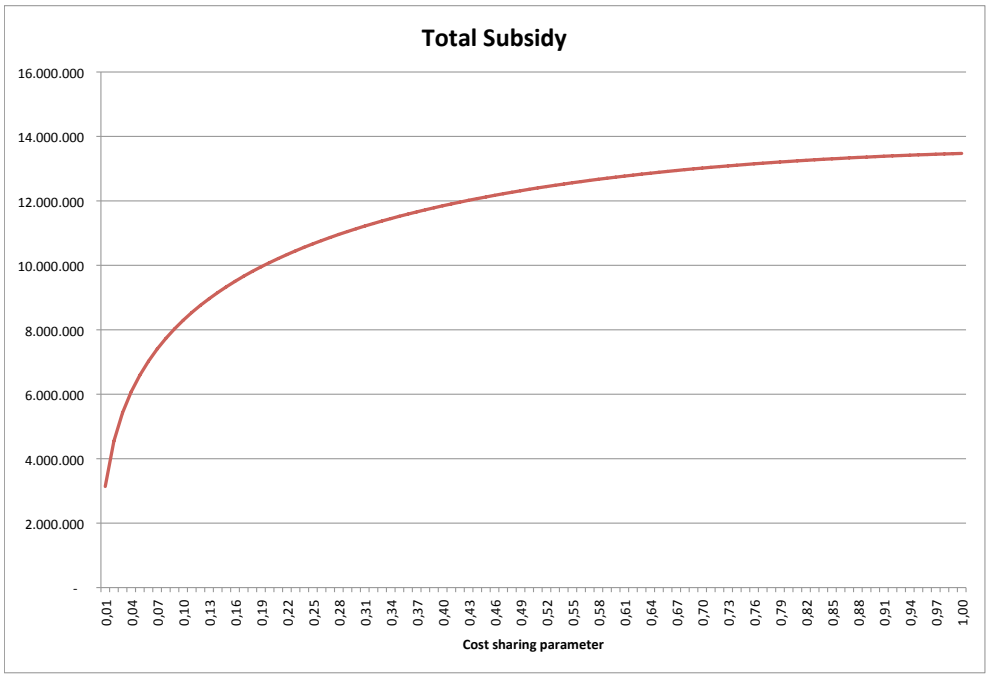

Figure 3: Total subsidy: model 1

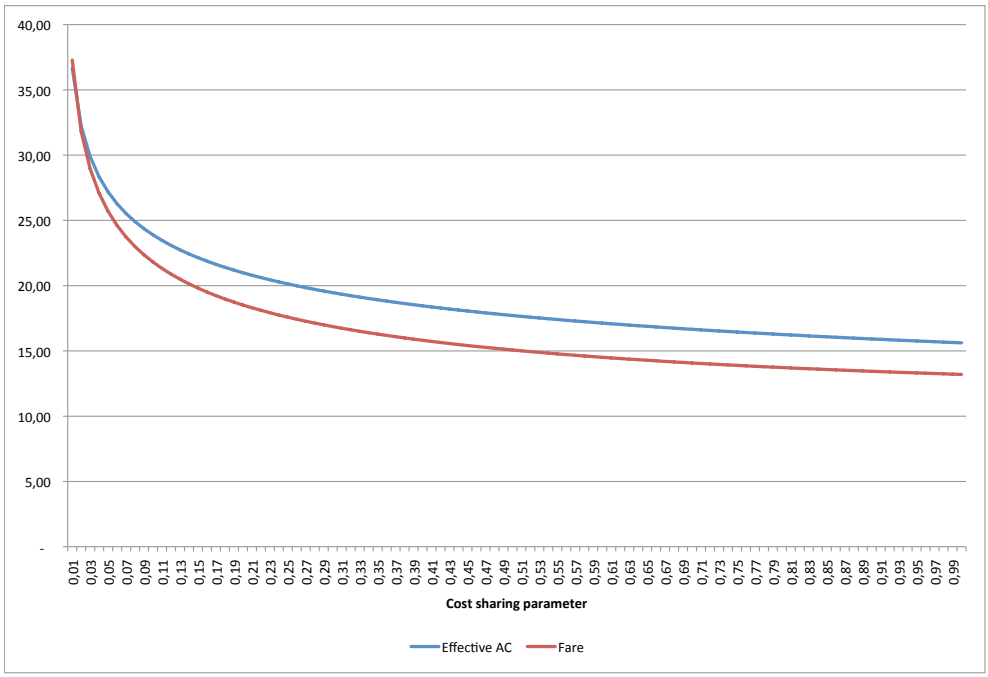

Figure 4: Average cost and optimal fare: model 2

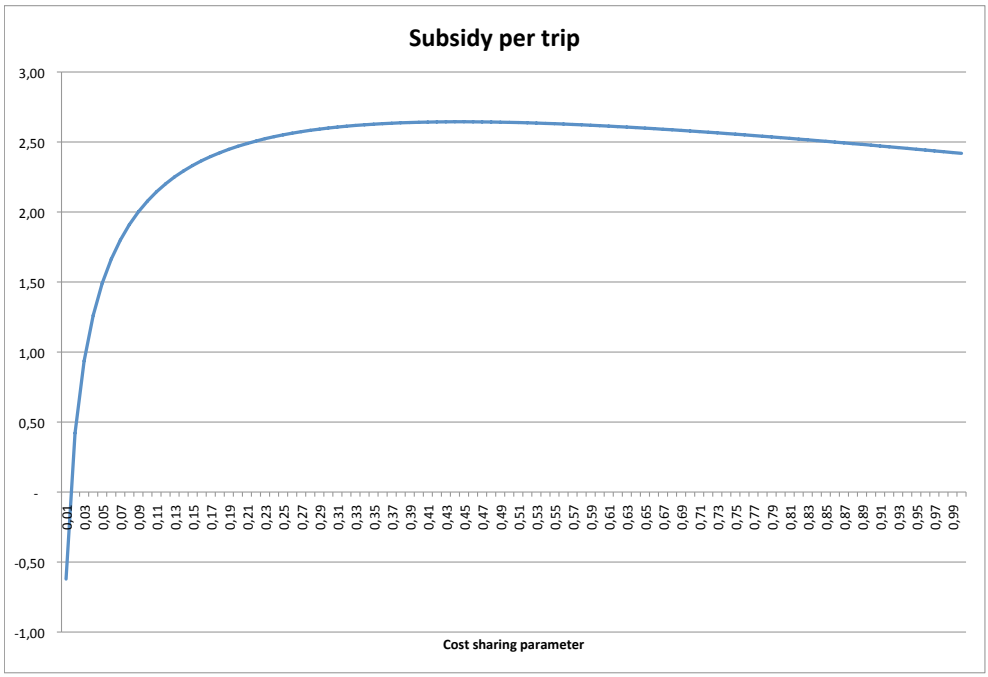

Figure 5: Subsidy per trip: model 2

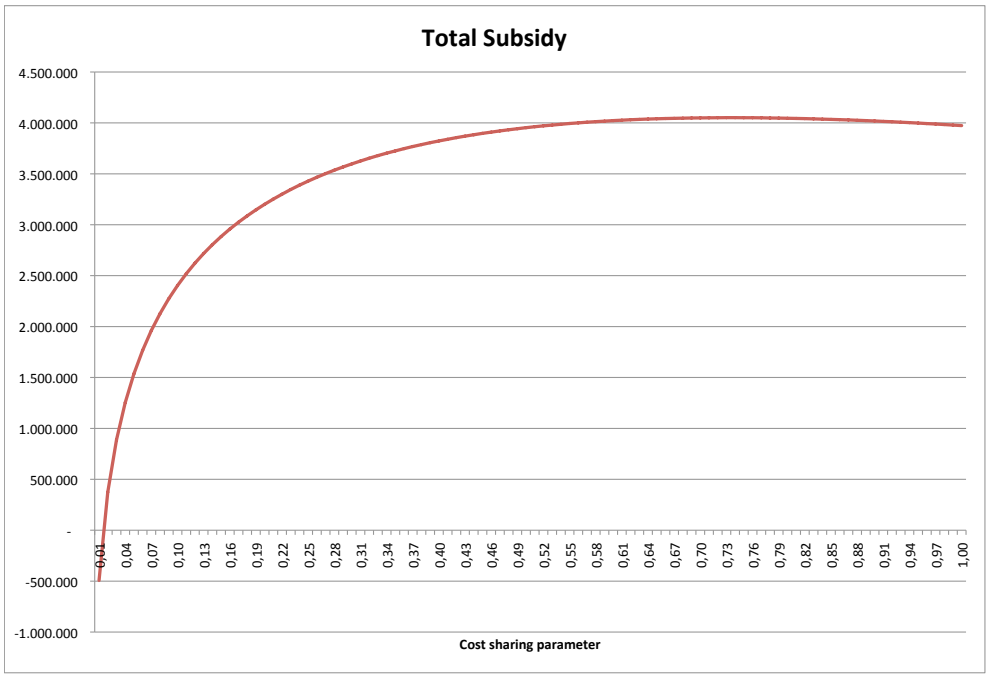

Figure 6: Total subsidy: model 2