Duration dependence and timevarying variables in discrete time duration models*

Anna Cristina D'Addio
OECD

Bo E. Honoré
Princeton University

June, 2006

Abstract

This paper considers estimation of a dynamic discrete choice model with second order state dependence in the presence of strictly exogenous time–varying explanatory variables. We propose a new method for estimating such models, and a small Monte Carlo study suggests that the method performs well in practice. The method is used to test for duration dependence in labour market spells for young people in France. The novelty in the application is that we are able to control for time–varying explanatory variables.

In a discrete time duration model, duration dependence will result in second order state dependence, and the paper therefore also adds to the literature on estimation of duration models with unobserved heterogeneity.

Keywords: Panel Data, Duration Models, Discrete Choice, Unemployment.
JEL Classification: C23, C41, C25, J60

*Mailing addresses: Anna Cristina D'Addio, OECD, Social Policy Division, Directorate for Employment, Labour and Social Affairs, 2 rue André-Pascal, F-75775 Paris, Cedex 16, France(D'Addio) and Princeton University, Department of Economics, Princeton NJ 08544-1021 (Honoré). The authors gratefully acknowledge financial support from the Marie-Curie fellowships scheme of the European Commission, the National Science Foundation, The Gregory C. Chow Econometric Research Program at Princeton University, and the Danish National Research Foundation (through CAM at The University of Copenhagen). We also thank two anonymous referees, Hidehiko Ichimura, Nandita Gawada, Aureo de Paula, Marina Sallustro and participants at the Tenth International Conference on Panel Data and at the 2002 European Meetings of the Econometric Society as well an numerous seminar participants for comments.

1 Introduction

This paper is concerned with estimation and testing of dynamic discrete choice panel data models with second order state dependence. These models are closely related to two–state, discrete time duration models with duration dependence.

An individual who has experienced an event in the past, is frequently more likely to experience the same event in the future than an individual who has not experienced the event. Examples where one might expect this include unemployment, union participation, accident occurrence, purchase decisions, etc. Heckman (1981a, 1981b, 1981c) discusses two explanations for this serial correlation in the context of standard discrete choice threshold crossing models. The first explanation is the presence of *true state dependence*, in which case lagged choices/decisions enter the model in a structural way as explanatory variables. For example, second order state dependence, which is the topic of this paper, refers to the case where the choice probability is allowed to depend on whether the event happened in the two most recent periods. The second source of persistence is the presence of serial correlation in the unobserved error. Heckman calls this source of serial correlation *spurious state dependence.* The serial correlation in the unobserved error is frequently modelled by assuming that the error is composed of a time–invariant component (unobserved heterogeneity) and a time–specific, serially independent component.

Distinguishing between the two sources of persistence is important if one wants to evaluate the effect of economic policies that temporarily change the outcome of the dependent variable. If the serial correlation is due to unobserved heterogeneity, then such a policy will not change future choice probabilities, whereas these will change if the dependence is due to true state dependence.

In this paper, we consider estimation of a discrete choice model that accommodates both second order state dependence and unobserved heterogeneity. Specifically, we considered the model:

$$y_{it} = 1\left\{x_{it}\beta_{y_{i,t-1}} + \delta_{i1}y_{it-1} + \delta_{2,y_{i,t-1}}y_{it-2} + \alpha_i + \varepsilon_{it} > 0\right\} \tag{1}$$

where x_{it} is a vector of strictly exogenous variables for individual i in time–period t, ε_{it} is an unobservable error term, α_i and δ_{i1} are unobservable individual-specific effects, and β_0, β_1, $\delta_{2,0}$ and $\delta_{2,1}$ are the parameters of interest to be estimated or tested. In this model, we allow the effect of x_{it} and y_{it} to depend on the lagged value of y_{it}. This is natural in situations where y_{it} is an indicator for whether an individual is in one of two states at time t. In that case, it is natural to allow the transition probabilities (the hazards rates), $P\left(y_{it} = 1 | \, y_{it-1} = 0, x_{it}, y_{it-2}\right)$ and $P\left(y_{it} = 0 | \, y_{it-1} = 1, x_{it}, y_{it-2}\right)$ to depend differently on x_{it} and y_{it-2}. (1) allows for that. In other situations, for example if one thinks of (1) as the result of some structural model in the spirit of Chintagunta, Kyriazidou, and Perktold (2001), it is natural to restrict β_0 and $\delta_{2,0}$ to equal β_1

and $\delta_{2,1}$, respectively. Therefore, in some of our discussion in this paper, we emphasize this case. Moreover, we will focus on a logit specification in which ε_{it} is i.i.d and logistically distributed, but the approach can also be used to construct estimators and tests for more semiparametric versions of the model.

The hypotheses $\beta_0 = \beta_1 = 0$ and $\delta_{2,0} = \delta_{2,1} = 0$ are of particular interest, and we will discuss how to test these in the model given in (1). On one hand, when both β_0 and β_1 equal 0, it is known that a conditional likelihood approach (see e.g., Chamberlain (1985)) can be used to estimate $\delta_{2,0}$ and $\delta_{2,1}$ and to test hypotheses regarding them. The resulting estimator and tests will have all the usual asymptotic properties such as consistency and root-n asymptotic normality (where n denotes the number of cross sectional units and the number of time periods, T, is assumed to be fixed). On the other hand, when $\delta_{2,0} = \delta_{2,1} = 0$, the model becomes consistent with a two–state duration model with no duration dependence.[1] In that case, the probability that an individual is in a given state at time t ($y_{it} = 1$) depends on whether the individual was in that state in the previous period ($y_{it-1} = 1$), but not whether she/he was in the state in periods before to that. Hence, that probability does not depend on the duration of time spent in the state.

The contribution of this paper is twofold. First, in section 2, we propose econometric methods for estimating the model in (1). As mentioned, when $\beta = 0$ a conditional likelihood approach can be used to estimate δ_2. See e.g. Magnac (2000) for a discussion of this. When $\delta_{2,0} = \delta_{2,1} = 0$ and δ_{i1} is constant, the model is similar to the model with first order state dependence discussed in Honoré and Kyriazidou (2000b).

The approach proposed in our paper generalizes the suggestions in Honoré and Kyriazidou (2000b). Like the estimator proposed by them, our estimator will depend on a bandwidth, which must shrink to zero (as $n \to \infty$) for the estimator to be consistent. As in Honoré and Kyriazidou (2000b), this will prevent our estimator from being root–n consistent. This is in contrast to the conditional likelihood method that one would use to estimate $\delta_{2,0}$ and $\delta_{2,1}$ in (1) if one knows that $\beta = 0$. That approach does not depend on a bandwidth, and it generally leads to a root–n consistent estimator. It is therefore interesting to note that the Wald–test of the hypothesis that $\beta = 0$ in (1) will have its usual χ^2–distribution (under the null), even if one considers asymptotics that holds the bandwidth fixed. A small scale Monte Carlo study presented in section 4 of this paper suggests that our estimator performs well in practice. A feasible bandwidth selection rule is also described in that section and its justification is given in Appendix 2.

The second contribution of this paper is to reconsider the issue of second order state depen-

[1] Alternatively, one might estimate $\delta_{2,0}$ and $\delta_{2,1}$ to be nonzero even if there is no true duration dependence, but the unobserved errors are serially correlated in a way that is more complicated than the one assumed here.

dence in youth unemployment by estimating (1) using the difference in the (monthly) number of unemployed between t and $t-1$, as the time–varying explanatory variable. We interpret this variable as a proxy for business cycle effects, and finding that it has a significant effect in (1) should be considered as evidence that *some* time varying variable plays a role in (1). A recent paper by Magnac (2000) estimates a model like (1), as well as more complicated multi–state models, to study the dynamics of youth labour market behavior. Unfortunately, existing methods did not allow that paper to control for business cycle effects by including time–varying (macroeconomic) variables. It is exactly this problem that motivates the econometric developments in the next section. We therefore estimate model (1) using the same data as Magnac (2000), but also including the first difference of the number of unemployed in a month as explanatory variable, and distinguishing according to the gender and the age. We find that the macroeconomic variable is statistically significant in some of the sub-samples considered, and that its effects vary according to the state occupied on the labour market. Specifically, we find that the probability of remaining unemployed is lower when the aggregate number of unemployed is falling, but that this variable has no or little effect on the transition probabilities for employed individuals. We also find that second order state dependence is important and our results point in the direction of negative duration dependence in both the duration of unemployment and the duration of employment (although the effect is statistically insignificant in some of the subsamples).

2 Estimation

As explained in Chamberlain (1985), one can test for duration dependence in a duration model with point sampling by considering the hypothesis that $\delta_2 = 0$ in the model

$$P\left(y_{i,t} = 1 \middle| \, y_{i,t-1}, y_{i,t-2}, \alpha_i\right) = \frac{\exp\left(\alpha_i + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)}{1 + \exp\left(\alpha_i + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)} \tag{2}$$

This model has been used, for example, by Chay, Hoynes, and Hyslop (2001) to estimate welfare participation, and a multinomial version of it has been used by Magnac (2000) to estimate labour market transitions. In Magnac (2000), the coefficient measuring the duration dependence, δ_2, is allowed to depend on which state the individual is in. In other words, δ_2 in (2) is replaced by $\delta_{2,y_{it-1}}$. This is the model given in (1), but without the explanatory variables, x_{it}.

When $\delta_2 = 0$, (2) is a Markov chain for a given individual, i. Both α_i and δ_{1i} are allowed to differ across individuals in an arbitrary way, which implies that the logit assumption places no restrictions on the transition probabilities. (2) is therefore a nonparametric Markov chain when $\delta_2 = 0$. However, the logit assumption *does* impose restrictions when δ_2 does not equal 0, and the

model given by (2) is therefore semiparametric. A parametric version of (2) can be obtained by making distributional assumptions on α_i and δ_{1i}.

The absence of time-varying effects in (2) is a limitation, which is sometimes undesirable in empirical applications. In order to allow for such time–varying effects, it is natural to generalize (2) by allowing the probability to also depend on a set of strictly exogenous time–varying explanatory variables, x_{it}. One way to do this, is to introduce x_{it} as additional explanatory variables in (2)

$$P\left(y_{i,t}=1|\, y_{i,t-1}, y_{i,t-2}, x_i, \alpha_i\right) = \frac{\exp\left(\alpha_i + x_{i,t}\beta + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)}{1+\exp\left(\alpha_i + x_{i,t}\beta + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)} \tag{3}$$

where x_i denotes the set of all the explanatory variables in all time periods for individual i. While (3) is a natural extension of (2), it differs from (2) in that it imposes parametric assumptions even when $\delta_2 = 0$. The model can therefore no longer be considered nonparametric in this case.

It seems natural to let $x_{i,t}$ and $y_{i,t-2}$ appear as in (3) because that makes it a simple generalization of a logit model. However, this functional form is unnatural if one interprets the model as a discrete time, two–state duration model. In that case (3) imposes restrictions between the exit rate from state 1 to state 0**,** and the exit rate from state 0 to state 1. For example, the same explanatory variables affect the two exit rates, and the relative importance of different explanatory variables is the same in the two rates (because they both depend on $x_{i,t}\beta$). Similarly, the relative importance of duration dependence is the same for the two states. To overcome this, it is therefore interesting to also consider a generalization of (3) that allows β and δ_2 to depend on the $y_{i,t-1}$,

$$P\left(y_{i,t}=1|\, y_{i,t-1}, y_{i,t-2}, x_i, \alpha_i\right) = \frac{\exp\left(\alpha_i + x_{i,t}\beta_{y_{i,t-1}} + \delta_{1i} y_{i,t-1} + \delta_{2,y_{i,t-1}} y_{i,t-2}\right)}{1+\exp\left(\alpha_i + x_{i,t}\beta_{y_{i,t-1}} + \delta_{1i} y_{i,t-1} + \delta_{2,y_{i,t-1}}\, y_{i,t-2}\right)} \tag{4}$$

The objects of interest in this paper are δ_2 and β in (3) (or $\delta_{2,0}$, $\delta_{2,1}$, β_0 and β_1 in (4)). The δ_2's capture the degree of duration dependence, and the β's account for the effect of time–varying explanatory variables. Since the point of departure is (2), it is of particular interest to test $\beta = 0$.

The paper by Honoré and Kyriazidou (2000b) considered a model with first order state dependence which is the same across individuals, i.e. δ_1 is not individual specific. Their paper emphasized the case where β does not depend on the $y_{i,t-1}$. However, their discussion of generalizations to multinomial models implicitly applies to models with first order state dependence in which β is allowed to depend on the $y_{i,t-1}$.

Non–Bayesian estimation of non-linear models like (3) and (4) is usually justified by asymptotic arguments. These asymptotic arguments can be based on letting either the number of individuals, n, or the number of time periods, T, (or both) increase to infinity. Since most relevant data sets have many more individuals than time periods, the asymptotic arguments used to justify the

proposed estimators of δ_2 and β will be based on letting the number of individuals, n, increase for fixed number of time periods, T.

When specifying and estimating (3) and (4) one has the choice of whether to take a random effects approach in which one specifies a distribution for (α_i, δ_{1i}), or a fixed effects approach which attempts to estimate the β's and the δ_2's without making any assumptions on the distribution of (α_i, δ_{1i}) and on the way this distribution relates to x_{it}. There is a trade-off between these approaches. The main advantage of the random effects approach is that it delivers a completely specified model. This means that one can calculate all probabilities of interest under any "what–if" scenario, provided, of course, that the model remains true. One disadvantage of the random effect approach is that it requires one to specify the distribution of (α_i, δ_{1i}) conditional on the time–varying explanatory variables in all the time–periods. If one assumes that the basic structure of the model is correct no matter how many time–periods one observes, this often leads to specifications that are inconsistent with the observed distribution of the time–varying explanatory variables, unless one assumes that (α_i, δ_{1i}) is independent of the time–varying explanatory variables. See Honoré (2002) for a discussion of this point. A second, and possibly more severe, disadvantage of the random effect approach is the initial conditions problem, since it also needs to specify either the distribution of (α_i, δ_{1i}) conditional on (y_{i1}, y_{i2}, x_i), or the distribution of (y_{i1}, y_{i2}) conditional on (α_i, x_i). If values of y_{it} before the start of the sample were also generated by (3) or by (4), then the relationship between (α_i, δ_{1i}) and (y_{i1}, y_{i2}) would depend on the time–varying variables before the start of the sample in a complicated way. Hence, if one models either the distribution of (α_i, δ_{1i}) conditional on (y_{i1}, y_{i2}, x_i), or the distribution of (y_{i1}, y_{i2}) conditional on $(\alpha_i, \delta_{1i}, x_i)$, one is implicitly modelling the behavior of the time–varying explanatory variables.

There are also severe drawbacks associated with a fixed effects approach like the one proposed in this paper. First, it will not always be possible to estimate a nonlinear model with fixed effects. For example, the approach used here places restrictions on the behavior of support of the time–varying explanatory variable that may not be satisfied, and it is not known how to estimate the model without these restrictions. Secondly, the semiparametric nature of fixed effects models may lead to estimates that are much less precise than the corresponding random effects estimates. Thirdly, and perhaps most seriously, the parameters estimated by the fixed effects approach often do not allow one to calculate objects such as the average effect of the explanatory variable on the probability that y_{it} equals 1 (because this will depend on the distribution of (α_i, δ_{1i})).

In this paper we pursue the fixed effects approach to estimate (3) and (4) , but it should be clear from the discussion above that this will only provide a partial answer to the question of how one should estimate such models.

The key idea behind the construction of estimators of fixed effects panel data models is to find some characteristic of the distribution of some random variable (which can be constructed from the data) that does not depend on the fixed effects. In a textbook linear panel data model with strictly exogenous explanatory variables, this is the conditional mean of the dependent variable minus the individual specific averages of the dependent variable. In the conditional likelihood approach, it is the distribution of the dependent variable conditional on the sufficient statistic for the fixed effects.

Our proposed methods for analyzing (3) and (4) are based on the following expressions which are derived in Appendix 1. They are all probability statements which are satisfied at the true parameter values and which do not depend on the fixed effects.

Define Ξ_{its} to be the sequence of all the y's for individual i, except for y_{it} and y_{is}, $\Xi_{its} = \{y_{i,1}, ..., y_{i,T}\} \setminus \{y_{i,t}, y_{i,s}\}$, where $3 \leq t < s \leq T-2$.

For $t = 3, ..., T-3$, we have

$$\begin{aligned} &P(\, y_{i,t} = 1 |\, \Xi_{it,t+1}, y_{i,t} \neq y_{i,t+1}, y_{i,t-1} = y_{i,t+2}, \\ &\qquad x_{i,t+1}\beta_1 = x_{i,t+2}\beta_1, x_{i,t+1}\beta_0 = x_{i,t+2}\beta_0, x_{i,t+2}\beta_{y_{i,t-1}} = x_{i,t+3}\beta_{y_{i,t-1}}) \end{aligned}$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)} \tag{5}$$

For $t = 3, ..., T-4$

$$\begin{aligned} &P(\, y_{i,t} = 1 |\, \Xi_{it,t+2}, y_{i,t} \neq y_{i,t+2}, y_{i,t-1} = y_{i,t+1} = y_{i,t+3}, \\ &\qquad x_{i,t+1}\beta_1 = x_{i,t+3}\beta_1, x_{i,t+1}\beta_0 = x_{i,t+3}\beta_0, x_{i,t+2}\beta_{y_{i,t-1}} = x_{i,t+4}\beta_{y_{i,t-1}}) \end{aligned}$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)} \tag{6}$$

Finally for $t = 3, ..., T-5$ and $s = t+3, ..., T-2$

$$\begin{aligned} &P(\, y_{i,t} = 1 |\, \Xi_{its}, y_{i,t} \neq y_{i,s}, y_{i,t-1} = y_{i,s-1}, y_{i,t+1} = y_{i,s+1}, \\ &\qquad x_{i,t+1}\beta_1 = x_{i,s+1}\beta_1, x_{i,t+1}\beta_0 = x_{i,s+1}\beta_0, x_{i,t+2}\beta_{y_{i,t+1}} = x_{i,t+3}\beta_{y_{i,t+1}}) \end{aligned}$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)} \tag{7}$$

Although equations (5), (6) and (7) are derived by brute force in appendix 1, they are motivated by a conditional likelihood approach. Consider a version of (4) with no exogenous variables. Such

a model could be estimated by the conditional likelihood approach (see Chamberlain (1985)). This would involve conditioning on the first two and the last two observations, the sum of all the observations, as well as $\sum_t y_{it}y_{it-1}$. Equations (5), (6) and (7) are derived by considering any subset of the data starting at time period $t-2$ and ending in period $s+2$. We then first condition on the first two and the last two observations as well as on the sum of all the observations and $\sum_t y_{i,t}y_{i,t-1}$, as one would do in a model with no exogenous variables. Because of the exogenous variables, this leads to expressions that depend on (α_i, δ_{1i}). To eliminate the terms depending on (α_i, δ_{1i}) one must condition on events related to the exogenous variables. As will be seen shortly, this is costly in terms of the rate of convergence of the proposed estimator. It is therefore desirable to minimize this type of conditioning. The smallest amount of conditioning done in this way is obtained if one further conditions on all but two of the dependent variables (recall that we also condition on the sum of all of the dependent variables, so this is the most conditioning that one can do without making the distribution degenerate). This is why the probabilities in (5), (6) and (7) condition on all the values of y_{it}, except for two. This gives rise to conditioning on events of the type $\Xi_{its}, y_{i,t} \neq y_{i,s}$. When there are no time–varying explanatory variables, one would want to be sure to be implicitly conditioning on $\sum_t y_{i,t}y_{i,t-1}$ (because then one would have conditioned on the sufficient statistic for (α_i, δ_{1i})). When $s = t+1$, this is achieved by conditioning on $y_{i,t-1} = y_{i,t+2}$. When $s = t+2$, one would need to condition on $y_{i,t-1} = y_{i,t+3}$. However, for the case where $y_{i,t-1} = y_{i,t+3} \neq y_{i,t+1}$ to deliver expressions that do not depend on (α_i, δ_{1i}), one needs to condition on $x_{i,t+1} = x_{i,t+2} = x_{i,t+3} = x_{i,t+4}$. The curse of dimensionality suggests that this is undesirable (relative to the conditioning in (6)) when x is continuously distributed, and we therefore ignore these terms. When $s > t+2$, conditioning on $\sum_t y_{i,t}y_{i,t-1}$, implies that $y_{i,t-1} + y_{i,t+1} = y_{i,s-1} + y_{i,s+1}$. However, conditioning on either $\{y_{t-1} = 1, y_{t+1} = 0, y_{s-1} = 0, y_{s+1} = 1\}$ or on $\{y_{t-1} = 0, y_{t+1} = 1, y_{s-1} = 1, y_{s+1} = 0\}$ leads to expressions that do not involve (α_i, δ_{1i}) (for all values of β_0, β_1, $\delta_{2,0}$ and $\delta_{2,1}$) only if $x_{i,t+1} = x_{i,t+2} = x_{i,s+1} = x_{i,s+2}$. Since these are based on three equalities rather than two (as in the other expressions), we will not use these expressions[2]. The probabilities in (5), (6) and (7) are therefore the only cases in which α_i and/or δ_{1i} do not appear (for all values of β_0, β_1, $\delta_{2,0}$ and $\delta_{2,1}$), and which require conditioning on only two events of the type $x_{it} = x_{is}$.

If one assumes that $\beta = 0$, then the $x\beta$ terms in (5), (6) and (7) are all equal by construction and one can use them to set up a "partial" conditional likelihood function in order to estimate δ_2. This

[2]In the case when x_{it} has a continuously distributed element, there is no loss (asymptotically) in dropping these terms from the objective functions below. However if all of the elements of x_{it} are discrete, then matching on events like $x_{i,t+1} = x_{i,t+2} = x_{i,s+1} = x_{i,s+2}$ could potentially improve on the efficiency of the estimators.

would be inefficient relative to the conditional maximum likelihood approach because the additional conditioning eliminates variability which is informative about δ_2. If β does not equal zero, then one can mimic the line of argument in Honoré and Kyriazidou (2000b) and use nonparametric regression techniques to essentially construct a sample analog of the conditional likelihood function based on (5), (6) and (7). The problem with this is that the conditioning sets in these equations depend on β. This can be overcome by noting that the calculations in the appendix 1 also imply the following probability statements.

For $t = 3, ..., T-3$, we have

$$P\left(y_{i,t} = 1 | \, \Xi_{it,t+1}, y_{i,t} \neq y_{i,t+1}, y_{i,t-1} = y_{i,t+2}, x_{i,t+1} = x_{i,t+2} = x_{i,t+3}\right)$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)} \tag{8}$$

For $t = 3, ..., T-4$

$$P\left(y_{i,t} = 1 | \, \Xi_{it,t+2}, y_{i,t} \neq y_{i,t+2}, y_{i,t-1} = y_{i,t+1} = y_{i,t+3}, x_{i,t+1} = x_{i,t+3}, x_{i,t+2} = x_{i,t+4}\right)$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)} \tag{9}$$

Finally for $t = 3, ..., T-5$ and $s = t+3, ..., T-2$

$$P\left(y_{i,t} = 1 | \, \Xi_{its}, y_{i,t} \neq y_{i,s}, y_{i,t-1} = y_{i,s-1}, y_{i,t+1} = y_{i,s+1}, x_{i,t+1} = x_{i,s+1}, x_{i,t+2} = x_{i,t+3}\right)$$

$$= \frac{\exp\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)} \tag{10}$$

When x_{it} is continuously distributed, we therefore propose to estimate (β, δ_2) by maximizing[3]

$$\sum_i q_i\left(b_0, b_1, d_{2,0}, d_{2,1}\right) \tag{11}$$

[3]When elements of x_{it} are discrete, it is not necessary to use a kernel for those components of x_{it}.

where

$$q_i = \sum_{t=3}^{T-3} 1\{y_{i,t} \neq y_{i,t+1}\} 1\{y_{i,t-1} = y_{i,t+2}\} K\left(\frac{x_{i,t+1} - x_{i,t+2}}{h}\right) K\left(\frac{x_{i,t+2} - x_{i,t+3}}{h}\right) \tag{12}$$

$$\log\left(\frac{\exp\left(y_{i,t}\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+1}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+3}\right)\right)}\right)$$

$$+ \sum_{t=3}^{T-4} 1\{y_{i,t} \neq y_{i,t+2}\} 1\{y_{i,t-1} = y_{i,t+1} = y_{i,t+3}\} K\left(\frac{x_{i,t+1} - x_{i,t+3}}{h}\right) K\left(\frac{x_{i,t+2} - x_{i,t+4}}{h}\right)$$

$$\log\left(\frac{\exp\left(y_{i,t}\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,t+2}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,t+4}\right)\right)}\right)$$

$$+ \sum_{t=3}^{T-5} \sum_{s=t+3}^{T-2} 1\{y_{i,t} \neq y_{i,s}\} 1\{y_{i,t-1} = y_{i,s-1}\} 1\{y_{i,t+1} = y_{i,s+1}\} K\left(\frac{x_{i,t+1} - x_{i,s+1}}{h}\right) K\left(\frac{x_{i,t+2} - x_{i,s+2}}{h}\right)$$

$$\log\left(\frac{\exp\left(y_{i,t}\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)\right)}{1 + \exp\left(\left(x_{i,t} - x_{i,s}\right) b_{y_{i,t-1}} + d_{2,y_{i,t-1}}\left(y_{i,t-2} - y_{i,s-2}\right) + d_{2,y_{i,t+1}}\left(y_{i,t+2} - y_{i,s+2}\right)\right)}\right),$$

$K(\cdot)$ is a kernel and h is a bandwidth that approaches 0 as the number of observations increase to ∞. In the empirical application in the next section we choose K to be an Epanichnikov kernel[4] and we experiment with the bandwidth h.

It is interesting to note that the first two sums in q_i separate into a sum of two terms, one that depends on $(b_0, d_{2,0})$ and one that depends on $(b_1, d_{2,1})$. This means that one can estimate the parameters of the two transition probabilities (from state 0 to state 1, and from state 1 to state 0) separately by considering only the first two sums in q_i. On the other hand, if one wants to use all the terms in (12) that are informative about $(\delta_{2,0}, \beta_0)$ (or$(\delta_{2,1}, \beta_1)$), then one must simultaneously estimate $(\delta_{2,1}, \beta_1)$ (or $(\delta_{2,0}, \beta_0)$).

By arguments similar to those in Honoré and Kyriazidou (2000b),

$$\sqrt{nh^{2k}}\left(\widehat{\theta} - \theta\right) \widetilde{\Longrightarrow} N\left(0, \Gamma^{-1} V \Gamma^{-1}\right) \tag{13}$$

[4]This kernel is efficient (in a particular sense) in other settings, here we use it primarily because of its simplicity and because it is continuous and has finite support (which means that many of the terms in the objective function are 0).

under suitable regularity conditions, where $\widehat{\theta}$ and θ denote $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_{2,0}, \widehat{\delta}_{2,1}\right)$ and $(\beta_0, \beta_1, \delta_{2,0}, \delta_{2,1})$, respectively, and k is the dimensionality of x_{it}. $Avar\left(\widehat{\theta}\right) = \frac{1}{nh^{2k}}\Gamma^{-1}V\Gamma^{-1}$ can be estimated by

$$\left(\sum_i q_i''\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_{2,0}, \widehat{\delta}_{2,1}\right)\right)^{-1} \tag{14}$$
$$\left(\sum_i q_i'\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_{2,0}, \widehat{\delta}_{2,1}\right)^T q_i'\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_{2,0}, \widehat{\delta}_{2,1}\right)\right)\left(\sum_i q_i''\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_{2,0}, \widehat{\delta}_{2,1}\right)\right)^{-1}.$$

Since the contribution of this paper is to allow for time–varying explanatory variables in models like (2), it is useful to consider a test of $\beta_0 = \beta_1 = 0$. A natural test would be the Wald test based on (13) and (14). Such a test will be justified in the sense that it will have the usual χ^2 –distribution (under the null) even if the bandwidth is a fixed constant (i.e., does not decrease to 0 as the sample size increases). The reason for this is that if the true β_0 and β_1 equal 0, then (for fixed h), $E[q_i]$ would be maximized by making the terms inside the log's equal to the probability that $y_{it} = 1$ (conditional on the events in the indicator function for each of the terms in the sums in q_i). This happens at $b_1 = b_0 = 0$, $d_{2,0} = \delta_{2,0}$ and $d_{2,1} = \delta_{2,1}$. It therefore follows from standard asymptotics for M–estimators that if β equals 0, then the proposed estimator will be consistent and asymptotically normal as $n \to \infty$, but with h (and T) fixed. This in turn implies that the Wald test of the hypothesis $\beta = 0$ that uses $\widehat{Avar}\left(\widehat{\theta}\right)$ in (14) will have the usual asymptotic χ^2 –distribution under the null (for h fixed). In other words, under the null, the Wald test has the same asymptotic distribution as it would in a parametric model. Unfortunately, it does not follow that the Wald test is consistent for h fixed. This is because the estimator of (β_0, β_1) is not guaranteed to be consistent under the alternative.

A random effects approach to estimating (3) or (4) has to deal with the fact that the model does not specify the distribution of the first two observations conditional on $(x_{i1}, ..., x_{iT}, \alpha_i, \delta_{1i})$. If the first two observations are also generated from (3) or (4), then their distribution (given $(x_{i1}, ..., x_{iT}, \alpha_i, \delta_{1i})$) will depend on the distribution of the explanatory variables in time periods prior to the sample (given $(x_{i1}, ..., x_{iT}, \alpha_i, \delta_{1i})$) which is typically unspecified. This is the initial conditions problem. It is therefore important to note that the probability statements in (5), (6) and (7) all condition on the first two observations for an individual. This is because Ξ_{its} always contains y_{i1} and y_{i2}. This means that the estimators based on (5), (6) and (7) do not suffer from the initial conditions problem.

The difference between basing an estimator on (8), (9) and (10) rather than on (5), (6) and (7) is bigger than it might appear. The reason is that in the latter case, the dimensionality of the nonparametric problem is $2k$, whereas it would be either 2 or 3 in the former (depending on

whether $k = 1$ or $k > 1$). The curse of dimensionality implies that this can be very important. One way to exploit (5), (6) and (7) is to note that they can be used to construct moment conditions that are based on conditional probabilities given values of $x_{it}\beta$. However, it is not clear whether those moment conditions identify β. One should therefore combine them with the moment conditions based on conditional probabilities given values of x_{it}. If the former satisfy a local identification condition, then this would lead to an estimator whose rate of convergence is driven by the moment conditions based on the conditional probabilities given values of $x_{it}\beta$ (by an argument similar to that in Honoré and Hu (2004)). Since the application discussed in the next section has only one time–varying explanatory variable, we do not pursue this here.

3 The Empirical Application

In this section we use the methods outlined above to estimate equation (4) for a dataset composed of employment outcomes for French youth. This is the same data set used by Magnac (2000). Since one of the specifications used by him is similar to (2), it is especially interesting to test $\beta_0 = \beta_1 = 0$.

The data are extracted from the 1990-1992 waves of the French Labour Force Survey and an additional survey held in 1992 (Module Jeunes) focusing on individuals and their family background since they were 16 years old up to the survey date.

The French Labour Force Survey is a rotating panel on three years concerning approximately 60,000 households. One third of the sample is renewed each year implying that 20,000 households are present in the survey at three successive dates. The sub–sample used here consists of 5,824 young individuals aged between 18 and 29 in 1992. More specifically, it contains information about their histories on the labour market for the period going from January 1989 to March 1992. Surveys took place at three dates, January 1990, March 1991 and March 1992. At each survey date, the interviewer attempted to rebuild the individuals' labour market history through questions about their activities in each month of the previous year. The interviewer also asked about the current labour market activity of the individual. As a result of this, the information about the month of February 90 is missing.

The survey sampling scheme makes spells in various states left–censored at the beginning of the observation period and right–censored at the end. This complications matters if one models the durations using standard continuous time duration models. See d'Addio and Rosholm (2002). However, as discussed in the previous section, the initial conditions problem (which is similar to the problem of left censoring in duration models) plays no role for the approach proposed here. Right–censoring is also not a problem, provided that the censoring time is exogenous. This is

clearly the case here, since censoring time is the final survey date.

For the empirical application, the dependent variable is defined to be 1 if the individual is unemployed or out of the labour force, and 0 otherwise. This differs from Magnac (2000) who estimated a multinomial model with a more disaggregated definition of the labour market states. If an individual is in school at the start of the survey, we ignore the data until the moment she/he enters the labour market. Later periods of schooling are treated as employment.

Table 1 presents summary statistics regarding the dynamic behavior of the dependent variable, non-employment,

[Insert Table 1 here]

Although Table 1 does not control for individual–specific heterogeneity, it appears that state dependence of order 1 (see the first two lines of Table 1) and order 2 (the last four lines of Table 1) is important in the non-employment behaviour of young individuals. Without first order state dependence, $\Pr(y_t = 1|y_{t-1} = 0)$ would equal $\Pr(y_t = 1|y_{t-1} = 1)$, but, as one would expect, the probability of nonemployment is much higher for an individual who was not employed in the previous month. Without second order state dependence, $\Pr(y_t = 1|y_{t-1} = 0, y_{t-2} = 0)$ would equal $\Pr(y_t = 1|y_{t-1} = 0, y_{t-2} = 1)$, and $\Pr(y_t = 1|y_{t-1} = 1, y_{t-2} = 0)$ would equal $\Pr(y_t = 1|y_{t-1} = 1, y_{t-2} = 1)$. However, the data clearly suggest that the employment status in time period $t-2$ plays a role for the employment history in period t, and this role is consistent with a decreasing hazard in both employment and unemployment.

As mentioned earlier, it is interesting to control for macro–economic business cycle effects when studying unemployment. For this purpose, we use the difference in the number of French unemployed between t and $t-1$ as the time–varying explanatory variable. Strictly speaking, this variable will not satisfy the regularity conditions needed for the consistency of the estimator discussed in the previous section because it does not vary by individual. However, the Wald tests will have the correct asymptotic distribution (under the null).

Figure 1 shows the time–path for the difference between t and $t-1$ in the number of French unemployed over the relevant time–period.

[insert Fig. 1 here]

We think of this variable as a proxy for broad business cycle effects and more particularly for labour market conditions.

The French debate on youth unemployment has focused mainly on two issues, the costs and the inadequate qualifications of young workers. The high level of youth unemployment has also fed the debate about the need for structural adjustments in the labour market. It has been argued that the high rate of youth unemployment is the result of the high level of their wages Moghadham (1993)). Proposals are therefore regularly advanced for the establishment of a "youth minimum wage", like in Belgium and the Netherlands. However, others have argued that labour costs are not sufficient to explain the high rates of unemployment existing in France (Bruno and Cazes (1997)). One reason is that there is a variety of mechanisms bringing youth wages below the level of the minimum wage (SMIC)[5]. For instance, the wages of people aged between 16 and 18 can be fixed on the basis of a floor equal to 80 per cent of the SMIC. Moreover, young people can be employed under "entry" contracts that enable firms to reduce their labour costs[6].

Several arguments suggest that macroeconomic effects, frequently summarized by the overall unemployment rate, may have a more important impact on the unemployment of young individuals than of adults. One of the reasons is that turnover is related to the cost of investment in human capital. For young people this cost is lower compared to that for the adults. Therefore firms are more likely to separate from younger workers. The same is true because of the weaker protection offered by the employment legislation to people hired under an apprenticeship or a limited term contract. It therefore seems that young people are more vulnerable than the adults to unemployment. The flow of young people into unemployment is often considered as a mechanism of adjustment: during a recession firms first cut jobs held by young people to protect those of adults. This is consistent with the LIFO criterion in the firing decisions of young individuals as suggested by Layard, Nickell, and Jackman (1991). Young individuals are also more mobile than the adults. Over the period 1970-1994, their inflow into employment from unemployment or from outside the labour force is higher than that of the adults. The same is true for the outflow from employment into unemployment or into inactivity. Young people, whether economic conditions allow for it, are more likely than adults to "shop around" before finding a stable job and to quit voluntarily (see e.g. Blanchflower (1996)) because their opportunity cost of changing jobs is lower than the adults'. This phenomenon has been found to be especially important for very young workers and to become less important with age.

It is widely recognized that labour market opportunities increase with qualifications. However,

[5]The SMIC, established in January 1970, is a gross hourly wage indexed to the consumer price index; in July of each year it is adjusted by at least half the increase in the average hourly wage (Bruno and Cazes (1997)).

[6] There were almost 600.000 such contracts in 1996, corresponding to 20% of the labour force aged under 26; see Cette, Cuneo, Eyssartier, and Gautié (1996)

several studies have shown that it is difficult for young people, whatever their educational attainments, to achieve a rapid and direct entry into the labour market (CSERC (1996)). The entry into professional life often starts with an unemployment spell. In 1995, 31.1% of young unemployed people in France were looking for a first job (Eurostat (1997)). Since the labour market experience of the young candidates is an important factor in hiring decisions, many young people start with a limited term contract and temporary jobs. On average in the European Union, 35% of employed young people (aged less than 25) are employed under such a contract (Eurostat (1997)). The distribution of these contracts appears to be related to the age of the workers.

In summary, there are many reasons to think that the employment dynamics for young workers is substantially different from adults'. We will therefore consider subsets of the sample based on age. Since the behaviour of young individuals has been shown to differ between men and women, we will also consider subsets of sample based on gender. Specifically we consider separately samples of men and women aged less than 25 and over 25 (but under 30), respectively.

All the data used in this paper were collected by the French National Institute of Statistics (INSEE).

3.1 Results

In this section we present the results from estimation of (3) and (4) using the data described above. As previously mentioned, it is likely that duration dependence and the effect of the business cycle differ across age groups. One might also expect these to differ according to gender. In order to account for this, we estimate the models using the total sample as well as four different sub–samples. Specifically, we consider the following samples: (1) the overall sample; (2) men in the sample aged less than 25;[7] (3) women in the sample aged less than 25; (4) men in the sample aged 25 or more; and (5) women in the sample aged 25 or more. The individuals who turn 25 during the sampling period will be used in both the under 25 and the over 25 samples, but only for the periods for which they were under 25 or over 25, respectively. This makes some of the samples unbalanced, but the number of observations for an individual is exogenous.

The contribution of a single individual to the objective function is a sum of $\binom{T_i-4}{2} = \frac{(T_i-4)(T_i-5)}{2}$ terms (although many of them will be 0), where T_i is the number of observations for individual i. Heuristically, this seems to give too much weight to individuals with a large value of T_i. Following Honoré and Kyriazidou (2000a), we therefore use a weighted version of (3) with weights given by $\frac{1}{T_i-4}$, with the standard errors adjusted appropriately. This does not affect the consistency of

[7] This corresponds to the ILO definition of young people.

the estimator, but while we do not claim the weights to be optimal, we suspect that they will result in an improvement in efficiency over the unweighted estimator. Our motivation for using these weights is that by conditioning on the first two and last two observations, we essentially have $T_i - 4$ observations for each individual. The contribution to the objective function is then defined by all pairwise comparisons of observations taken from those. Hence, we have $\binom{T_i-4}{2}$ terms. The deviations from mean estimator of the standard linear panel data model can also be written as the minimizer of a weighted sum (across individuals) of terms that are defined by all pairwise comparisons of observations for that individual. The weight given to an individual in that case is the inverse of the number of observations for that individual minus 1. This is the reason for using $\frac{1}{T_i-4}$ as the weight.

The estimator defined by maximizing (11) depends on a bandwidth and a kernel to be chosen by the researcher. The choice of kernel is usually less critical than the choice of bandwidth in applications of semi– and nonparametric methods. We therefore use only one kernel, which is the Epanichnikov kernel given by

$$K(u) = \max\left\{0, 1 - u^2\right\} \tag{15}$$

The fact that $K(\cdot)$ has bounded support implies that many of the terms in the objective function are 0. This makes it computationally much more tractable than, say, a normal kernel. Since we expect that the choice of bandwidth is more important than the choice of kernel, we experiment with different values of the bandwidth h.

The results from estimating (4) are presented in Table 2 and in Figures 2–6. The first column of Table 2 presents results from a linear probability model with individual specific intercepts, individual specific coefficients on $y_{i,t-1}$, and with y_{it-2} and x_{it} interacted with y_{it-1}. The second column presents the results from the corresponding logit model that treats α_i and δ_{i1} as parameters to be estimated and allows the coefficients on y_{it-2} and x_{it} to depend on y_{it-1},

$$\begin{aligned}\left(\widehat{\delta}_{20}, \widehat{\delta}_{21}, \widehat{\beta}_0, \widehat{\beta}_1\right) = \underset{d_{20},d_{21},b_0,b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \max_{a_i,\delta_{1i}} \sum_t \big\{ & y_{it} \log\left(\Lambda\left(x_{it} b_{y_{i,t-1}} + \delta_{1i} y_{i,t-1} + \delta_{2,y_{i,t-1}} y_{i,t-2} + a_i\right)\right) \\ & + (1 - y_{it}) \log\left(1 - \Lambda\left(x_{it} b_{y_{i,t-1}} + \delta_{1i} y_{i,t-1} + \delta_{2,y_{i,t-1}} y_{i,t-2} + a_i\right)\right)\big\}\end{aligned} \tag{16}$$

As mentioned in the previous section, y_{it} is missing for the month February 1990. Since this is exogenous, the sums in (12) can be replaced by the same sums excluding terms that involve February of 1990 without affecting the asymptotic properties of the estimator (except for the loss of efficiency). For the linear probability model and the maximum likelihood estimator defined in (16), we also ignore terms that involve February of 1990. This also will not affect the interpretation

of our results, although it does mean that (16) will not yield the maximum likelihood estimators (but rather quasi maximum likelihood estimators with the same properties).

[Table 2 to be inserted here]

We also calculated the estimator defined by maximizing (11). Since one might worry that the estimator proposed here is sensitive to the bandwidth, we present the results from it by plotting the 95% confidence intervals for each parameter as a function of the bandwidth. These are given in Figures 2–6.

The results that emerge from Figures 2–6 can be summarized as follows. $\delta_{2,0}$ and $\delta_{2,1}$ are both estimated to be positive for all the samples. This is consistent with negative duration dependence in both the duration of employment and unemployment[8]. The estimates are significantly different from 0 when one uses the whole sample. The coefficient on the strictly exogenous explanatory variable is positive or zero for people who are currently employed and essentially zero for individuals who are unemployed. These findings are fairly robust to the choice of bandwidth although the actual magnitude of the estimates do depend on it. The fixed effects estimates in Table 2 are somewhat different when it comes to the duration dependence parameters. There the estimates are statistically significant and suggest positive duration dependence. This is consistent with findings of the Monte Carlo results reported in the next section. They suggest that the incidental parameters problem which is associated with the linear probability model and the estimates based on (16), results in severe bias in the direction of positive duration dependence.

[insert Fig. 2–6 here]

4 Monte Carlo Investigation

In this section we report the results of a small Monte Carlo study designed to investigate the behaviour of the estimator defined in equation (11). In the first part of the Monte Carlo study, we focus on the models (and estimators) that restrict the coefficients on x_{it} and y_{it-2} to not depend on y_{it-1}. The reason for this is that it makes the model more similar to a standard logit model and that it makes the Monte Carlo study presented here more comparable to the one in Honoré and

[8]The interpretation of $\delta_{2,0}$ and $\delta_{2,1}$ is somewhat counterintuitive as positive values imply negative duration dependence. The reason is that $\delta_{2,1} > 0$ means that somebody who is current unemployed is more likely to be unemployed next period if she/he was also unemployed last period. On the other hand $\delta_{2,0} > 0$ means that somebody who is currently employed is more likely to be unemployed next period if she/he was unemployed last period. Both are consistent with negative duration dependence.

Kyriazidou (2000b). In the second part we investigate the behaviour of the estimator in a situation where the explanatory variable is the one used in the application above, and the degree of serial correlation is the same as that found in the full sample used there. We also investigate an ad hoc bandwidth selection method.

4.1 Performance of the Estimator in a Relatively Simple Situation

We consider twenty versions of the model

$$P\left(y_{i,t}=1|\, y_{i,t-1}, y_{i,t-2}, x_i, \alpha_i\right) = \frac{\exp\left(\alpha_i + x_{i,t}\beta + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)}{1+\exp\left(\alpha_i + x_{i,t}\beta + \delta_{1i} y_{i,t-1} + \delta_2 y_{i,t-2}\right)} \tag{17}$$

All designs have x_{it} distributed as independent normal random variables with mean zero and variance 2. All the data sets have $n = 1000$ and $T = 10$, and data is generated from (17) for time periods 1 to $T+10$, where $y_{i,1}$ and $y_{i,2}$ are generated from (17) with $y_{i,0} = 0$ and $y_{i,-1} = 0$. Only the last T time periods are used for the estimation. This essentially means that the initial observations are drawn from the stationary distribution of (y_{it}, y_{it+1}). The difference in the designs comes from the distributions of α_i and δ_{1i} and the values of β and δ_2. We consider five distributions of α_i and δ_{1i}:

Design 1	$\alpha_i = 0$	$\delta_{1i} = 1$
Design 2	$\alpha_i = 0$	$\delta_{1i} \sim N(1,1)$
Design 3	$\alpha_i = 0$	$\delta_{1i} \sim N(1,4)$
Design 4	$\alpha_i \sim N(1,1)$	$\delta_{1i} = 1$
Design 5	$\alpha_i \sim N(1,4)$	$\delta_{1i} = 1$

For each of these we consider four combinations of β and δ_2: $(1,1)$, $(2,1)$, $(1,2)$ and $(2,2)$. These designs are chosen to reflect a mix of situations ranging from no unobserved heterogeneity in Design 1, to heterogeneity of magnitude similar to the underlying logit error in Designs 3 and 5. This is important as we expect that the amount of heterogeneity will interact with the precision with which the duration dependence is estimated. The values of β and δ_2 vary the importance of the explanatory variable and the duration dependence. Since no kernel–matching is needed when $\beta = 0$, we expect the estimator proposed here to do less well when β is larger.

For each dataset, we estimate β and δ_2 using the estimator defined by maximizing (11) as well as the maximum likelihood estimator that treats the α_i's and a common δ_1 as parameters to be estimated. Since one might suspect that the estimator (11) is sensitive to the choice of bandwidth, we calculate the estimator using 251 values of h evenly spaced between 0.52 and 5.52. We use the Epanichnikov kernel in (15).

Figures 7–11 summarize the results for the estimator defined by maximizing (11). The relationship between the absolute value of the median biases and the bandwidths are depicted by the thinner lines. The thicker lines show the median absolute deviations (from the median). *Ex-ante* one would expect this for the estimator proposed here, the trade-off is that larger bandwidths should result in more biased estimators with less variability. The second part of this is confirmed by the downward slope in median absolute deviations in Figures 7–11.

[insert Fig. 7–11 here]

Somewhat surprisingly, the absolute value of the biases tends to be U–shaped. This is especially true for the estimator of δ_2. We attribute this behaviour to a trade-off between two different sources of bias. The first is that large bandwidth will result in large biases (in absolute values). This explains the upward slope in the bias as a function of bandwidth. We attribute the initial downward slope to the fact that even though small bandwidths will result in precise matching, it will also lead to a small number of non–negative terms in the objective function for the estimator. This means that the well–known small sample biases in the logit maximum likelihood estimator are likely to be relevant for small bandwidths.

It is also surprising that the bias in the estimator of β is so small, and that it only increases slightly, if at all, as a function of the bandwidth. This suggests that a very large bandwidth is appropriate if one is only interested in the coefficient on the strictly exogenous explanatory variables.

Table 3 presents the results for the estimator that is based on treating the individual specific effects as constants to be estimated. For small T, this estimator should be subject to the incidental parameters problem and for designs 2 and 3 it is also misspecified because it treats the individual specific effect as constant (which is not interacted with the lagged dependent variable). It is clear from Table 3 that this estimator does not perform well for the design considered here.

4.2 Performance of the estimator when the design mimics the real data

The Monte Carlo study in the previous sub–section is designed to illustrate the behaviour of the estimator proposed in this paper in an ideal setting like the ones that are often considered when evaluating newly proposed procedures. However, it is clear that such a design is much "nicer" than what one would expect in many applications. In particular, the explanatory variable in the paper's application is a macroeconomic variable which is the same for all individuals in the sample. Moreover, the dependent variable used there, exhibits serial correlation that is much more extreme

than the designs in the previous subsection. In this subsection, we therefore present the results of a Monte Carlo experiment in which the data generating process much more closely resembles the data used in the empirical application. Specifically, we generate 500 dataset as follows

- The sample size is 3590 (equaling the total sample sized used in the application)
- $\delta_{2,0} = 0.6$, $\delta_{2,1} = 0.6$, $\beta_1 = 0$ and $\beta_0 = 0.6$. These values are similar to the estimates found in Figure 2.
- The number of observations used for each individual is the same as in the application. This results in an unbalanced panel and as in the application, we do not use the data for one period (corresponding to February 1990 in the application.)
- The initial observations are drawn by actually simulating the model for 49 time–periods for each individual (with the last 39 representing the period considered in the application). The first two are drawn by

$$y_{i1} = 1\left\{x_{i1}\left(\beta_1 + \beta_0\right)/2 + \alpha_i + \varepsilon_{i1} > 0\right\}$$

$$y_{i2} = 1\left\{x_{i2}\beta_{y_{i1}} + y_{i1}\delta_{1i} + \alpha_i + \varepsilon_{i2} > 0\right\}$$

  while the others are drawn from (1). The explanatory variable for the first ten periods is defined by $x_{i,11-t} = x_{i,11+t}$. After that, the explanatory variable is the same than that used in the empirical application.
- $\alpha_i \sim N\left(-5.5, 9\right)$ and $\delta_{1i} \sim N\left(2, 9\right)$. With these, we have $P\left(y_{it} = 1 \middle| y_{it-1} = 1\right) = 0.926$ and $P\left(y_{it} = 1 \middle| y_{it-1} = 0\right) = 0.027$. These values are close to the sample averages of 0.927 and 0.024 found in the data used in the previous section's application.

[insert Fig. 12–13 here]

The results are presented in Figures 12 and 13. Figure 12 reports the (absolute) median bias and the median absolute deviation (from the median) for thirteen different bandwidths. The figure suggests that the estimator is essentially unbiased for the design considered here. This is consistent with the fact that the coefficient on the strictly exogenous variable is quite small. The results also suggest that the choice of bandwidth has a very small effect on the performance of the estimator.

Figure 13 illustrates the estimated small sample probabilities that 10% and 5% t–test will reject the true parameter value. These are very close to 0.10 and 0.05, especially in light of the fact that they are estimated from 500 replications.

4.3 Bandwidth Selection

Appendix 2 describes a feasible bandwidth selection rule which attempts to minimize the mean squared errors measure, $E\left[\left(\widehat{\theta}-\theta_0\right)' A \left(\widehat{\theta}-\theta_0\right)\right]$, for some weighting matrix A. In this section we report the results from applying this rule in the simple Monte Carlo design described in Section 4.1. The bandwidth selection rule requires an initial estimate of the parameter vector. We use the estimate based on a fixed bandwidth of 1. Judging from Figures 7–11 this bandwidth is of the right order of magnitude, but it appears not be close to the best bandwidth. We use the identity matrix as the weighting matrix.

The results are presented in Table 4. The column labelled MAE reports the median absolute errors[9] of the estimator based on the automatic bandwidth selection, whereas the column labelled "min MAE" reports the minimum of the median absolute errors over the bandwidths used in section 4.1. We observe that the bandwidth selection rule works well in this case. First of all, it is clear from Figures 7–11 that different bandwidths are optimal for the two parameters. Since the bandwidth selection rule attempts to minimize the sum of the mean squared errors, it is not too surprising that the estimator based on it is less precise than the infeasible estimator that picks the minimizer of the absolute error for each parameter. However, by comparing the results in Table 4 to Figures 7–11 it is clear that the selection rule improved over the initial bandwidth of 1.

Table 4 suggests that the bandwidth selection rule does less well relative to the infeasible choice when β and δ_2 are large. However, it is worth noting that this is also the case in which the bandwidth matters the most (cf. Figures 7–11) and it is also the case in which the gain from using the selection rule is largest compared to using the initial bandwidth of 1.

5 Conclusion

Existing methods do not allow one to estimate and test for second order state dependence in dynamic discrete choice models with unrestricted individual–specific effects. Building on Honoré and Kyriazidou (2000b), this paper proposes methods for doing this in the context of a logit model. We discuss the large sample properties of the estimator and a small Monte Carlo study illustrates its performance in finite samples. A feasible bandwith selection rule is also described. An extension to the semiparametric case following the logic of Honoré and Kyriazidou (2000b) is relatively straightforward, and the resulting maximum score estimator based on Manski (1987)

[9]It would have been more natural to report mean squared errors of the estimators. However, these are very sensitive to outliers and there is no guarantee that any of the estimators presented here have finite moments in finite sample.

would be consistent.

The paper also applies the method to estimate a simple dynamic discrete choice model of youth unemployment which allows for a time–varying macroeconomic explanatory variable. The results suggest that such variables are indeed important in practice.

6 Appendix 1: Derivation of objective function

Recall that the model is

$$P\left(y_{i,t}=1|\, y_{i,t-1}, y_{i,t-2}, x_i, \alpha_i\right) = \frac{\exp\left(\alpha_i + x_{i,t}\beta_{y_{t-1}} + \delta_{1i}y_{t-1} + \delta_{2,y_{t-1}}y_{t-2}\right)}{1+\exp\left(\alpha_i + x_{i,t}\beta_{y_{t-1}} + \delta_{1i}y_{t-1} + \delta_{2,y_{t-1}}y_{t-2}\right)}$$

an that the estimation of $(\beta_0, \beta_1, \delta_{2,0}, \delta_{2,1})$ is based on minimizing

$$\sum_i q_i\left(b_1, b_2, d_{2,0}d_{2,1}\right)$$

where $q_i\left(b_1, b_2, d_{2,0}d_{2,1}\right)$ is defined in (12).

The objective function is defined by considering two sequences, A and B, each of which is of length $T \geq 6$. The two sequences differ only in the t^{th} and s^{th} coordinate, where $2 < t < s < T-1$.

We will now justify the objective function above by considering three cases based on whether t and s differ by one, two or more than two. In each case we will compare $P\left(A|\, x_{i,1}, ..., x_{i,T}, \alpha_i, \delta_{1i}\right)$ to $P\left(B|\, x_{i,1}, ..., x_{i,T}, \alpha_i, \delta_{1i}\right)$. For notational convenience, we will denote these by $P(A)$ and $P(B)$, and we will drop the subscript i on x, y, α and δ_1.

6.1 Case 1. $s = t+1$

Without loss of generality assume that A has $y_t = 1$, $y_{t+1} = 0$ (otherwise switch A and B)

$$\frac{P(A)}{P(B)} = \frac{F\left(\alpha + x_t\beta_{y_{t-1}} + \delta_1 y_{t-1} + \delta_{2,y_{t-1}}y_{t-2}\right)}{1-F\left(\alpha + x_t\beta_{y_{t-1}} + \delta_1 y_{t-1} + \delta_{2,y_{t-1}}y_{t-2}\right)} \times \frac{1-F\left(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1}y_{t-1}\right)}{F\left(\alpha + x_{t+1}\beta_0 + \delta_{2,0}y_{t-1}\right)} \tag{18}$$

$$\times \frac{F\left(\alpha + x_{t+2}\beta_0 + \delta_{2,0}\right)^{y_{t+2}}\left(1-F\left(\alpha + x_{t+2}\beta_0 + \delta_{2,0}\right)\right)^{1-y_{t+2}}}{F\left(\alpha + x_{t+2}\beta_1 + \delta_1\right)^{y_{t+2}}\left(1-F\left(\alpha + x_{t+2}\beta_1 + \delta_1\right)\right)^{1-y_{t+2}}}$$

$$\times \frac{F\left(\alpha + x_{t+3}\beta_{y_{t+2}} + \delta_1 y_{t+2}\right)^{y_{t+3}}\left(1-F\left(\alpha + x_{t+3}\beta_{y_{t+2}} + \delta_1 y_{t+2}\right)\right)^{1-y_{t+3}}}{F\left(\alpha + x_{t+3}\beta_{y_{t+2}} + \delta_1 y_{t+2} + \delta_{2,y_{t+2}}\right)^{y_{t+3}}\left(1-F\left(\alpha + x_{t+3}\beta_{y_{t+2}} + \delta_1 y_{t+2} + \delta_{2,y_{t+2}}\right)\right)^{1-y_{t+3}}}$$

If $y_{t-1} = y_{t+2} = 1$, $x_{t+1}\beta_0 = x_{t+2}\beta_0$ and $x_{t+1}\beta_1 = x_{t+2}\beta_1 = x_{t+3}\beta_1$ then (18) becomes

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}{1 - F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})}{F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})}$$

$$\times \frac{F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})}{F(\alpha + x_{t+1}\beta_1 + \delta_1)} \times \frac{F(\alpha + x_{t+1}\beta_1 + \delta_1)^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_1 + \delta_1))^{1-y_{t+3}}}{F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1}))^{1-y_{t+3}}}$$

or

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}{1 - F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}$$

$$\times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})}{F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1}))^{1-y_{t+3}}}$$

$$\times \frac{F(\alpha + x_{t+1}\beta_1 + \delta_1)^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_1 + \delta_1))^{1-y_{t+3}}}{F(\alpha + x_{t+1}\beta_1 + \delta_1)}$$

There are then two cases. If $y_{t+3} = 1$ then

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}{1 - F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})}{F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1})}$$

If $y_{t+3} = 0$ then

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}{1 - F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1)}{F(\alpha + x_{t+1}\beta_1 + \delta_1)}$$

Either way

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})}{1 - F(\alpha + x_t\beta_1 + \delta_1 + \delta_{2,1}y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1}y_{t+3})}{F(\alpha + x_{t+1}\beta_1 + \delta_1 + \delta_{2,1}y_{t+3})} \tag{19}$$

On the other hand, if $y_{t-1} = y_{t+2} = 0$ and $x_{t+1}\beta_0 = x_{t+2}\beta_0 = x_{t+3}\beta_0$ and $x_{t+1}\beta_1 = x_{t+2}\beta_1$then (18) becomes

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_0 + \delta_2 y_{t-2})}{1 - F(\alpha + x_t\beta_0 + \delta_2 y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1)}{F(\alpha + x_{t+1}\beta_0)}$$

$$\times \frac{1 - F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})}{1 - F(\alpha + x_{t+1}\beta_1 + \delta_1)} \times \frac{F(\alpha + x_{t+1}\beta_0)^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_0))^{1-y_{t+3}}}{F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_0 + \delta_{2,0}))^{1-y_{t+3}}}$$

or

$$\frac{P(A)}{P(B)} = \frac{F(\alpha + x_t\beta_0 + \delta_{2,0}y_{t-2})}{1 - F(\alpha + x_t\beta_0 + \delta_{2,0}y_{t-2})} \times \frac{1 - F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})}{F(\alpha + x_{t+1}\beta_0)}$$

$$\times \frac{F(\alpha + x_{t+1}\beta_0)^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_0))^{1-y_{t+3}}}{F(\alpha + x_{t+1}\beta_0 + \delta_{2,0})^{y_{t+3}}(1 - F(\alpha + x_{t+1}\beta_0 + \delta_{2,0}))^{1-y_{t+3}}}$$

There are again two cases. If $y_{t+3}=0$

$$\frac{P(A)}{P(B)}=\frac{F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}{1-F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}\times\frac{1-F(\alpha+x_{t+1}\beta_0)}{F(\alpha+x_{t+1}\beta_0)}$$

If $y_{t+3}=1$

$$\frac{P(A)}{P(B)}=\frac{F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}{1-F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}\times\frac{1-F(\alpha+x_{t+1}\beta_0+\delta_{2,0})}{F(\alpha+x_{t+1}\beta_0+\delta_{2,0})}$$

Either way

$$\frac{P(A)}{P(B)}=\frac{F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}{1-F(\alpha+x_t\beta_0+\delta_{2,0}y_{t-2})}\times\frac{1-F(\alpha+x_{t+1}\beta_0+\delta_{2,0}y_{t+3})}{F(\alpha+x_{t+1}\beta_0+\delta_{2,0}y_{t+3})} \tag{20}$$

Combining (19) and (20) we get that if $y_{t-1}=y_{t+2}$, $x_{t+1}\beta_{1-y_{t-1}}=x_{t+2}\beta_{1-y_{t-1}}$ and $x_{t+1}\beta_{y_{t-1}}=x_{t+2}\beta_{y_{t-1}}=x_{t+3}\beta_{y_{t-1}}$ then

$$\frac{P(A)}{P(B)} \tag{21}$$

$$=\frac{F\left(\alpha+x_t\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t-2}\right)}{1-F\left(\alpha+x_t\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t-2}\right)}\times\frac{1-F\left(\alpha+x_{t+1}\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t+3}\right)}{F\left(\alpha+x_{t+1}\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t+3}\right)}$$

The logit assumption, $F(\eta)=\frac{\exp(\eta)}{1+\exp(\eta)}$, implies $\frac{F(\eta)}{1-F(\eta)}=\exp(\eta)$, and therefore (21) becomes

$$\frac{P(A)}{P(B)}=\frac{\exp\left(\alpha+x_t\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t-2}\right)}{\exp\left(\alpha+x_{t+1}\beta_{y_{t-1}}+\delta_1y_{t-1}+\delta_{2,y_{t-1}}y_{t+3}\right)}=\exp\left((x_t-x_{t+1})\beta_{y_{t-1}}+\delta_{2,y_{t-1}}(y_{t-2}-y_{t+3})\right)$$

In other words,

$$P(A\,|\,A\cup B)=\frac{P(A)}{P(A)+P(B)}=\frac{\exp\left((x_t-x_{t+1})\beta_{y_{t-1}}+\delta_{2,y_{t-1}}(y_{t-2}-y_{t+3})\right)}{1+\exp\left((x_t-x_{t+1})\beta_{y_{t-1}}+\delta_{2,y_{t-1}}(y_{t-2}-y_{t+3})\right)}$$

It is easy to see that $y_{t-1}=y_{t+2}$, $x_{t+1}\beta_{1-y_{t-1}}=x_{t+2}\beta_{1-y_{t-1}}$ and $x_{t+1}\beta_{y_{t-1}}=x_{t+2}\beta_{y_{t-1}}=x_{t+3}\beta_{y_{t-1}}$ is the only case in which α and $\delta_1$cancel. In particular without $y_{t-1}=y_{t+2}$, the sum $\sum y_ty_{t-1}$ would differ for the sequences A and B, so in that case conditioning on $A\cup B$ will not condition on what would be the sufficient statistics for α and δ_1 in a model without time–varying explanatory variables.

6.2 Cases 2 ($s=t+2$) and 3 ($s>t+2$)

The cases where $s=t+2$ and $s>t+2$ are dealt with in the same manner. The calculations are space–consuming and no new insights emerge from them. They are therefore not presented here.

7 Appendix 2: Heuristic Justification for a Bandwidth Selection Procedure.

This appendix mimics the calculations in, for example, Horowitz (1992) or Kyriazidou (1997) in order to derive a rule for selecting the bandwidth. The generic setup for the estimator proposed here is as follows. The p–dimensional parameter, θ is estimated by

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i}}{h}\right) q_{\ell}\left(z_i, \theta\right)$$

where $x_{\ell i}$ is K dimensional. The first order condition is

$$0 = \sum_{i=1}^{n} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(z_i, \widehat{\theta}\right)$$

a first order Taylor expansion yields

$$0 \approx \sum_{i=1}^{n} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(z_i, \theta_0\right) + \sum_{i=1}^{n} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(z_i, \theta_0\right)\left(\widehat{\theta} - \theta_0\right)$$

or

$$\begin{aligned} 0 \quad &\approx \quad \frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} \left(\sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(z_i, \theta_0\right) - E\left[\sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(Z_i, \theta_0\right)\right] \right) \\ &\quad + \sqrt{nh^k} E\left[\frac{1}{h^k} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q'_{\ell}\left(Z_i, \theta_0\right)\right] + \left[\frac{1}{nh^k} \sum_{i=1}^{n} \sum_{\ell=1}^{L} K\left(\frac{x_{\ell i1}}{h}\right) q''_{\ell}\left(z_i, \theta^*\right)\right] \sqrt{nh^k}\left(\widehat{\theta} - \theta_0\right) \\ &\equiv \quad Z_n + \sqrt{nh^k} B_n - J_n \sqrt{nh^k}\left(\widehat{\theta} - \theta_0\right) \end{aligned}$$

or

$$\left(\widehat{\theta} - \theta_0\right) = \frac{1}{\sqrt{nh^k}} J_n^{-1}\left(Z_n + \sqrt{nh^k} B_n\right) = \frac{1}{\sqrt{nh^k}}\left(J_n^{-1} Z_n + \sqrt{nh^k} J_n^{-1} B_n\right)$$

Under suitable regularity conditions, $Z_n \to N\left(0, \Sigma\right)$ and $J_n \to J$.

Now let $r_{\ell}\left(x\right) = E\left[q'_{\ell}\left(Z_i, \theta_0\right) \middle| X_i = x\right] f_{X_i}\left(x\right)$, and note that B_n can be written as the sum of

L terms, $B_{n\ell}$. Write the j'th coordinate of $B_{n\ell}$ as

$$\begin{aligned}
B_{n\ell,j} &= E\left[\frac{1}{h^k}K\left(\frac{x_{\ell i1}}{h}\right)q'_{\ell j}\left(Z_i,\theta_0\right)\right] \\
&= \int \frac{1}{h^k}K\left(\frac{x_{12}}{h}\right)r_{\ell j}\left(x_{12}\right)dx_{12} \\
&= \int K\left(\eta\right)r_{\ell j}\left(h\eta\right)d\eta \\
&= \int K\left(\eta\right)\left\{r_{\ell j}\left(0\right)+hr_{\ell j}^{(1)}\left(0\right)'\eta+\frac{1}{2}h^2\eta' r_{\ell j}^{(2)}\left(0\right)\eta\right\}d\eta+O\left(h^3\right) \\
&= \int K\left(\eta\right)r_{\ell j}\left(0\right)d\eta+\int K\left(\eta\right)hr_{\ell j}^{(1)}\left(0\right)'\eta d\eta+\frac{1}{2}\int K\left(\eta\right)h^2\eta' r_{\ell j}^{(2)}\left(0\right)\eta d\eta+O\left(h^3\right) \\
&= 0+0+h^2\frac{1}{2}\int K\left(\eta\right)\eta' r_{\ell j}^{(2)}\left(0\right)\eta d\eta+O\left(h^3\right) \\
&\equiv h^2B_{\ell j}^*+O\left(h^3\right)
\end{aligned}$$

and

$$B_j^*=\sum_{\ell=1}^{L}B_{\ell j}^*$$

and let B^* be a vector with j'th element equal to then B_j^*.

For a weighting matrix, A, we then have (approximately)

$$\begin{aligned}
MSE\left(\widehat{\theta}\right) &\equiv E\left[\left(\widehat{\theta}-\theta_0\right)'A\left(\widehat{\theta}-\theta_0\right)\right] \\
&= E\left[\text{trace}\left(\left(\widehat{\theta}-\theta_0\right)'A\left(\widehat{\theta}-\theta_0\right)\right)\right] \\
&= \text{trace}\left(AE\left[\left(\widehat{\theta}-\theta_0\right)\left(\widehat{\theta}-\theta_0\right)'\right]\right) \\
&= \frac{1}{n}\text{trace}\left(h^{-k}AJ^{-1}\Sigma J^{-1}+nh^4AJ^{-1}B^*B^{*\prime}J^{-1}\right) \\
&= \frac{1}{n}\left(h^{-k}\,\text{trace}\left(AJ^{-1}\Sigma J^{-1}\right)+nh^4\,\text{trace}\left(AJ^{-1}B^*B^{*\prime}J^{-1}\right)\right) \\
&\equiv \frac{1}{n}\left(h^{-k}a_1+nh^4a_2\right)
\end{aligned}$$

Minimizing this with respect to h yields $-kh^{-k-1}a_1+4nh^3a_2=0$ or $ka_1=4na_2h^{k+4}$ or

$$h=\left(\frac{ka_1}{4na_2}\right)^{1/(k+4)}$$

It seems natural to choose the bandwidth to be an estimate of the right hand side. J and Σ are easily estimated by sample analogs. The issue is B^*

Recall that

$$B_{\ell j}^*=\frac{1}{2}\int K\left(\eta\right)\eta' r_{\ell j}^{(2)}\left(0\right)\eta d\eta$$

where

$$r_j^{(2)}(x) = \frac{\partial^2 E\left[q'_{\ell j}(Z_i,\theta_0)\middle| X_i = x\right] f_{X_i}(x)}{\partial x \partial x'}$$

$E\left[q'_j(Z_i,\theta_0)\middle| X_i = x\right] f_X(x)$ can be estimated by

$$\frac{1}{nh^k}\sum_{i=1}^{n} \widetilde{K}\left(\frac{x_i - x}{h}\right) \frac{\partial q_\ell(z_i,\theta)}{\partial \theta_j}\bigg|_{\widehat{\theta}}$$

and so one could estimate $r_j^{(2)}(0)$ by

$$\widehat{r}_j^{(2)}(0) = \frac{1}{nh^{k+2}}\sum_{i=1}^{n} \widetilde{K}''\left(\frac{x_i}{h}\right) \frac{\partial q_\ell(z_i,\theta)}{\partial \theta_j}\bigg|_{\widehat{\theta}}$$

This suggests

$$\begin{aligned}
\widehat{B}^*_{\ell j} &= \frac{1}{2}\int K(\eta)\,\eta' \left[\frac{1}{nh^{k+2}}\sum_{i=1}^{n} \widetilde{K}''\left(\frac{x_i}{h}\right) \frac{\partial q_\ell(z_i,\theta)}{\partial \theta_j}\bigg|_{\widehat{\theta}}\right] \eta d\eta \\
&= \frac{1}{2nh^{k+2}}\sum_{i=1}^{n}\sum_{i_1=1}^{k}\sum_{i_2=1}^{k} s_{i_1 i_2}\, \widetilde{K}''\left(\frac{x_i}{h}\right)_{i_1 i_2} \frac{\partial q_\ell(z_i,\theta)}{\partial \theta_j}\bigg|_{\widehat{\theta}}
\end{aligned}$$

where$\{s_{i_1 i_2}\}$ denotes the elements of the covariance matrix associated with K. Finally

$$\widehat{B}^*_j = \sum_{\ell=1}^{L} \widehat{B}^*_{\ell j} = \frac{1}{2nh^{k+2}}\sum_{i=1}^{n}\sum_{i_1=1}^{k}\sum_{i_2=1}^{k} s_{i_1 i_2}\sum_{\ell=1}^{L} \widetilde{K}''\left(\frac{x_i}{h}\right)_{i_1 i_2} \frac{\partial q_\ell(z_i,\theta)}{\partial \theta_j}\bigg|_{\widehat{\theta}}$$

References

Blanchflower, D. G., 1996. Youth labour markets in twenty three countries: A comparison using micro data. CEP Discussion Paper n. 284, LSE, London.

Bruno, C., Cazes, S., 1997. Le chômage des jeunes en france: un état des lieux. Révue de l'OFCE, (62).

Cette, G., Cuneo, P., Eyssartier, D., Gautié, J., 1996. Coût du travail et emploi des jeunes. Révue de l'OFCE, (56).

Chamberlain, G., 1985. Heterogeneity, omitted variable bias, and duration dependence. in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman, and B. Singer, no. 10 in Econometric Society Monographs series, pp. 3–38. Cambridge University Press, Cambridge, New York and Sydney.

Chay, K. Y., Hoynes, H., Hyslop, D., 2001. A non–experimental analysis of true state dependence in monthly welfare participation sequences. University of California, Berkeley.

Chintagunta, P., Kyriazidou, E., Perktold, J., 2001. Panel data analysis of household brand choices. Journal of Econometrics, 103(1-2), 111–53.

CSERC, 1996. L'allégement des charges sociales sur les bas salaires. La Documentation Française.

d'Addio, A. C., Rosholm, M., 2002. Left-censoring in duration data: Theory and applications. mimeo, University of Aarhus.

Eurostat, 1997. *Youth in the European Union: From Education to Working Life*. The Statistical Office of the European Communities.

Heckman, J. J., 1981a. Heterogeneity and state dependence. Studies in Labor Markets, S. Rosen (ed).

Heckman, J. J., 1981b. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. Structural Analysis of Discrete Panel Data with Econometric Applications, C. F. Manski and D. McFadden (eds), pp. 179–195.

Heckman, J. J., 1981c. Statistical models for discrete panel data. Structural Analysis of Discrete Panel Data with Econometric Applications, C. F. Manski and D. McFadden (eds), pp. 114–178.

Honoré, B. E., 2002. Nonlinear models with panel data. Portuguese Economic Journal, 1, 163–179.

Honoré, B. E., Hu, L., 2004. Estimation of cross sectional and panel data censored regression models with endogeneity. Journal of Econometrics, 122(2), 293–316.

Honoré, B. E., Kyriazidou, E., 2000a. Estimation of tobit-type models with individual specific effects. Econometric Reviews, 19, 341–66.

Honoré, B. E., Kyriazidou, E., 2000b. Panel data discrete choice models with lagged dependent variables. Econometrica, 68, 839–874.

Horowitz, J. L., 1992. A smoothed maximum score estimator for the binary response model. Econometrica, 60, 505–531.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica, 65, 1335–1364.

Layard, R., Nickell, S., Jackman, R., 1991. *Unemployment, Macroeconomic Performance and the Labour Market*. Oxford University Press.

Magnac, T., 2000. State dependence and unobserved heterogeneity in youth employment histories. Economic Journal, 110, 805–837.

Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica, 55, 357–62.

Moghadham, R., 1993. Les causes du chômage en france. IMF.

Table 1: Transition probabilities

$\Pr(y_t = 1 \mid y_{t-1} = 0)$	2.38
$\Pr(y_t = 1 \mid y_{t-1} = 1)$	92.68
$\Pr(y_t = 1 \mid y_{t-1} = 0, y_{t-2} = 0)$	2.26
$\Pr(y_t = 1 \mid y_{t-1} = 0, y_{t-2} = 1)$	7.61
$\Pr(y_t = 1 \mid y_{t-1} = 1, y_{t-2} = 0)$	84.90
$\Pr(y_t = 1 \mid y_{t-1} = 1, y_{t-2} = 1)$	93.10

Table 2: Linear Probability and Logit Model
(estimating the individual–specific effects)

	Everybody	**Young Men**	**Young Women**	**Men over 25**	**Women over 25**
	Linear Probability Model with Fixed Effects				
	$(n = 3590)$	$(n = 1625)$	$(n = 1436)$	$(n = 831)$	$(n = 894)$
$\delta_{2,0}$	-0.062*	-0.066*	-0.080*	-0.055*	-0.071*
	0.006	0.009	0.010	0.018	0.016
$\delta_{2,1}$	-0.067*	-0.039*	-0.103*	-0.093*	-0.088*
	0.008	0.014	0.014	0.025	0.019
β_0	0.004*	0.005*	0.011*	0.000	-0.001
	0.001	0.002	0.002	0.001	0.002
β_1	0.007*	0.009	0.011*	0.003	-0.003
	0.003	0.006	0.005	0.013	0.004
	Logit				
$\delta_{2,0}$	-0.890*	-1.079*	-1.402*	-0.796*	-0.936*
	0.124	0.206	0.236	0.350	0.335
$\delta_{2,1}$	-0.594*	-0.322*	-1.251*	-1.111*	-1.067*
	0.106	0.151	0.217	0.440	0.309
β_0	0.297*	0.257*	0.628*	0.053	0.013*
	0.058	0.090	0.109	0.184	0.033
β_1	0.012	0.031	0.194	0.010	-0.149
	0.028	0.120	0.115	0.127	0.165

Table 3: Estimating the fixed effects

	β			δ_2		
	True	Med Bias	MAE	True	Med Bias	MAE
Design 1	1.000	0.238	0.238	1.000	0.612	0.612
	2.000	0.749	0.749	1.000	0.432	0.432
	1.000	0.264	0.264	2.000	0.647	0.647
	2.000	0.809	0.809	2.000	0.172	0.175
Design 2	1.000	0.196	0.196	1.000	0.507	0.507
	2.000	0.656	0.656	1.000	0.374	0.374
	1.000	0.230	0.230	2.000	0.556	0.556
	2.000	0.703	0.703	2.000	0.144	0.153
Design 3	1.000	0.140	0.140	1.000	0.355	0.355
	2.000	0.447	0.447	1.000	0.255	0.255
	1.000	0.178	0.178	2.000	0.419	0.419
	2.000	0.517	0.517	2.000	0.096	0.121
Design 4	1.000	0.238	0.238	1.000	0.619	0.619
	2.000	0.772	0.772	1.000	0.440	0.440
	1.000	0.258	0.258	2.000	0.611	0.611
	2.000	0.796	0.796	2.000	0.173	0.183
Design 5	1.000	0.247	0.247	1.000	0.663	0.663
	2.000	0.795	0.795	1.000	0.505	0.505
	1.000	0.259	0.259	2.000	0.619	0.619
	2.000	0.822	0.822	2.000	0.179	0.191

Table 4: Automatic Bandwidth Selection

	β				δ_2				
	True	median bias	MAE	min MAE	True	median bias	MAE	min MAE	median h_{opt}
Design 1	1.000	0.036	0.047	0.045	1.000	0.133	0.153	0.136	5.120
	2.000	0.082	0.121	0.113	1.000	0.242	0.248	0.223	5.107
	1.000	0.036	0.057	0.055	2.000	0.210	0.221	0.174	5.349
	2.000	0.076	0.113	0.112	2.000	0.431	0.431	0.262	5.101
Design 2	1.000	0.034	0.054	0.046	1.000	0.115	0.136	0.131	5.177
	2.000	0.060	0.101	0.105	1.000	0.235	0.241	0.205	5.117
	1.000	0.019	0.048	0.051	2.000	0.214	0.225	0.197	5.233
	2.000	0.048	0.100	0.104	2.000	0.443	0.443	0.280	5.198
Design 3	1.000	0.013	0.045	0.042	1.000	0.111	0.135	0.129	5.139
	2.000	0.008	0.098	0.096	1.000	0.227	0.234	0.193	5.138
	1.000	0.006	0.045	0.045	2.000	0.205	0.218	0.179	5.183
	2.000	0.020	0.106	0.100	2.000	0.466	0.467	0.274	5.160
Design 4	1.000	0.036	0.048	0.044	1.000	0.131	0.147	0.141	5.290
	2.000	0.081	0.120	0.116	1.000	0.226	0.236	0.199	5.197
	1.000	0.043	0.056	0.055	2.000	0.212	0.229	0.192	5.265
	2.000	0.073	0.122	0.121	2.000	0.414	0.414	0.294	5.185
Design 5	1.000	0.032	0.051	0.050	1.000	0.123	0.153	0.145	5.339
	2.000	0.077	0.115	0.110	1.000	0.254	0.267	0.227	5.185
	1.000	0.034	0.061	0.059	2.000	0.222	0.237	0.205	5.270
	2.000	0.077	0.122	0.115	2.000	0.432	0.434	0.291	5.144

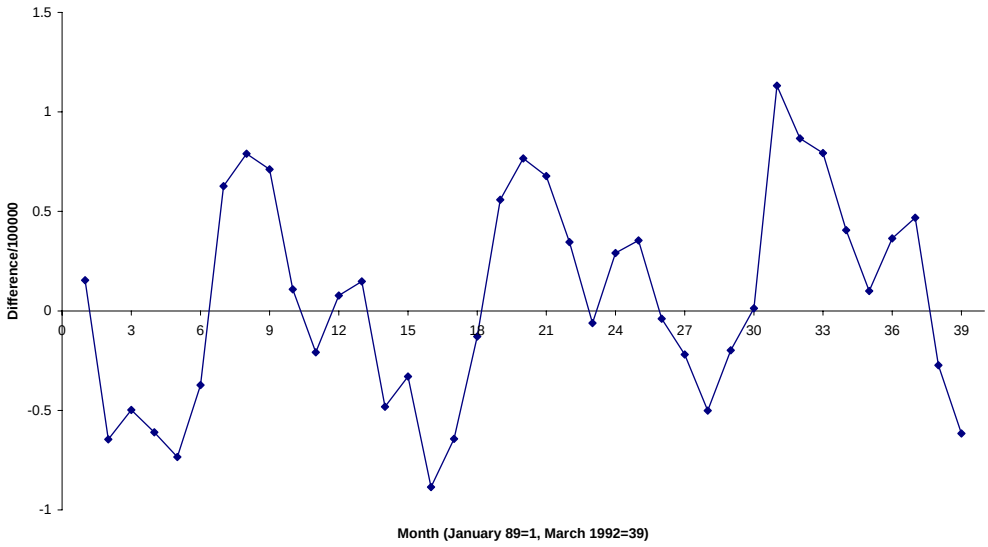

Figure 1: Difference in the number of unemployed

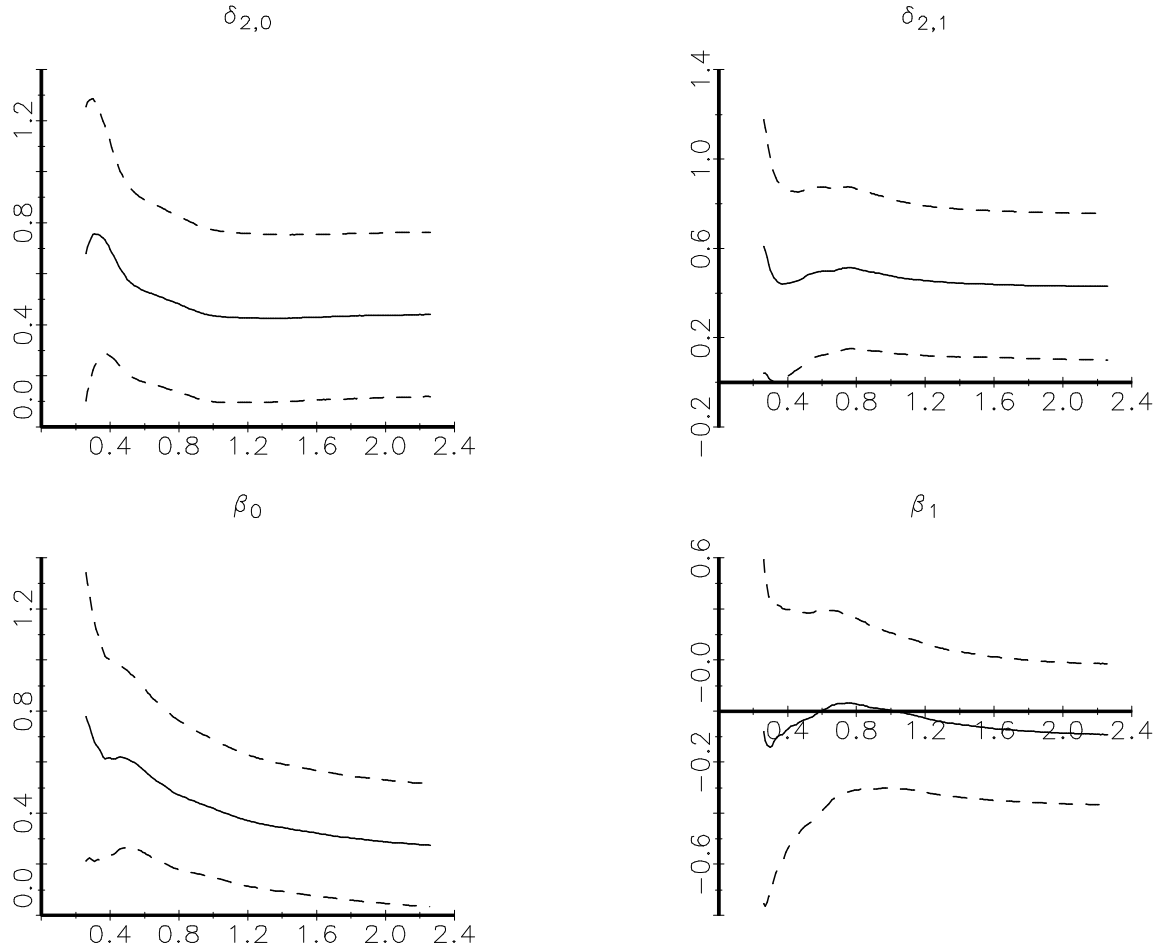

Figure 2: 95% confidence intervals as a function of the bandwidth (all)

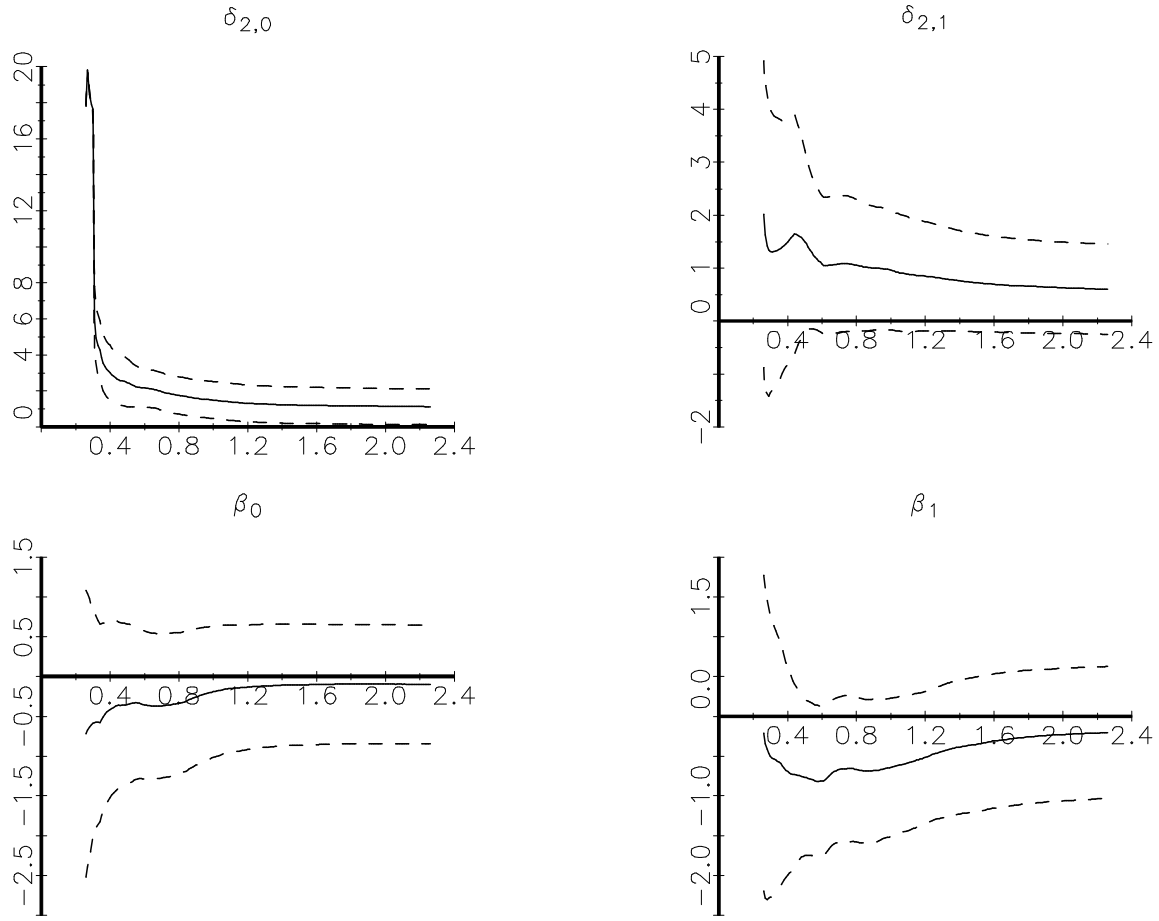

Figure 3: 95% confidence intervals as a function of the bandwidth (older men)

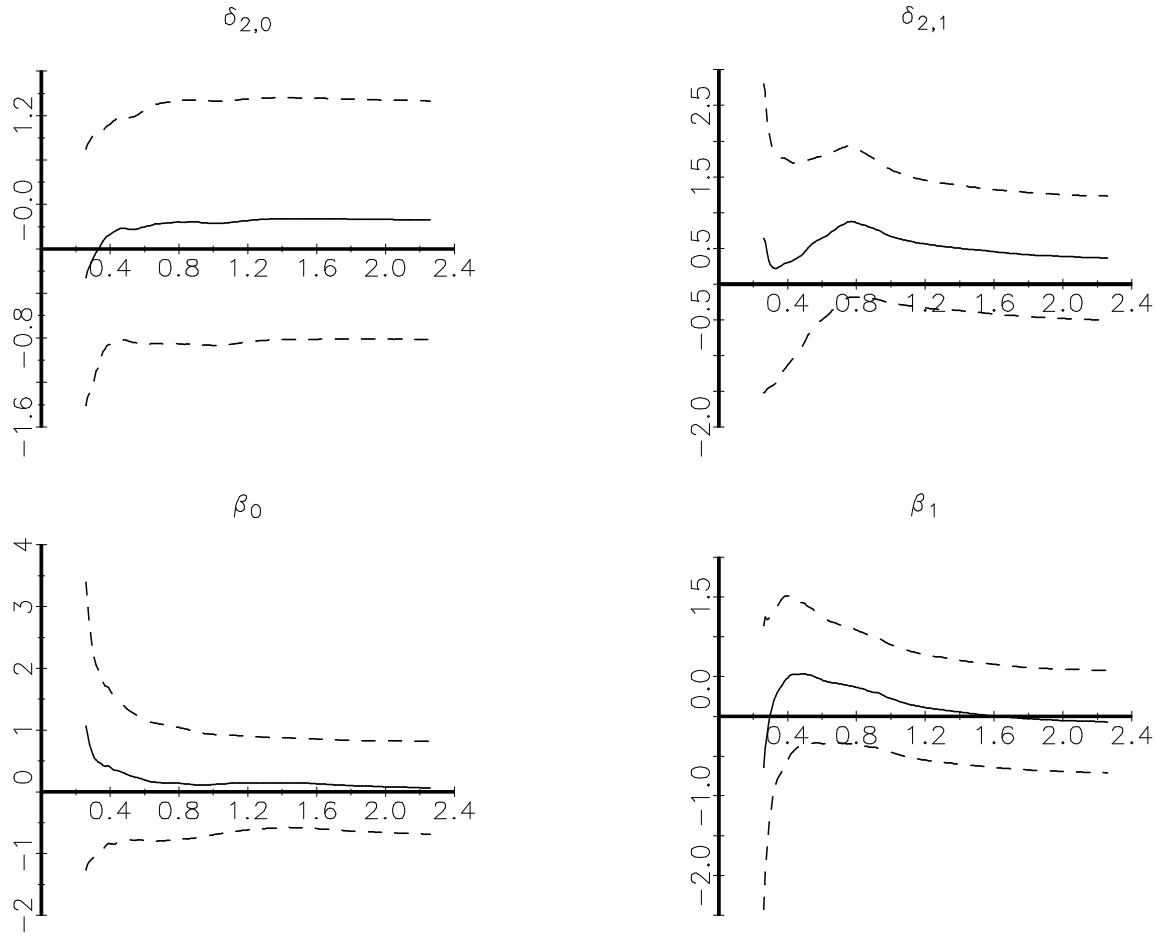

Figure 4: 95% confidence intervals as a function of the bandwidth (older women)

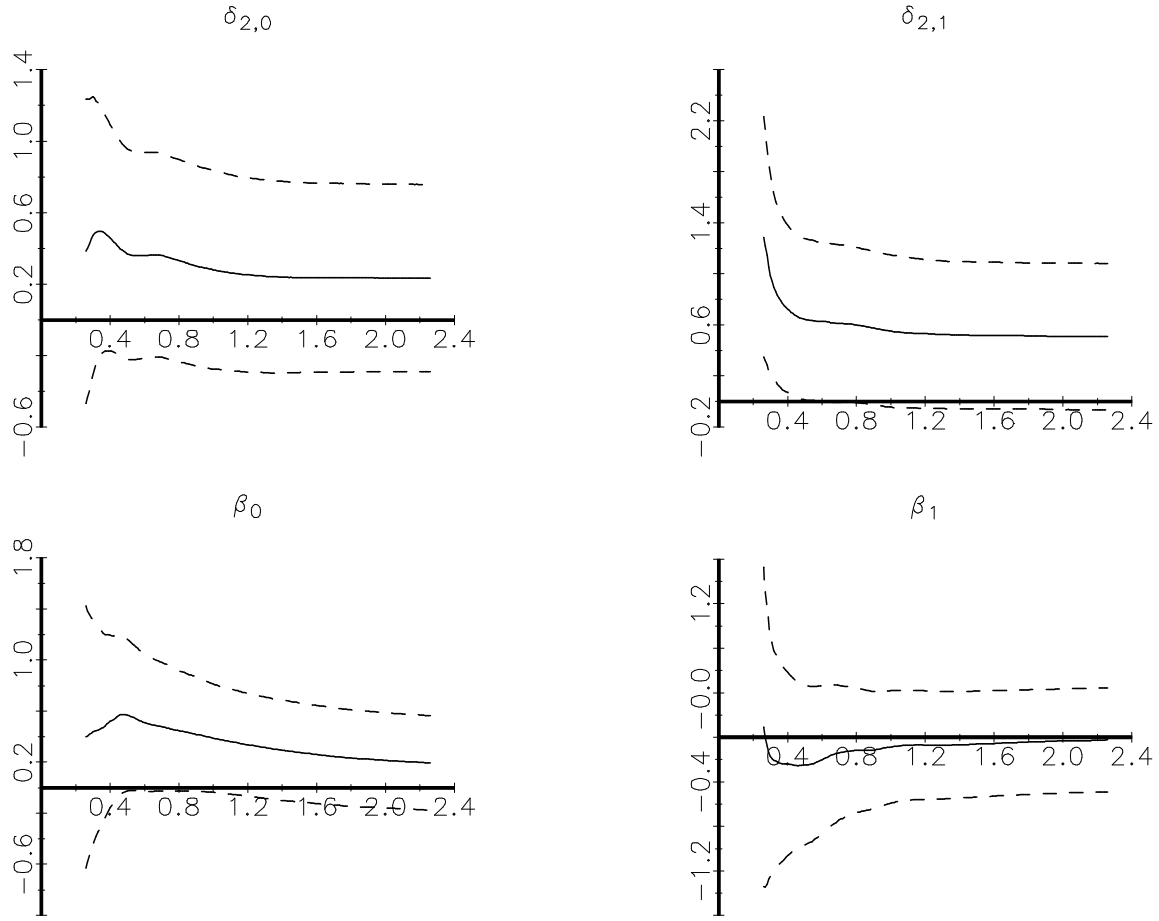

Figure 5: 95% confidence intervals as a function of the bandwidth (young men)

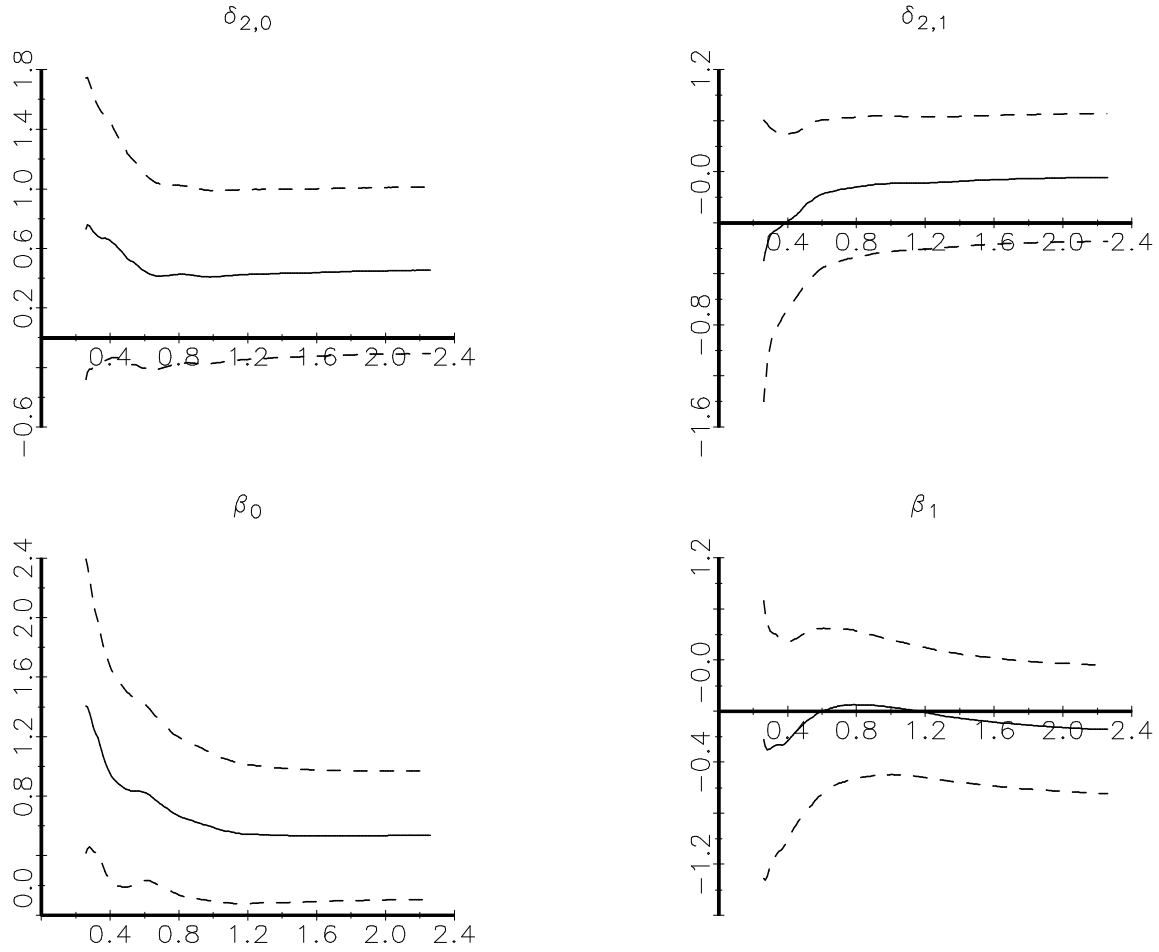

Figure 6: 95% confidence intervals as a function of the bandwidth (young women)

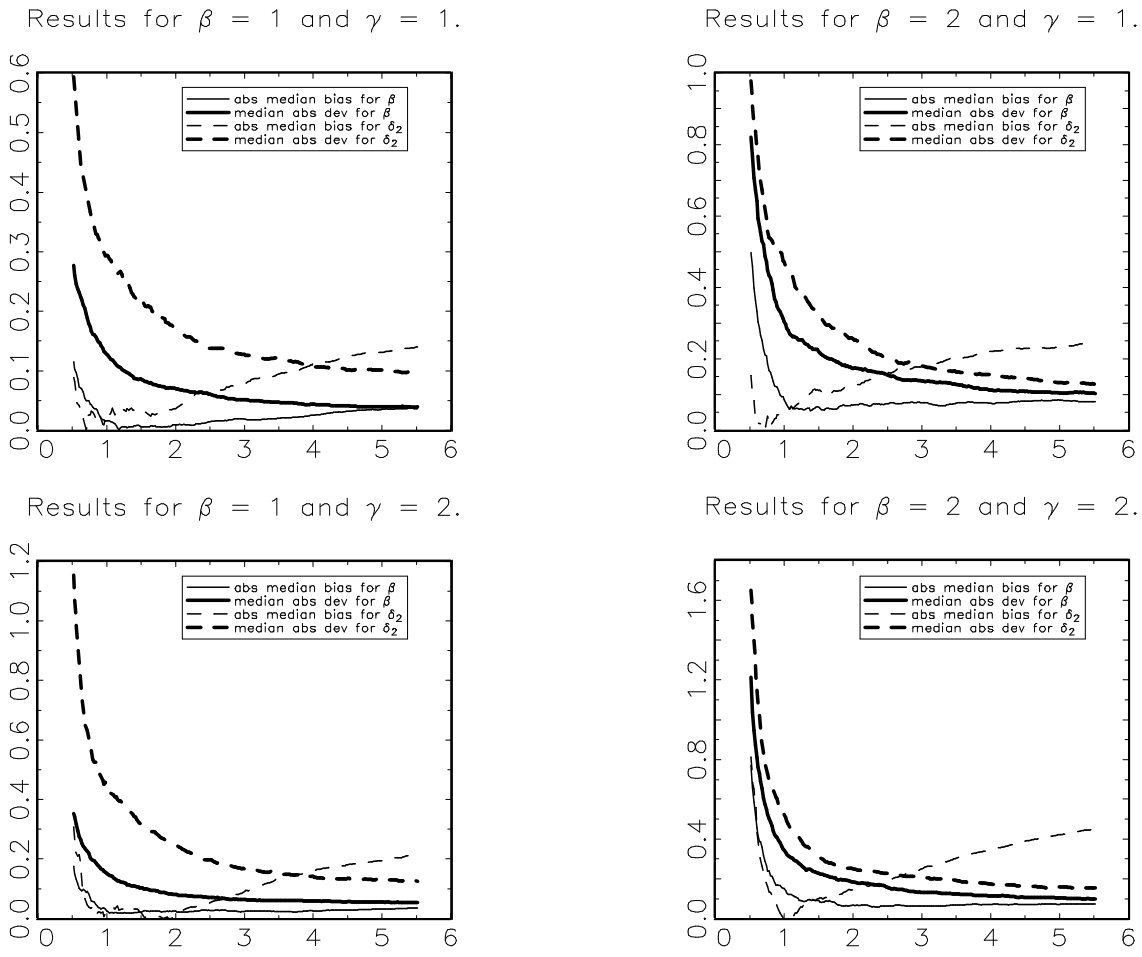

Figure 7: Results for Design 1

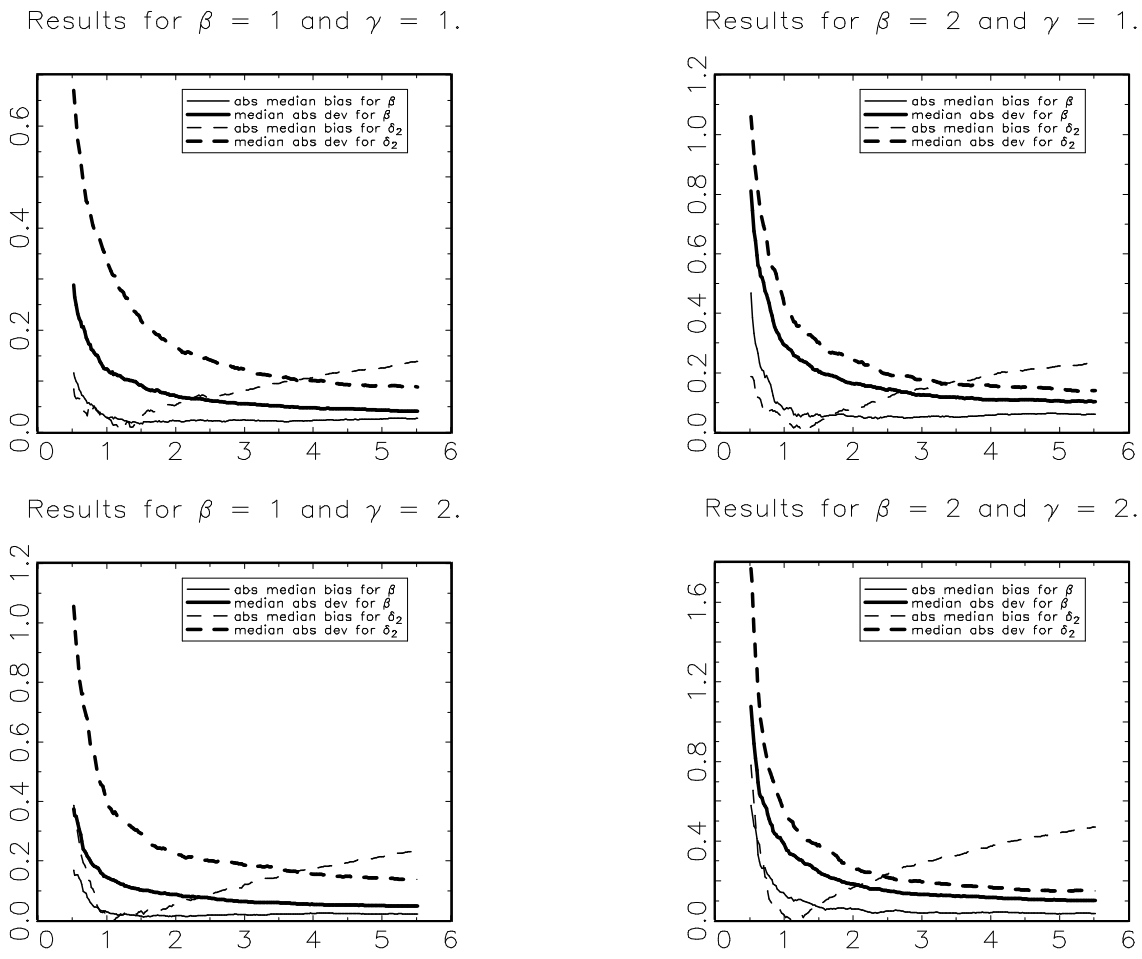

Figure 8: Results for Design 2

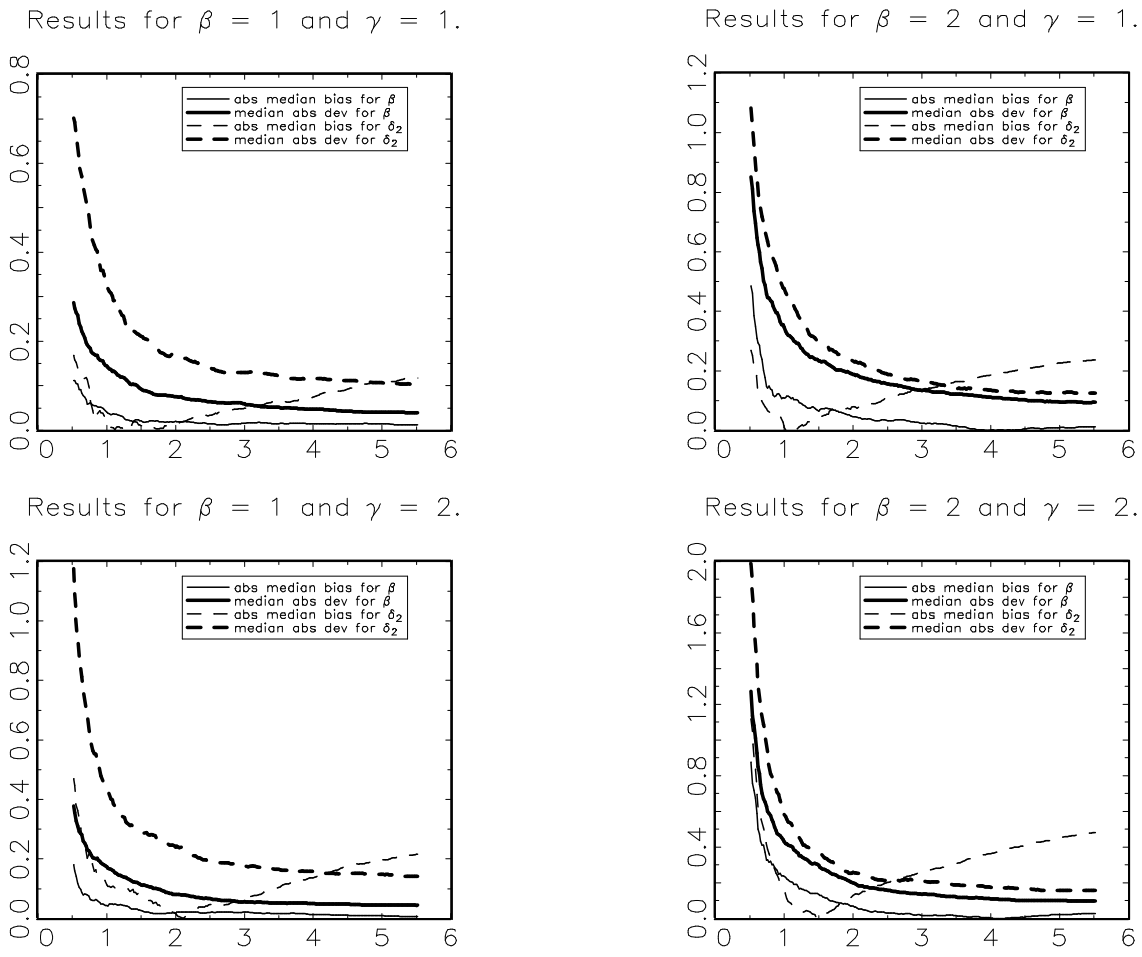

Figure 9: Results for Design 3

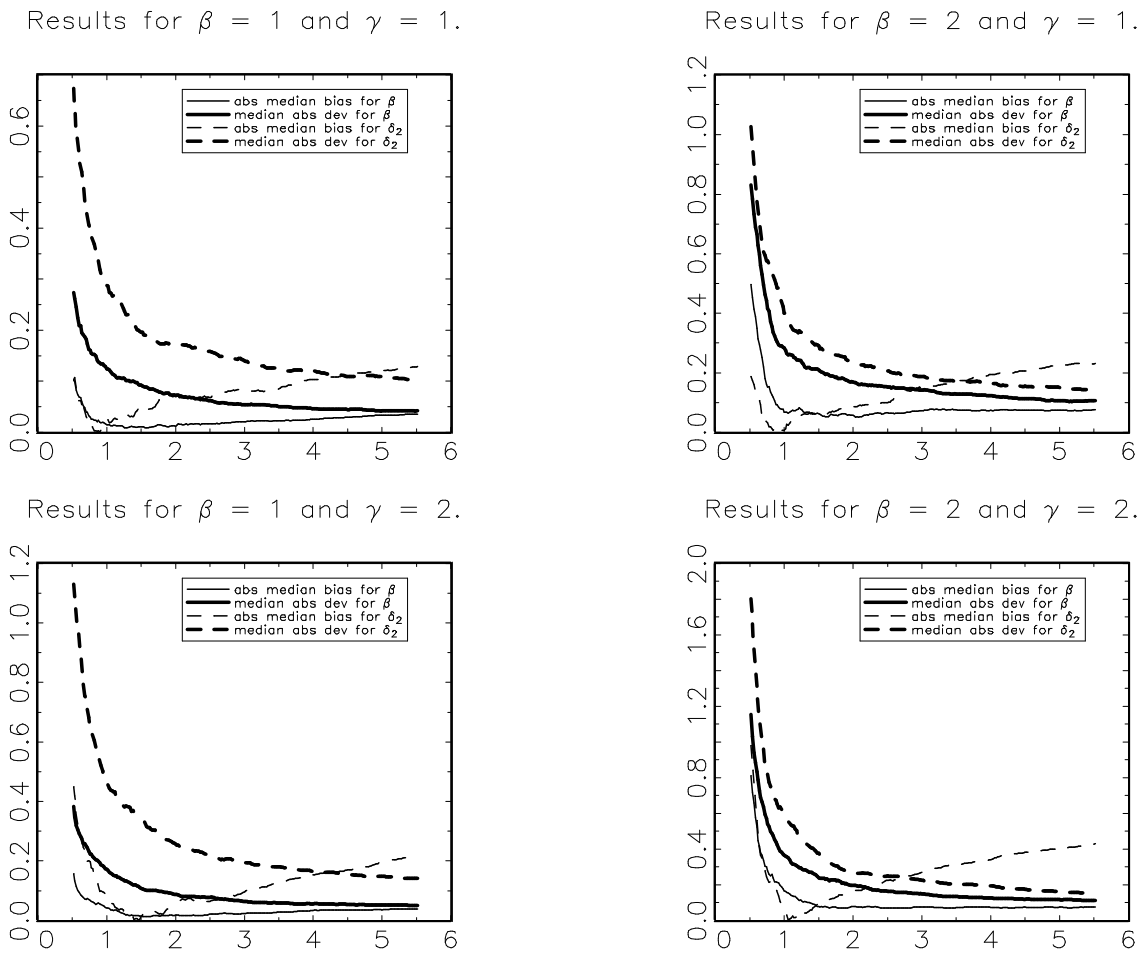

Figure 10: Results for Design 4

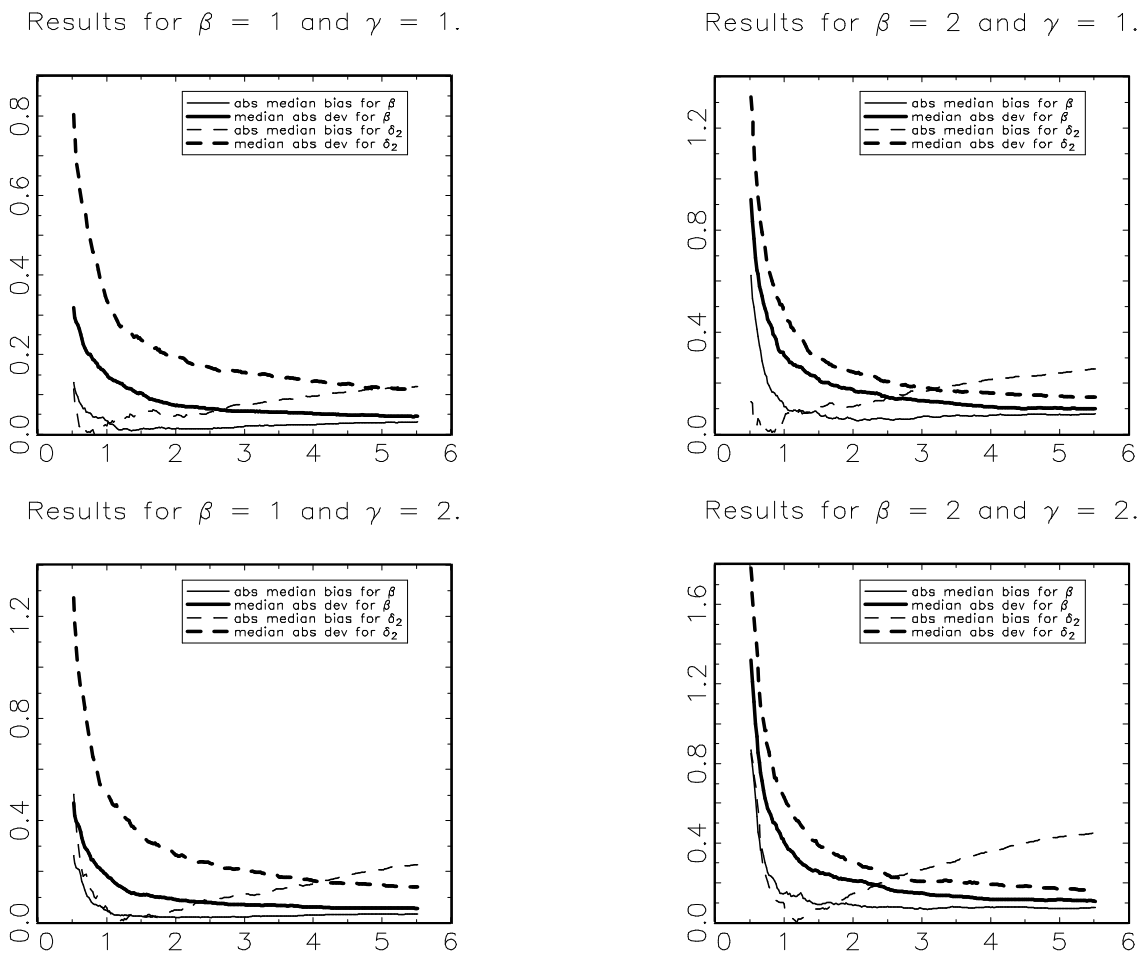

Figure 11: Results for Design 5

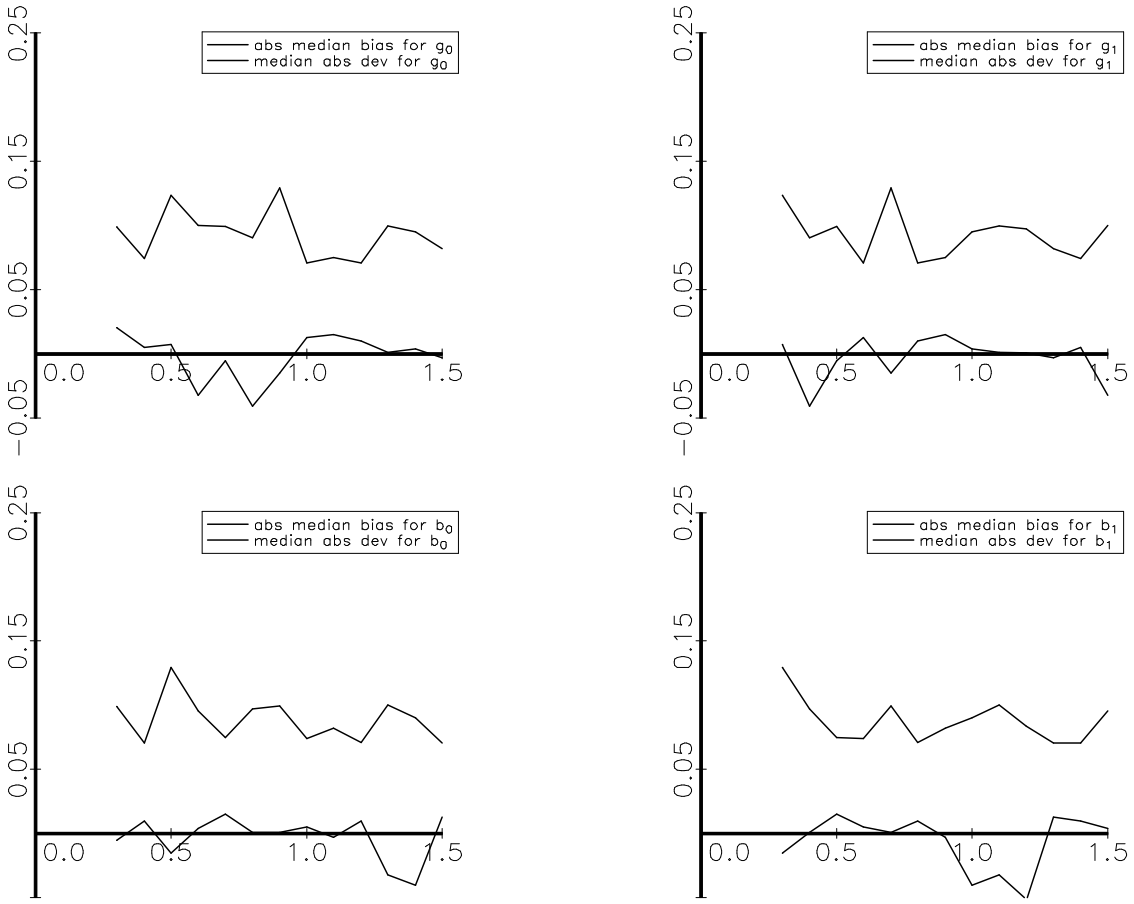

Figure 12: Results using real data

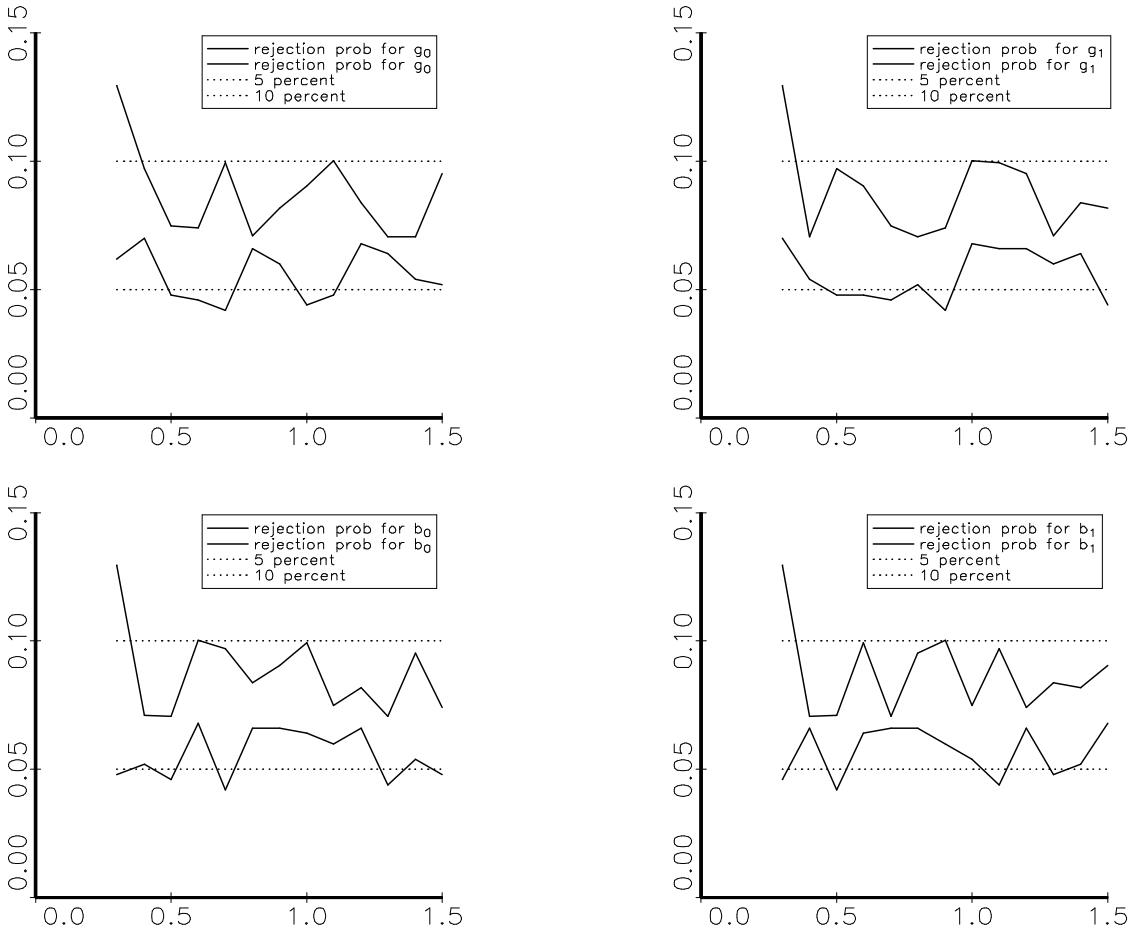

Figure 13: Significance Level for T–statistics