Econometric Game 2012:

How does maternal smoking during pregnancy affect infants' birthweight?

Case A

April 18, 2012

1 Introduction

Low birthweight is associated with adverse health related and economic outcomes. In the short run, newborns with low birthweights have higher mortality rates and higher chances of birth defects. Various studies have shown that, in the long run, babies with low birthweights will have lower earnings and education attainment.

There are many determinants of birthweight which range from genetic factors, demographic and psychological factors, obstetric factors and mother behaviors such as cigarette smoking, alcohol consumption, caffeine intake, and narcotic addictions. Such behaviors are associated with lower birthweigh, but the extent to which they are causal needs further examination.

In this paper, we analyze a cross-section data from 141929 observations to estimate the effects of maternal smoking during pregnancy on birthweight. We will perform four different sets of analyses to address a variety of issues. First, we use a simple OLS framework to select and specify a reduced form model that gives the linear effects of smoking on birthweight.

Second, we perform a quantile regression to understand what features of the conditional distribution of weights depend on smoking. We complement the previous analysis with a probit regression, which allows us to estimate the effect of smoking on the probability of low, intermediate, and high birthweight outcomes. In the last section, we discuss the limitations of the results, in particular, we discuss the extent to which the results can be interpreted as causal. We propose a simple instrumental variable model to address this problem.

2 Data

We analyze the marginal empirical distributions of the different variables in the data sets. We noticed there are outliers present in the dependent variable (y_i) and the independent variable $cigs_i$.

The empirical distribution for the dependent variable is very wide. More formally, the sample excess kurtosis is high (slightly above 5), but there is no reasonable ground to beileve that this is driven by measurement error. High kurtosis is usually an indicator that "heavy tails" may be a potential concern, specially in the inference exercise we are interested in.[1] The WHO classification for low birthweight uses the 2500 grams threshold. The empirical cumulative distribution at this point contains sufficient observations to claim that standard asymptotic theory will be a reasonable guide for inference.

The outliers in the empirical distribution of $cigs_i$ limits how smoking can be measured. With few observations from high values of cigarettes smoked per-day, inference may be heavily influenced by the extreme observations, even though they are small relative to the sample size. Therefore, we prefer to measure smoking using dummy variables that preserve the measure of intensity. Specifically, we *distinguish the intensity of cigarette consumption by enriching the definition of smoking status.* We create two additional dummy variables.

- Moderate smoking dummy: Indicates cigarette consumption between half a pack to a

[1] This is specifically adressed by **?** in a quantile regression set-up.

pack a day.

- Heavy smoking dummy: Indicates cigarette consumption of more than one pack per-day.

3 Model Selection

The first challenge is to select an appropriate specification to perform the analysis. We will do the selection in two stages: first, to select the appropriate measures of cigarette consumption, and then to decide between a multiplicative and additive model.

- **Step 1**: For both of the outcomes birthweight and ln(birthweight), we estimate ordinary least squares specifications of the form:

$$E^*[y_i|X_i] = \gamma_1' X_i$$

We compare models that use three different measures of cigarette consumption. The first is similar to that of **?** and includes only a dummy for reporting having smoked cigarettes. The second includes additional indicators for moderate smoking (10-20 cigarettes per day) and heavy smoking (20+ cigarettes per day). The third includes the smoking dummy, cigarettes consumed, and cigarettes squared. In all regressions we control for marital status, race, male child, education dummies, and the five measures of pre-natal care as well as current state fixed effects. Standard errors are clustered by current state.

- **Step 2**: We next use cross-validation to compare the out of sample fits of our chosen linear and log specifications. We split our sample randomly in to 10 equally-sized groups and perform "leave one group out" cross validation for each of the ten groups. For each model we compute the normalized mean squared error as follows:

$$MSE = \frac{(Y_i - \hat{Y}_i)^2}{var(Y_i)}$$

Results for the level specification are attached at the end of the paper.

For both birth weight and log birthweight (regression results not reported), the model including continuous measures of cigarettes has the lowest Bayesian information criterion, indicating the best in-sample fit. However, the estimates in this model are disproportionately impacted by outliers, as discussed in the data section. We thus proceed with the second specification for both outcomes, which has the second-lowest BIC and which we believe is more robust to outliers and mismeasurement.

Our cross-validation indicates that the linear specification has the lower mean sqaured error (0.903 versus 0.913) and thus we proceed with our analysis in levels.

Our OLS estimation makes clear that including measures of intensity of smoking, above and beyond a dummy for having smoked, improves model fit and is an important addition to the analysis of Abrevaya (2006).

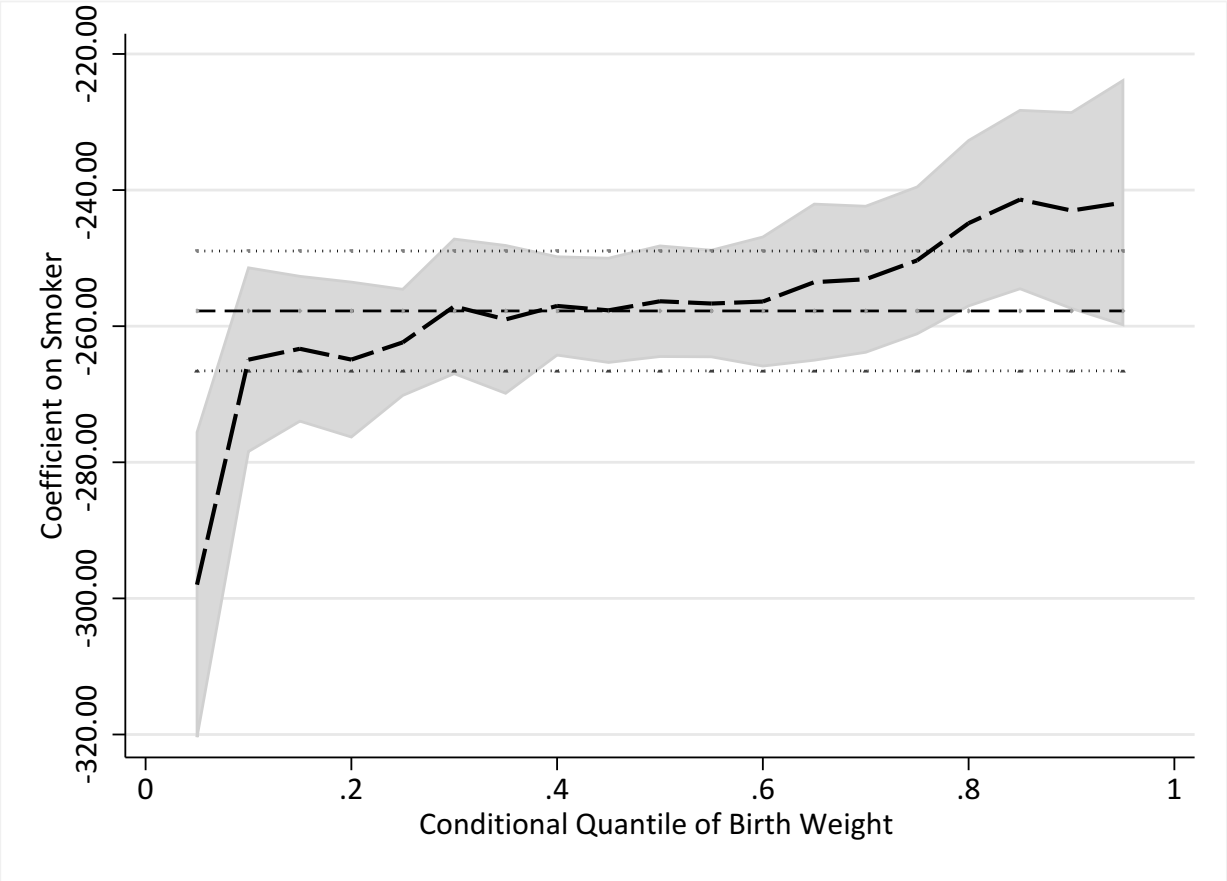
To check for heterogeneity in the treatment effect, we add to our preferred specification interactions between the smoking dummy and demographic and education controls (See table at the end of the paper, Column 4). The coefficients on the interactions are jointly significant at 1 %, confirming the findings of Abrevaya (2006). In particular, smoking x age and age squared, smoking x married, and smoking x high school graduate are significant at 5%.

4 Quantile Regression

As indicated in the analysis of **?** analysis of the conditional mean in OLS does not fully capture the correlation between cigarette consumption and the distribution of birth weights. We thus use quantile regression to assess the impact of smoking on birth weight.

Our quantile regression specifications have the form

Figure 1: Quantile Regression for birthweight: Partial Effect of Smoking Status



$$Q_\tau(Y_i|X_i) = \phi_\tau + f(smoking) + \beta X_i$$

We estimate quantile regression functions for the 10th through the 90th quantiles in 10 percentile intervals. All specifications include the demographic, education, and prenatal care controls as in our preferred OLS specifications and include heteroskedasticity-robust standard errors. Note that given the number of computed quantiles, uniform confidence bands as discussed in Hardle and Song (2010) would be preferred; we leave this improvement for future analysis.

Figure plots the coefficient on "smoking" in the model

$$Q_\tau(Y_i|X_i) = \phi_\tau + smoking_i + \beta X_i$$

along with pointwise 95% confidence intervals. The coefficient from the analogous OLS regression and its confidence intervals are also given in the solid and dotted horizontal lines, respectively. The heterogeneity in the correlation between reporting having smoked and birthweight is immediately apparent: smoking has a disproportionately large, negative coefficient for the bottom quantiles of the distribution and smaller (less negative) coefficients for higher quantiles.

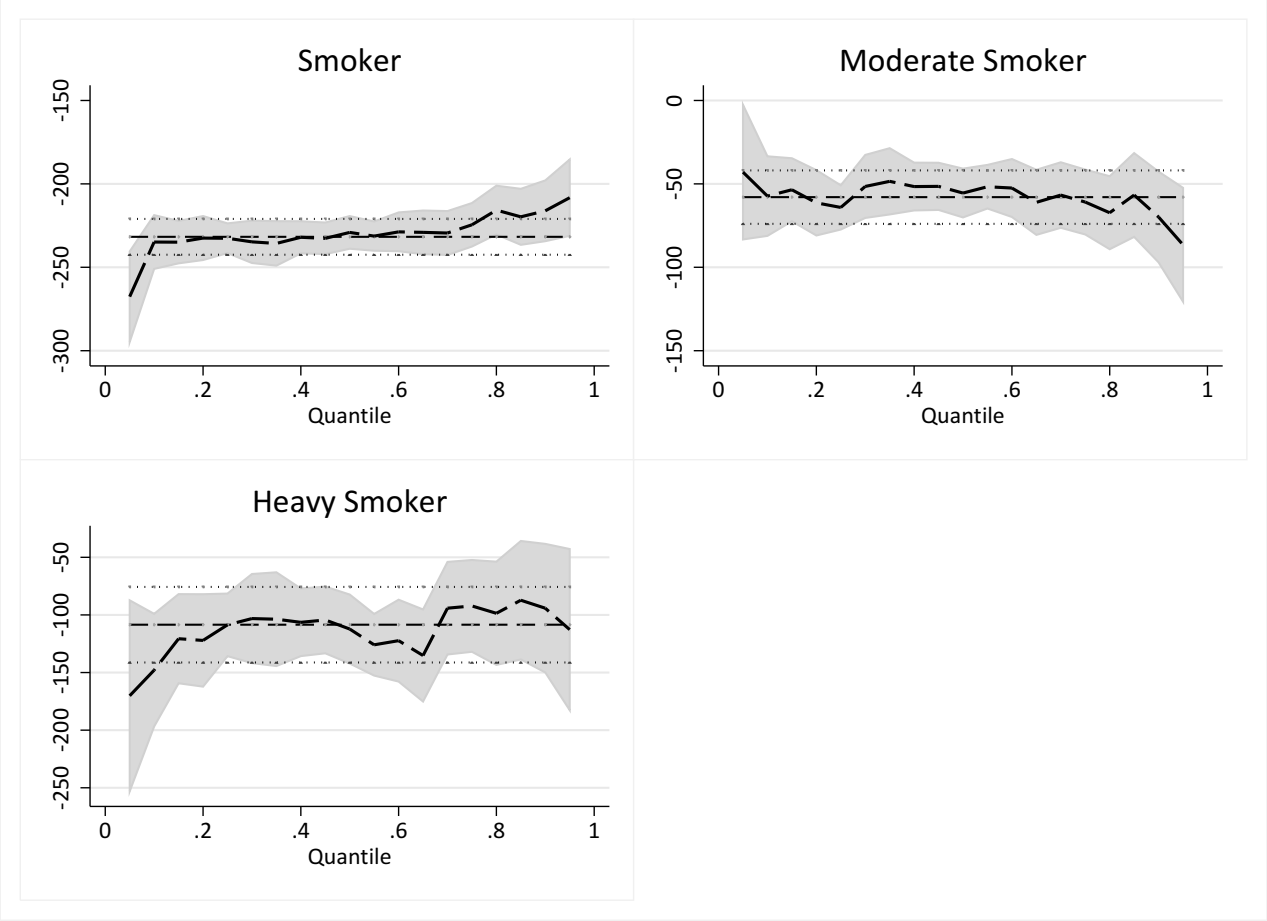
Figure **??** plots the coefficients from the quantile regression including dummies for smoking, moderate smoking, and heavy smoking, along with confidence intervals and OLS estimation results. The pattern in the coefficient on smoking from the previous specification persists. In addition, we see that heavy smoking (defined as smoking 20 or more cigarettes per day) has a disproportionately large negative effect on the bottom quantiles of the conditional birthweight distribution.

Finally, following the OLS finding that interactions between demographics and the smoking dummy are statistically significant, we interact the smoking dummy with various controls and plot the coefficients in Figure Z. There appears to be a strong interaction between black and smoking at the bottom of the outcome distribution, and evidence of some interplay between some college and smoking.

5 Probabilities of low and high birthweights

We now shift our focus to assessing the differential effect of maternal smoking on the relative probabilities that the child's birth weight is high or low. Medical studies show that the health-related and economic costs of abnormal birth weight is largest at extreme birth weights. In particular, the WHO defines low birth weight as less than 2500 g due to the spike in infant mortality rates below this threshold. High birth weights, however, are also associated with

Figure 2: Quantile Regression for birthweight: Partial Effects of Smoking Variables (All Included)



health problems.

We seek to gauge the effect of smoking over a range of birth weight bins. Specifically, we focus on the birth weight intervals

$$I_1 = [0, 2250), I_2 = [2250, 2500), I_3 = [2500, 2750), \ldots,$$

$$I_{10} = [4250, 4500), I_{11} = [4500, 4750).$$

These bins allow us to not only analyze the cut-offs 2500 g and 4000 g for low and high birth weight, respectively, they also make it possible for us to consider intermediate cases which may further inform the inference. Note that we do not include a bin for the highest birth weights. There are only 1105 observations in the dataset with birth weights greater than

4750 g, and so to aid the following graphical exposition we omit those observations for the present purposes. For each of the 11 bins we run a probit regression

$$\text{Prob}(y_i \in I_n | X_i) = \Phi(X_i'\beta_n), \quad n = 1, \ldots, 11 \tag{1}$$

using left-hand side dummies $1_{\{y_i \in I_n\}}$. We then store estimates of the average partial effects (APEs)

$$APE_{mn} = E\left[\frac{\partial}{\partial X_{im}}\Phi(X_i'\beta_n)\right], \quad m = 1, \ldots, M, \; n = 1, \ldots, 11.$$

Here M is the total number of covariates. The list of covariates does not vary with n. Specifically, we use the same list of covariates as in the linear model presented earlier (see Table at the end of the paper, column 2). The exception is that the only smoking-related variables are the smoking dummy and the heavy-smoking dummy (i.e., more than 20 cigarettes smoked per day). We implement the probit regressions in Stata using the commands `probit` and `margins`. The exercise is carried out separately on male and female babies.

Figures **??**,**??** display the estimated APEs for the smoking dummy graphically. The horizontal axis lists the *upper* bound of the relevant bin. The dashed lines indicate *pointwise* (i.e., bin-by-bin) 95% heteroskedasticity robust (White) confidence bands, clustering by the current state of residence.[2] We immediately observe the differential (average partial) effect of smoking on the probability of the birthweight being high, low or intermediate. The results are very similar for the two genders. Clearly, maternal smoking has a significantly negative effect on the relative probabilities of giving birth to a child with high birth weight compared to low birth weight. For example, using data on male children, the APE for the smoking dummy in the $[3000, 3250)$ bin is estimated at 0.047 ± 0.009, i.e., switching from non-smoking to smoking increases the probability of falling in this bin by about 5%,

[2]Ideally, we would like to have reported *uniform* confidence bands that, under standard frequentist assumptions, would cover the entire graph (n, APE_{nm}) with 95% probability in repeated samples. For a finite number of bins, this could be accomplished by a Bonferroni correction, although the band would be very conservative for a modest number of bins. We are not aware of asymptotic theory providing tight uniform confidence bands in the present setting.

averaging over the observed covariates. In comparison, the APE for the $[4000, 4250)$ bin is estimated at -0.059 ± 0.012, i.e., smoking decreases the probability of having a child with a relatively high birth weight in this bin.[3]

One noticeable and perhaps surprising finding is that the absolute value of the APE of the smoking dummy is economically small in the extreme bins (although the APEs are statistically significant from 0 and go in the intuitively right direction). Thus, it appears that the mere fact that a pregnant woman is smoking should not lead us to substantially revise our forecast of how likely it is that the child is born with abnormal birth weight. A priori, it seems possible that only heavy cigarette consumption could substantially affect the tails of the birth weight distribution. Figures **??**,**??** graphically depict the APEs for the heavy smoking dummy, again separately for each gender. We conclude that this dummy is statistically insignificant at the 95% level for almost all bins and both genders.

We are led to conclude that while the binary smoking status of the mother significantly affects the relative probabilities of intermediate birth weights (with the effect going in the direction predicted by biological considerations), we do not find an economically substantial effect at either extreme of the birth weight distribution.

The preceding analysis comes with a major caveat. It is hard to provide a structural interpretation for the probability model (**??**). An alternative approach would be to continue working with the linear model

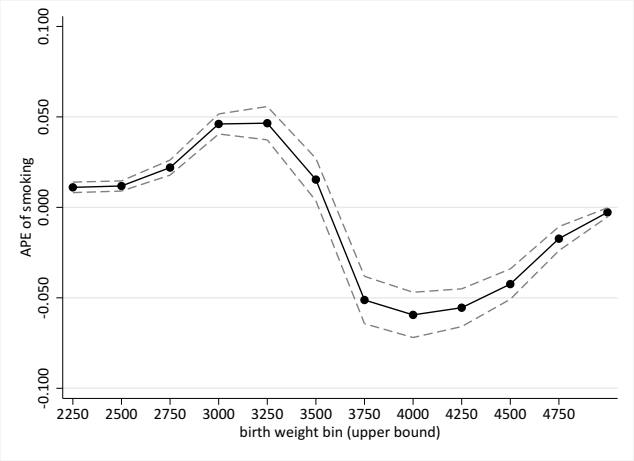
$$y_i = X_i'\beta + \varepsilon_i, \tag{2}$$

where $\varepsilon_i | X_i \sim N(0, \sigma^2)$ and (y_i, X_i) are i.i.d. Then $\text{Prob}(y_i \in (a, b)|X_i) = \Phi((b - X_i'\beta)/\sigma) - \Phi((a - X_i'\beta)/\sigma)$. Hence, given an OLS estimate of β, it would be straight-forward to estimate $\text{Prob}(y_i \in (a, b)|X_i)$ (at some X_i) and obtain asymptotic standard errors using the delta

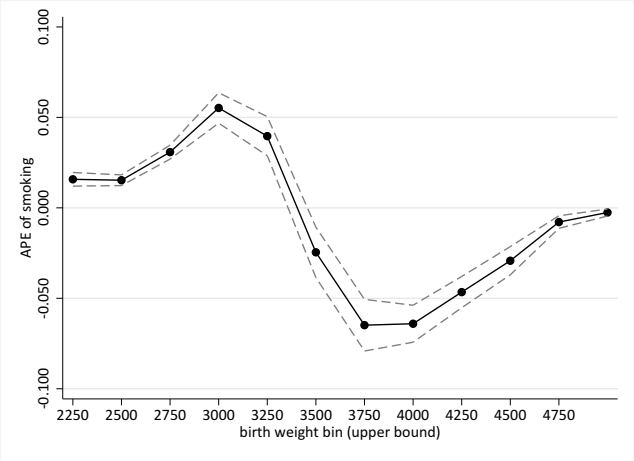
[3] Again, a potential problem is that the standard errors are not uniform, so we technically are not able to compare the magnitude of APEs across bins. One way to make our analysis formally sound, is to restrict attention to two bins of interest (corresponding to high and low weight, respectively) and test the joint hypothesis that both estimates equal zero against a one-sided alternative. The critical values for the two sub-hypotheses would have to be Bonferroni-adjusted. Evidently, any such adjustment would only slightly change the critical values and so our qualitative conclusions would go through.

method. However, such an approach would impose more linearity and homoskedasticity than we are ready to assume. Due to their reduced-form nature, the results presented in this section should be more robust to various departures from the model (**??**), even if they do not lend themselves to causal interpretation.

Figure 3: APEs for Smoking Dummy, Probit Regression

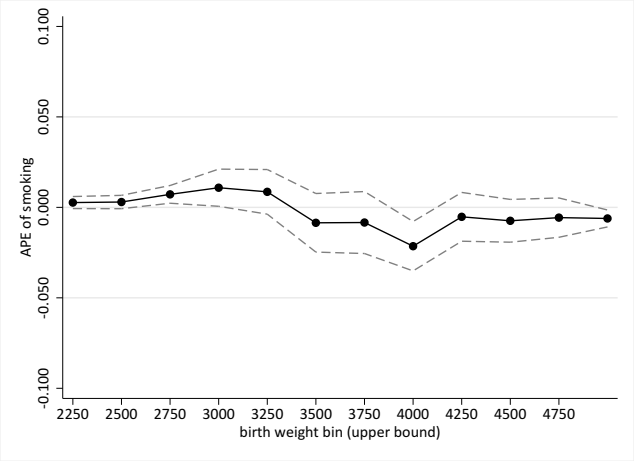


(a) Male

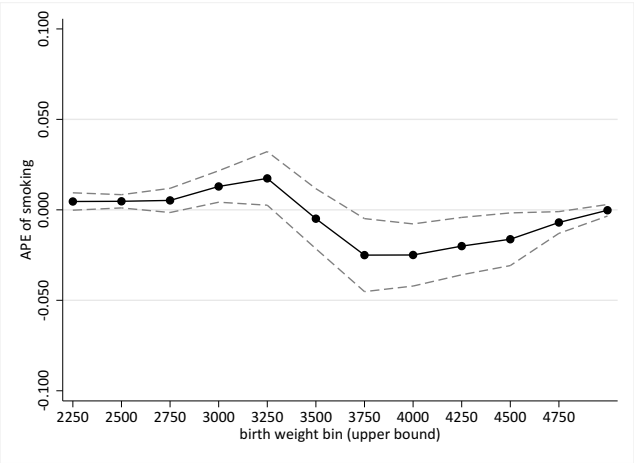


(b) Female

Figure 4: APEs for Heavy Smoking Dummy, Probit Regression



(a) Male



(b) Female

6 Endogeneity analysis

The previous analysis has been of a reduced form nature. In particular, we have not considered the causal interpretation of the smoking variables used in our specification. It is important to establish causality, if the interest is to obtain policy recommendations such as imposition of smoking taxes. For such questions, a specification that does well in a reduced form or forecasting sense will not suffice. Consider the model given by

$$\mathbb{E}^*[y_i|1, s_i, h_i, hc_i, \tilde{X}_i] = \alpha + \beta s_i + \gamma_1 h_i + \gamma_2 hc_i + \gamma_3' X_i \tag{3}$$

where y_i is birthweight, s_i is the smoking indicator variable and $\mathbb{E}^*$ denotes the population best linear predictor of birthweight given the covariates $(1, s_i, pc_i, h_i, hc_i, X_i)$. The random vector X_i contains a set of observed covariates for mother i, including education, race and some others. The random variables h_i and hc_i are two unobserved covariates that we would like to include in our specification to assess the partial effect of smoking over birthweight. These unobserved components will give us a concrete framework to model endogeneity. For simplicity, the two unobserved characteristics considered are the following: personal health/health status of the mother during the gestation period (h_i) and the health consciousness during pregnancy (hc_i).

As it is well known, in the context of model (**??**) the standard Ordinary Least Squares regression analysis based on the linear regression of y_i on the observed variables (s_i, X_i) will deliver a biased estimate for the parameter of interest, β. The omitted variable bias is a consequence of the remaining correlation between the smoking behavior s_i and the unobserved variables (h_i, hc_i) after controlling for the covariates X_i.

In this section, we will propose a simple Instrumental Variable (IV) strategy to address the endogeneity problem. We will start by assuming that the health consciousness (hc_i) component of the unobserved error is captured by the indicator variable of inadequate prenatal care (pc_i).

Assumption 1. $\mathbb{E}^*[y_i|1, s_i, ipc_i, h_i, h_{ic}\tilde{X}_i] = \tilde{\alpha} + \beta s_i + \tilde{\gamma}_1 h_i + \tilde{\gamma_3}'\tilde{X}_i$

This is, the bias in the OLS estimate for β arises only through the correlation between the mother's smoking status and her personal health status. We propose the use of *state of birth* as instrumental variable for the endogenous regressor s_i. Since the omitted variable in the specification is given by the unobserved personal health (h_i) a necessary condition for estimation using an instrumental variable Z_i, requires the instrument to be uncorrelated with personal health. We will assume that the state of birth variable is uncorrelated with the personal health component. Since the state of birth does not seem to have a direct effect over the outcome variable, we use it as an instrument.

The first-stage statistic is significant for conventional levels (3.9). However, it is "weak" under the definitions of **?**. A simple way to perform inference in this model, despite of the problems with the instrument is to use robust methods (like the robust extensions of the CLR)

	(1)	(2)	(3)
	birwt	birwt	birwt
smoker	-252.0***	-257.4***	-254.6***
	[-263.5,-240.5]	[-268.9,-246.0]	[-262.7,-246.5]
N	72076	69099	141175

95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)
	birwt	birwt	birwt
smoker	-216.6*	-216.6*	-226.9**
	[-418.5,-14.83]	[-418.5,-14.83]	[-391.1,-62.73]
N	72076	72076	141175

95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

7 Discussion

The reduced form results that we have obtained indicate an important relation between smoking and birthweight, not only for the conditional mean, but for the different quantiles of the distribution. Our OLS estimate of the partial effect of smoking is in line with estimates in the literature. Our quantile regression results indicated that the effect may be more severe for extremely low birthweights. Hence, it would be interesting to focus future

research on the lower tails of the distribution. The probit specification, on the other hand, found substantial negative effects of smoking in intermediate ranges of birth weights but not at the extremes. Future research should attempt to sort out why the two specifications reach disparate conclusions. Preliminary evidence suggests that the interaction of smoking behavior with socioeconomic and educational indicators is relevant.

As outline in the previous section, our analysis suffers from potential endogeneity problems. We attempted a linear IV specification as a first cut and found that the point estimate decreases in absolute value, as expected. The standard errors are large, however. Future research could attempt to estimate treatment effects based on matching on observables, although the large differences in covariates across smoking status makes such an approach questionable. Ideally, one could incorporate a stronger instrument such as smoking taxes in the relevant states. Another approach to address the endogeneity problem would be to obtain panel data.

References

Abrevaya, J. (2006): "Estimating the effect of smoking on birth outcomes using a matched panel data approach," *Journal of Applied Econometrics*, 21, 489–519.

Abrevaya, J. and C. Dahl (2008): "The effects of birth inputs on birthweight," *Journal of Business and Economic Statistics*, 26, 379–397.

Black, S., P. Devereux, and K. Salvanes (2007): "From the cradle to the labor market? The effect of birth weight on adult outcomes," *Quarterly Journal of Economics*.

Chernozhukov, V. and I. Fernandez-Val (2011): "Inference for Extremal Conditional Quantile Models with an application to Market and Birthweight Risks," *Review of Economic Studies*.

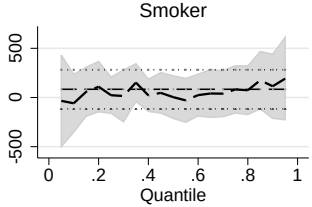
Koenker, R. and K. Hallock (2001): "Quantile regression," *The Journal of Economic Perspectives*, 15, 143–156.

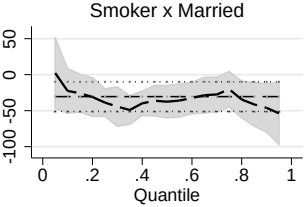
Yogo, M. (2004): "Estimating the elasticity of intertemporal substitution when instruments are weak," *Review of Economics and Statistics*, 86, 797–810.

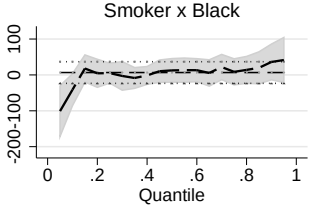
Regressions on Birth Weight (in grams)

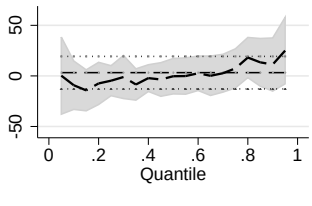
	(1) est1	(2) est2	(3) est3	(4) est4
Smoker	-260.58***	-234.67***	-164.81***	49.89
	(5.77)	(5.97)	(10.69)	(127.45)
Moderate Smoker		-57.36***		-52.54***
		(8.91)		(9.01)
Heavy Smoker		-110.45***		-97.92***
		(14.78)		(15.45)
Cigarettes Per Day			-9.73***	
			(1.26)	
Cigarettes Squared			0.12***	
			(0.03)	
Married	61.40***	61.38***	61.30***	70.91***
	(5.96)	(5.97)	(5.98)	(7.19)
Black	-249.12***	-251.18***	-252.81***	-244.23***
	(5.85)	(5.87)	(5.86)	(7.25)
Male	130.70***	130.83***	130.77***	130.10***
	(3.32)	(3.34)	(3.34)	(3.44)
High School	45.97***	44.18***	43.12***	49.93***
	(6.43)	(6.41)	(6.43)	(7.68)
Some College	70.50***	68.14***	66.77***	68.47***
	(7.94)	(7.91)	(7.93)	(8.60)
College Graduate	69.83***	67.17***	65.74***	62.53***
	(7.44)	(7.43)	(7.47)	(8.45)
No Prenatal Visits	-16.04	-14.31	-13.55	-12.89
	(26.59)	(26.68)	(26.75)	(26.54)
First Visit 2nd Tri	83.06***	83.19***	83.26***	83.98***
	(9.44)	(9.40)	(9.38)	(9.37)
First Visit 3rd Tri	148.64***	148.37***	148.48***	148.67***
	(23.19)	(23.42)	(23.50)	(23.38)
Intermediate Care	-75.35***	-75.12***	-74.99***	-75.36***
	(9.78)	(9.72)	(9.71)	(9.66)
Inadequate Care	-147.18***	-146.55***	-146.14***	-146.81***
	(18.21)	(18.29)	(18.36)	(18.36)
Age	22.65***	23.40***	23.61***	29.87***
	(2.86)	(2.89)	(2.89)	(3.02)
Age Squared	-0.32***	-0.33***	-0.33***	-0.42***
	(0.05)	(0.05)	(0.05)	(0.05)
Interactions with Smoker	No	No	No	Yes
State Fixed Effects	Yes	Yes	Yes	Yes
Observations	141175	141175	141175	141175
Adjusted R^2	0.097	0.097	0.097	0.098
BIC	2162267.46	2162212.07	2162152.18	2162179.31

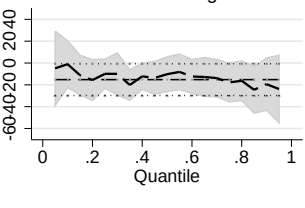
Standard errors in parentheses. Regression coefficients are estimated using OLS. Robust standard errors are clustered by the state of current residence. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

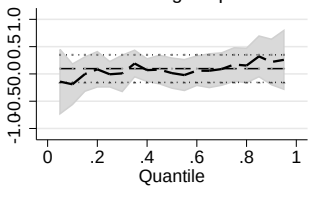

Smoker
-500
0
500
0
.2
.4
.6
.8
1
Quantile

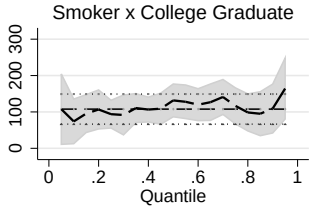

Smoker x Married
-100
-50
0
50
0
.2
.4
.6
.8
1
Quantile

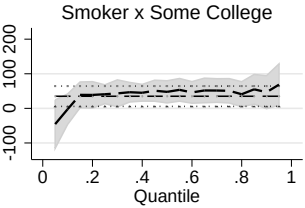

Smoker x Black
-200
-100
0
100
0
.2
.4
.6
.8
1
Quantile

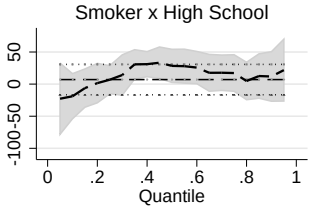

Smoker x Male
-50
0
50
0
.2
.4
.6
.8
1
Quantile


Smoker x Age
-60
-40
-20
0
20
40
0
.2
.4
.6
.8
1
Quantile


Smoker x Age Squared
-1.0
-.5
0
.5
1.0
0
.2
.4
.6
.8
1
Quantile


Smoker x College Graduate
0
100
200
300
0
.2
.4
.6
.8
1
Quantile


Smoker x Some College
-100
0
100
200
0
.2
.4
.6
.8
1
Quantile


Smoker x High School
-100
-50
0
50
0
.2
.4
.6
.8
1
Quantile