ISER Working Paper Series

Quantile regression with aggregated data

Cheti Nicoletti
Institute for Social and Economic Research
University of Essex

Nicky G. Best
School of Public Health
Imperial College London

No. 2011-12
May 2011

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non technical summary

Administrative data can contain a wealth of information for empirical research. Just to cite two examples, administrative data on schools can be used to study pupils' educational attainments while hospital data can be useful for health research. However, access to administrative information is often restricted to aggregated data and this can lead to biased results. The estimation bias caused by using aggregated rather than individual data is known as the ecological bias.

In this paper we consider for the first time this issue in the context of quantile regressions. We show how data can be aggregated to obtain unbiased estimation of quantile regressions with categorical covariates and how the bias can be reduced when researchers are interested to estimate quantile regression where some of the covariates are continuous.

Quantile regression with aggregated data

Cheti Nicoletti

ISER, University of Essex

Nicky Best

Imperial College London

2011

Abstract

Analyses using aggregated data may bias inference. In this work we show how to avoid or at least reduce this bias when estimating quantile regressions using aggregated information. This is possible by considering the unconditional quantile regression recently introduced by Firpo et al (2009) and using a specific strategy to aggregate the data.

Key words: quantile regression, ecological inference, aggregation bias.

JEL: C18, C21.

Corresponding Author: Cheti Nicoletti, ISER, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, e-mail: nicolet@essex.ac.uk

Acknowledgements: We would like to thank participants at the Workshop on Measurement Errors in Administrative Data, Mannheim, June 2010 where a previous version of this paper was presented. This work was supported by the Economic and Social Research Council through MiSoC and BIAS which is a node of the Economic and Social Research Council's National Centre for Research Methods.

1. Introduction

One of the main advantages of administrative and census data is that they usually cover the whole population of interest and have substantially larger sizes than sample surveys, therefore providing more precise estimation. The use of administrative and census data in applied research has increased in recent years, but the access to individual information is still frequently limited because of confidentiality reasons. The question is then: how can we make correct inference on individual behaviour when data are available only at aggregated level? This is the fundamental question posed by the literature on ecological inference. Most of the research has focused on methods providing point identification of the parameters (or distribution) characterizing individual behaviour, but this comes at the cost of imposing untestable assumptions (see for example King 1997 and King et al 2004). On the contrary, some research has focused on partial identification, i.e. on the identification of bounds on the parameters of interests relaxing any untestable assumption (see Duncan and Davis 1953 and Cho and Manski 2008). This paper adds to the literature on ecological inference by looking for the first time at the aggregation problem for quantile regressions. Nevertheless, we do not suggest new methods to point or partially identify the parameters of interest, but rather a strategy to aggregate data to minimize the potential ecological bias.

Let us consider an administrative dataset with information on individual categorical variables and assume we are interested in the regression of Y on a set of variables X, where both Y and X are categorical variables. Then individual data can be aggregated without any loss of information by simply considering the frequency of individuals for each of the possible combinations of values taken by the categorical variables Y and X. This way to aggregate administrative data allows preserving the whole information provided by individual data and avoiding any confidentiality issue, as long as the number of all possible combinations of values taken by Y and X is small. We propose an extension of this type of aggregation to the case where Y is a continuous variable and we are interested in the quantile regression of Y on X. To make possible this extension we utilize the unconditional quantile regression recently proposed by Firpo et al (2009). Additionally, we consider the case where the explanatory variables are a mix of categorical variables X and continuous variables Z. In this case aggregation always implies a loss of information. We suggest some methods to aggregate the

continuous variables *Z* and a test to verify which of these methods minimize the potential aggregation bias for the X-coefficients.

The paper structure is organized as follows. Section 2 defines the unconditional quantile regression and the unconditional partial quantile effect. Section 3 shows how to aggregate data to produce unbiased estimation of unconditional quantile regressions when using categorical covariates; while section 4 shows how to reduce the aggregation bias when the covariates also include continuous variables. In section 5 we suggest a test to verify whether the aggregation bias cancels out. Finally, we draw some conclusion in section 6.

2. Unconditional quantile regression

Researchers are often interested in evaluating the effect of a variable *T*, e.g. an intervention or an individual decision, on a continuous outcome variable *Y*. Examples of evaluation studies include the effect of smoking during pregnancy on birth weight and of school programs on exam scores. Most of the empirical research focuses on the average effect of *T*, i.e. on the effect at the mean. But, since the effect of a variable can vary across the *Y*-distribution and very low (or high) levels of outcomes can be associated with especially negative consequences, it is important to study the effect also at lower (higher) quantiles. For this reason recent research has begun to estimate quantile effects rather than only mean effects. For example, Bitler et al (2006) have analysed the effect of welfare reforms on income and earnings allowing for a heterogeneous effect across the Y-distribution, while Abrevaya and Dahl (2008) have evaluated the effect of birth inputs on birth weight at different quantiles. More generally, the evaluation of quantile effects is important every time there is a concern that low (or high) levels of outcomes may have negative consequences, as in the case of low birth weight, poor educational attainments and low income. In all these cases, the evaluation of the effects at low quantiles helps in understanding what can cause a change in *Y* for people who are at the low (upper) end of the Y-distribution, i.e. for people who are more at risk of negative consequences.

Since the effect of a variable *T* on *Y* could be due to confounding, empirical researchers usually estimate quantile effects of *T* on *Y* by controlling for potential cofounding variables, *W*, using conditional quantile regressions (see Koenker and Bassett, 1978 and Koenker and

Hallock, 2001). In a conditional quantile regression the τ-quantile of the conditional distribution of Y given $X=(W,T)$, y_τ, is usually expressed as a linear function of these variables and a set of parameters θ and α,

$$y_\tau = X\,\theta + U,$$

where U is an error term independent of X and with τ-quantile equal to zero.

This conditional quantile regression allows estimating the effect of a variable T on the conditional quantile, but it does not allow inferring its effect on the unconditional quantile,[1] i.e. the effect of a change in T on the marginal quantile of Y if all other variables, W, were kept unchanged. Firpo et al (2009) propose a method to estimate the unconditional quantile effect and it is based on what they call unconditional quantile regression.

This method consists of the regression of the recentered influence function (*RIF*) for the unconditional τ-quantile, q_τ, on the explanatory variables X. The *RIF* for the τ-quantile is given by $RIF(Y,q_\tau)= q_\tau+[\tau-d_\tau]/f_Y(q_\tau)$, where $f_Y(q_\tau)$ is the density distribution function of Y computed at the quantile q_τ, and d_τ is a dummy variable taking value one if $Y \leq q_\tau$ and zero otherwise. The $RIF(Y,q_\tau)$ satisfies the following properties: [2]

- its mean is equal to the actual τ-quantile, $E_y[RIF(Y,q_\tau)]= q_\tau$;
- the mean of its conditional expectation, $E_y[RIF(Y,q_\tau)|X]$, is again equal to the actual statistic q_τ, i.e. $E_x\{E_y[RIF(Y,q_\tau)|X]\}=q_\tau$.

The conditional expectation $E_y[RIF(y,q_\tau)|X]$ is a function of X and it is what Firpo et al (2009) define as the unconditional quantile regression.

Assuming a linear relationship between $RIF(Y,q_\tau)$ and X, we have a linear regression model

$$RIF(Y,q_\tau)=X\,\beta+u, \qquad (1)$$

where u is an error term, which we assume to be identically and independently distributed with mean zero and variance σ_u^2 and independent of X, and β is a vector of coefficients which

[1] The unconditional τ-quantile is the quantile of the marginal distribution of Y.
[2] For a more detailed definition of the recentered influence function and a full list of properties we refer to Firpo et al (2009).

can be estimated by ordinary least squares (RIF-OLS regression)[3]. β is equal to the unconditional quantile partial effect of the variables X, i.e. $E[dE[RIF(Y,q_\tau)|X]/dx]$.

3. Unconditional quantile regression with categorical covariates

Assuming that we can observe Y_i and X_i, for each individual i ($i=1,\ldots,N$) in the population (using register or census data) and that X_i is a vector of categorical variables $X_{k,i}$ with $k=1,\ldots,K$, we can use these N individual observations to estimate the unconditional quantile regression,

$$RIF(Y_i,q_\tau)=X_i\,\beta+u_i, \tag{2}$$

where $RIF(Y_i,q_\tau)= q_\tau+[\tau-d_{\tau,i}]/f_Y(q_\tau)$ and $d_{\tau i}$ is a dummy variable taking value one if $Y_i \leq q_\tau$ and zero otherwise.

On the contrary, let us assume that we are unable to observe $RIF(Y_i,q_\tau)$ and X_i at individual level, but we observe their average values over individuals belonging to each of S groups,[4] which are mutually exclusive and collectively exhaustive ($s=1,\ldots,S$), i.e. we observe

$$\overline{RIF}_s=1/N_s \sum_{i=1,\ldots N} RIF(Y_i,q_\tau)d_{s,i}= q_\tau+[\tau-\bar{d}_{\tau s}]/f_Y(q_\tau)$$

$$\bar{X}_s=1/N_s \sum_{i=1,\ldots N} (X_i d_{s,i}),$$

where $d_{s,i}$ is equal to one if individual i belongs to the group s and zero otherwise, N_s is the number of individual in group s and $\sum_{s=1}^{N} N_s=N$, $\bar{d}_{\tau s}=[1/N_s \sum_{i=1,\ldots Ns} d_{s,i}\, d_{\tau i}]$ is the proportion of individuals with values of Y_i equal or below q_τ in group s. With these aggregated data we can estimate the following regression, which is usually called an ecological regression,

$$\overline{RIF}_s=\bar{X}_s\tilde{\beta}+\varepsilon_s,\ s=1,\ldots,S; \tag{3}$$

But the estimated $\tilde{\beta}$ is generally a biased estimation of the parameter of interest β in equation (2). This bias is known as aggregation or ecological bias.

---

[3] Firpo et al (2009) show also two alternative methods to estimate the relationship between the RIF and the covariates: the RIF-logit regression and nonparametric-RIF regression.

[4] For example, data can be aggregated by geographical areas.

We can avoid this ecological bias if we divide individuals into groups based on their observed X-values, i.e. if we consider separate groups for each of the possible combinations of the X-values. Let the number of possible combinations (groups) be S, and assume that we observe q_τ and $f_Y(q_\tau)$,[5] and $\bar{d}_{\tau s}$, $\bar{X}_s$ and N_s for each $s=1,...,S$. Then, for any group s we know

- the exact values of X, which we denote $\bar{X}_s = [X_{1,s}, X_{2,s}, \ldots, X_{K,s}]$,
- the number of individuals who have values of Y equal or below q_τ, which is $n_s = N_s \bar{d}_{\tau s}$,
- the number of individuals who have values of Y above q_τ, i.e. $(N_s - n_s)$.

In other words, we know that in group s there are

- n_s individuals with $\text{RIF}(Y_i, q_\tau) = q_\tau + [\tau - 1]/f_Y(q_\tau)$ and $X_i = [X_{1,s}, X_{2,s}, \ldots, X_{K,s}]$ and
- $(N_s - n_s)$ individuals with $RIF(Y_i, q_\tau) = q_\tau + \tau/f_Y(q_\tau)$ and $X_i = [X_{1,s}, X_{2,s}, \ldots, X_{K,s}]$.

By pooling together the information on individuals from each of the S groups, we can reproduce the complete dataset with observations on $RIF(Y_i, q_\tau)$ and X_i for all N individuals, and we can use this dataset to estimate the unconditional quantile regression without any ecological bias.

To summarize, it is possible estimate the unconditional quantile regression using aggregated data without any loss of information or ecological bias if we can observe:

- q_τ, the τ-quantile for the whole population;
- $f_Y(q_\tau)$, the density of Y at the τ-quantile again computed using the whole population;
- the percentage of individuals with a value of Y below the τ-quantile for each of the possible combinations of values of the set of explanatory variables X;
- the absolute frequency of individuals for each of the possible combinations of values of X.

When the numbers of variables X and combinations of their possible values are small, then this aggregation method helps in avoiding both confidentiality issues and ecological bias. Nevertheless, there can be situations where the variables X are large in number or contain continuous and categorical variables which can take many or even infinite different values. When the number of possible combinations of the variables X is too large to preserve confidentiality, then we need to discretise the continuous variables and to group the

[5] q_τ and $f_Y(q_\tau)$ are constant across individuals and can be estimated using the sample quantile and the non-parametric (kernel) estimation of the density distribution of Y computed at the sample quantile.

categorical variables in fewer categories. In this situation aggregating the data implies a trade-off between between estimation bias and confidentiality.

4. Unconditional quantile regression with categorical and continuous covariates

Let us consider an administrative dataset with individual information on a continuous variable *Y*, a set of categorical variables *X* and an additional set continuous variables *Z*, and let us be interested in estimating the coefficients β_0 in the following individual regression

$$RIF(Y_i,q_\tau)=X_i\,\beta_0+Z_i\,\gamma_0+u_{0i},\ i=1,\ldots,N, \tag{4}$$

Assuming that for each individual i we can observe Y_i, X_i and Z_i, we can estimate β_0 by simply regressing $RIF(Y_i,q_\tau)= q_\tau+[\tau-d_{\tau,i}]/f_Y(q_\tau)$ on the covariates X_i and Z_i. On the contrary, if we can access only aggregated data, then the estimation of β_0 will be potentially biased. The question is then how to aggregate data to minimize this bias.

An aggregation method often used to release administrative data is the averaging of each variable by geographic areas, i.e. the computation of $\overline{RIF}_j$, $\bar{X}_j$ and $\bar{Z}_j$ for each area j, where $j=1,\ldots,J$. These aggregated data can be used to estimate the following ecological regression,

$$\overline{RIF}j\ = \bar{X}_j\,\beta_1 + \bar{Z}_\mathrm{j}\,\gamma_1 + v_{1j},\quad j = 1,\ldots,J; \tag{5}$$

but the estimated β_1 and γ_1 are generally a biased estimation of the parameters β_0 and γ_0 in (4).

An alternative aggregation method consists of the following steps:

- dividing individuals into groups by considering the set of *S* possible combinations of values for *X* for each of the *J* possible geographic areas;
- for each of these (*S* x *J*) groups computing the percentage of individuals with a value of Y_i below the τ-quantile, the absolute frequency of individuals, the actual values assumed by X_i and the average value assumed by Z_i for individuals belonging to the corresponding area *j*, $\bar{Z}_j$.

These observations together with the knowledge of q_τ and $f_Y(q_\tau)$ allows us the reconstruct the data necessary to compute an unbiased estimation of the following semi-individual regression[6]

$$RIF(Y_i,q_\tau)=X_i\,\beta_2+[\sum_{j=1}^{J}\bar{Z}_j\,d_{ij}]\,\gamma_2+v_{2i},\ i=1,\ldots,N, \tag{6}$$

where j indexes the geographic areas, d_{ij} is a dummy variable taking value one if individual i lives in area j and zero otherwise, and $\bar{Z}_j$ is the average of the characteristics Z_i observed in areas j where the individual i lives.

A further possible aggregation method consists in discretising or grouping the continuous variables Z_i. For example, let us consider a variable measuring individual income, then we can discretise it into a categorical variable indicating whether the individual income is above or below the 25th percentile, between the 25th and 75th percentiles, or above the 75th percentile (where the percentiles refer to the whole population). Given D possible values of the discredited variable, we can consider D corresponding dummy variables which we denote with Z_{di}, $d=1,\ldots,D$. Observations on Y_i, X_i and Z_{di} do not allow to estimate regression (4), but they allow to estimate without bias the following regression

$$RIF(Y_i,q_\tau)=X_i\,\beta_3+\sum_{d=1}^{D}(Z_{di}\,\gamma_{3d})+v_{3i},\ i=1,\ldots,N. \tag{7}$$

Furthermore, we can estimate without bias equation (7) using aggregated data or more specifically information on:

- q_τ, the τ-quantile for the whole population in the administrative data;
- $f_Y(q_\tau)$, the density of Y at the τ-quantile again computed using the whole administrative data;
- the percentage of individuals with a value of Y below the τ-quantile for each of the possible combinations of values of the set of explanatory variables X and Z_d;
- the absolute frequency of individuals for each of the possible combinations of values of X and and Z_d.

The last two aggregation methods allows for better estimation because they allow estimating models (6) and (7), where only Z is measured with aggregation error. On the contrary, in the ecological model (5) the variables Y, X and Z are all observed with aggregation error (see

[6] A semi-individual regression is a model with all variables observed at individual level except for some of the covariates (see Kunzli and Tager, 1997).

Kunzli and Tager 1997). The fact that we use $\bar{Z}_j$ (Z_1…,Z_D) rather than Z can bias the estimation of the coefficient Z, γ_0, as well as of the coefficients of the remaining explanatory variables X, β_0. In the following section we prove that this last bias cancels when X and Z are uncorrelated conditional on $\bar{Z}_j$ (the set of dummy variables [Z_1…,Z_D]).

5. Testing aggregation bias

As in the last section, assume we are interested in estimating the coefficient β_0 in the regression

$$RIF=X\beta_0+Z\gamma_0+u_0, \tag{8}$$

where u_0 is assumed to be uncorrelated with X and Z. Model (8) is identical to model (4) except for the fact that we have dropped the subscript i to simplify notation. Let Z_a be the aggregated or grouped variable Z, corresponding to $\bar{Z}_j$ or the set of dummy variables [Z_1…,Z_D], and let us consider the regression of Z on Z_a

$$Z=Z_a\rho+v, \tag{9}$$

then Z can be written as the sum of its projection in the space generated by the columns Z_a, $P_{Za}Z= Z_a\hat{\rho}$, and its projection in the orthogonal space, $M_{Za}Z=\hat{\varepsilon}$. If we regress $RIF(Y,q_\tau)$ on X and Z_a then equation (8) becomes

$$RIF=X\beta_0+Z_a\hat{\rho}\gamma_0+\hat{\varepsilon}\gamma_0+u_0 \tag{10}$$

where $\hat{\varepsilon}$ is uncorrelated with Z_a by construction and β_0 is consistently estimated if $\text{Cov}(\hat{\varepsilon},X)=0$. Since $\text{Cov}(\hat{\varepsilon},X)=\text{Cov}(M_{Za}Z,X)=\text{Cov}(M_{Za}Z, M_{Za}X)$, regressing $RIF(Y,q_\tau)$ on X and Z_a produces a consistent estimation for β_0 if $\text{Cov}(M_{Za}Z, M_{Za}X)=0$ i.e. if Z and X are uncorrelated conditioning on Z_a, $\text{Cov}(Z,X|Z_a)$.

To test the assumption that $\text{Cov}(Z,X|Z_a)=0$ we can consider the following regression

$$Z=X\eta+Z_a\theta+v \tag{11}$$

and check whether $\eta=0$ using a Wald test. Comparing different methods to aggregate or discretise the variables Z, we can choose the one which minimizes the Wald test and presumably reduces the bias of β_0.

A similar testing procedure has been proposes also by Geronimus et al (1996) for the case where the model of interest is a mean regression and the continuous covariates are approximated using geocoded variables.

6. Conclusions

In this paper we show how to aggregate individual register or census data to estimate unconditional quantile regressions avoiding both the confidentiality issue and ecological bias. This is feasible when the covariates are categorical variables with a small number of possible values. On the contrary, when some of the covariates are continuous any aggregation strategy leads to some loss of information and a potential ecological bias. However, it is still possible to aggregate the data in a way such that we can estimate without bias a semi-individual quantile regression model, i.e. a regression where the dependent variable and categorical variables are measured without aggregation error, while the continuous variables are approximated by their area-mean or by a set dummy variables corresponding to each possible value assumed by their discretised version. Finally, we suggest a test to check the potential bias caused by approximating these continuous variables.

References

Abrevaya J.D., Dahl C.M. (2008), The Effects of Birth Inputs on Birthweight, *Journal of Business & Economic Statistics*, 26, 379-397.

Bitler M. P., Gelbach J. B., Hoynes H. W. (2006), What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96, 988–1012

Cho W.K.T., Manski C.F. (2008), Cross Level/Ecological Inference, in H. Brady, D. Collier, and J. Box-Steffensmeier (editors), *Oxford Handbook of Political Methodology*, Oxford: Oxford University Press, Chapter 22, 547-569.

Duncan O. D., Davis B. (1953), An Alternative to Ecological Correlation, *American Sociological Review*, 18, 665–666.

Firpo S., Fortin N.M., Lemieux T. (2009), Unconditional Quantile Regressions, *Econometrica*, 77, 3, 953-973

Geronimus A.T., Bound J., Neidert L.J. (1996), On the Validity of Using Census Geocode Characteristics to Proxy Individual Socioeconomic Characteristics, *Journal of the American Statistical Association*, 91, 434, 529- 537

King G. (1997), *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregated data*, Princeton: Princeton University Press.

King G., Rosen O., Tanner M.A., Eds., (2004), *Ecological Inference: New Methodological Strategies*, New York: Cambridge University Press.

Koenker R., Bassett Jr. G. (1978), Regression Quantiles, *Econometrica*, 46, 33–50.

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

Kunzli N., Tager I.B. (1997), The Semi-individual Study in Air Pollution Epidemiology: A Valid Design as Compared to Ecologic Studies, *Environmental Health Perspectives*, 105, 10, 1078-1083.