Honey, I shrunk the kids: Effects of Maternal Smoking on Birthweight with Panel Methods

Econometric Game 2012

April 19, 2012

Abstract

We assess the significance of the effect of maternal smoking on birthweight. Our main results suggest that there is a significantly negative effect of smoking on the mean birthweight. Furthermore, we find that both the decision to smoke or not - but also the decision on how much to smoke has significant effects. From extending the analysis to quantile panel methods, we see that there is a relatively constant penalty for being a smoker across quantiles, independently of the amount smoked. This does not hold in the extreme quantiles, where infants in the hazard zone are shown to be more severely affected by smoking than other parts of the distribution, regardless of the quantity of cigarettes. Lastly, using shrinkage FE estimation for robustness, we find that OLS with many predictors gives a numerical upper bound on the effect of smoking, while pure FE estimation gives a lower bound.

1 Introduction

Low birthweight increases child mortality and has long term effects through increased morbidity. These are obviously undesirable outcomes for the child, the families and society as a whole. For example, it increases costs for health, education and social services occur (see, e.g., Petrou et al., 2001). While the consequences of low birthweight have received a considerable amount of attention in both the medical and the economic literature it is perhaps more interesting from a policy perspective to consider the determinants of birthweights. After all, once a child is born, the child, its families and society bear the adverse health consequences of low birthweight.

Maternal smoking has long been identified as a key determinant of birthweight, see e.g. Kramer (2001) Abrevaya (2006), Abrevaya and Dahl (2008) and Bache et al. (2012), and all results, despite being obtained through different methods, point towards a negative effect of smoking on birthweight.

To elaborate on the findings in the literature, we investigate the relation between smoking and birthweight, trying unveil the ”true” causal effect of smoking on birthweight. This is a Herculean task, but we try to accomplish it by presenting results from both standard mean regressions in a panel setting and using quantile panel methods. Ideally, we would like to have a natural experiment to determine the causal effects of smoking on birthweight. Even with the ethical implication in mind, this data generating process is completely unrealistic. We believe the data that are available to us, allows us to get as close as possible to finding causality, if carefully analyzed given the above constraints.

Our main findings from standard mean panel regressions illustrate the problem of omitted variables. Our baseline model gives us an estimated impact of smoking of around -160g. Furthermore, we find that both the decision to smoke or not - but also the decision how much to smoke has significant effects. From a policy perspective these effects are interesting as they happen from behaviour policy makers can actually change.

We extend the analysis from analyzing the mean to analyzing the whole distribution of birthweight using quantile panel methods Our main finding are that there is a relatively constant penalty for being a smoker across quantiles, independently of the amount smoked. This does not hold in the extreme quantiles, where infants in the hazard zone are shown to be more severely affected by smoking than other parts of the distribution, regardless of the quantity of cigarettes.

All-in-all we find that smoking shrinks the birthweight, and this result is robust to analyzing the mean and the entire distribution of birthweights. Furthermore, our analysis using shrinkage FE estimation for robustness suggest that OLS with many predictors gives a numerical upper bound on the effect of smoking, while pure FE estimation give a lower bound. This also shows the gains from using a panel data structure compared with standard cross-section methods as we have the tools to deal with some of the heterogeneous effects.

The outline of the paper is as follows; in section 2 we describe the sample, furthermore we perform mean regressions using the panel dimensions of our data. Section 3 extends the analysis to quantile regressions. In section 4 we elaborate on the quantile model even further and finally section 5 concludes.

2 Data and Preliminary Analysis

2.1 Data Description and Transformations

The raw dataset contains 37080 observations. We have initially tabulated descriptive statistics for the explanatory variables to screen for outliers and determine compositional effects in the sample. Since our primary interest is in the effects of smoking, we have also tabulated descriptives conditional on smoking status. The tables with summary statistics in the appendix, shows that on average smokers are younger mothers, they have a shorter education, are less likely to be married and they have had a larger share of intermediate and inadequate cases of prenatal care. Thus, we see that there is important differences between smokers and non-smokers.

The average daily consumption of cigarettes has missing observations in about 179 cases. These mothers have been removed along with observations of mothers with a consumption above 60 cigarettes per day, which we deem a highly unlikely intake. The reason that we doubt the reported cigarette intake is that we suspect that the numbers are self-reported. This amounts to 483 observations or 161 mothers. This reduced the sample size to 36597 in total, which still represents the vast majority of the data. Regarding birthweight, the question comes up whether extreme low and extreme high weights should be treated as outliers or not, a quick web search showed that even though these extreme cases are rare they are indeed plausible, therefore we chose to leave all the observations in the sample.

Finally, we transform the data into a panel, where we use the unique mother identifier and the birth number as the time dimension. We have 12199 mothers and a strongly balanced panel, where each mother gives birth to 3 babies. Since our identification strategy will build on panel estimators, we also need to comment on the time variation that we have available in the data. This has important effects for identification. Obviously, the variation in the number of cigarettes and smoking status are central variables to analyze. In the appendix we decompose the variation of central explanatory variables into variation within and between, where variation within denotes variation due to changes in behaviour for a given mother and between variation denotes variation between different mothers. Since especially FE estimators rely on variation within, it is important to asses whether this variation is present and ”large” enough to create credible results. From the tables with summary statistics in the appendix, we see that there will on average be 2 years between a birth for a given mother.

The panel consists of 3829 observations of mothers smoking during pregnancy. Of these smoking mothers there are 1661 who are not smoking in all three waves, that is, they are changing their behaviour from either smoking to non-smoking or non-smoking to smoking. It is the variation in these 1661 mothers that will allow us to identify the effect of smoking during pregnancy.

Predetermined variables and also variables like marriage do not change for a given mother in our panel and therefore these will be dropped when we do FE estimation.

2.2 Panel Data Estimation Results

Let Y be the dependent variable, birthweight in our case, X is a vector explanatory variables. We consider the following linear model

$$y_{i,b} = \beta' x_{i,b} + c_i + \epsilon_{i,b}, \quad \forall i = 1, \ldots, n; b = 1, 2, 3 \tag{1}$$

where i indexes the mothers, b is the birth index and n is the number of observations. Importantly, c denotes the mother specific effects, which is of fundamental importance in the following analysis, as it might be correlated with smoking, preventing us from identifying the causal parameters.

This familiar problem of omitted variable bias exists due to compositional differences between the group of smokers and non-smokers. Therefore we ideally need to control for all variables that are different between the two group, and at the same time are considered an important determinant for birthweight. To mitigate the omitted variable bias, we exploit the panel structure of the data, in particular using random-effects and fixed-effects type models. We do not present the well-known standard random-effects model.

The *Correlated Random-Effects (CRE) Model* of Chamberlain (1982, 1984) models the unobservables c_i as a linear projection on X with an additional disturbance, denoted by v_i. The linear structure implies that the endogenous variable, smoking, in X impacts birthweight through two channels: 1) A direct effect modelled as $\beta' X$ and 2) through the unobservable effect c. Under the assumption $E(v_i|x_i) = 0$, the CRE model allows us to identify the direct effect of smoking, thereby removing the indirect effect through the unobservables. This is indeed the parameter of interest in the analysis.

The *Fixed-Effect (FE) Model* allows the unobserved heterogeneity c_i to be correlated with observables X in a arbitrary way. Under the assumption $E(\epsilon_{i,b}|x_i, c_i) = 0$ for all i and b, we are able to identify the partial effect of smoking on birthweight conditional on both the observables and the unobserved heterogeneity. For the standard FE estimation case, we can estimate β consistently using first differences, see Wooldridge (2010), but for quantile panel methods, things are not as straightforward, see Section 3 for details. There are both pros and cons from this method. One argument, clearly, is that allowing for fixed individual heterogeneity reduces the omitted variable bias even further, but since smoking behaviour is likely to be rather constant it could potentially mean that important variation is removed from the estimation process.

In order to reliably assess the significance of smoking for maternal decisions, we need to identify causal effects. That is the "pure" effect obtained from smoking status. However, the existence of unobserved variables that are correlated with smoking and, at the same time, impacts birthweight makes identification troublesome. More formally, let Y be the dependent variable, birthweight in our case, $X = (X_1', X_2')'$ is a vector explanatory variables, where X_1 contains smoking related variables (i.e, smoking indicator and/or average number of cigarettes smoked) and X_2 contains the remaining variables and a constant. We consider the following linear model

$$y_i = \beta_1' x_{1,i} + \beta_2' x_{2,i} + \epsilon_i, \quad \forall i = 1, \ldots, n \tag{2}$$

where n is the number of observations and ϵ is the error term satisfying $E(\epsilon) = 0$ and $E(\epsilon X_2) = 0$. However, X_1, the variables of interest, may potentially be correlated with ϵ, preventing us from identifying β_1. This familiar problem of omitted variable bias exists due to compositional differences between the group of smokers and non-smokers. Therefore we ideally need to control for all variables that are different between the two groups, and at the same time are considered an important determinant for birthweight.

The panel data we have available allows us to try to solve this problem by extending the model by accounting for unobserved heterogeneity. Therefore to the ommitted variable bias, we sequentially control for confounding effects extending the model from a naive regression of birthweight on smoking in a pooled OLS model, into a RE setting and lastly a FE setting. The results are presented in Table 1 below. In the table we see in column 1 that our initial raw estimate of the effect of smoking is -291g. This estimate does not change when we allow for random effects which is also what we could expect since pooled ols results are still consistent under the assumption of random effects. Interestingly the effect from smoking falls dramatically when we estimate the model in an FE setting (we remove the unobserved heterogeneity term by estimating the model in first differences and also in the usual dummy approach - the estimates do not differ from each other suggesting that either way is appropiate). This illustrates the importance of unobservables (ommitted variable bias). The effect is now -158g which is a lot smaller than the results we have obtained using a cross section data set. We stress again that it is important to keep in mind that the huge difference is likely also to be caused by the fact that we remove important variation from the explanatory variables. We will investigate wheter this is relevant below. Lastly we perform a Hausman test to test the endogeneity problems that our estimates between the RE and FE model suggest. The teststatistic is 179.30 and this clearly illustrates that the panel dimension (or many explanatory variables) is crucial in order for us to identify causal effects.

As a next step we control even further for omitted variable bias by including our baseline model from our earlier paper. Generally, determinants of birthweight can be divided into predetermined variables like genetic effects etc. (variables determined prior to pregnancy) and variables linked to the maternal behavior during the pregnancy. Of course in our dataset a strict distinction between such variables are likely not to be possible. For instance, the variable education can be thought to determine both the maternal behavior and it is likely also to be an effect of other "predetermined" variables. To keep the number of estimated parameters at a minimum exclude regional dummies and dummies for years, we have tested wheter these change any of our results and it does not.

Table 1: Fixed effects and Random effects models

	POLS	RE	FE	FE	FE
	(1)	(2)	(3)	(4)	(5)
	birwt	birwt	birwt	birwt	birwt
smoker	-291.0***	-291.0***	-163.9***	-158.4***	-166.0***
	(-29.86)	(-29.86)	(-12.05)	(-12.98)	(-12.45)
age				-0.326	3.480
				(-0.05)	(0.49)
agesq				0.313**	0.175
				(2.93)	(1.50)
male				147.5***	136.7***
				(31.27)	(26.56)
parity1				70.45***	68.83***
				(12.22)	(10.94)
parity2				63.01***	58.45***
				(9.69)	(8.23)
parity3				42.62***	35.22***
				(4.97)	(3.76)
nopnv				42.88	29.19
				(1.33)	(0.83)
pnv2t				28.21*	56.28***
				(2.50)	(4.58)
pnv3t				75.88**	119.5***
				(3.03)	(4.37)
pnc_inter				-49.96***	-73.97***
				(-5.18)	(-7.03)
pnc_inad				-92.99***	-128.4***
				(-4.51)	(-5.71)
gest				84.69***	
				(68.46)	
_cons	3520.9***	3520.9***	3507.6***	-188.6	3157.9***
	(925.46)	(925.46)	(1370.32)	(-1.73)	(29.62)
N	36597	36597	36597	36597	36597

Column 4 in Table 1 reports the estimated coefficients of the FE model including all explanatory variables that vary over time. We have also tried the same model and a model including variables that doesn't change over time, but again the hausman test clearly prefers the FE specification. The results in column 4 reflect the fact we reduce the omitted variable bias which in this case overestimates the negative effect of smoking. Finally we include a variable for gestation and report the estimated coefficients in column 5. Notice that the effect of smoking changes very little as we include gestation, supporting the robustness of our estimation procedure.

Our final estimated coefficient on smoking is -158.4g. This result is consistent with the findings from the FE results reported in Table IV in Abrevaya (2006). From our baseline model (column 4) we also see that: 1) The effects from the mothers age is non-linear. 2) Male babies are generally bigger. 3) The positive coefficient on parity (1st, 2nd and 3rd child from the parity variable) is likely to reflect feedback effects, see e.g. the argument in Abrevaya and Dahl (2008). The estimated effects suggest a decreasing relation very likely to be due to the fact that the mother is getting older. 4) The variables middling prenatal care shows that prenatal visits generally affects birthweight in a positive direction, but the quality and intensity has to be sufficient. 5) Lastly, and not surprisingly, gestation has a large positive effect. It is interesting to notice that the partial correlation between gestation and smoking is relatively weak since the estimated coefficient is hardly affected by the inclusion of the former variable.

The specification estimated above is an additive specification with respect to the effects from smoking. This is of course an approximation of the true conditional mean model, and in the next subsection we extend this model by considering additive effects and we also include the number of cigarettes smoked as an explanatory variable[1].

2.3 Extending the baseline model

Our baseline model analyzed above assumes that differences in birthweight can be explained by whether or not the mother smokes and the number of cigarettes is less relevant. In column 1 of Table 2 we extend the analysis by including the number of cigarettes smoked.

[1] We have tried several other combinations of other non-linearities between other explanatory variables but these have no effect on our findings

Table 2: Fixed effect - extension to the baseline model

	FE	FE	FE
	(1)	(2)	(3)
	birwt	birwt	birwt
smoker	-99.08***		
	(-5.64)		
cigs	-5.265***	.	-4.565***
	(-4.69)	.	(-4.04)
smokerblack		78.11	78.11
		(1.94)	(1.94)
smokerage		-1.683	-1.683
		(-0.74)	(-0.74)
smokerparity		-7.390	-7.390
		(-0.92)	(-0.92)
smokergest		-1.574	-1.574
		(-1.11)	(-1.11)
age	0.290	-15.47	-15.47
	(0.04)	(-1.85)	(-1.85)
agesq	0.304**	0.271*	0.271*
	(2.84)	(2.51)	(2.51)
male	147.7***	147.5***	147.5***
	(31.32)	(31.27)	(31.27)
parity1	70.47***	69.77***	69.77***
	(12.23)	(12.09)	(12.09)
parity2	63.32***	61.48***	61.48***
	(9.74)	(9.43)	(9.43)
parity3	43.03***	41.43***	41.43***
	(5.02)	(4.82)	(4.82)
nopnv	46.75	48.10	48.10
	(1.45)	(1.50)	(1.50)
pnv2t	28.55*	28.37*	28.37*
	(2.54)	(2.52)	(2.52)
pnv3t	75.19**	75.14**	75.14**
	(3.00)	(3.00)	(3.00)
pnc_inter	-50.08***	-50.18***	-50.18***
	(-5.20)	(-5.21)	(-5.21)
pnc_inad	-92.32***	-92.34***	-92.34***
	(-4.48)	(-4.48)	(-4.48)
gest	84.68***	84.96***	84.96***
	(68.49)	(67.50)	(67.50)
year		18.59***	18.59***
		(3.39)	(3.39)
_cons	-197.7	201.0	201.0
	(-1.81)	(1.18)	(1.18)
N	36597	36597	36597

t statistics in parentheses

Interestingly, the inclusion of the latter reduces the point estimate of smoking dramatically. As expected the number of cigarettes have a negative impact on birthweight, ie the more you smoke the less your baby weights at the time of birth. Quite surprisingly, the comprehensive empirical studies in e.g. Abrevaya (2006), Abrevaya and Dahl (2008) and Bache et al. (2012) have not considered this effect. This is potentially caused by the lack of or reluctance among the authors to use survey data. For obvious reasons, the number of cigarettes might be under-reported, which will bias our estimates. Arguably, the under-reporting leads to estimated effects that are ”too negative”, but given significance at a 1% level, we still think this result opens the scope for further research in this area. It is interesting from an economic perspective to compare the effects of smoking status against the amount smoked. Our estimates suggests that there is a large punishment from being a smoker alone (99g) while the punishment for an average smoker in the sample (13 cigarettes) is about three times smaller (approximately 65g).

The above analysis, along with its counterparts in the literature, relies on an assumption on homogeneous effects of smoking, i.e. that smoking affects birthweight linearly. This is of course an approximation and in columns 2-3, we relax this assumption by interacting smoking status with race, parity, gestation and age. We see that none of the interaction terms are statistically significant. Note that this reflacts that it is a very small subsample for example both black and change smoking behavior, therefore are the interaction terms very imprecisly estimated. For that reason it is not feasible to determine multiplicative effects. In the following we will thus focus on additive effects.

3 Quantile-Based Panel Estimation

3.1 Birthweight Distributions

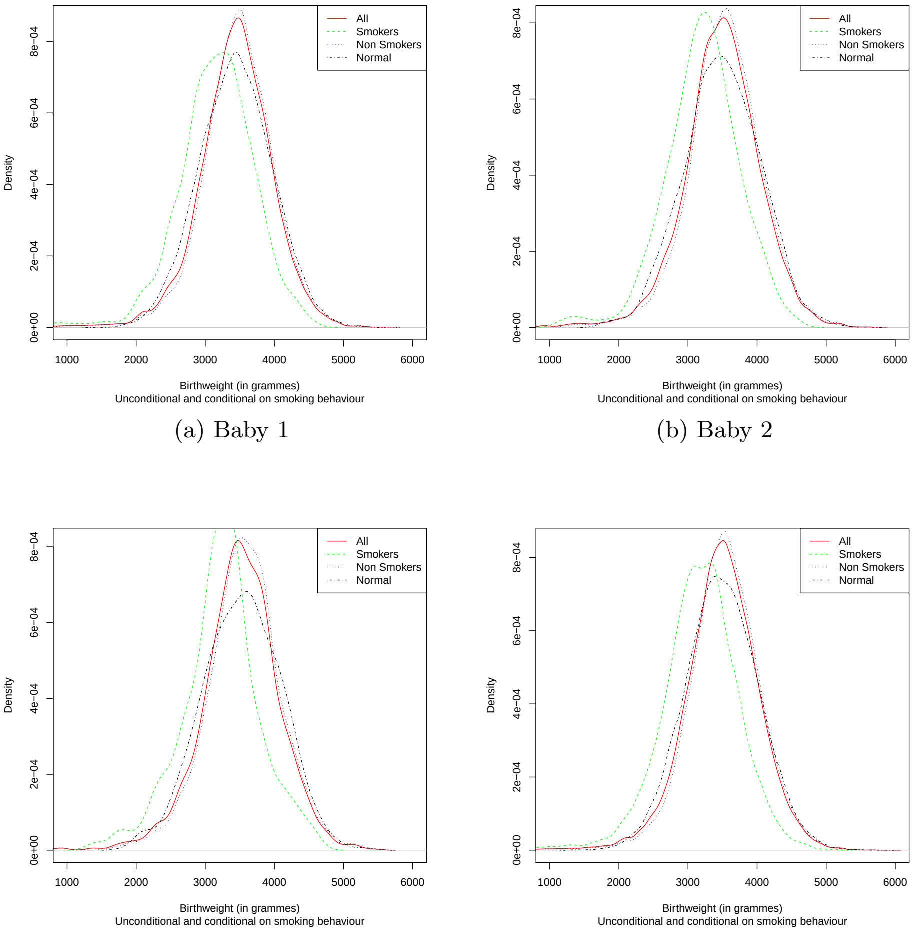
While Section 2 provides information on the average mother and birthweight, we need to consider the entire density for a more comprehensive understanding of the data. The average birthweight is not in the critical region below 2500g and, hence, the average is not of primary interest from a child health, maternal decision and policy perspective. To enhance our understanding of birthweights, we plot the unconditional density (using kernel smoothing techniques) along with the corresponding conditional densities for smokers and non-smokers, respectively, in Figure 1 below. The Figure is partitioned into four panels reflecting the first, second and third birth wave, respectively, and all waves. The partition is made to reflect the waves in the panel data and *not* between parity.

A few general comments are in place. The most striking feature, in Figure 1, is that both the mode and the mean of birthweight is lower for smokers than for non-smokers. Compared to the normal distribution, the left tail is more pronounced for the unconditional distribution of birthweights. This could be explained by the fact that is hard to prevent preterm births, but the birth can be forced (e.g., medication and C-section) and this is usually done relative shortly after the due date. Thus, high birthweight can and are prevented, while low birthweights cannot be prevents, such that we would expect such a skewed distribution. On the left tail, left of 2500g, the distribution is much fatter for smokers than for non-smokers. It is particularly illustrating to consider birthweights less than 1500g, where there is essentially zero probability mass for non-smokers and considerable mass for smokers

An important feature of Figure 1, motivating the further analysis, is that smoking does not seem to affect all regions on the distributions in the same way. In the left tail, the conditional distribution for smokers is fatter and more wiggly than for non-smokers, suggesting the imposition of a constant effect of maternal smoking on birthweights will be insufficient.

Lastly, when considering the differences across waves in the panel data, we see that the distribution of later (2 and 3) waves are more fat tailed compared with the first wave. This clearly relates to mother age, since by construction she must be older in later waves. This relates to the findings in Section 2, where age had a multiplicative negative effect on birthweight in connection with smoking and, and similarly for parity.

This visual inspection suggest that it is fruitful to analyze the whole distribution of birthweight rather than focusing on averages. The standard approach to allow for varying effects of smoking on birthweights across the distribution is quantile regression (Koenker and Bassett, 1978).



(a) Baby 1

(b) Baby 2

(c) Baby 3

(d) All

Figure 1: Density plot of birthweight

3.2 Quantile-Based Panel Estimation Methods

Let $\tau \in (0, 1)$ be the quantile of interest Define the conditional distribution function of Y given $X = x$ as $F_Y(y|x) = Pr[Y \leq y|X = x]$. Then, the conditional τ-quantile of Y given $X = x$ is $Q_Y(\tau|x) = \inf\{y : F_Y(y|x) > \tau\}$.

Instead of specifying a data generating process (DGP) for Y and derive its implied quantile function, we take a reduced form approach by directly specifying the quantile regression function as an approximation to the true quantile function. This is standard in the literature, because deriving the implied quantile functions are intractable for most specifications of the DGP (see Abrevaya and Dahl, 2008).The parametric quantile regression specifies the conditional τ-quantile as a parametric functions of the covariates X, and estimates the parameters. More formally, we consider panel extensions of the classical linear conditional function of Y given $X = x$:

$$Q_Y(\tau|x) = x'\beta(\tau).$$

The linear functional form is flexible in the sense that is accommodates a set of regressors $x \in \mathbf{X}$, the support of X, the includes power functions, splines, etc. of the original variables.

3.2.1 Quantile CRE

A limitation of the classical linear quantile regression model above is that it cannot accommodate unobservable effects, which impact the estimated effect of smoking, as discussed in Section 3. To extend the panel data analysis to accommodate unobservables, we follow Abrevaya and Dahl (2008) and Bache et al. (2012), who account for unobservables by augmenting the quantile function with constructed covariate(s), s_i. This means that we can write the reduced form quantile for Y as

$$Q_Y(\tau|x, s) = x'\beta(\tau) + s'\pi(\tau).$$

This implies that we may used the standard quantile regression of Y on X and S as

$$(\hat{\pi}(\tau), \hat{\beta}(\tau)) = \arg\min_{\pi,\beta} \sum_{i=1}^{n} \sum_{b=1}^{3} \rho_\tau(y_{i,b} - s_i'\pi - x_{i,b}'\beta),$$

where $\rho_\tau(u) = (\tau - (u < 0))$ (see Koenker and Bassett, 1978; Fernandez-Val and Chernozhukov, 2011; Koenker and Hallock, 2001). For identification of $\beta(\tau)$, s has to sufficiently capture the effects of unobservables on the quantile of interest, see Abrevaya and Dahl (2008); Bache et al. (2012) for a formal definition of sufficiency. It readily follows that identification hinges on whether or not sufficiency holds, which we will assess this critically in the application. Note that $\beta(\tau)$ has interpretation of marginal effects in a world without heterogeneity, i.e. a counterfactual effect.

3.2.2 Quantile FE

The validity of quantile CRE hinges on the sufficiency assumption on s. Hence, if one knows the functional form (linear projection) of c, it is straightforward to implement. However, its simplicity also makes it prone to misspecification. An alternative is quantile FE. Following Koenker (2004a) and Bache et al. (2012), we consider the following reduced-form quantile specification:

$$Q_Y(\tau|x, s) = x'\beta(\tau) + a_i,$$

where $a_i = c_i \delta(\tau)$ is the impact of unobservables on the τ-quantile. The standard approach to FE estimation in linear model, differencing the data, is not feasible in a quantile setting. Let $\tau_1, \dots, \tau_k$ be k distinct quantiles and define $w_1, \dots, w_k$ as the weights of these indices on estimation. The model parameters are estimated solving the following minimization problem.

$$(\hat{\beta}(\tau_1), \dots, \hat{\beta}(\tau_k), \hat{a}_1, \dots, \hat{a}_n) = \arg \min_{\beta_1, \dots, \beta_k, a_1, \dots, a_n} M(\tau, w, X, y, \lambda), \tag{3}$$

$$M(\tau, w, X, y, \lambda) = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{b=1}^{3} w_j \rho_{\tau_j}(y_{i,b} - x'_{i,b}\beta - a_i) + \lambda \sum_{i=1}^{n} |a_i|. \tag{4}$$

The special feature of this minimization problem is the regularization parameter λ. This method was introduced by Tibshirani (1996) and Koenker (2004b) for quantile regressions. To enhance the understanding of the estimator, consider two special cases. First if $\lambda \to 0$, the estimator collapses to a weighted dummy-variable regression (i.e., a_i acts as dummy for mother i). Secondly, if $\lambda \to \infty$, the above estimation problem becomes equivalent to a weighted cross-sectional regression. The appeal of the above estimator is that it allows for a continuum of cases in between these two extremes, and mitigates the problem of overparametrization encountered in panels with a short time dimension. The parameter λ penalizes absolute values of a_i, shrinking the resulting estimates towards 0. It essentially lets us control how much heterogeneity we want to allow for in the estimation. Thus, we can trade off bias from imposing homogeneity against imprecision from estimating many estimators. The interpretation of the FE results is subject to the same limitations discussed in Section 2.

3.3 Empirical Results

All panel quantile regressions are computed using the `quantreg` package for `R 2.15.0`. In this section we first consider different model specifications, then turn our focus to the tails of the birthweight distribution. [2]

Our main interest is measuring the causal impact of smoking during pregnancy on the infant's birthweight. The two variable describing the smoking behaviour of mothers are complementary. The first is a dummy variable reporting whether the mother smoked during pregnancy or not, while the other one reports the average number of cigarettes smoked per day. We included the following variables in all regressions: male, all maternal behaviour and age. As these coefficients are not of interest, we will not report them. For the variable of smoking behaviour, we show results for the inclusion of both the smoking dummy and the number of cigarettes, while results using only cigarettes or only the smoking dummy are reported in the appendix.

We look at two different sets of quantiles. For the standard specification, the use $\tau = (0.1, 0.2, \dots, 0.9)$, i.e. steps of 0.1. The second set of quantiles looks closer at the tails of the distribution by using the values

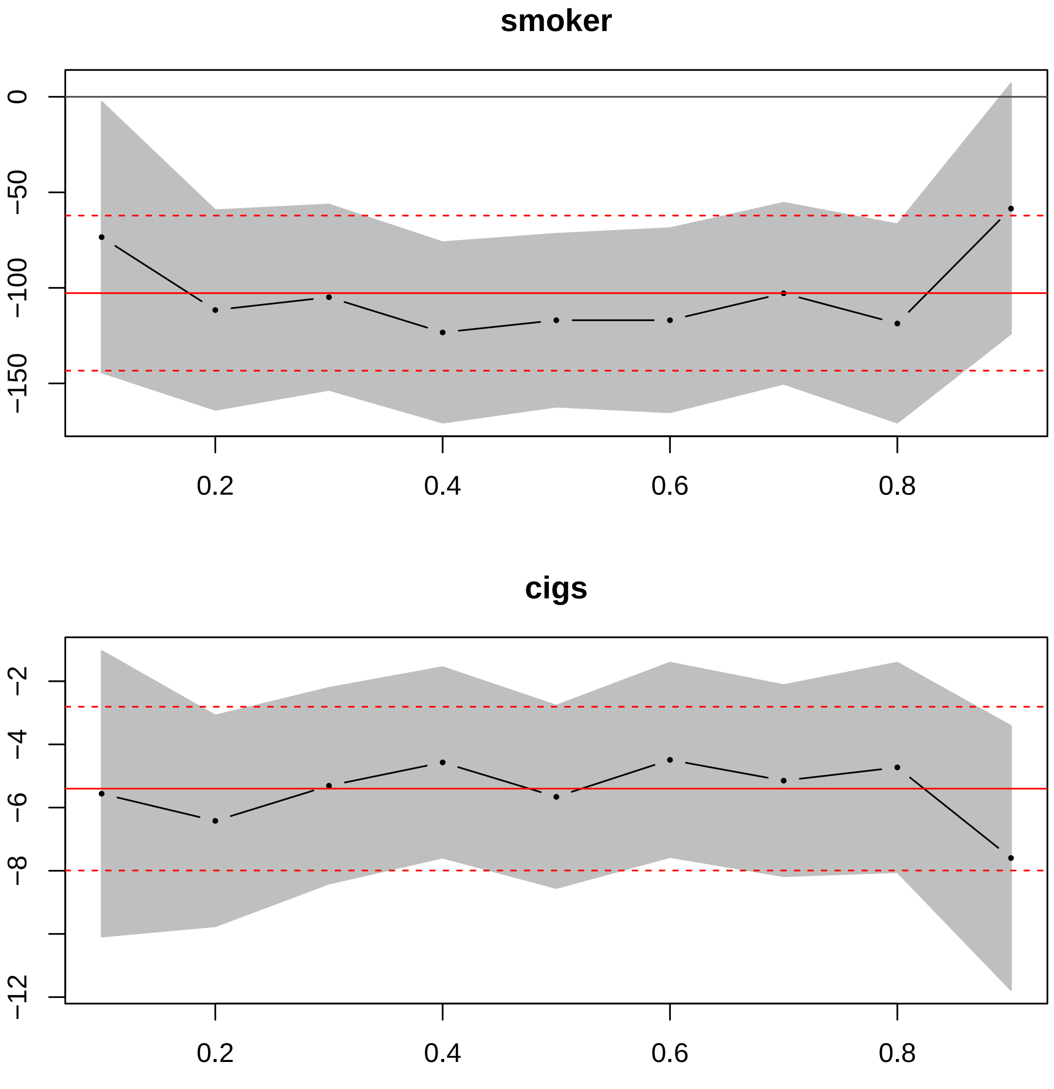
$$\tau = (0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99).$$

[2] If the explanatory variables included in regressions are not reported, then they are the same as in the baseline model presented above.

3.3.1 Results for Panel Quantile CRE

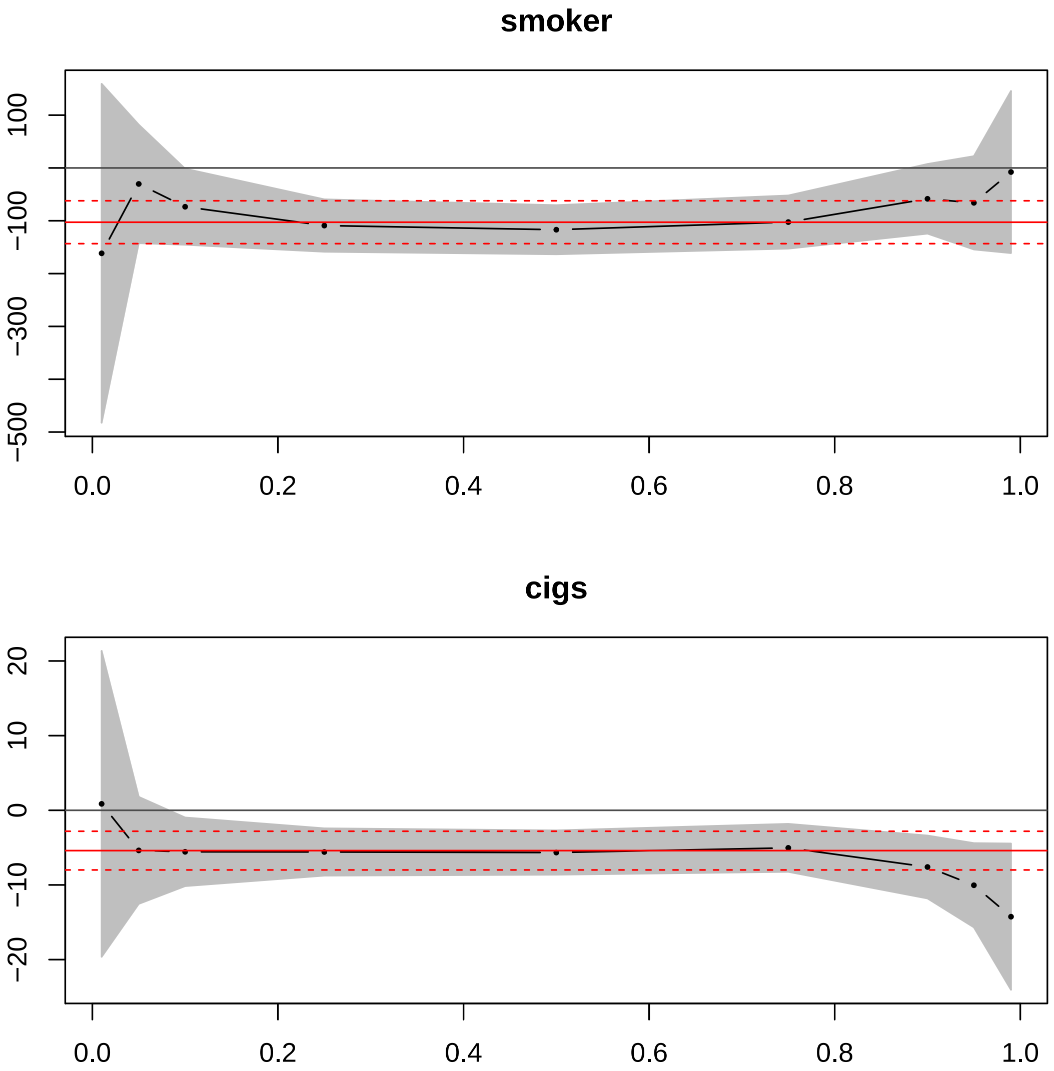
First, we show the estimates using panel quantile CRE, where we have computed s_i as in Abrevaya and Dahl (2008) and Bache et al. (2012), and the results are presented in Figure 2 for the standard quantiles.

Figure 2: Quantile Regression Estimates for Model with both smoking variables



In the model where both cigarette consumption and smoking status are included, there is a relatively constant penalty for being a smoker across quantiles, independently of the amount smoked. This suggests that smoking is ”in some sense” rank preserving on the distribution of birthweights, given that you smoke. The effect of additional cigarettes is however far to be negligible even though the amplitude of the effect is smaller than that found in Figure 9.Comparing these numbers to their counterparts in the appendix illustrates the need to include both the smoking dummy and the cigarette variable to capture the direct effect of each. The effects are very close to those estimated using FE on the mean. Despite the relatively constancy of both smoking as a dummy and cigarettes, there are indications that we see different effects in the tails, and this is elaborated upon in Figure 3

Figure 3: Tail Quantile Regression Estimates for Model with both smoking variables



From Figure 3, we observe some interesting results. We see that the tails estimates (in both tails) both for the smoking dummy and for cigarettes lies outside the FE confidence bands. For the smoking dummy, this implies that infants in the hazard zone (below the 0.05 quantile) are more severely affected than other parts of the distribution, regardless of the quantity of cigarettes. In relative birthweight, this effect is even larger as the children are, by definition, small. This has clear policy implications and suggest physicians should carefully identify risk groups and advice them to stop smoking. The results of this paper, along with those of e.g. Abrevaya and Dahl (2008), suggest that age, education marital status, drinking behaviour and prenatal care are important in identifying these risk groups, but a detailed study of risk groups are beyond the scope of this paper. A possible direction for further research in this area could be to design threshold probit models to determine risk factors.

3.3.2 Results for Quantile FE estimation

As Koenker (2004b) mentions, the choice of the regularization parameter λ is an open research question. Cross validation or information criterion are often used to select among a set of possible regularization parameters. Another method is to use a consistent estimator of the parameter to penalize raised to a negative power, this approach was coined adaptive LASSO by **?**. However the computing power required to estimate a large number of models and select among them is too big and available time too short to implement any of these approaches. We experimented with a handfull of values of λ and settled for $\lambda = 1$. The choice of regularization parameter didn't appear to be crucial. Also due to the computational burden of this method, the results were estimated only on a randomly selected subset of mothers corresponding to 10% of the available sample.

Fixed effect parameters where considered to be equal to zero if their absolute value was below 10^{-7}. Because of the penalty, the parameters value solving the penalized objective function are biased. The model was reestimated in a second step using only the fixed effects that passed the first step screening. We find that very few (or none in most cases) of the fixed effects were non-zero. These results would tend to indicate that unobserved heterogeneity is not a major issue in these data, keeping in mind the limitation of the method discussed above. These results in the quantile setting contradict the test results reported in the panel-OLS estimation part of the paper. This could be due to the very different nature of the estmation problem in both mean regressions and quantile regressions.

Figure 4: Quantile Regression Estimates for Model with both smoking variables

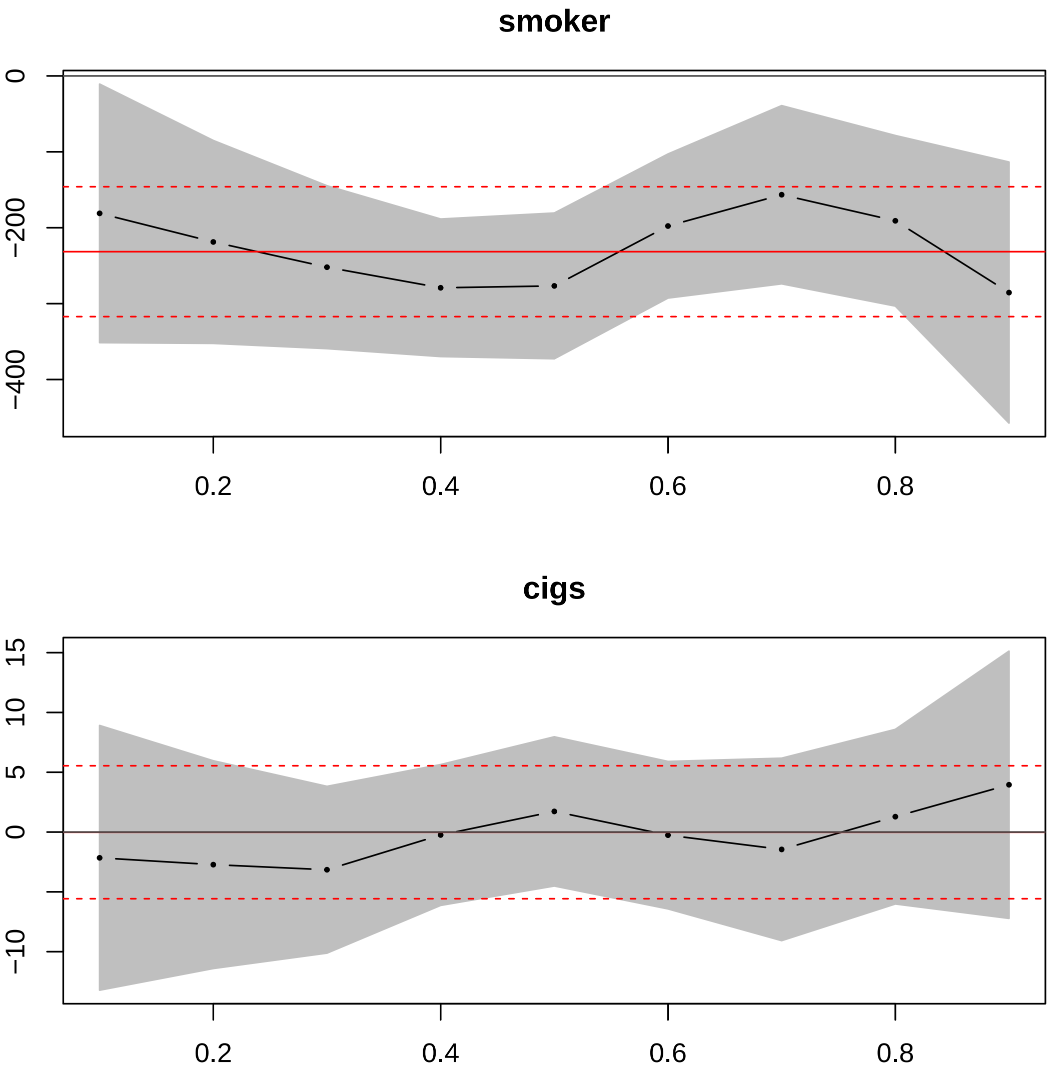


Figure 4 reports the results from this estimation. Notice that the effects on the quantiles of the distribution lies somewhere above the earlier reported CRE results, although the effects from smoking is not as large as cross sectional quantile regression results. The relative effect on quantiles is similar to earlier findings throughout the distribution. The impact of the quantity of cigarettes smoked is not significantly different from 0, this indicates that some caution is needed when extrapolating these results. Keeping this in mind we nevertheless use the results above as an indication on the fact that normal FE methods could exacerbate the importance on unobserved heterogeneity by removing relevant variation from the data.

4 Causality Discussion

In this section we discuss some of the assumptions made in the above analasis. First of all, the methods are based on a strict exogeneity assumotion. This assumption might be violated for three reasons. First it would be violated if there is some sort of feedback effect, such that smoking behavior during pregnancy can be influenced by previous birth outcomes. Abrevaya (2006) suggest an IV approach for controling for this problem. The basic idea is to use of, for example, the outecome of the first wave as an instrument for the change in the smoking behavior in betwwen the seccond and the third wave. This instrument should not be correlated with birth specific unobservables in the seccond and third wave but might be correlated with the smoking behaviour. Our sample size of 12199 mothers is too small for the IV approach to identify the smokings effect on birthweight according to Abrevaya (2006). Seccond, we would have a problem, if the smoking status is misreported in the status, especially if the misreporting are sytematically correlated with unobserved charcteristics. Third, it would be a problem if the change in smoking status is correlated with changes in the unobservables, as both our fixed effects approach and correlated random effects approach would only capture time invariant unobservables.

Another pitfall in both the fixed effects approach and correlated random effects approach, is that we difference out all time invariant variables, thus removing a lot of potentially important variation in the data.

So, whether the results are causal or not depends on whether the above mentioned assumptions holds. But given the robustness of our results and the similarity to the existing literature on the subject, we deem that this is the closest we can come to finding a causal relationship between maternal smoking and birthweight.

5 Conclusion

The purpose of this paper is to assess the significance of the effect of maternal smoking on birthweight. We find a statistical significant effects, which is robust across various panel estimation methods and specifications. We are not only interested in statistical significance, but also the significance for maternal decision making. Our main results suggest that there is a significantly negative effect of smoking on the mean birthweight. Furthermore, we find that both the decision to smoke or not - but also the decision how much to smoke has significant effects. From extending the analysis to quantile panel methods, we see that there is a relatively constant penalty for being a smoker across quantiles, independently of the amount smoked. This does not hold in the extreme quantiles, where infants in the hazard zone are shown to be more severely affected by smoking than other parts of the distribution, regardless of the quantity of cigarettes. Lastly, by using shrinkage FE estimation for robustness suggest, we find that OLS with many predictors gives a numerical upper bound on the effect of smoking, while pure FE estimation gives a lower bound. Pooled OLS works as a cross-section, and we see by utilizing the panel structure of the data, we are able to reduce the omitted variable bias considerably.

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. M. Dahl (2008): "The Effects of Birth Inputs on Birthweight," *Journal of Business & Economic Statistics*, 26, 379–397.

Bache, S. H., C. M. Dahl, and J. T. Kristensen (2012): "Headlights on tobacco road to low birthweight outcomes: Evidence from a battery of quantile regression estimators and a heterogeneous panel," *Empirical Economics*, forthcomming.

Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

——— (1984): "Panel data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, Elsevier, vol. 2, chap. 22, 1247–1318.

Fernandez-Val, I. and V. Chernozhukov (2011): "Inference for Extremal Conditional Quantile Models, with an Application to Market and Birthweight Risks," *Review of Economic Studies*, 78, 559–589.

Koenker, R. (2004a): "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91, 74–89.

——— (2004b): "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91, 74–89.

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

Koenker, R. W. and J. Bassett, Gilbert (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Kramer, M. S. (2001): "Determinants of low birth weight : methodological assessment and meta-analysis," *Bulletin of the World Health Organization (WHO)*, 65, 663–737.

Petrou, S., T. Sach, and L. Davidson (2001): "The long-term costs of preterm birth and low birth weight: results of a systematic review," *Child: Care, Health and Development*, 27, 97–115.

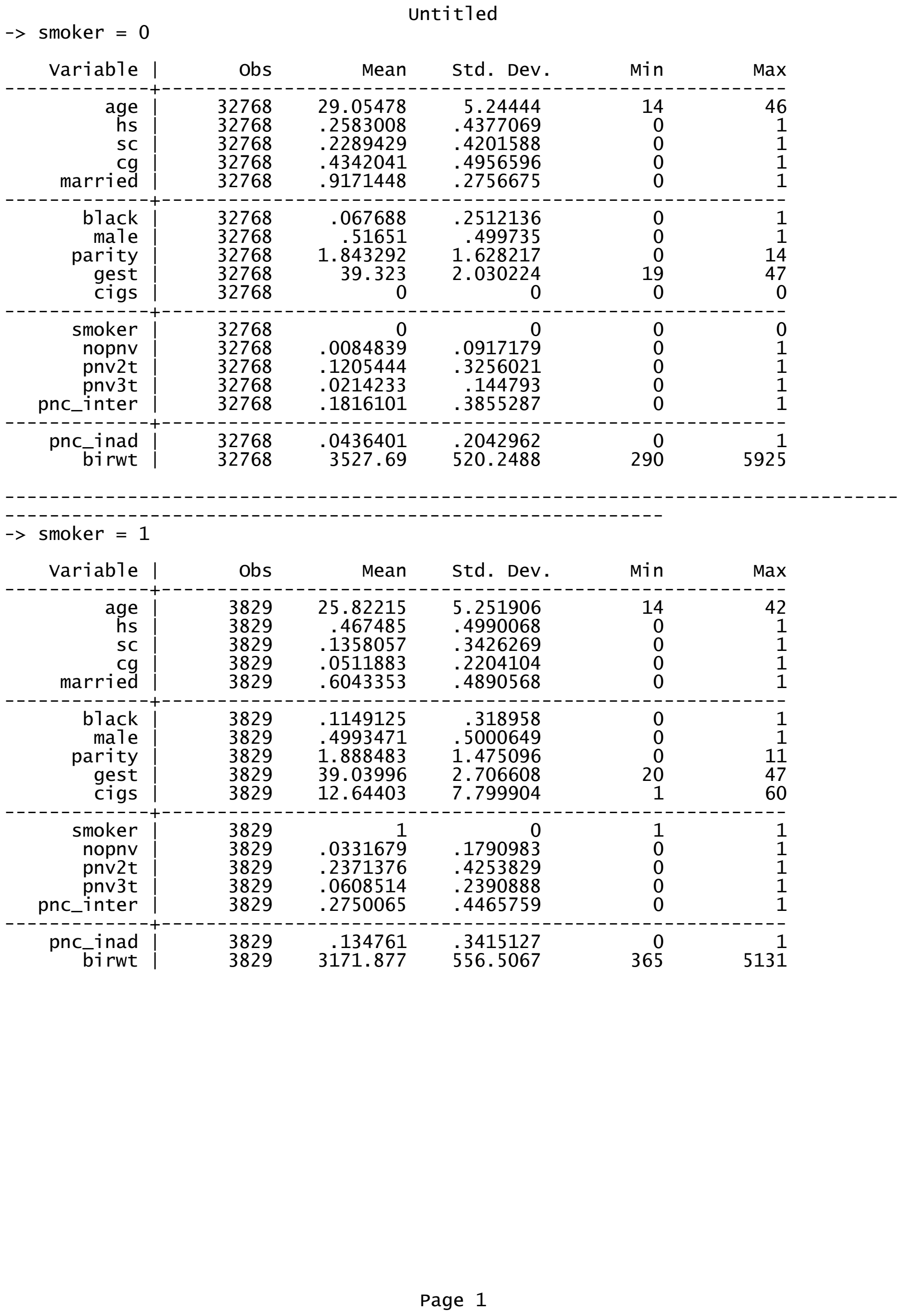
Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, vol. 1 of *MIT Press Books*, The MIT Press.

Appendix Supplementary Figures

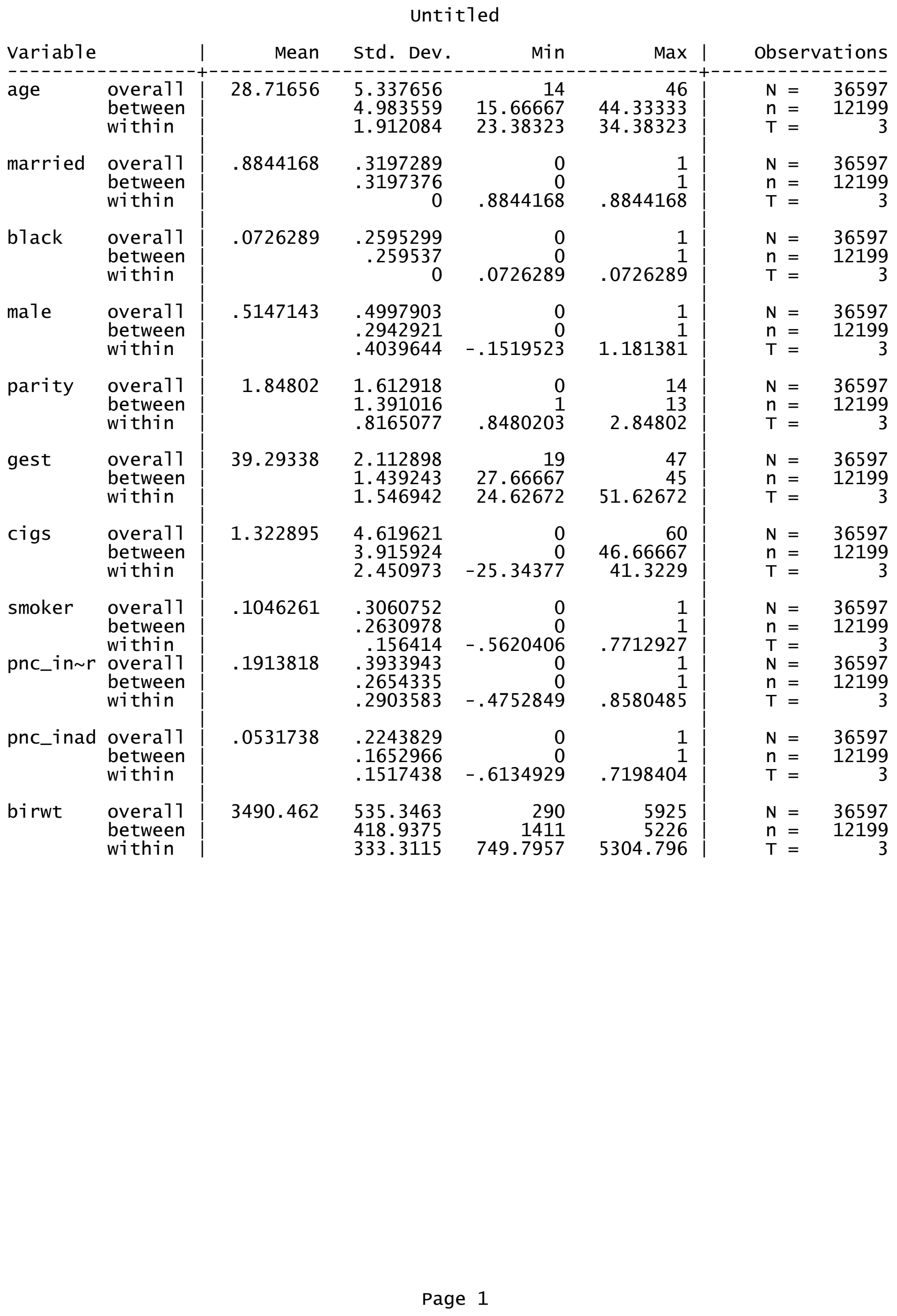
Descriptive Statistics for Section 2

Figure 5: Descriptive Statistics for Smokers and Non-Smokers



```
                                  Untitled
-> smoker = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     32768    29.05478     5.24444         14         46
          hs |     32768    .2583008    .4377069          0          1
          sc |     32768    .2289429    .4201588          0          1
          cg |     32768    .4342041    .4956596          0          1
     married |     32768    .9171448    .2756675          0          1
-------------+--------------------------------------------------------
       black |     32768     .067688    .2512136          0          1
        male |     32768      .51651     .499735          0          1
      parity |     32768    1.843292    1.628217          0         14
        gest |     32768      39.323    2.030224         19         47
        cigs |     32768           0           0          0          0
-------------+--------------------------------------------------------
      smoker |     32768           0           0          0          0
       nopnv |     32768    .0084839    .0917179          0          1
       pnv2t |     32768    .1205444    .3256021          0          1
       pnv3t |     32768    .0214233     .144793          0          1
   pnc_inter |     32768    .1816101    .3855287          0          1
-------------+--------------------------------------------------------
    pnc_inad |     32768    .0436401    .2042962          0          1
       birwt |     32768     3527.69    520.2488        290       5925

--------------------------------------------------------------------------------
-----------------------------------------------------------
-> smoker = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      3829    25.82215    5.251906         14         42
          hs |      3829     .467485    .4990068          0          1
          sc |      3829    .1358057    .3426269          0          1
          cg |      3829    .0511883    .2204104          0          1
     married |      3829    .6043353    .4890568          0          1
-------------+--------------------------------------------------------
       black |      3829    .1149125     .318958          0          1
        male |      3829    .4993471    .5000649          0          1
      parity |      3829    1.888483    1.475096          0         11
        gest |      3829    39.03996    2.706608         20         47
        cigs |      3829    12.64403    7.799904          1         60
-------------+--------------------------------------------------------
      smoker |      3829           1           0          1          1
       nopnv |      3829    .0331679    .1790983          0          1
       pnv2t |      3829    .2371376    .4253829          0          1
       pnv3t |      3829    .0608514    .2390888          0          1
   pnc_inter |      3829    .2750065    .4465759          0          1
-------------+--------------------------------------------------------
    pnc_inad |      3829     .134761    .3415127          0          1
       birwt |      3829    3171.877    556.5067        365       5131

                                  Page 1
```

Figure 6: Summary Statistics for all Explanatory Variables



```
                                 Untitled

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
age      overall |  28.71656   5.337656         14         46 |     N =   36597
         between |             4.983559   15.66667   44.33333 |     n =   12199
         within  |             1.912084   23.38323   34.38323 |     T =       3
                 |                                            |
married  overall |  .8844168   .3197289          0          1 |     N =   36597
         between |             .3197376          0          1 |     n =   12199
         within  |                    0   .8844168   .8844168 |     T =       3
                 |                                            |
black    overall |  .0726289   .2595299          0          1 |     N =   36597
         between |              .259537          0          1 |     n =   12199
         within  |                    0   .0726289   .0726289 |     T =       3
                 |                                            |
male     overall |  .5147143   .4997903          0          1 |     N =   36597
         between |             .2942921          0          1 |     n =   12199
         within  |             .4039644  -.1519523   1.181381 |     T =       3
                 |                                            |
parity   overall |   1.84802   1.612918          0         14 |     N =   36597
         between |             1.391016          1         13 |     n =   12199
         within  |             .8165077   .8480203    2.84802 |     T =       3
                 |                                            |
gest     overall |  39.29338   2.112898         19         47 |     N =   36597
         between |             1.439243   27.66667         45 |     n =   12199
         within  |             1.546942   24.62672   51.62672 |     T =       3
                 |                                            |
cigs     overall |  1.322895   4.619621          0         60 |     N =   36597
         between |             3.915924          0   46.66667 |     n =   12199
         within  |             2.450973  -25.34377    41.3229 |     T =       3
                 |                                            |
smoker   overall |  .1046261   .3060752          0          1 |     N =   36597
         between |             .2630978          0          1 |     n =   12199
         within  |              .156414  -.5620406   .7712927 |     T =       3
pnc_in~r overall |  .1913818   .3933943          0          1 |     N =   36597
         between |             .2654335          0          1 |     n =   12199
         within  |             .2903583  -.4752849   .8580485 |     T =       3
                 |                                            |
pnc_inad overall |  .0531738   .2243829          0          1 |     N =   36597
         between |             .1652966          0          1 |     n =   12199
         within  |             .1517438  -.6134929   .7198404 |     T =       3
                 |                                            |
birwt    overall |  3490.462   535.3463        290       5925 |     N =   36597
         between |             418.9375       1411       5226 |     n =   12199
         within  |             333.3115   749.7957   5304.796 |     T =       3



                                  Page 1
```

Panel Quantile Regression Results

Figure 7: Quantile Regression Estimates for Model with cigs variable

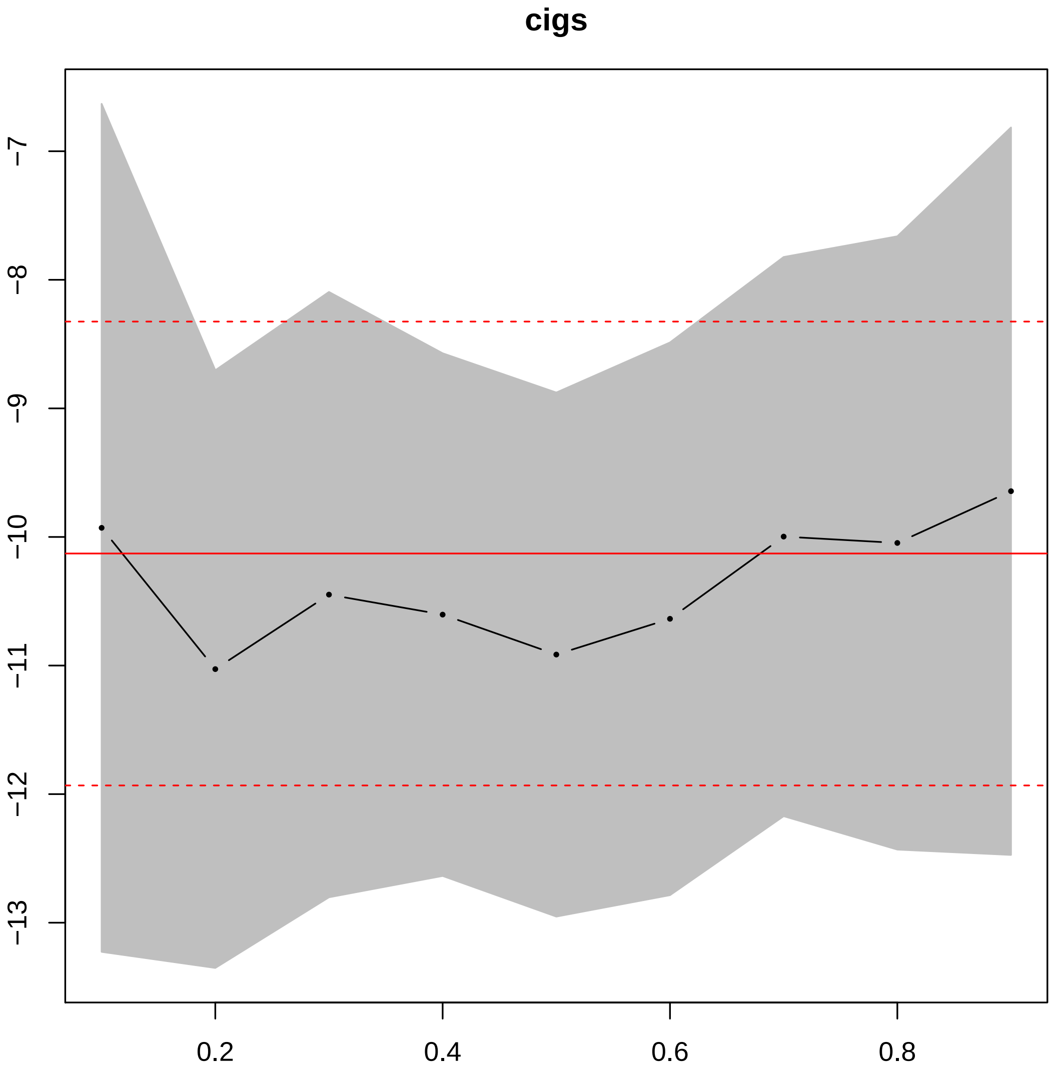


Figure 8: Tail Quantile Regression Estimates for Model with cigs variable

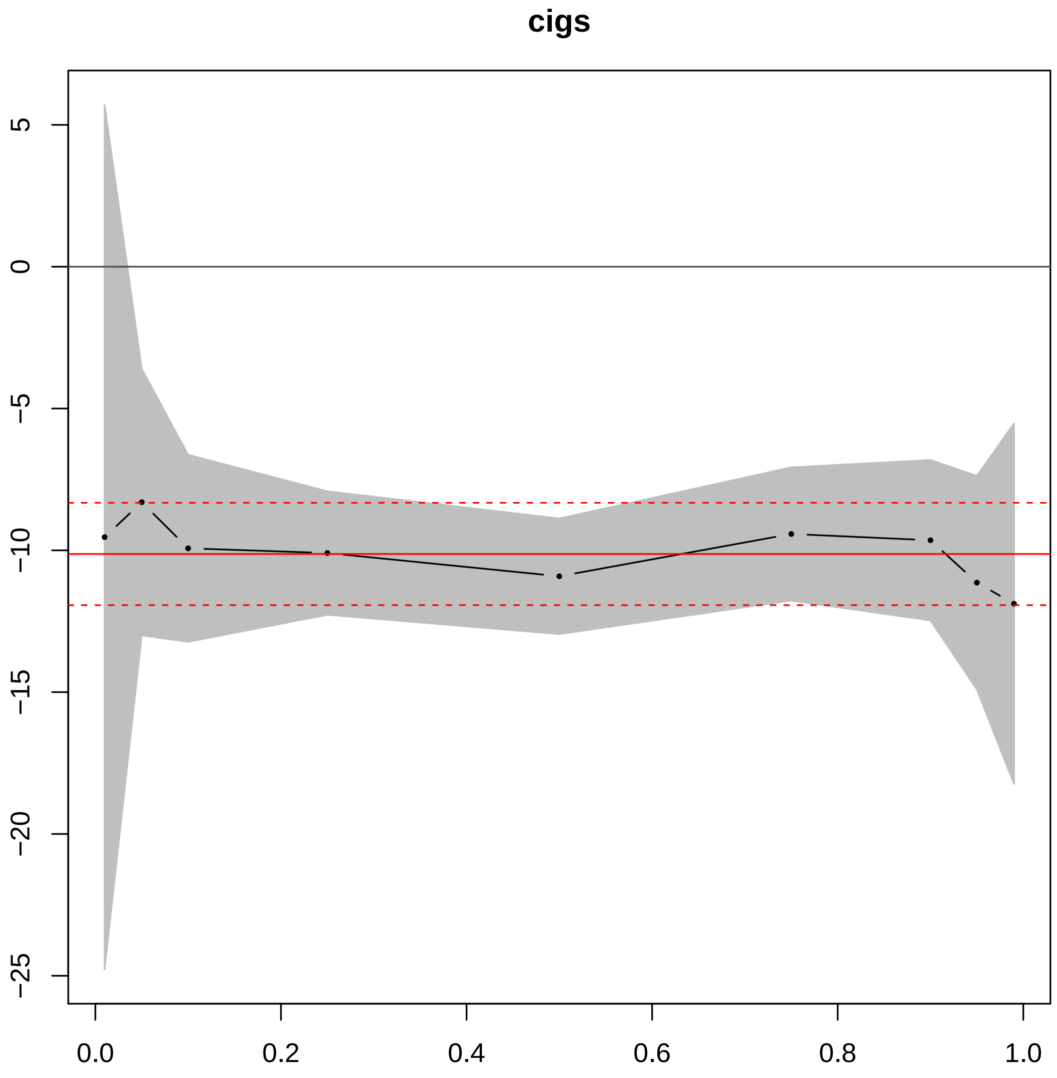


Figure 9: Quantile Regression Estimates for Model with smoking dummy

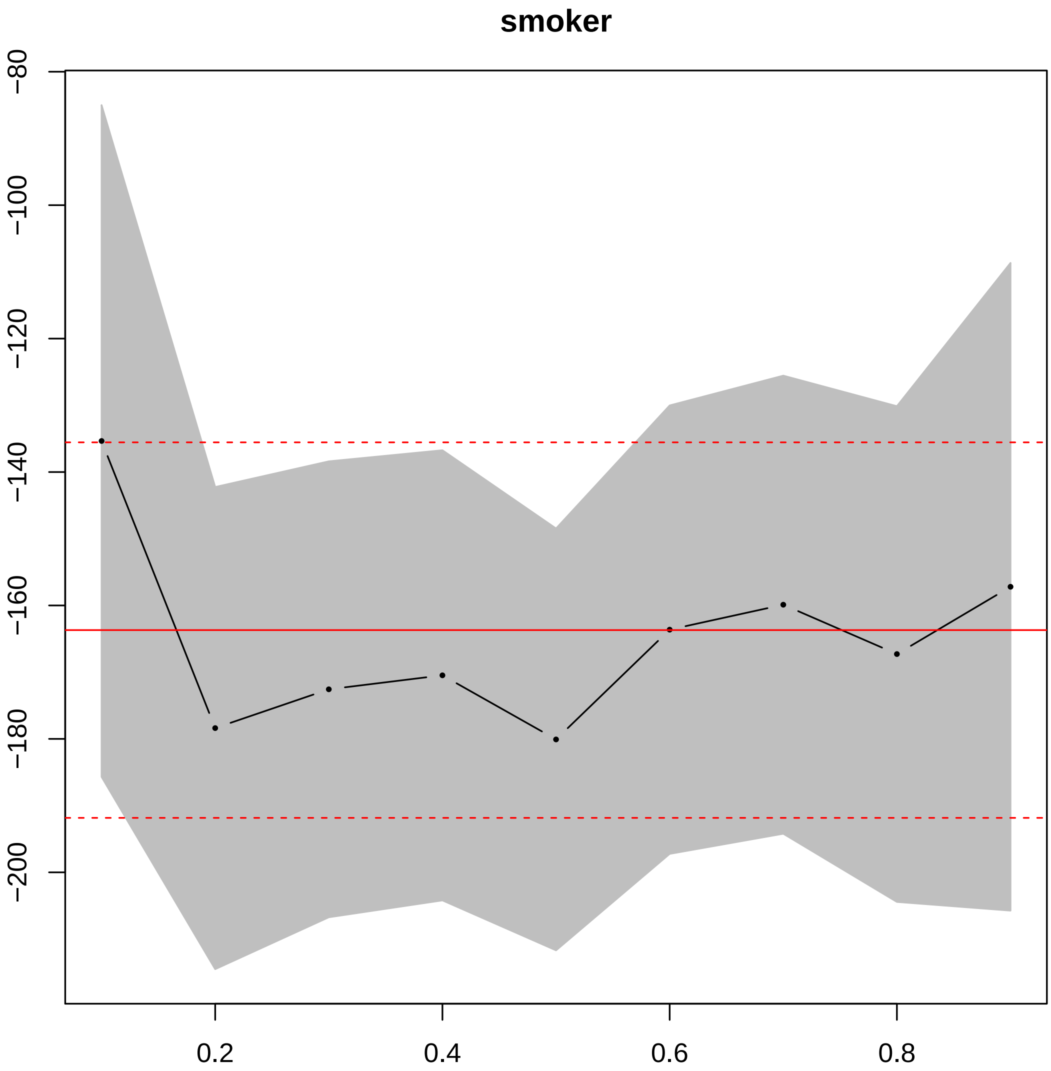
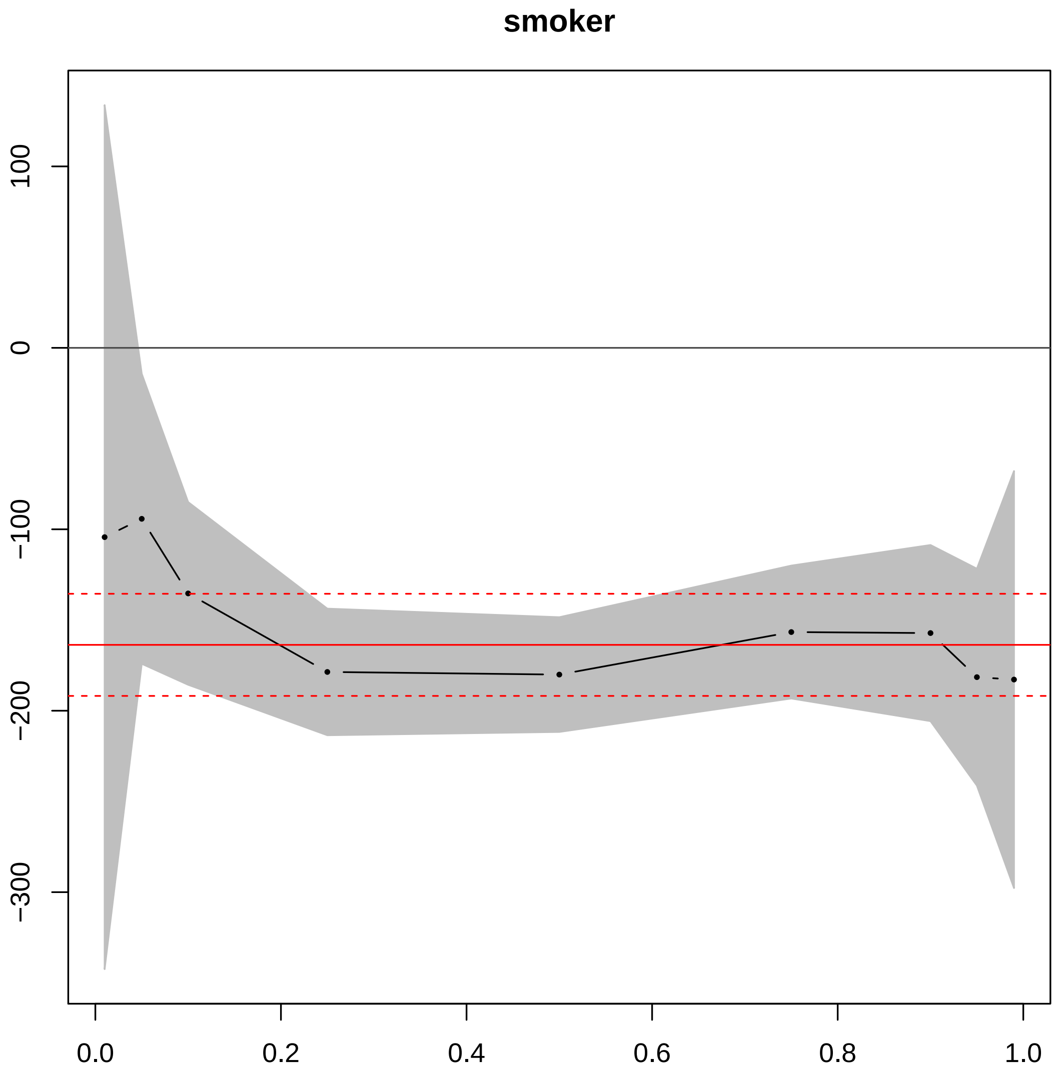


Figure 10: Tail Quantile Regression Estimates for Model with smoking dummy


smoker