CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2008-20

Determinants of Birthweight Outcomes: Quantile Regressions Based on Panel Data

Stefan Holst Bache, Christian M. Dahl and Johannes Tang Kristensen

School of Economics and Management
University of Aarhus
Building 1322, DK-8000 Aarhus C
Denmark

Aarhus School of Business
University of Aarhus
Handelshøjskolen
Aarhus Universitet

UNIVERSITY OF
COPENHAGEN

Determinants of Birthweight Outcomes: Quantile Regressions Based on Panel Data

Stefan Holst Bache*, Christian M. Dahl† and Johannes Tang Kristensen‡

May, 2008

Abstract. Low birthweight outcomes are associated with large social and economic costs, and therefore the possible determinants of low birthweight are of great interest. One such determinant which has received considerable attention is maternal smoking. From an economic perspective this is in part due to the possibility that smoking habits can be influenced through policy conduct. It is widely believed that maternal smoking reduces birthweight; however, the crucial difficulty in estimating such effects is the unobserved heterogeneity among mothers. We consider extensions of three panel data models to a quantile regression framework in order to control for heterogeneity and to infer conclusions about causality across the entire birthweight distribution. We obtain estimation results for maternal smoking and other interesting determinants, applying these to data obtained from Aarhus University Hospital, Skejby (Denmark). We examine the use of both balanced and unbalanced panels. In conclusion, our results show the importance of considering conditional quantiles and controlling for unobserved heterogeneity when estimating determinants of birthweight outcomes. An example of this is the change in magnitude and significance of prenatal smoking. Controlling for unobserved effects does not change the fact that smoking reduces birthweight, but it shows that the effect is primarily a problem in the left tail of the distribution on a slightly smaller scale.

Keywords: Random Correlated Effects, Fixed Effects, Cross Section, Quantile Regression, Maternal Smoking, Birthweight.

JEL Classifications: C13, C23, I10.

*Department of Business and Economics, University of Southern Denmark. stefan@sdu.dk.

†Corresponding author, CREATES, School of Economics and Management, University of Aarhus. cdahl@econ.au.dk. The author gratefully acknowledges financial support from the Danish Social Sciences Research Council (grant 275-06-0105) and the Center for Research in Econometric Analysis of Time Series, CREATES, funded by The Danish National Research Foundation.

‡Department of Business and Economics, University of Southern Denmark. johannes@sdu.dk.

1 Introduction

The potential adverse health consequences of low birthweight outcomes, along with the considerable economic burden they are believed to impose on society, have attracted much attention by researchers in both medical and economic literature. The use of birthweight as a proxy for the general health condition of infants is commonplace, as it has been linked to a vast array of health related complications, both short- and long-term.

The most severe event, perinatal mortality, has been found to be more likely in the event of a low birthweight outcome. Several studies find statistical evidence of this linkage, see e.g. Bernstein et al. (2000), Almond et al. (2005), and Black et al. (2005). Furthermore, it is believed that low birthweight may lead to complications such as epilepsy, mental retardation, blindness, and deafness. For a review and references, see Hack et al. (1995). While many of these complications are directly observable, some studies also consider less obvious socio-economic implications of low birthweight, a very popular topic being school performance. Kirkegaard et al. (2006) find a graded relationship between birthweight and school performance. In a follow-up study with 5.319 children aged 9–11, they conclude that the risk of reading, spelling, and arithmetic disabilities is greater with low birthweight children. Similarly, Corman and Chaikind (1998) find that repeating a grade, or special class attendance is more likely among low birthweight children. This may suggest that even future earnings and labour market outcomes may be affected by birthweight. According to Black et al. (2005) this is indeed the case.

The strong evidence that low birthweight has adverse effects has naturally led to substantial efforts towards identifying the determinants of these undesirable outcomes. One such determinant which has received much attention in the literature is maternal smoking habits during pregnancy. Statistical efforts suggest a strong correlation between birthweight and maternal smoking, see e.g. Bernstein et al. (1978) and Permutt and Hebel (1989). Other studies examine the effect of smoking on some of the above-mentioned complications directly, e.g. Wisborg et al. (2000), who find that smoking increases the risk of the sudden infant death syndrome, Wisborg et al. (2001), who find an increased risk of still birth and infant mortality from maternal smoking, and Linnet et al. (2006), who link hyperactive-distractible behaviour in preschool children to intrauterine exposure to tobacco smoke. Medical research gives several reasons why cigarette smoking may affect birthweight. An explanation that seems to stand out is that the foetus can suffer from chronic hypoxic stress as a consequence of smoking. DiFranza et al. (2004) and Hofhuisi et al. (2003) explain this phenomenon in part by a lowered maternal uterine blood flow and a reduction in oxygen diffusion across the placenta. An interesting observation is that smoking does not seem to have a significant adverse effect on all birth outcomes. Wang et al. (2002) conclude that the association between maternal cigarette smoking and reduced birthweight is modified by maternal genetic susceptibility, after having considered two specific gene polymorphisms.

From an economic perspective, interest lies not with the individual as such, but rather with society as a whole. Maternal smoking habits are thus an especially interesting determinant since it is modifiable through policy conduct, e.g. by regulating taxes on tobacco products. While medical research gives much attention to why smoking causes low birthweight, the above has led economists to focus primarily on the extent of this effect, and the associated costs. This perspective has the advantage of allowing analysts to disregard the specific medical links between maternal smoking and low birthweight, when using appropriate methods.

In an attempt to estimate the direct costs associated with birthweight, Almond et al. (2005) use data from hospitals in New York and New Jersey to find that the costs peak at \$150.000 (in year 2000 dollars) for newborns weighing 800 grams. In contrast, an infant weighing 2000 grams has an estimated associated cost of \$15.000. The soaring costs at the low end of the birthweight distribution highlight an important point. Using traditional mean regression will only uncover effects on the birthweight mean, i.e. infants weighing around 3500 grams. One way to overcome this problem is to use a quantile regression approach, which can provide estimation results across the entire distribution. This is done by Abrevaya (2001) and Koenker and Hallock (2001), who find justification for the quantile approach since regression estimates vary throughout the distribution. It is, however, troublesome to consider the estimated effects as causal, because the analyses do not account for unobserved heterogeneity. Not only is the susceptibility of smoking effects among mothers different, as noted above, but there are undoubtedly many other individual characteristics which cannot be accounted for.

Econometric panel data models allow controlling for unobserved heterogeneity. However, their extension to a quantile regression framework is still somewhat limited. In a recent paper, Abrevaya and Dahl (2005) consider the extension of the "correlated random effects" model by Chamberlain (1984) to a quantile regression framework, and estimate the effects of various birth inputs on birthweight, using data from Arizona and Washington. Their results indicate that the negative effects of smoking, albeit present, are significantly lower in magnitude across all quantiles than the corresponding cross-sectional estimates.

This paper extends the results of Abrevaya and Dahl, using Danish data. Furthermore, we consider a similar extension of the model by Mundlak (1978) which, at the cost of a more restricted specification, allows for the use of an unbalanced dataset. Finally, we consider the fixed effects quantile specification of Koenker (2004) and the two-stage fixed effects approach recently suggested by Arulampalam et al. (2007).

The outline of the paper is as follows. In Section 2 we first review three panel data models specified for conditional means. We then discuss their extension to a quantile regression framework. In Section 3 we give a description of the data we will use for our estimations. Section 4 provides results and interpretations. Concluding remarks are given in Section 5.

2 The Models

The main difficulty of examining the causal effect of prenatal smoking, and other relevant observable variables, on birthweight outcomes is the possible existence of infinitely many other influential determinants. The identification and measurement of all these determinants is an impossible task, and thus it is necessary to control for unobserved effects. Panel data models provide various clever ways of dealing with this issue, three of which will be our point of departure in this analysis. First we will introduce the basic conditional expectation models, and then discuss their extension to a quantile estimation setup.

Consider a setup in which data are available on a large sample of mothers, each contributing with two or more births. The basic linear panel data model can then be specified as

$$y_{mb} = \boldsymbol{x}'_{mb}\boldsymbol{\beta} + c_m + u_{mb}, \quad m = 1, \dots, M;\ b = 1, \dots, B_m, \tag{2.1}$$

where the subscripts m and b index mothers and births respectively. The dependent variable y denotes birthweight in grams, the vector $\boldsymbol{x}$ contains the observable variables, c is the unobserved "mother effect", and u is a birth-specific disturbance. Stacking observations and sorting by mothers, we can rewrite the model more compactly on matrix form as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zc} + \boldsymbol{u}, \tag{2.2}$$

Here the matrix $\boldsymbol{Z}$ is a block diagonal incidence matrix of the form $\operatorname{diag}(\boldsymbol{e}_{B_1}, \dots, \boldsymbol{e}_{B_M})$, where $\boldsymbol{e}_i$ is an i-vector of ones. In the case of a balanced panel, i.e. $B_m \equiv B$, this simplifies to $\boldsymbol{I}_M \otimes \boldsymbol{e}_B$.

In the following we shall consider three different panel data models, i.e. three approaches on how to treat the unobserved mother-specific elements $\boldsymbol{c}$. The specifications we will discuss are two "random effects"-type models, provided by Chamberlain (1984) and Mundlak (1978) respectively, along with the fixed effects model.

In a classic random effects version of (2.2) it is assumed that the unobserved effects are random and uncorrelated with the observable variables. This assumption will most likely be violated in the current application, and thus this model will not be applicable. It is therefore interesting to consider some variations of the random effects model where some correlation is allowed for.

Chamberlain (1984) assumes that the unobserved effects can be viewed as linear projections onto the observables. The model requires a balanced panel, i.e. $B_m \equiv B$, and the unobserved elements are specified as

$$\begin{aligned} c_m &= \psi + \boldsymbol{x}'_{m1}\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{x}'_{mB}\boldsymbol{\lambda}_B + v_m, \\ &= \psi + \boldsymbol{x}'_m\boldsymbol{\lambda} + v_m, \end{aligned} \tag{2.3}$$

where $\boldsymbol{x}_m \equiv (\boldsymbol{x}_{m1}, \dots, \boldsymbol{x}_{mB})'$ and $\boldsymbol{\lambda} \equiv (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_B)'$. The scalar ψ is constant across mothers, and v is a mother-specific disturbance, which by definition of linear projections satisfies $\mathbb{E}(v_m) = 0$. This specification allows for an individual intercept which is dependent on the observed variables from all births for a given mother. The model consists of equations (2.2) and (2.3), and the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \psi)$ can be estimated by least squares given $\mathbb{E}(u_{mb}|\boldsymbol{x}_m, c_m) = 0, \forall m, b$ (Wooldridge, 2002, p. 324).

A similar approach by Mundlak (1978) assumes a slightly more restrictive projection, where the unobserved effects are specified as projections onto an average of the observables, taken over birth inputs for a given mother. This is given by

$$\begin{aligned} c_m &= \psi + B_m^{-1}(\boldsymbol{x}_{m1} + \cdots + \boldsymbol{x}_{mB_m})'\boldsymbol{\pi} + v_m, \\ &= \psi + (B_m^{-1}\boldsymbol{e}'_{B_m}\boldsymbol{x}_m)\,\boldsymbol{\pi} + v_m. \end{aligned} \tag{2.4}$$

The advantage of this approach, in contrast to the Chamberlain model, is that it allows for the use of an unbalanced panel, possibly adding several observations to the sample. In the case of a balanced panel the model can be seen as a restricted version of Chamberlain's approach, with $\boldsymbol{\lambda}_1 = \cdots = \boldsymbol{\lambda}_B = B^{-1}\boldsymbol{\pi}$. The parameters $(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\pi})$ can likewise be estimated using OLS.

Finally, we consider the fixed effects model, which by subtracting an average across observations from the standard model (2.1), i.e. a within transformation, eliminates the (constant) unobserved mother-specific term, and thus allows for correlation with the observable variables. Letting $\bar{y}_m$, $\bar{\boldsymbol{x}}_m$, and $\bar{u}_m$ denote averages across b, the equation to be estimated is

$$y_{mb} - \bar{y}_m = (\boldsymbol{x}_{mb} - \bar{\boldsymbol{x}}_m)\boldsymbol{\beta} + u_{mb} - \bar{u}_m. \tag{2.5}$$

It is well known that $\boldsymbol{\beta}$ can be consistently estimated using OLS. The fixed effects estimator is equivalent to the one obtained from a least squares dummy variable regression (LSDV), i.e. an estimation of (2.2) directly. However, this can become a cumbersome task when the number of cross-sectional units, and thus the number of dummy variables included, becomes large.

Quantile Estimation Framework

In some applications, focusing on conditional means may provide the desired insight on causality. However, many economic policy concerns should focus on entire distributions. For example, it will not be sufficient to consider averages when conducting policy targeting wage or wealth distribution. Likewise, in the case of low birth weight outcomes, interest lies in the left tail of the distribution. Quantile regression estimation allows for a more complete picture of cause and effect throughout the distribution.

Quantile regression can be compared to the classical least squares minimization problem, with the conditional mean specified as $\mathbb{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ can be estimated by solving

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2. \tag{2.6}$$

Similarly, if specifying the conditional quantile function as $\mathbb{Q}_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$ for the quantile τ, then $\boldsymbol{\beta}_\tau$ can be estimated by finding the $\boldsymbol{\beta}$ that solves

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{2.7}$$

where

$$\rho_\tau(u) = \begin{cases} u\tau & \text{if } u > 0 \\ u(\tau - 1) & \text{if } u < 0 \end{cases}, \tag{2.8}$$

is a check function that ensures non-negativity and scales according to the desired quantile. A thorough review of the subject can be found in Koenker (2005).

One hurdle which must be addressed is the common assumption of a "nice" linear data-generating process, e.g. as (2.1). Such an assumption, however, will rarely imply a corresponding linear specification of the conditional quantile function. Consider an arbitrary specification of the unobserved effect

$$c_m = \phi(\boldsymbol{x}_m) + v_m, \quad \mathbb{E}(v_i|\boldsymbol{x}_m) = 0. \tag{2.9}$$

This, along with (2.1), implies the conditional quantile function

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \boldsymbol{x}'_{mb}\boldsymbol{\beta} + \phi(\boldsymbol{x}_m) + \mathbb{Q}_\tau(u_{mb} + v_m|\boldsymbol{x}_m). \tag{2.10}$$

The function ϕ is often assumed to be linear, and the essential difficulty is how to handle the last term in the expression properly. It is not straight-forward to go from the assumption of a linear data-generating process to a conditional quantile function that is directly applicable. On the other hand, if we specified a linear conditional quantile function, we would have similar difficulties determining the functional form of the data-generating process. We thus need some simplifying assumptions in order to extend the models above to a quantile framework. We follow Abrevaya and Dahl (2005) and assume the following

$$v_m \text{ is independent of } \boldsymbol{x}_m, \tag{2.11}$$

$$\mathbb{Q}_\tau(u_{mb}|\boldsymbol{x}_m, v_m) = \mathbb{Q}_\tau(u_{mb}|\boldsymbol{x}_{mb}). \tag{2.12}$$

Using these assumptions, we can simplify the conditional quantile function of the error terms to

$$\mathbb{Q}_\tau(u_{mb} + v_m|\boldsymbol{x}_m) = \mathbb{Q}_\tau(u_{mb} + v_m|\boldsymbol{x}_{mb}) \equiv f_{\tau,b}(\boldsymbol{x}_{mb}), \tag{2.13}$$

and write (2.10) as

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \boldsymbol{x}'_{mb}\boldsymbol{\beta} + f_{\tau,b}(\boldsymbol{x}_{mb}) + \phi(\boldsymbol{x}_m). \tag{2.14}$$

In general $f_{\tau,b}(\boldsymbol{x}_{mb})$ will be a non-linear function. A simple example illustrating this is given in Abrevaya and Dahl (2005). Thus in order to obtain a linear expression for the conditional quantile function, we need to make assumptions about $f_{\tau,b}(\boldsymbol{x}_{mb})$. We will assume that $\boldsymbol{x}'_{mb}\boldsymbol{\beta} + f_{\tau,b}(\boldsymbol{x}_{mb})$ can be approximated by $\boldsymbol{x}'_{mb}\boldsymbol{\beta}_\tau$, giving us the conditional quantile function

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \boldsymbol{x}'_{mb}\boldsymbol{\beta}_\tau + \phi(\boldsymbol{x}_m). \tag{2.15}$$

Under these simplifying assumptions, Abrevaya and Dahl (2005) have extended Chamberlain's model to quantiles. The result, which we will henceforth refer to as the AD-model, is given by

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \psi_\tau + \boldsymbol{x}'_{mb}\boldsymbol{\beta}_\tau + \boldsymbol{x}'_m\boldsymbol{\lambda}_\tau, \tag{2.16}$$

This model can be estimated, by stacking observations and sorting by mothers, to obtain a right-hand side design matrix $D = [D'_1, \ldots, D'_M]'$, where

$$D_m = \begin{bmatrix} 1 & \boldsymbol{x}'_{m1} & \boldsymbol{x}'_{m1} & \cdots & \boldsymbol{x}'_{mB} \\ 1 & \boldsymbol{x}'_{m2} & \boldsymbol{x}'_{m1} & \cdots & \boldsymbol{x}'_{mB} \\ & & \vdots & & \\ 1 & \boldsymbol{x}'_{mB} & \boldsymbol{x}'_{m1} & \cdots & \boldsymbol{x}'_{mB} \end{bmatrix}. \tag{2.17}$$

If the model contains time-invariant variables or dummy variables, i.e. variables that do not belong in $\boldsymbol{x}$, then these should be included as columns in (2.17).[1]

[1] Our notation differs from the one of Abrevaya and Dahl (2005), in that we have not let the mother-specific effects depend on the specific birth through ψ. Thus to capture this effect it would be required to include dummies in the main equation.

Analogously, by making the same assumptions for the Mundlak specification, we can extend this to a quantile framework, where (2.15) becomes

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \psi_\tau + \boldsymbol{x}'_{mb}\boldsymbol{\beta}_\tau + (B_m^{-1}\boldsymbol{e}'_{B_m}\boldsymbol{x}_m)\,\boldsymbol{\pi}_\tau, \tag{2.18}$$

The block elements of D corresponding to (2.17) becomes

$$D_m = \left[\boldsymbol{e}_{B_m} \vdots\, \boldsymbol{x}_m \vdots\, \boldsymbol{e}_{B_m} \otimes (B_m^{-1}\boldsymbol{e}'_{B_m}\boldsymbol{x}_m)\right], \tag{2.19}$$

where we have used $\vdots$ as a separator for readability. When the data consist of a balanced panel, the entire design matrix can be written as

$$D = \left[\boldsymbol{e}_{M\times B} \vdots\, \boldsymbol{X} \vdots\, [\boldsymbol{I}_M \otimes (B^{-1}\boldsymbol{e}_B\boldsymbol{e}'_B)]\boldsymbol{X}\right]. \tag{2.20}$$

As with the AD-model, any time-invariant or dummy variables should be included in the main equation, and thus added as columns in D.

Finally, we consider the extension of the fixed effects model. As noted above, this can be estimated either by differencing out the unobserved effects or by a dummy variable regression when considering mean regression. However, differencing is not an option, since conditional quantiles are not linear operators. The dummy variable regression, on the other hand, is feasible and is described by Koenker (2004). One downside of the dummy variable regression is the enormous amount of parameters to estimate, especially when M is large and B_m is small. This is typically the case when considering birth outcomes. Koenker mitigates the problem by estimating all desired quantiles simultaneously and holding unobserved effects dummies constant across quantiles. Furthermore, he adds a penalty term to control the variability introduced by the large number of estimated parameters.

Assuming again that the conditional quantile function can be specified as (2.15), the fixed effects model is

$$\mathbb{Q}_\tau(y_{mb}|\boldsymbol{x}_m) = \psi_\tau + \boldsymbol{x}'_m\boldsymbol{\beta}_\tau + c_m. \tag{2.21}$$

The simultaneous estimation of quantiles and the penalty term alter the minimization problem. In contrast to (2.7), Koenker proposes to solve

$$\min_{\boldsymbol{c},\boldsymbol{\beta}} \sum_{k=1}^{q}\sum_{m=1}^{M}\sum_{b=1}^{B_m} w_k\rho_{\tau_k}(y_{mb} - c_m - \boldsymbol{x}'_{mb}\boldsymbol{\beta}_{\tau_k}) + \lambda\sum_{m=1}^{M}|c_m|, \tag{2.22}$$

where q is the number of quantiles to estimate, w_k is the weight assigned to quantile k in order to control the relative influence of the quantile on the estimation of $\boldsymbol{c}$. The parameter λ controls the impact of the penalty term. In the limit as $\lambda \to 0$, the penalty term disappears, and at the other extreme with $\lambda \to \infty$, we have that $\boldsymbol{c} \to \boldsymbol{0}$, i.e. a model with no fixed effects.

Using the above specification of the minimization problem, the design matrix can be written as

$$D = \begin{bmatrix} \operatorname{diag}(\boldsymbol{w}) \otimes \boldsymbol{X} & \boldsymbol{w} \otimes \boldsymbol{Z} \\ \boldsymbol{0}_M & \lambda\boldsymbol{I}_M \end{bmatrix}, \tag{2.23}$$

where $\boldsymbol{0}_M$ is an M-vector of zeros, $\boldsymbol{w}$ is the vector of weights, and $\boldsymbol{Z}$ is defined as in (2.2). The corresponding response vector is $[(\boldsymbol{w} \otimes \boldsymbol{y})' \vdots\, \boldsymbol{0}'_M]'$.

Recently, Arulampalam et al. (2007) have suggested an alternative quantile regression fixed effects approach. In particular, they propose – in the first stage – to obtain a consistent estimator of the fixed effects from applying LSDV on equation (2.1). Denote this estimator $\hat{c}_m$. In the second stage c_m in equation (2.21) is substituted with $\hat{c}_m$. Subsequently, estimates of β_τ can be obtained from standard quantile regression on the observables and $\hat{c}_m$. Although the statistical properties of the two stage fixed effects estimator are unknown the approach is quite appealing due to its simplicity.

Owing to our simplifying assumptions regarding the conditional quantile functions, the design matrices for these three models can be used to obtain point estimates for both quantile and mean regression, using linear quantile regression and OLS, respectively. However, because of dependence between each mother's births, the standard asymptotic-variance formulas cannot be used to obtain standard errors. Furthermore, a standard bootstrapping approach is not applicable either. Instead, a subsampling bootstrap approach should be used where random subsamples of mothers (including all their births) are drawn repeatedly with replacement (Abrevaya and Dahl, 2005).

3 Data

The data used in this study were obtained from Aarhus University Hospital, Skejby, in Denmark. In the Aarhus region this hospital is the only one with a maternity ward, and thus the data in fact represent a broad population group, i.e. all economic and social classes. The models considered in this paper require a panel of mothers with two or more registered births. For the variables of interest the data offer a panel consisting of 20.407 births and 9.629 mothers. The descriptive statistics for the dataset are given in **Table 3.1**.

	1st child		2nd child		3rd child		4th child	
Variable	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
Birthweight	3504.33	(532.82)	3647.79	(531.88)	3673.48	(555.05)	3622.97	(535.56)
Doctor visits	3.05		2.99		2.93		2.79	
Complications	0.22		0.25		0.28		0.27	
Test tube baby	0.02		0.01		0.00		0.01	
Birth control pills	0.27		0.16		0.14		0.11	
Weight	63.99	(10.90)	65.25	(12.01)	65.82	(12.35)	65.70	(12.18)
Prenatal visits	5.10		4.59		4.41		4.38	
Age	28.20	(3.78)	30.94	(3.83)	33.11	(3.92)	34.64	(4.16)
Smoked before	0.28		0.23		0.24		0.28	
Smoked during	0.13		0.12		0.15		0.21	
Drink	0.03		0.03		0.03		0.03	
Male child	0.51		0.50		0.51		0.50	
Married	0.43		0.65		0.76		0.72	
Height	168.62	(6.11)	168.60	(6.14)	168.20	(6.07)	167.41	(6.01)
Diabetes	0.01		0.01		0.02		0.02	

	Birthweight quantiles			
Quantile	1st child	2nd child	3rd child	4th child
10%	2900	3030	3040	3009
25%	3200	3330	3350	3280
50%	3500	3650	3670	3650
75%	3840	3990	4020	3980
90%	4150	4300	4350	4291
Observations	8276	9003	2678	450

Table 3.1: Descriptive statistics for the Aarhus Birth Cohort.

Some of the variables in the table are self-explanatory, while others require a brief description. *Doctor visits* refers to the number of consultations with a doctor during the pregnancy. *Complications* is a binary variable indicating whether or not there were any kind of complications during the delivery. *Test tube baby* is a binary variable indicating whether the child was conceived using artificial insemination. *Birth control pills* is also a binary variable which indicates whether or not the mother has used these within a 4 month period prior to the pregnancy. *Prenatal visits* is the number of consultations with a midwife during the pregnancy. *Smoked before* and *smoked during* are both binary variables for the smoking status of the mother before and during the pregnancy, respectively. *Drink* is a binary variable indicating whether or not the mother has consumed alcohol during the pregnancy.[2] Finally, *diabetes* is a binary variable which indicates whether or not the mother had diabetes at one or more of the registered pregnancies.

The choice of variables is highly motivated by the study of Abrevaya and Dahl (2005). In order to reproduce their results we need the same, or at least similar explanatory variables. The variables for which this was possible are *weight*, *age*, *drink*, *male child*, and *married.* Furthermore, prenatal smoking status and prenatal care are represented, although in a slightly different way. We use data on smoking both during and before pregnancy, allowing for smoke to have causal effect in different ways. Prenatal care is not represented by dummies for the occurrence of the first prenatal care visit, but only as the number of actual consultations with a midwife. Even though the data sets display similarities, there are also a few differences. One variable, or a proxy, could not be obtained, namely the mother's education, and some additional variables have been included. These are the mother's height, birth complications, artificial insemination, usage of birth control pills, and diabetes.

The descriptive statistics show some interesting differences to the data from Arizona and Washington. First, it can be noticed that, on average, the birthweights are around 100 grams higher both for first and second child observations. Furthermore, it seems that mothers in Denmark are almost three years older on average. This is also the case both for the first and second child. Finally the data indicate that it is more commonplace to have children out of wedlock in Denmark, than it is in Arizona and Washington.

The models discussed in the previous section do not all allow the use of an unbalanced panel, which opts for a division of the data. This study will make use of three sub data sets, a balanced panel containing the two first observed births of each mother (19.528 observations), and an unbalanced panel containing all observations (20.407). Furthermore, for a more precise comparison, we include a balanced panel with the restriction that the two observed births must actually be the mother's two first births (15.620 observations). We will refer to this as the restricted balanced panel.

[2] We define alcohol consumption as more than 1 Danish standard drink (12 grams of pure alcohol) per week.

4 Estimation Results

Applying the models from Section 2 to the data described in Section 3, the causal effects of maternal smoking and the other observed variables on birthweight are estimated. The model specification includes the variables summarized in **Table 3.1**, dummy variables indicating whether an observation is a mother's second, third, or fourth birth, along with dummy variables for birth years. In order to allow for non-linear effects in *weight* and *age* these are also included in quadratic forms.

The choice of birhtweight as the dependent variable gives rise to an important question: should gestational age be included as an explanatory variable? There is no doubt that gestational age is correlated with birthweight and that it is not captured in the unobserved heterogeneity. However, in the present analysis the interest lies in the total effect of maternal smoking on birthweight, including any effects propagated through gestational age. Therefore it is not necessary to include gestational age as an explanatory variable, it might even be inappropriate as it could have undesirable effects due to multicollinearity. In this context it should also be emphasized that on the matter of not including gestational age as explanatory variable we follow recent leading econometric studies on birthweight, see e.g. Abrevaya (2001, 2006), Abrevaya and Dahl (2005), Chernozhukov (2005), and Koenker and Hallock (2001).

All estimations agree on the presence of a significant adverse effect of maternal smoking *during* pregnancy, leaving little doubt that smoking may reduce birthweight to some extent. This, along with other interesting findings from each model, will be discussed in the following. First, in the interest of comparison with Abrevaya and Dahl (2005), we examine the AD-model. Second, the pros and cons of the extended version of the Mundlak model are presented, along with a discussion of the balanced and unbalanced datasets. Finally, we comment on the usability of the fixed effects model in the present analysis. Providing a complete set of results would be too extensive and unnecessary for our purpose. The assessment is therefore restricted to results yielding main insights and differences between the models. A complete set of results is, however, available upon request from the corresponding author.

The AD-model

The results for the AD-model estimation and a comparable cross-sectional estimation, i.e. a pooled regression ignoring all unobserved heterogeneity, are presented in **Tables 4.1–4.2**. These will serve as a good frame of reference for our discussion since they are comparable to Abrevaya and Dahl (2005, tables 2–3) and illustrate important points regarding the analysis.

The variables *second child* and *male child* are significant and positive across all quantiles in both the cross-section and AD-model estimations. This is to be expected since it is generally acknowledged that the birthweight of male children is higher on average, and that the birthweight increases with the parity of the mother. This is generally in line with Abrevaya and Dahl's results except that their estimates of the *second child*-coefficients are smaller.

The interpretation of the variables *doctor visits* and *prenatal visits* is somewhat difficult. The main problem with *prenatal visits* is that there may be two reasons for consulting a midwife, either as a routine/precautionary measure or because of complications. It is not possible to distinguish between these two effects of the variable. In spite of this, the estimates are significant, positive, and slightly decreasing across the quantiles, and clearly capture some effect. However, it is difficult to say

	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
Second child	189.755 ***	194.697 ***	180.800 ***	154.846 ***	142.358 ***	163.750 ***
	(23.750)	(16.637)	(13.678)	(25.372)	(38.138)	(21.215)
Height	9.043 ***	9.429 ***	10.485 ***	12.016 ***	11.227 ***	11.122 ***
	(1.444)	(1.054)	(0.945)	(1.037)	(1.268)	(0.861)
Diabetes	208.023 **	166.742 **	258.316 ***	307.097 ***	346.674 ***	259.728 ***
	(105.063)	(72.229)	(50.345)	(60.695)	(107.671)	(54.452)
Doctor visits	9.080	11.546	14.837 *	15.242 *	7.884	15.895
	(14.641)	(10.602)	(8.014)	(8.199)	(10.184)	(9.939)
Complications	-107.590 ***	-69.679 ***	-34.493 **	-35.107 **	-20.853	-68.540 ***
	(26.905)	(17.739)	(15.457)	(17.628)	(27.202)	(12.466)
Test tube baby	-58.471	-39.007	9.673	-25.361	-245.637 **	-51.388
	(125.002)	(84.679)	(51.349)	(70.056)	(108.613)	(47.038)
Birth control pills	-34.941	-30.793	-22.358	-6.226	4.336	-16.644
	(23.717)	(18.873)	(16.013)	(16.754)	(23.662)	(11.500)
Weight	21.008	1.484	4.390	16.663 *	7.021	11.087
	(14.566)	(11.126)	(8.744)	(9.691)	(14.654)	(7.202)
Weight^2	-0.140	-0.009	-0.041	-0.109 *	-0.014	-0.069
	(0.097)	(0.073)	(0.058)	(0.064)	(0.096)	(0.047)
Prenatal visits	98.923 ***	82.284 ***	73.899 ***	71.518 ***	73.636 ***	59.392 ***
	(12.795)	(7.779)	(5.468)	(5.793)	(7.735)	(14.269)
Age	-19.790	-7.228	-19.173	-33.333	-36.239	-29.097
	(36.486)	(26.033)	(22.047)	(25.435)	(35.063)	(18.757)
Age^2	0.294	0.118	0.286	0.674 *	0.827	0.553 *
	(0.586)	(0.421)	(0.357)	(0.399)	(0.567)	(0.293)
Smoked before	-9.656	-10.147	-2.290	-17.972	-19.368	-17.993
	(40.758)	(34.466)	(27.275)	(30.851)	(39.134)	(22.182)
Smoked during	-199.603 ***	-175.664 ***	-97.892 ***	-61.750	-44.052	-91.221 ***
	(55.014)	(39.147)	(31.715)	(39.620)	(41.454)	(26.386)
Drink	-0.283	5.985	-16.021	-59.112	-90.497 *	-54.434 *
	(56.018)	(51.755)	(40.477)	(41.017)	(54.334)	(29.501)
Male child	146.448 ***	137.056 ***	142.064 ***	166.553 ***	185.202 ***	150.489 ***
	(20.990)	(13.946)	(11.113)	(12.939)	(17.596)	(8.961)
Married	15.998	-12.503	4.690	8.819	-7.206	9.084
	(29.768)	(20.529)	(17.250)	(20.335)	(28.925)	(14.457)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.
Bootstrapped standard errors are given in parentheses. The bootstrapping was done using samples of 10.000 births and 1.000 iterations.

Table 4.1: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the AD-model applied to the restricted balanced dataset. The dependent variable is birthweight in grams.

whether it is the prenatal visits as such that increases birthweight or if it is simply a proxy for something else. One could, for instance, imagine that mothers of lower social class would be less inclined to attend prenatal visits due to either economic or social reasons. If, say, lower class mothers in general give birth to less healthy babies, then this could account for a significant part of the observed effect. If this is the case, we would expect the estimates from the AD-model to be lower than the cross-section estimates, because the mother's social class would be part of the unobserved heterogeneity. When not controlling for this, the effect could be captured in *prenatal visits*. The estimates of the AD-model are, however, only slightly smaller than the cross-section estimates.

The interpretation of *doctor visits* is also difficult since the need for a doctor visit may or may not be related to the pregnancy, and the assessment of when a doctor visit is actually necessary is very individual. Thus we are again faced with a problem of unobserved heterogeneity. Our estimates, however, show only little significance, so judging whether the AD-model handles this is impossible. A comparison with Abrevaya and Dahl is troublesome for these two variables since they account for

	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
Second child	180.412 ***	200.236 ***	175.475 ***	177.905 ***	174.807 ***	180.189 ***
	(16.554)	(10.846)	(9.403)	(11.439)	(16.311)	(9.532)
Height	8.883 ***	9.619 ***	10.248 ***	11.675 ***	11.007 ***	10.931 ***
	(1.405)	(0.970)	(0.929)	(1.018)	(1.228)	(0.845)
Diabetes	35.502	125.087 *	264.538 ***	274.123 ***	288.340 ***	224.711 ***
	(106.122)	(70.557)	(46.208)	(53.998)	(105.230)	(53.099)
Doctor visits	14.270	12.286	14.102 **	11.308	1.219	17.498 *
	(10.530)	(7.776)	(6.699)	(7.868)	(10.192)	(9.359)
Complications	-221.364 ***	-126.828 ***	-78.780 ***	-66.966 ***	-58.654 ***	-129.241 ***
	(22.370)	(13.909)	(11.313)	(11.913)	(16.762)	(11.476)
Test tube baby	-9.895	-7.795	-9.278	-22.087	-31.191	-19.736
	(73.345)	(46.189)	(29.450)	(37.648)	(48.922)	(32.649)
Birth control pills	-14.933	-14.974	-20.171 *	-12.654	-8.765	-12.284
	(17.279)	(11.985)	(11.964)	(13.836)	(13.729)	(10.030)
Weight	22.821 ***	24.132 ***	22.685 ***	20.730 ***	21.981 ***	23.043 ***
	(5.927)	(3.886)	(3.308)	(4.066)	(4.646)	(3.159)
Weight^2	-0.125 ***	-0.124 ***	-0.109 ***	-0.089 ***	-0.089 ***	-0.111 ***
	(0.040)	(0.027)	(0.022)	(0.028)	(0.031)	(0.021)
Prenatal visits	109.217 ***	96.544 ***	92.358 ***	94.920 ***	88.196 ***	72.130 ***
	(12.436)	(7.172)	(5.400)	(4.814)	(6.615)	(15.870)
Age	29.311	13.737	2.735	-20.771	6.849	8.149
	(18.129)	(16.294)	(12.283)	(13.504)	(18.272)	(11.879)
Age^2	-0.494	-0.248	-0.057	0.348	-0.102	-0.143
	(0.304)	(0.271)	(0.204)	(0.226)	(0.298)	(0.195)
Smoked before	7.580	24.190	-1.754	4.349	15.803	0.843
	(21.009)	(16.820)	(14.697)	(18.916)	(26.121)	(15.619)
Smoked during	-204.451 ***	-207.515 ***	-171.138 ***	-163.841 ***	-136.874 ***	-166.515 ***
	(31.334)	(23.073)	(19.527)	(28.523)	(39.527)	(23.651)
Drink	10.210	-3.196	9.895	-52.831 *	-57.418 *	-10.840
	(44.174)	(28.541)	(24.523)	(30.981)	(34.270)	(22.964)
Male child	115.649 ***	113.763 ***	121.437 ***	149.842 ***	165.181 ***	133.374 ***
	(13.979)	(9.640)	(8.504)	(9.871)	(13.576)	(7.725)
Married	-19.810	-16.545	-8.808	-1.487	10.025	-8.323
	(16.037)	(11.098)	(9.788)	(11.077)	(14.541)	(8.944)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.
Bootstrapped standard errors are given in parentheses. The bootstrapping was done using samples of 10.000 births and 1.000 iterations.

Table 4.2: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the cross-sectional estimation based on the restricted balanced dataset. The dependent variable is birthweight in grams.

prenatal care with different variables, but in general they also find that prenatal care is significant.

In the cross-sectional case, the estimates of *smoked during* are significant across the entire distribution, suggesting an effect of 140–200 grams reduction in birthweight. It seems plausible, however, that the variable also captures the effect of an overall unhealthy lifestyle. This should be controlled for in the AD-model producing a different picture, and surely this is the case. The AD-model shows a similarly large and significant effect at the far left part of the distribution. This effect, however, diminishes more rapidly and loses significance when we move towards the right tail. This is in contrast to of Abrevaya and Dahl's findings, where there is a pronounced reduction in the effect across the *entire* distribution. Our model specification also includes the variable *smoked before.* Interestingly, this variable is not significant, indicating no direct effects from maternal smoking prior to pregnancy on birth outcomes.

The remaining variables comparable to Abrevaya and Dahl's study are *married*, *age*, and *age^2*. Here our results differ substantially. In the study by Abrevaya and

Dahl all three variables are significant in both estimations, whereas our results show almost no significance. A reason for this could be that America has a substantial social gap compared to Denmark. This will undoubtedly have consequences for young and/or unmarried mothers in America, who do not have the same social benefits as offered in Denmark.

Our model includes three variables that describe the mother's physical size, *weight*, *weight^2*, and *height*. If we consider the cross-sectional results first, we find that the estimates are significant across the distribution, which indicates that they affect birth weight. However, the same results are not present in the AD-model. The conclusions of the weight coefficients change dramatically. While the mother's weight significantly explains some of the unobserved heterogeneity through $\boldsymbol{\lambda}$, e.g. genes or an unhealthy lifestyle, which may have causal effect on the birthweight outcome, it does not have direct effect through the point estimates. The two estimations obtain similar results for *height*.

Diabetes is commonly thought to increase the birthweight of babies. This is also seen in both estimations. The AD-model shows a positive and significant estimate across all quantiles, where the cross-sectional estimation lacks significance in the left tail. *Complications* shows a decreasing effect, in absolute terms, across the quantiles. This seems plausible since birth complications are more common in cases of very low birthweights. After controlling for unobserved effects, the coefficients are approximately halved and even become insignificant at the 90% quantile.

The last three variables, *test tube baby*, *birth control pills*, and *drink*, are all mostly insignificant in both estimations. It seems puzzling that *drink* has very little significance. In view of the negative health implications attributed to alcohol consumption, it could be expected to have a negative effect on birthweight. There could be many reasons why this does not show up in our results, one of which could be measurement or reporting errors.

The Extended Mundlak Model

The motivation for considering an extension of the Mundlak model to quantiles is that, at the expense of a restriction on the projection of c_m, it allows for unbalanced datasets. As argued in Section 2, the Mundlak model can be viewed as a restricted version of the AD-model when considering balanced datasets. Therefore, it would be tempting to expect the results of the two models to be similar, which in fact they turn out to be.

If we assume the AD-model is a well specified model for estimating effects on birthweight, it seems that the restriction imposed by using the extended Mundlak model does not invalidate the results, and this justifies its use when beneficial. To explore this further we will now consider the balanced dataset, i.e. the dataset without the restriction that only first and second children are allowed. The estimation results for the extended Mundlak model are given in **Table 4.3**.

Comparing the results to those of the corresponding AD-model estimation reveals only few differences, and this further supports the idea that the models reach similar conclusions. These findings suggest that, at least for our specification, we can extend the AD-model analysis to unbalanced datasets by using the extended Mundlak model. The obvious reason for using an unbalanced panel is that it increases our sample size. There is, however, also a more subtle reason. The restricted balanced dataset used for the AD-model estimation in **Table 4.1** only includes the first two births for a given mother. This is a sample selection we have not justified. Restrict-

	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
Height	7.978 ***	9.248 ***	10.345 ***	11.604 ***	10.524 ***	10.580 ***
	(1.364)	(0.904)	(0.866)	(0.992)	(1.276)	(0.818)
Diabetes	227.743 ***	233.965 ***	264.737 ***	320.381 ***	333.542 ***	287.184 ***
	(78.996)	(63.938)	(43.952)	(45.670)	(86.036)	(44.881)
Second child	159.972 ***	165.738 ***	146.458 ***	161.901 ***	146.070 ***	152.574 ***
	(20.503)	(13.641)	(14.239)	(12.274)	(20.485)	(13.645)
Third child	190.762 ***	181.595 ***	165.659 ***	217.390 ***	221.757 ***	187.903 ***
	(31.421)	(25.315)	(21.456)	(23.266)	(28.936)	(20.110)
Fourth child	124.089 *	183.398 ***	194.305 ***	167.819 ***	221.412 ***	171.732 ***
	(63.518)	(48.495)	(43.837)	(45.089)	(55.041)	(33.644)
Doctor visits	5.323	6.708	16.147 ***	11.461 *	1.767	13.929 **
	(10.983)	(8.566)	(6.111)	(6.939)	(9.545)	(6.892)
Complications	-126.132 ***	-72.458 ***	-48.998 ***	-38.661 **	-17.035	-77.132 ***
	(24.225)	(16.424)	(13.320)	(15.879)	(21.691)	(10.996)
Test tube baby	-40.857	-34.183	32.675	-32.220	-183.582 **	-43.127
	(108.593)	(79.441)	(55.083)	(57.907)	(82.413)	(40.005)
Birth control pills	-47.132 *	-51.985 ***	-25.949 *	-5.240	2.269	-27.736 **
	(25.968)	(16.642)	(13.609)	(16.969)	(22.594)	(11.810)
Weight	4.171	0.058	1.766	4.208	7.319	6.388
	(13.149)	(9.136)	(7.791)	(7.250)	(11.752)	(5.587)
Weight^2	-0.022	0.009	-0.019	-0.025	-0.011	-0.035
	(0.085)	(0.058)	(0.053)	(0.047)	(0.078)	(0.036)
Prenatal visits	106.033 ***	88.020 ***	72.258 ***	73.366 ***	65.808 ***	63.255 ***
	(9.229)	(5.066)	(4.430)	(5.063)	(7.072)	(13.384)
Age	-31.579	-9.213	-0.366	-35.329 *	4.155	-10.110
	(29.321)	(19.721)	(16.621)	(20.443)	(27.703)	(14.503)
Age^2	0.625	0.236	0.113	0.666 **	0.076	0.280
	(0.466)	(0.311)	(0.256)	(0.326)	(0.437)	(0.223)
Smoked before	-41.428	-8.077	-13.029	-4.285	-17.518	-21.043
	(40.110)	(27.490)	(25.356)	(25.926)	(36.435)	(19.314)
Smoked during	-154.714 ***	-145.682 ***	-105.910 ***	-118.914 ***	-29.055	-111.761 ***
	(54.337)	(36.197)	(31.635)	(35.411)	(49.585)	(25.267)
Drink	37.855	-39.930	-41.613	-60.300	-45.709	-45.961 *
	(55.184)	(39.995)	(33.765)	(36.719)	(43.853)	(25.302)
Male child	132.131 ***	130.009 ***	135.214 ***	163.873 ***	182.605 ***	149.059 ***
	(16.549)	(11.625)	(9.596)	(12.618)	(15.593)	(8.295)
Married	14.053	-0.299	8.881	1.107	-9.919	6.347
	(26.185)	(17.923)	(15.248)	(17.731)	(28.270)	(12.808)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.

Bootstrapped standard errors are given in parentheses. The bootstrapping was done using samples of 10.000 births and 1.000 iterations.

Table 4.3: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the Mundlak-model applied to the balanced dataset. The dependent variable is birthweight in grams.

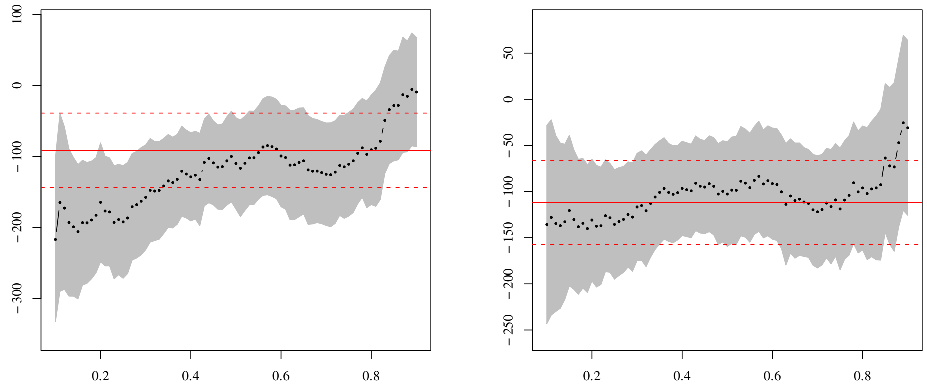
	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
Height	8.017 *** (1.381)	9.329 *** (0.881)	10.345 *** (0.854)	11.783 *** (0.945)	10.640 *** (1.226)	10.609 *** (0.769)
Diabetes	214.311 *** (78.770)	237.259 *** (65.245)	268.596 *** (42.555)	313.671 *** (47.417)	365.995 *** (89.944)	286.970 *** (46.080)
Second child	162.594 *** (17.867)	163.974 *** (14.272)	146.047 *** (14.192)	167.760 *** (11.937)	151.549 *** (18.573)	158.344 *** (11.854)
Third child	191.163 *** (26.972)	185.084 *** (22.762)	174.986 *** (19.963)	232.630 *** (21.385)	225.607 *** (26.919)	200.432 *** (17.404)
Fourth child	149.724 *** (52.256)	174.944 *** (34.491)	176.530 *** (38.501)	194.989 *** (38.876)	212.862 *** (45.732)	176.321 *** (29.381)
Doctor visits	5.223 (9.385)	6.812 (7.524)	15.735 *** (5.979)	11.922 * (6.302)	0.917 (7.955)	12.991 ** (6.511)
Complications	-141.136 *** (21.490)	-77.764 *** (14.846)	-54.438 *** (11.708)	-36.287 ** (14.139)	-13.061 (19.629)	-78.756 *** (10.038)
Test tube baby	-54.188 (100.113)	-14.497 (74.896)	12.181 (55.983)	-35.263 (55.975)	-169.351 ** (79.365)	-47.760 (39.067)
Birth control pills	-47.964 ** (22.014)	-48.141 *** (15.761)	-26.857 ** (12.948)	-2.581 (16.523)	-1.977 (20.412)	-28.424 ** (11.312)
Weight	2.530 (11.288)	-1.276 (9.119)	2.758 (7.827)	4.114 (7.118)	17.714 (11.029)	7.857 (5.680)
Weight^2	-0.012 (0.074)	0.019 (0.059)	-0.021 (0.053)	-0.023 (0.046)	-0.079 (0.072)	-0.042 (0.037)
Prenatal visits	105.219 *** (8.102)	88.322 *** (4.991)	72.275 *** (4.504)	72.906 *** (4.895)	65.419 *** (6.440)	64.910 *** (12.127)
Age	-13.588 (24.921)	-12.287 (16.166)	-5.671 (17.833)	-30.932 * (17.144)	3.166 (24.278)	-9.703 (13.137)
Age^2	0.334 (0.392)	0.327 (0.250)	0.183 (0.271)	0.551 ** (0.269)	0.038 (0.386)	0.251 (0.199)
Smoked before	-29.795 (35.511)	-7.694 (25.186)	-15.973 (23.699)	-4.508 (25.682)	-23.288 (33.935)	-15.269 (18.088)
Smoked during	-135.706 ** (54.916)	-135.621 *** (30.292)	-102.690 *** (28.994)	-118.826 *** (33.754)	-30.931 (48.382)	-112.046 *** (23.212)
Drink	10.109 (53.593)	-31.438 (35.293)	-37.624 (35.422)	-49.840 (36.005)	-47.725 (41.802)	-42.651 * (24.605)
Male child	119.863 *** (15.397)	126.476 *** (10.807)	135.747 *** (9.161)	161.193 *** (11.461)	192.018 *** (140)	145.486 *** (7.466)
Married	1.389 (24.369)	-12.561 (16.776)	1.967 (14.179)	-4.695 (16.298)	-16.802 (24.628)	-2.811 (11.225)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.
Bootstrapped standard errors are given in parentheses. The bootstrapping was done using samples of 10.000 births and 1.000 iterations.

Table 4.4: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the Mundlak-model applied to the unbalanced dataset. The dependent variable is birthweight in grams.

ing the data in such a way induces a potential risk of biased estimates. By using an unbalanced panel we can reduce this sample selection problem, as all available observations are included. The estimates based on this panel, using the extended Mundlak model, are presented in **Table 4.4**.

The balanced and unbalanced panels, in this case, do not differ substantially in terms of the number of observations. Therefore, it is hardly surprising that the estimations based on these produce almost identical results. The restricted balanced panel, on the other hand, contains almost 5.000 observations less than the unbalanced panel, a difference which has an impact on some of the estimation results. The most conspicuous difference is the point estimates of *smoked during*, which after the inclusion of the remaining observations, becomes smaller in magnitude, and shows less variation across quantiles. However, the significance is hardly affected, cf. This is conveniently illustrated in **Figure 4.1**.



(a) Mundlak, restricted balanced.

(b) Mundlak, unbalanced.

Figure 4.1: Difference in point estimates of *smoked during* for the extended Mundlak model, including 95% confidence bands, when applied to the restricted balanced and unbalanced datasets. The straight lines are from the OLS estimation.

This naturally leads to the very interesting question of which estimates are more "correct". It would be tempting to conclude that the restricted balanced dataset is biased and that the "correct" effect is smaller in magnitude. However, this is not necessarily true. It is important to realize is that we obtain two different estimates from the restricted balanced and balanced/unbalanced datasets, since the estimates are assumed to be constant across births. When estimations are based on the restricted balanced dataset we obtain an estimate of how *smoked during* affects first and second borns, whereas the balanced and unbalanced datasets give an estimate of causality for first through fourth borns. Thus, if we cannot assume that the estimates are constant across births, then the estimates are not directly comparable.

The Fixed Effects Model

The fixed effects model provides a quite different way of controlling for the unobserved heterogeneity. While the two models above view the heterogeneity as projections onto the observables, the fixed effects model directly estimates these effects, i.e. an estimate for each c_m in (2.21) is obtained. The model proposed by Koenker (2004) assumes these to be constant across quantiles, and thus requires all quantiles to be estimated simultaneously. For this purpose we need to define the weights assigned to each quantile. In his paper, Koenker only considers three quantiles, using Tukey's trimean as a prototype for his weights. This is one way of providing estimates that reflect the central tendency of the distribution of c_m. In our estimation we are not interested in using the estimates of the unobserved heterogeneity directly, their only purpose is to control for the unobserved effects. We therefore simply weigh all quantiles equally. Furthermore, we choose the penalty parameter in the same way as Koenker, i.e. $\lambda = 1$. The estimation results for the restricted balanced dataset are provided in **Table 4.5**.

The table also provides OLS estimates of the model which, however, do not include the penalty. The variables *height* and *diabetes* are for a given mother constant across births, and can therefore not be estimated by means of OLS. The results are

	Quantile Regressions					
	10 %	25 %	50 %	75 %	90 %	OLS
Height	13.304 (218.617)	13.001 (218.648)	13.051 (218.636)	12.607 (218.673)	12.452 (218.648)	
Diabetes	204.062 (44941.512)	215.103 (44938.923)	247.081 (44939.356)	277.842 (44939.642)	270.466 (44939.678)	
Doctor visits	16.249 * (8.732)	15.989 ** (7.655)	16.309 ** (7.182)	14.947 ** (6.736)	13.753 * (7.331)	15.058 (11.093)
Complications	-103.455 *** (27.931)	-80.241 *** (30.384)	-65.932 *** (22.987)	-59.095 *** (21.334)	-58.157 (50.580)	-201.472 *** (44.196)
Test tube baby	-26.827 (71.322)	-19.604 (57.075)	-7.193 (48.216)	-11.861 (52.322)	-28.208 (58.609)	22.162 (42.654)
Birth control pills	-26.219 (16.210)	-23.001 * (13.004)	-20.115 (12.945)	-14.788 (14.562)	-13.085 (16.258)	-13.681 (14.007)
Weight	21.638 * (11.056)	21.776 (13.360)	19.504 (12.657)	19.448 (13.821)	20.162 (12.524)	35.684 *** (4.241)
Weight^2	-0.110 * (0.057)	-0.108 * (0.065)	-0.089 (0.059)	-0.084 (0.065)	-0.087 (0.061)	-0.177 *** (0.029)
Prenatal visits	78.235 *** (9.516)	78.558 *** (8.713)	78.478 *** (8.196)	78.829 *** (8.090)	78.263 *** (7.617)	96.201 *** (22.178)
Age	-2.213 (29.522)	-5.751 (35.504)	-10.954 (34.353)	-17.389 (32.249)	-14.261 (30.816)	63.678 *** (18.684)
Age^2	-0.016 (0.392)	0.054 (0.517)	0.156 (0.488)	0.277 (0.443)	0.236 (0.430)	-1.069 *** (0.297)
Smoked before	-20.371 (30.076)	-11.101 (22.240)	-4.500 (21.819)	-2.090 (22.851)	-2.306 (26.582)	12.217 (20.522)
Smoked during	-164.691 *** (62.883)	-160.723 ** (64.972)	-153.053 ** (60.259)	-143.558 *** (55.380)	-131.266 *** (40.121)	-193.477 *** (29.453)
Drink	-27.749 (44.219)	-30.080 (40.223)	-46.020 (34.820)	-45.556 (36.064)	-53.730 (44.272)	19.880 (32.622)
Male child	126.667 *** (14.287)	133.984 *** (15.215)	146.549 *** (10.395)	153.923 *** (11.437)	157.357 *** (16.738)	116.775 *** (14.750)
Married	7.324 (18.363)	6.297 (16.137)	3.024 (16.281)	1.900 (16.869)	2.688 (17.472)	-32.659 ** (14.055)
Second child	175.163 *** (42.410)	176.564 *** (18.290)	176.647 *** (18.686)	175.074 *** (18.552)	173.241 *** (35.592)	177.532 *** (24.682)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.
Bootstrapped standard errors are given in parentheses. The bootstrapping was done using samples of 10.000 births and 1.000 iterations.

Table 4.5: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the FE-model applied to the restricted balanced dataset. The dependent variable is birthweight in grams.

roughly comparable to those of the AD- and extended Mundlak models. Except for *height* and *diabetes*, the variables show similar significance levels and point estimates. The standard errors for *height* and *diabetes* are, however, quite unsettling. Applying the model to the two larger datasets further adds to our discomfort as the results are altered dramatically. **Figure 4.2** illustrates this by comparing two sets of estimates for *smoked during* using the unbalanced dataset, where the only difference is the omission of the insignificant variable *smoked before*. Obviously, Koenker's method might behave in this fashion because of the lack of meaningful variation in "problem" variables such as *height* and *diabetes*.

However, we believe that an estimation problem of this magnitude causes great numerical stability issues. In particular, it turns out that estimation results based on the balanced and unbalanced panels vary unpleasantly with the choice of computer processor. For some estimates differences even appear on the first digit. This phenomenon is present, but not as pronounced, for the restricted balanced panel. This indicates that numerical errors accumulate drastically with the proportions of the minimization problem.

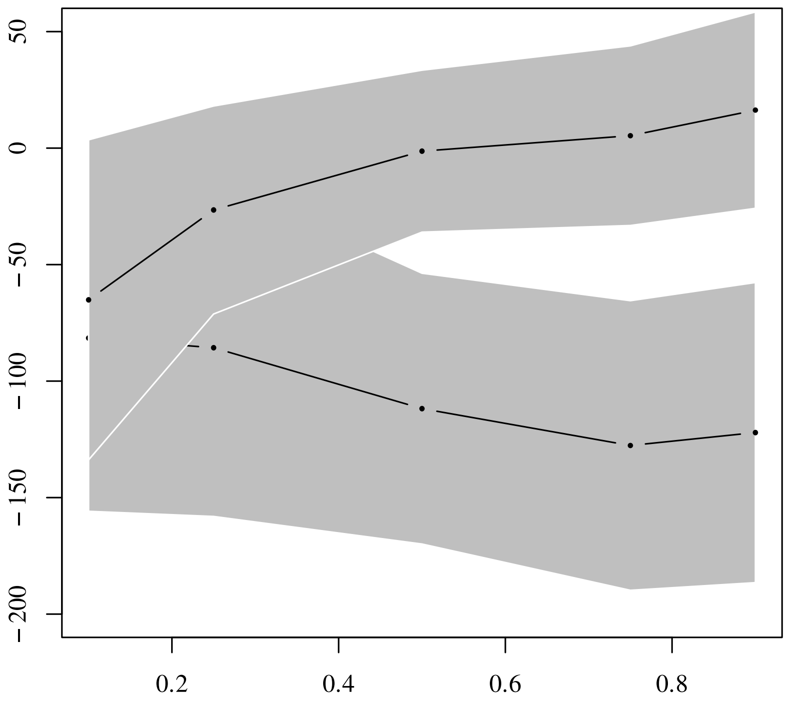


Figure 4.2: Difference in point estimates of *smoked during* for the FE-model, including 95% confidence bands, when removing the insignificant variable *smoked before*.

Assuming that the problem is the dimensions of the panel, we seek to overcome this by employing a simple two-stage estimation procedure along the lines of Arulampalam et al. (2007). The results from using this procedure on the restricted balanced dataset are given in **Table 4.6**. Here the coefficient $\hat{c}_m$ is associated with the estimated unobserved mother-specific effect. With the exception of the coefficients to the variables *height* and *diabetes* that still seem to be problematic (as in Koenker's method), the signs of the estimated effects are generally as expected hereby confirming our previous findings. In addition, there are a few other noticeable features. First, the estimated effects of *smoked during* are of a smaller magnitude numerically and are less than 100 grams uniformly across quantiles. Second, *smoked before* comes out significantly negative at the 10% quantile. Third, *drink* has a significantly negative effect on the right-hand side of the conditional birthweight distribution.

5 Concluding Remarks

In this paper we have used the AD-model to investigate causal effects of birth input on birthweight outcomes. Like those of Abrevaya and Dahl (2005), our results show the importance of considering conditional quantiles and controlling for unobserved heterogeneity when estimating determinants of birthweight outcomes in a Danish

	Quantile Regressions					
	10 %	25 %	50 %	75 %	90 %	OLS
Height	-1.031 *	-0.922	0.148	0.768	0.792	
	(0.608)	(0.752)	(0.119)	(0.537)	(0.571)	
Diabetes	-72.418 *	-53.042	4.062	48.756	100.171 **	
	(43.892)	(33.662)	(3.250)	(31.952)	(50.980)	
Doctor visits	15.008	15.299 *	16.026 *	15.887 *	12.258	16.160 *
	(9.957)	(9.235)	(9.453)	(8.861)	(9.298)	(9.413)
Complications	-94.608 ***	-69.085 ***	-54.620 ***	-36.477 *	-23.680 *	-65.549 ***
	(20.580)	(14.171)	(15.949)	(20.564)	(14.362)	(13.114)
Test tube baby	-82.054	-59.978	-37.668	-47.442	-26.815	-51.721
	(58.974)	(51.942)	(48.787)	(50.519)	(54.297)	(48.180)
Birth control pills	-18.737	-15.467	-19.528 *	-14.214	-12.229	-16.850
	(14.470)	(12.182)	(11.847)	(12.241)	(13.083)	(11.688)
Weight	9.839	12.672 *	11.030	10.582	14.013 *	11.638
	(8.059)	(7.332)	(7.440)	(7.564)	(7.772)	(7.409)
Weight^2	-0.070	-0.084 *	-0.069	-0.060	-0.079	-0.072
	(0.053)	(0.049)	(0.050)	(0.051)	(0.051)	(0.050)
Prenatal visits	70.206 ***	66.021 ***	64.351 ***	64.611 ***	66.923 ***	58.637 ***
	(13.705)	(13.736)	(14.971)	(12.848)	(10.655)	(15.558)
Age	-45.808 *	-39.803	-37.273	-32.939	-26.179	-34.806
	(25.818)	(24.821)	(24.766)	(24.735)	(24.510)	(24.721)
Age^2	0.632 **	0.535 *	0.520 *	0.468 *	0.378	0.477 *
	(0.302)	(0.284)	(0.284)	(0.284)	(0.285)	(0.284)
Smoked before	-53.044 **	-31.555	-17.527	-5.696	7.902	-19.561
	(26.112)	(23.617)	(21.045)	(23.044)	(23.142)	(21.040)
Smoked during	-86.668 ***	-97.116 ***	-90.409 ***	-91.580 ***	-89.592 ***	-92.724 ***
	(30.385)	(26.322)	(26.311)	(26.581)	(28.957)	(25.904)
Drink	-39.208	-38.222	-53.992 *	-64.381 **	-73.152 **	-55.082 *
	(36.464)	(33.790)	(29.967)	(32.516)	(33.813)	(29.948)
Male child	138.050 ***	146.482 ***	149.144 ***	159.331 ***	153.927 ***	150.052 ***
	(11.515)	(9.899)	(9.428)	(11.601)	(11.749)	(9.389)
Married	8.045	14.965	8.424	9.063	9.313	10.921
	(16.711)	(15.038)	(14.282)	(14.117)	(14.781)	(14.089)
Second child	164.540 ***	162.812 ***	162.711 ***	160.855 ***	156.740 ***	160.170 ***
	(20.179)	(19.266)	(18.964)	(19.010)	(18.617)	(19.001)
$\hat{c}_m$	1.038 ***	1.009 ***	1.000 ***	0.986 ***	0.951 ***	1.000 ***
	(0.018)	(0.008)	(0.000)	(0.010)	(0.023)	(0.000)

Asterisks denote the significance level (double-sided). *: 10%, **: 5%, ***: 1%.
Bootstrapped standard errors are given in parentheses. The bootstrap was done using a sample size of 10.000 births and 1.000 iterations.

Table 4.6: Quantile regression estimates of $\boldsymbol{\beta}_\tau, \tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and OLS estimates of $\boldsymbol{\beta}$ for the two-step plug-in FE-model applied to the restricted balanced dataset. The dependent variable is birthweight in grams.

dataset. An example of this is the change in magnitude and significance of prenatal smoking. Controlling for unobserved effects does not change the fact that smoking reduces birthweight, but it shows that the effect is primarily a problem in the left tail of the distribution on a slightly smaller scale. We have argued that a restriction based on the model by Mundlak (1978) may allow the use of unbalanced datasets, without this affecting conclusions. We find that including more births is not trivial and changes both the value and the interpretation of the estimates. Thus, while the Mundlak-type extension seems to hold for an unbalanced dataset, the estimation objective should be clear when deciding upon the data to be used. Finally, we consider two fixed effects estimation approaches applicable to quantile regression models proposed by Koenker (2004) and Arulampalam et al. (2007), respectively. With respect to Koenker's approach, we find the estimations to be numerically unstable and not very suitable for our specific purpose. The results based on the two-stage fixed effects estimation approach proposed by Arulampalam et al. (2007) are overall in overall agreement with the AD and Mundlak estimations, but with the important exception that the estimated effects of *smoking* are reduced significantly in magnitude at the lower quantiles.

The possible weakness of our methodology is the simplifying assumptions that were needed in order to go from a linear specification of the data-generating process to a linear conditional quantile model. This is a challenging problem for further research, as also noted by Abrevaya and Dahl (2005).

Acknowledgments

We are grateful to Jason Abrevaya, Tine B. Henriksen and Kirsten Wisborg for helpful comments. Lise V. Hansen offered invaluable assistance by providing access to the data from Aarhus University Hospital, Skejby. The terms of the data-sharing agreements do not allow release of these data. However, codes for estimation in Ox/R are available from the corresponding author upon request. We would also also like to thank Roger Koenker and Daniel Morillo, who wrote the RQ 1.0 package for Ox/R that was used as the starting point for the quantile estimation in this paper. The second author gratefully acknowledges financial support from the Danish Social Sciences Research Council (grant 275-06-0105) and the Center for Research in Econometric Analysis of Time Series, CREATES, funded by The Danish National Research Foundation.

References

Abrevaya, J. (2001). The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes. *Empirical Economics 26*, 247–257.

Abrevaya, J. (2006). Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach. *Journal of Applied Econometrics 21*(4), 489–519.

Abrevaya, J. and C. M. Dahl (2005, May). The Effects of Birth Inputs on Birthweight: Evidence from Quantile Estimation on Panel Data. *Journal of Business and Economic Statistics (forthcoming)*.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The Costs Of Low Birth Weight. *The Quarterly Journal of Economics*, 1031–1083.

Arulampalam, W., R. A. Naylor, and J. Smith (2007). Am I Missing Something? The Effects of Absence From Class on Student Performance. *University of Warwick, Working Paper*.

Bernstein, I. M., J. D. Horbar, G. J. Badger, A. Ohlsson, and A. Golan (2000). Morbidity and Mortality Among Very-Low-Birth-Weight Neonates with Intrauterine Growth Restriction. *American Journal of Obstetrics and Gynecology 182*(1), 196–206.

Bernstein, I. M., J. A. Mongeon, G. J. Badger, L. Solomon, S. H. Heil, and S. T. Higgins (1978). Maternal Smoking and Its Association With Birth Weight. *Obstetrics & Gynecology 106*(5), 986–991.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005). From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes. *IZA Discussion Paper Series 1864*.

Chamberlain, G. (1984). Panel Data. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume 2, pp. 1247–1318. Elsevier Science B. V.

Chernozhukov, V. (2005, January). Inference for Extremal Conditional Quantile Models (Extreme Value Inference for Quantile Regression). *MIT, Working Paper*.

Corman, H. and S. Chaikind (1998). The Effect of Low Birthweight on the School Performance and Behavior of School-Aged Children. *Economics of Educations Review 17*(3), 307–316.

DiFranza, J. R., C. A. Aligne, and M. Weitzman (2004). Prenatal and Postnatal Environmental Tobacco Smoke Exposure and Children's Health. *Pediatrics 113*, 1007–1015.

Hack, M., N. K. Klein, and H. G. Taylor (1995). Long-Term Developmental Outcomes of Low Birth Weight Infants. *The Future of Children 5*(1), 176–196.

Hofhuisi, W., J. C. de Jongste, and P. J. F. M. Merkus (2003). Adverse Health Effects of Prenatal and Postnatal Tobacco Smoke Exposure on Children. *Arch. Dis. Child. 88*, 1086–1090.

Kirkegaard, I., C. Obel, M. Hedegaard, and T. B. Henriksen (2006). Gestational Age and Birth Weight in Relation to School Performance of 10-Year-Old Children: A Follow-up Study of Children Born After 32 Completed Weeks. *Pediatrics 118*, 1600–1606.

Koenker, R. (2004, May). Quantile Regression and Longitudinal Data. *University of Illinois, Working Paper*.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. and K. F. Hallock (2001). Quantile Regression. *The Journal of Economic Perspectives 15*(4), 143–156.

Linnet, K. M., C. Obel, E. Bonde, P. Hove, Thomsen, N. J. Secher, K. Wisborg, and T. B. Henriksen (2006). Cigarette Smoking During Pregnancy and Hyperactive-Distractible Preschooler's: A follow-up Study. *Acta Pædiatrica 95*, 694–700.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica 46*(1), 69–85.

Permutt, T. and J. R. Hebel (1989). Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight. *Biometrics 45*(2), 619–622.

Wang, X., B. Zuckerman, C. Pearson, G. Kaufman, C. Chen, G. Wang, T. Niu, P. H. Wise, H. Bauchner, and X. Xu (2002). Maternal Cigarette Smoking, Metabolic Gene Polymorphism, and Infant Birth Weight. *JAMA 287*(2), 195–202.

Wisborg, K., U. Kesmodel, T. B. Henriksen, S. F. Olsen, and N. J. Secher (2000). A Prospective Study of Smoking During Pregnancy and SIDS. *Arch. Dis. Child. 83*, 203–206.

Wisborg, K., U. Kesmodel, T. B. Henriksen, S. F. Olsen, and N. J. Secher (2001). Exposure to Tobacco Smoke in Utero and the Risk of Stillbirth and Death in the First Year of Life. *American Journal of Epidemiology 154*(4), 322–327.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press.

Research Papers 2008

CREATES

Center for Research in Econometric Analysis of Time Series

2008-07 Changli He, Annastiina Silvennoinen and Timo Teräsvirta: Parameterizing unconditional skewness in models for financial time series

2008-08 Cristina Amado and Timo Teräsvirta: Modelling Conditional and Unconditional Heteroskedasticity with Smoothly Time-Varying Structure

2008-09 Søren Johansen and Bent Nielsen: An analysis of the indicator saturation estimator as a robust regression estimator

2008-10 Peter Christoffersen, Kris Jacobs, Christian Dorion and Yintian Wang: Volatility Components, Affine Restrictions and Non-Normal Innovations

2008-11 Peter Christoffersen, Kris Jacobs, Chayawat Ornthanalai and Yintian Wang: Option Valuation with Long-run and Short-run Volatility Components

2008-12 Tom Engsted and Stig V. Møller: An iterated GMM procedure for estimating the Campbell-Cochrane habit formation model, with an application to Danish stock and bond returns

2008-13 Lars Stentoft: Option Pricing using Realized Volatility

2008-14 Jie Zhu: Pricing Volatility of Stock Returns with Volatile and Persistent Components

2008-15 Jie Zhu: Testing for Expected Return and Market Price of Risk in Chinese A-B Share Market: A Geometric Brownian Motion and Multivariate GARCH Model Approach

2008-16: Jie Zhu: FIEGARCH-M and and International Crises: A Cross-Country Analysis

2008-17: Almut E. D. Veraart: Inference for the jump part of quadratic variation of Itô semimartingales

2008-18: Michael Sørensen: Parametric inference for discretely sampled stochastic differential equations

2008-19: Anne Péguin-Feissolle, Birgit Strikholm and Timo Teräsvirta: Testing the Granger noncausality hypothesis in stationary nonlinear models of unknown functional form

2008-20: Stefan Holst Bache, Christian M. Dahl and Johannes Tang Kristensen: Determinants of Birthweight Outcomes: Quantile Regressions Based on Panel Data