PEP-PMMA Technical Note Series (PTN 01)

Percentile Weighed Regression

Araar Abdelkrim
University of Laval, aabd@ecn.ulaval.ca
May 29, 2016

Abstract *- In this brief note, we review the quantile and unconditional quantile models, and then we propose a new Quantile Regression (QR) method that we call the Weighed Percentile Regression (WPR). Beforehand, recall that the main aim of these models is to estimate appropriate coefficients for given percentile, which is based on the rank of the dependent variable. Effectively, for a given explanatory variable, the impact can be heterogeneous on the outcome, and this depending on the rank (percentile) of outcome. For instance, the incidence of social assistance program can vary depending on the level of wellbeing or outcome. The QR models are assumed to be helpful to show such heterogeneity in impact.*

Index Terms - About four, alphabetical order, key words or phrases, separated by commas (e.g., Camera-ready, FIE format, Preparation of papers, Two-column format).

1- INTRODUCTION TO THE QUANTILE REGRESSION (QR)

Quantile regressions attempt to assess how the conditional quantile $Q_\tau(Y|X) = \inf\{y: F_{Y|X}(y) \geq \tau\}$, are modified when the determinants $X \in R^p$ of the outcome of interest vary. Remember that the definition of the quantile is:$q_\tau(Y) = \inf\{y: F_Y(y) \geq \tau\}$, or in some words, the lowest element (y) of the random variable Y among the elements that make the cumulative distribution $F_Y(y)$ higher than the percentile of interest τ. Also, it may be helpful to present another general formula for the estimation of the quantile. Formally, let:

- $\rho_\tau(\mu) = (\tau - I[\mu < 0])\mu.$
- $\mu = (y - b)$

Thus, we have that

- $\rho_\tau(\mu) = (\tau - 1)(y - b)$ if $y < b$
- $\rho_\tau(\mu) = (\tau)(y - b)$ if $y \geq b$

Note that

$$\mathrm{E}\,[\rho_\tau(\mu)] = \tau(\mu_y - b) - (\tau - 1)(\tau_b)(\mu_{y<b} - b)$$

As we can observe $\mathrm{E}\,[\rho_\tau(\mu)]$ reacts its minimum where $b = q_\tau$. Indeed, the first element is nil in this case: $\tau = \tau_b$. Also, the first component the predominant part. For instance, if $\tau = 0.5$, we wrongly select $b = \min(y)$ or $\max(y)$, we have that:

1. $\mathrm{E}\,[\rho_\tau(\mu)|b = y_{min}] = 0.50 * (\mu_y - y_{min})$;
2. $\mathrm{E}\,[\rho_\tau(\mu)|b = y_{max}] = 0.50\,(y_{max})$;
3. $\mathrm{E}[\rho_\tau(\mu)|b = q_\tau] = 0.50\,(\mu_y - y_{median}) + 0.25(\mu_{median} - y_{median})$.

Thus, minimizing $\mathrm{E}\,[\rho_\tau(\mu)]$ implies that: $b = q_\tau$.

$$\hat{q}_\tau = \arg\min_b : \sum_{i=1}^{n} \frac{1}{n} \rho_\tau(y_i - b)$$

For the simple classical quantile regression, we can be limited to the linear for as:

$$\hat{\beta}_\tau = \arg\min_\beta : \sum_{i=1}^{n} \frac{1}{n} \rho_\tau(y_i - X\beta)$$

Thus in the case where the predictive part gives $X\hat{\beta}_\tau \approx \hat{q}_\tau$ the found parameters can be interpreted as the impact of covariates on X at percentile τ.

2- INTRODUCTION TO THE UNCONDITIONAL QUANTILE REGRESSION (UQR)

Fripo et al. (2009) have proposed a new model in order to overcome the conditional estimates of the QR, since the practitioners are more interested to the impact of explanatory variables on the unconditional distribution of outcome. Remember that, with QR model, we estimate:

$$\beta_\tau = (F^{-1}(\tau|x + dx) - F^{-1}(\tau|x))/dx$$

The rank of the predicted component can differ from that of the outcome, or also, with and without the marginal change in x, especially if x is a dummy variable that varies from 0 to 1. However, with the UQR regression, we have that:

$$\beta_\tau = \frac{\partial q_\tau(p)}{dp}\frac{\partial p}{dx} = (Pr[y \geq q_\tau | x + dx] - Pr[y \geq q_\tau | x])/f_y(q_\tau)$$

The latter expression is tightly related to what we call the Influence Function (IF). The Re-centered Influence Function (RIF) is as follows:

$$RIF(y; q_\tau; F_y) = q_\tau + (\tau - I[y \geq q_\tau | x])/f_y(q_\tau)$$

Fripo et al. (2009) suggest to use this transformed dependant variable with the OLS regression to estimate the UQR coefficients.

3- *THE WEIGHED PERCENTILE REGRESSION (WPR)*

Starting from the fact that, for practitioners, the main aim by using the quantile regression models is to estimate the coefficients for the group of population with outcomes ranked closely to the percentile of interest, and consequently, it is trivial that these observations should contribute the most in the prediction of the model.

With the new proposed WPR model, the basic idea is to attribute large weights for those with levels of outcome that are close to the percentile/quantile of interest and low weights for those with fare levels of outcome. An easy way is to start by estimating a Gaussian normal distribution around the percentile of interest, and this, using the Kernel method.

Of course, the level of the bandwidth can control for the importance attributed to the observations around of the percentile of interest. In our illustrative examples that follow, the optimal bandwidth, suggested by Silverman (1986), is divided by 3 to increase the precision of the estimated coefficients. Thus, the steps of the PWR model are:

- Generating the percentiles of the dependent variable;
- Estimating the Gaussian density around the percentile of interest. These densities are what we call the percentile weights;
- Running the weighed OLS regression with percentile weights.

4- *ESTIMATIONS AND ARTIFICIAL EXAMPLES*

Mainly, to check the relevance of the different models, we propose to construct a set of artificial examples for which the true values of the coefficients are assumed to be known without regressions. For instance, let x be a random variable, if we construct: $y = 2x$, the relevant models must give a coefficient 2 for the explanatory variable x. This will be the case using the OLS model, or even with the QR model, and this, for any percentile.

Let x_1 be an explanatory variable generated with Stata, and this, based on the following rules:

```
/*Artificial Example A1*/
#delimit cr
set seed 1234
clear
set obs 1000
gen    x1= _n-1
gen p=_n/1000
gen income =  1000 + p*x1 + 0.00001*uniform()
```

As we can observe in this first easy example, we only use one continues explanatory variable. Remember that _n is the position of the observation in Stata. Thus, the percentile is equal to 1 at the last observation (we have 1000 observations). Assume that the percentile of interest p=0.15. Also, assume that the standardized expected change in income between percentiles 0.149 and 0.151 is used as a numerical proxy of the derivative of income with regards to x2. We have that:

- Change in income = (0.151*150 - 0.149*148) / (150-148) = 0.299.
- In general, we will find that the numerical derivative: $\partial income/\partial x_1\Big|_{p=p^*} = 2p^*$. Thus, in this artificial example, we have a clear idea on the accurate level of the coefficient of x_1.

The natural question that may raise at this stage is: Can each of the quantile regression, the unconditional quantile regression and the new proposed percentile weight regression models estimate accurately this coefficient at different levels of percentile? Further, how one can determines the true value of the coefficient for the different artificial examples?

Two methods are proposed to assess the true of coefficients for a given percentile of interest:

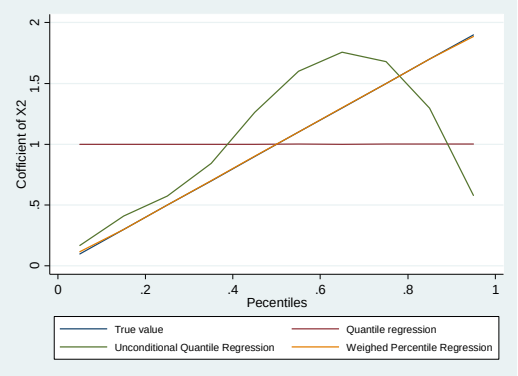
- The first is to use the numerical derivative (command dydx of Stata).
- The second is to use the *derivative locally non parametric regression*, available already in DASP.

In the following table, we show the results of estimates at different percentiles and using different methods.

Table 1: Estimated coefficients: Example A

Percentile of interest	True Value	QR	UQR	WPR
0.05	**0.1**	1	0.06	0.12
0.15	**0.3**	1	0.21	0.3
0.25	**0.5**	1	0.56	0.5
0.35	**0.7**	1	0.95	0.7
0.45	**0.9**	1	1.34	0.9
0.55	**1.1**	1	1.63	1.1
0.65	**1.3**	1	1.77	1.3
0.75	**1.5**	1	1.69	1.5
0.85	**1.7**	1	1.3	1.7
0.95	**1.9**	1	0.54	1.88

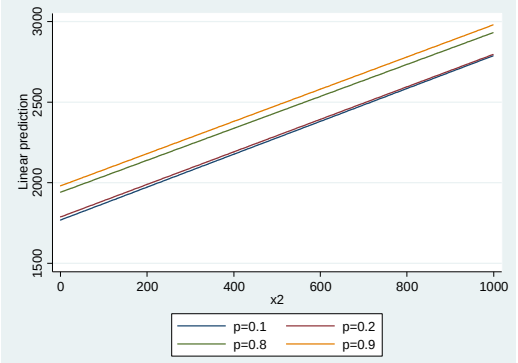
Figure 1: Estimated coefficients: Example A



Surprisingly, only the new proposed method (PWR) enables to estimate consistent coefficients, and this based on the expected true values of the coefficient. Note that, for the quantile regression method, the Stata ***qreg*** module is used. For the unconditional quantile regression (UQR), the Fripo et al. (2009) method is used (see also the attached do file) and precisely the ***rifreg*** Stata command. Obviously, among the questions that can raise is: Why the QR model gives a constant coefficient for x_1 that is equal to one? In reality this result is not strange for the QR model because of the form of the artificial example. Mainly, for QR models, when we have the predicted outcome: $q_\tau(Y|X) = C_\tau + \beta X \ \ \forall\tau \in [0,1]$, and when we have an homoscedastic error term: $V(\varepsilon|X) = \sigma^2$, the model corresponds to the case of *constant translation model* with homogenous slops. In this, case, only the constant coefficient (C_τ) changes with the change of the percentile of interest τ (See also P. Givord & X. Dhaultfoeuille, 2013).

```
/*Artificial Example A2*/
#delimit cr
set seed 1234
clear
set obs 1000
gen     x1= _n-1
gen p=_n/1000
gen income =  2000+ p*x1 + 60*uniform()
qreg income x1, quantile(0.1)
predict  q_1
qreg income x1, quantile(0.2)
predict  q_2
qreg income x1, quantile(0.8)
predict  q_3
qreg income x1, quantile(0.9)
predict  q_4
line q_* x1, legend(order(1 "p=0.1" 2 "p=0.2"  3 "p=0.8"  4 "p=0.9" ))
```

Figure 2: Estimated coefficients: Example A2

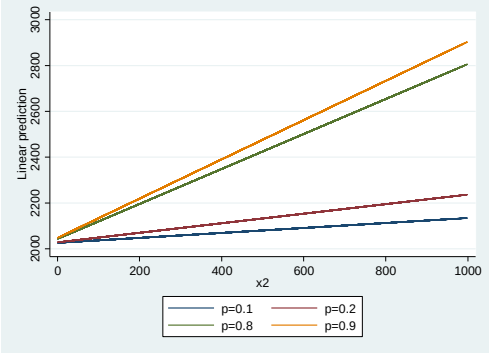


As we can observe starting from this example, the slope is equal to one for the different curves where each lies the x_1 to the predicted outcome. This also corresponds to: $\frac{\partial \hat{q}_\tau}{\partial \tau}\frac{\partial \tau}{\partial x_1}$.

Now we present another model where the correlation between the explanatory variable (x_1) and the percentile variable is nil. This model corresponds also to what we call the *scale translation model* (See P. Givord & X. Dhaultfoeuille, 2013). In this case, the coefficient of x_1 starts to be equal to τ .

```
/*Artificial Example A3*/
#delimit cr
set seed 1234
clear
set obs 1000
gen     x1= uniform()*1000
gen p=_n/1000
gen income =  2000+ p*x1 + 60*uniform()
qreg income x1, quantile(0.1)
predict  q_1
qreg income x1, quantile(0.2)
predict  q_2
qreg income x1, quantile(0.8)
predict  q_3
qreg income x1, quantile(0.9)
predict  q_4
line q_* x1, legend(order(1 "p=0.1" 2 "p=0.2"  3 "p=0.8"  4 "p=0.9" ))
```

Figure 3: Estimated coefficients: Example A3



Thus, the QR model absorbs the influence of the explanatory variable rank from the estimated coefficient, as we can deduce by comparing the artificial examples A2 and A3.

QR, low percentiles and biased coefficients

At this stage, we propose to discuss how estimates at low percentiles can depend on the true impact at highest percentiles. To show this in clear, in this artificial example, it is assumed that the level of the true values of the coefficient is always 0.8 when the percentile of interest is lower than 0.5 and 1.2 if the percentile is higher than the half. The values of x_1 are equal to _n (or: 1, 2, 3…, etc.).

Using the QR regression model, we observe that this this model overestimate the coefficient for low percentiles and is close to the true value for highest percentiles (see red line in the Figure 4). Why this is the case? Assume that the QR algorithm gives a coefficient of 0.1 for the percentiles between 0.1 and 0.5. The generated error with this coefficient will be high for the observations highest percentiles (higher

than 0.5), and it is equal approximately to: $0.5 * (0.8 - 1.2)\mu^*_{x_1}$ and $\mu^*_{x_1}$ is the average x_1 for the observations between 501 and 1000. As an approximation, the average of the absolute errors will be about 0.5*0.4*750=150. Instead of this, the QR algorithm, can force to reduce the error when the disturbance is high. For instance, if the algorithm gives a coefficient of 1.2 for the percentile 0.2, the average error is equal to: 0.5*0.4*250=50. This fact explains why the QR estimates can be biased in presence of heteroscedasticity. What can be the case if we reduce the disturbance of the observations with highest values of the x-axis? As we can observe, in data 2, the values of the explanatory variable are the similar to those of data 1 for the percentiles between zero and half. For the second half, the values of the explanatory variable have changed from (501,502…1000) to (500.02, 500.04, …, 510). Obviously, the income rank remain the same, as well as the true values of coefficients. As it is expected, for the lowest percentiles, The QR starts to give better results until the percentile 0.25. This is because the algorithm starts to estimate low errors for the second half group, and when it gives the true coefficient 0.8.

As a last investigation, we show how the QR coefficients well estimated when we weight the QR model by the inverse of the absolute distance between the dependent variable and the quantile of interest:$1/\text{abs}(y - q(\tau))$. As we can observe, when less importance is given to the error that is fare from the percentile of interest, the estimated coefficients are more accurate. Note that this application is simply for illustration, and it is not tested for general models.

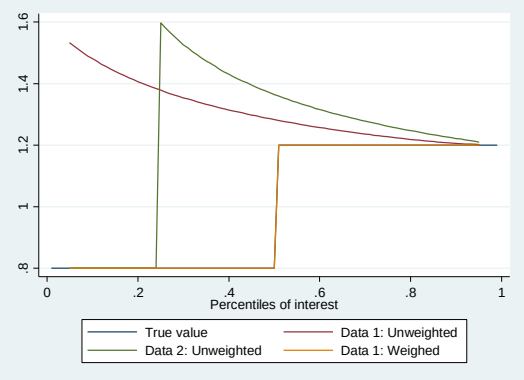
```
/*Artificial Example A4*/
#delimit cr
set seed 7777
clear
set obs 1000

/*DATA1*/
    gen      x1  = _n
    gen income1  = 20+ 0.8 * x1 if p <= 0.5
replace income1  = 20+ 1.2 * x1 if p >  0.5

/*DATA2*/
    gen        x2 = _n            in   1/500
replace        x2 =  500+_n/100    in 501/1000
    gen income2  = 20+ 0.8 * x2  if p <= 0.5
replace income2  = 20+ 1.2 * x2  if p >  0.5

gen p=_n/1000
```

Figure 4: Estimated coefficients: Example A4



Now, it may be helpful to discuss the difference between the PWR and the QR models. The coefficients of the PWR model are interested to assess exactly: $\frac{\partial y}{\partial x}\Big|_{p=\tau}$. Thus, with this model, we can avoid the influence of correlation between the explanatory variable and the rank of the predicted part. Indeed, mainly, the percentile of interest must be related to the depended variable and not to the predicted part. Further, the predicted component for the observations that are fare from the percentile of interest will not influence the estimated coefficients. Obviously, we cannot say that one statistical approach or model is wrong, except where the latter do estimate what we target. Basically, it may be helpful to be more careful when we interpret the coefficients of the QR model.

MULTIVARIATE ARTIFICIAL EXAMPLES

At this stage, we propose to increase partially the complexity of the artificial example, and this by adding another explanatory variable x_1 that can be partially correlated with x_2.

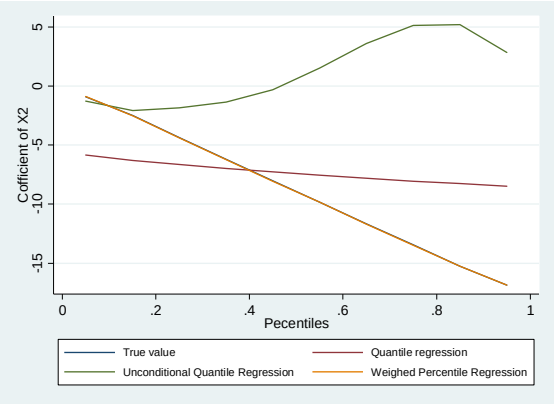
```
/*Artificial Example B1 */
#delimit cr
set seed 1234
clear
set obs 1000
gen     x1=  3* _n^1.1
gen     x2= _n-1
gen p=_n/1000
gen income =  1000 + x1 + p*x2 + 0.00001*uniform()
```

What is the trick to assess the accurate expected coefficient of x_1 when x_2 or a set of other covariates are in the model? Remember that, for the locally linear non parametric approach we cannot use more than one explanatory variable, and this is also the case for the numerical derivative approach. To overcome this difficulty, it is suggested the following trick:

- Estimating the model with both explanatory variables x_1 and x_2;
- Removing $\beta_2 x_2$ from the dependent variable;
- Performing the non-parametric regression of the residual on x_1.

Mayo, 2016, Araar, A.

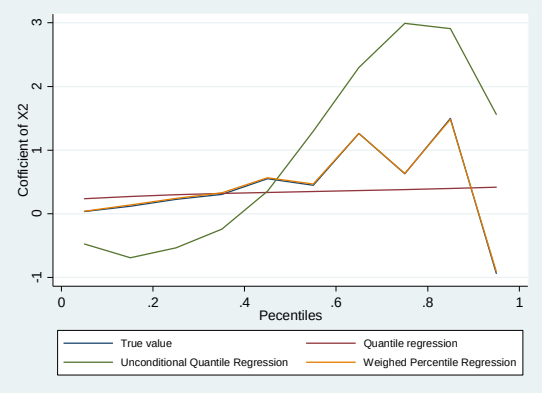
Figure 5: Estimated coefficients: Example B1



As we can observe, the new proposed PWR method continues to function well. Now, we propose another example, where v2 is not fully linear.

```
/*Artificial Example B2 */
#delimit cr
set seed 1234
clear
set obs 1000
gen     x1= -(_n-1)^0.5
gen     x2=  20* _n^2
gen p=_n/1000
gen income =  1000 + x1+ p*x2 + 0.00001*uniform()
```

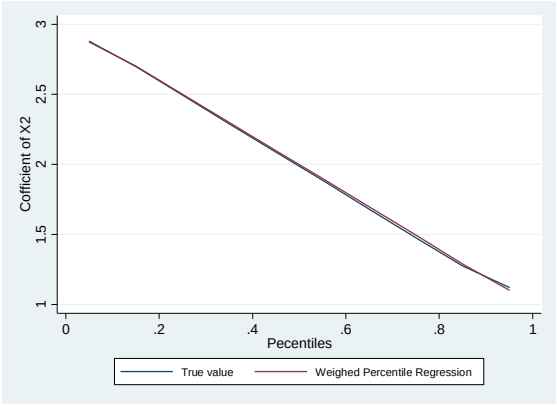
Figure 6: Estimated coefficients: Example B2



Among the main objectives of this note is to propose a new approach that can estimate the quantile regression model when the ranking variable is different from the dependent variable. For this end, we propose to simply ranking the observations according to the ranking variable of interest (for instancex_1), and then to run the PWR model.

```
/*Artificial Example B3 */
#delimit cr
set seed 1234
clear
set obs 1000
gen     x1=  3* _n^1.5
gen     x2= -(_n-1)
sort x2
gen p=_n/1000
gen income =  1000 + x1+ p*x2 + 0.00001*uniform()
```

Figure 7: Estimated coefficients: Example B3



DUMMY VARIABLE AND ARTIFICIAL EXAMPLES

At this stage, we focus on the consistence of the coefficients of the dummy explanatory variable. Mainly, for the artificial example, we assume that the treatment variable is a dummy variable. Further, we assume that, for each percentile of interest, we a population group that represents this percentile. Some individuals of this group are treated (dummy variable=1) and the rest are not treated (dummy variable=0). With a sample size of 4000 observations, the percentile 1/200 will be represented by the first 20 observations as an approximation, the percentile 2/200 by the next 20 observations, and so on. Thus, we have 200 percentile-groups. For simplicity, we assume that for each percentile group, 10 observations are treated and 10 are not treated. As we can observe in the Stata code above, the impact of the treatment can be summarized as follow:

- 10 if p in 0.0 to 0.2;
- 20 if p in 0.2 to 0.4;
- 26 if p in 0.4 to 0.6;
- 32 if p in 0.6 to 0.8;
- 44 if p in 0.8 to 1.0;

```
/*Artificial Example C1 */
#delimit cr
set seed 1234
clear
set obs 4000

gen     x1=  3*_n^1.2
gen     x2= (_n-1)

gen v=int((_n-1)/20)+1
gen p=min(1,v/200)

gen treatment = .
local cho=1
forvalues i=1(10)4000 {
local j=min(`i'+10, 4000)
dis `i' "  " `j'
replace treatment=( `cho' == 1 ) in `i'/`j'
local cho = `cho'*(-1)
}
```

```
gen         x3= treatment *10 if p>=0.0 & p<=0.2
replace     x3= treatment *20 if p> 0.2 & p<=0.4
replace     x3= treatment *26 if p> 0.4 & p<=0.6
replace     x3= treatment *32 if p> 0.6 & p<=0.8
replace     x3= treatment *44 if p> 0.8 & p<=1.0
gen income =  1000 + x1+ p*x2 + x3+  0.00001*uniform()
```

Table 2: Estimated coefficients: Example C

Percentile of interest	True Value	QR	UQR	WPR	WPR, low bandwidth
0.05	10	36.4	67.5	10.4	10.07
0.15	10	35.4	136.	11.4	10.21
0.25	20	33.5	202.	22.4	20.35
0.35	20	32.2	243.	23.4	20.49
0.45	26	34.9	257.	30.4	26.63
0.55	26	33.5	248.	31.4	26.77
0.65	32	30.2	218.	38.4	32.91
0.75	32	29.0	171.	39.4	33.05
0.85	44	26.6	109.	52.4	45.19
0.95	44	24.2	43.1	53.4	45.33

As we can observe, the WPR method continues to produce more consistent estimates of coefficients even with dummy variable (treatment). However, we can add the following remarks for the QR model. First, the results of the QR regression becomes similar to the expected or true values only for the case where the treatment variable is the explanatory variable of the model. Introducing other variables can affect two things: 1- The predictive power of the treatment variable, and this, if the latter is correlated with the rest of covariates. 2- Most important, is the marginal impact on the distribution of conditional quantile when the rest of explanatory variables are kept. Obviously, in our example, the impact of treatment was conceived in the artificial example to do not change the income rank.

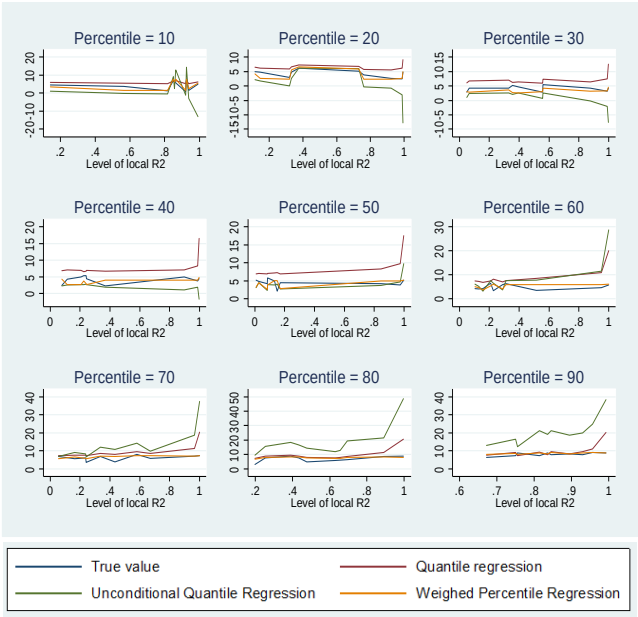
PREDICTIVE POWER AND ARTIFICIAL EXAMPLES

At this stage, we explore another important issue that concern the predictive power of the model. However, before going fare, it may be helpful to answer the following question: Is it the predictive power of local models around the percentile of interest the most important or that of the global model, and this, without focusing on a given percentile – simple OLS for instance?

Assume that the first median group (p: 0 to 0.5), the model fits perfectly the outcome, while for the rest (p: 0.5 to 1) the predictive power is practically nil.

- If the percentile of interest is 0.25 and the predictive power of the model is high between percentiles 0.24 and 0.26, we may be interest a local regression and a lower bandwidth to better estimate the coefficients of the model around the percentile of interest.
- In another case, if the predictive power of the model between percentiles 0.24 and 0.26 is very low, we may be interested to give more importance to the observations that are relatively fare from 0.25 to reduce the sub-sample bias: for instance the p in 0.2 to 0.3. This can be done by increasing the level of the bandwidth.

```
- /*Artificial Example G */
- #delimit cr
- set seed 1234
- clear
- set obs 1000
- local A= (`i'*10)// i varies from 1 to 10 : it
  can control the predictive power of the model
- gen      x1=  `i'* _n^0.8
- gen      x2= (2.5- 0.5*`i'*uniform())*_n
- gen  income =100+x1+_n/100*x2+uniform()*`A'
```



As a general rule, one can start by estimating the local model (keeping observation between p-0.01 to p+0.01, and p is the percentile of interest) and then, select an appropriate bandwidth that can have an inverse relationship with for instance the R2 of the local model. The proposed WPR approach is implicitly based on this idea. For the artificial example, we will control the level of the predictive power of the model by changing the importance of the added random component.

As we can observe from the figure above, WPR approach continues to show its relevance for the consistency of the estimated percentile coefficients.

BRIEF CONCLUSIONS

Based on these investigations we have that:

- Our analysis show that we need to do better to estimate consistent coefficients of the percentile models.

Mayo, 2016, Araar, A.

- Obviously, we recognize the limitation of the explorative approach used in this note, but we hope that other econometric researchers can help in developing a consistent framework.
- Meanwhile, these explorations helps to show the limitations of the usual QR and UQR models or the need of better interpreting their coefficients.

BASIC REFERENCES

- Koenker, Roger and Gilbert Bassett. 1978. “Regression Quantiles.” Econometrica. January, 46:1, pp. 33–50.
- Roger Koenker & Kevin F. Hallock, 2001. "Quantile Regression," Journal of Economic Perspectives, American Economic Association, vol. 15(4), pages 143-156, Fall
- Firpo, Sergio, Fortin, Nicole and Lemieux, Thomas, (2009), Unconditional Quantile Regressions, Econometrica, **77**, issue 3, p. 953-973.