Poster PO-12

SAS macro to obtain reference values based on estimation of the lower and upper percentiles via quantile regression.

Neeta Shenvi
Department of Biostatistics and Bioinformatics, Rollins School of Public Health

Amita Manatunga PhD
Department of Biostatistics and Bioinformatics, Rollins School of Public Health

Andrew Taylor MD
Department of Radiology and Imaging Sciences, Emory University School of Medicine, Emory University, Atlanta, Georgia 30322

ABSTRACT

Reference or normative values for disease or healthy populations are often required for medical studies. For example, in order to differentiate disease subjects from healthy subjects, the percentiles of the distribution of the disease marker among healthy subjects are frequently used. In addition, the disease marker may depend on certain covariates and there is a need to adjust for these covariates when establishing the reference values. Quantile regression methods (QR) have numerous advantages over existing least squared methods; for example, QR can deal with skewed data without usual distributional assumptions and is flexible enough to allow different regression coefficients for different percentiles. SAS code is available to program the QR and adjust for covariates for a given dataset; however the procedure is not automated to provide organized reports in the presence of many outcome variables. We describe a SAS macro that combines three SAS® procedures (Proc Quantreg, Proc Report and Graph template language) that provides the estimation of quantiles and regression equations with and without adjustments for possible covariates. In addition, the program provides informative plots to describe the relationship between the percentiles and covariate. We illustrate our program using a kidney study where reference values of renal area and length are determined for males and females from ^{99m}Tc- MAG3 renal scintigraphy.

INTRODUCTION

In many medical studies, reference values of biomarkers or outcomes are established to distinguish disease subjects from disease free (normal) subjects. Typically, outcomes are measured from healthy/normal subjects and the distribution of the outcome is examined to define the reference values. A value that is considered outside the range of the reference values is interpreted as an abnormal value.

This paper uses a dataset originated from a kidney study. The purpose of this study was to establish reference values for renal size determined from ^{99m}Tc-MAG3 renal scintigraphy and to derive regression equations to predict normal limits. A ^{99m}Tc-MAG3 renal scan consists of a series of images and time activity curves generated over a 20-30 minute period as the radioactive tracer, ^{99m}Tc-MAG3 , is injected intravenously, is removed from the blood by the kidneys and tranvels down the ureters to the bladder. The scan is usually obtained to evaluate relative function, renovascular hypertension and/or obstruction. The interpretation of renal images and curves is enhanced by having additional knowledge about the size of the kidney. Hence, the length, width and area of left and right kidneys of healthy subjects were measured to establish reference values. It is also known that some of these renal sizes may be correlated with the body surface area (BSA) (1). Therefore, it is important to establish reference values of outcomes after adjusting for the BSA since some renal sizes that would be normal is a short person will be abnormally small in a large person.

There are several commonly used approaches for establishing reference values but they have some known limitations. In order to define cut-points for abnormal values, a frequent method is to define the upper and lower bounds by assuming the data follow a normal distribution A normal distribution assumes that 95% of the data is centered around mean ± 2SD and therefore the 2.5% and 97.5% percentiles are chosen as possible cutpoints as reference values. This is intuitively straightforward but inappropriate for non-normally distributed outcomes. When possible, the data are transformed to reduce the degree of non-normality, but it does not address the limitation of assuming constant covariate effects. The constant covariate effect assumes that a covariate impacts the response of

interest to a similar extent at all percentiles, therefore, interpretation of reference values with original scale becomes difficult. As an alternative method, the percentiles of the distribution of outcomes such as 5^{th} and 95^{th} are also estimated based on the histogram. Without any information about the precision of this cut-point, the interpretation of the reference value is limited. Most importantly, the aforementioned methods do not adjust for covariates. For example, they can provide the reference values for the kidney length, however, the adjustment for BSA is not possible.
In order to adjust for possible covariates, linear regression is commonly used. Linear regression is most appropriate when there is interest in modeling the mean response of a variable for examining the possible influence of covariates. The inference about the regression coefficients naturally depends on the normal assumption of the error term in the regression model. However, when there is an interest in estimating extreme values of the distribution (5^{th} and 95^{th} percentiles) linear regression method is not robust to the departure of normality assumption (2).

Quantile regression approach is considered to provide robust estimates compared to ordinary least squares method when the research interest lies not in the mean response but rather in determining the extremes of response (e.g. 5th percentile) (3). When data are skewed with heavy tails, the linear regression coefficients may be less accurate, particularly in the presence of outliers which leads to a considerably large bias in estimating the most extreme percentiles. Quantile regression method does not require normal assumption and can be used to estimate reference values with and without covariate adjustments. When there are no covariates, it still provides a more robust estimation of extreme percentiles (5^{th} and 95^{th}) while utilizing the whole dataset. On the other hand, the direct percentile estimation based on the histogram relies on a few data points that lie in the extremes of the distribution. A key advantage of the quantile regression approach is that its ability to adjust for covariates for a given percentile. For example, the covariate effect may not be the same for 5^{th} and 95^{th} percentiles. Quantile regression is quite flexible to allow different regression equations for different percentiles (4,5).

In this paper, we present a SAS macro that combines three SAS procedures (Proc Quantreg, Proc Report and Graph template language) that provides the estimation of quantiles, regression equations with and without adjustments for possible covariates. In addition, the program provides plots for describing the relationship between the response outcome and the covariate for various percentiles. The details of the SAS program are described in Section 3. We illustrate our program using a kidney study where normal values of renal area and length are determined for males and females from ^{99m}Tc- MAG3 renal scintigraphy.

2. METHOD

2.1 GENERAL DEFINITION

Quantile regression is as an important extension of linear regression which models the outcome mean response in terms of the covariates. The focus of the quantile regression is to directly model one or multiple specified quantiles (percentiles) of a response without any distributional assumptions. Specifically, a linear quantile regression model assumes that

$$y_\tau = Q(_\tau|x) = x^\tau\beta(\tau) = \beta_0(\tau) + \beta_1(\tau)x1 + \beta_2(\tau)x2 + \ldots..$$

where $Q(_\tau|\, x)$ denotes the $100\tau^{th}$ percentile of y given covariate x, and $0 < \tau < 1$.

Regression coefficients in quantile regression are dependent on the percentile instead of having one set of regression coefficients that applies to the entire data range. Quantile regression allows for different regression coefficients at the median and at the tail end of the data. Different estimated regression equations are obtained at different percentiles. This feature allows more flexibility in modeling procedure.

3. COMPUTER PROGRAM

3.1 PROGRAM OVERVIEW

The macro is designed to perform quantile regression to estimate percentiles and the corresponding confidence intervals. The macro performs quantile regression for each dependent variable with and without covariate adjustment. The macro can also provide percentile plot describing the relationship of the response variable and the covariates for specified percentiles.

The program has flexibility to specify several response variables and their corresponding covariates simultaneously. The program also allows for the “grouping” variables. For example, the quantile regression for males and females can be calculated.

All macro functions are bundled into one sas file (quantile_reg_macro_v7.sas). The code is given at the end of this paper.

3.2 DATA PREPARATION

The dataset must have one line per measurement, and that line should contain the outcome variable, covariate variable, and gender and patcode as subject identification variable. The partial dataset is shown below.

Patcode	Age	BSA(m)	LK_AREA (cm^2)	RK_AREA (cm^2)	Gender
1	52	2.1	58.3	63.6	1
2	37	1.8	60.8	71.0	0
3	49	2.1	79.9	69.5	1
4	59	1.6	53.0	55.4	0
5	40	1.5	57.0	58.3	1

In this dataset, Left kidney area (LK_AREA, cm^2) and right kidney area (RK_AREA, cm^2) are the response outcomes. Subject's BSA(m) and age are the covariates.

3.3 INPUT PARAMETERS

In order to use quantile regression with covariate adjustment to estimate the reference values, the data needs to be prepared. The dataset should contain one row for each participant with continuous outcome and continuous covariates.

The user should specify following four global macro parameters

(i) reg_quantile: This variable stores a list of quantiles for which outcome reference values are computed after adjusting for the covariate. For example following %let statement is written to obtain 2^{nd},5^{th},10^{th},50^{th} ,75^{th}, and 95^{th} percentile reference value of the outcome,

```
%let reg_quantile=p2 p5 p10 p25 p50 p75 p95; **** Outcome quantiles ;
```

(ii) covarq: This variable stores a list of covariate quantiles. For example to compute the reference value for a given covariate value equals (x). Alternatively, user can provide a list of covariate percentile such as 25^{th} or 50^{th} percentile which are calculated from the dataset. The following statement specifies 5^{th},25^{th}, 50^{th},75^{th},95^{th} percentile for covariate.

```
%let covarq=p5 p25 p50 p75 p95;*** covariate percentiles;
```

Or user can provide numeric value of covariate.
For example, to specify covariate adjustment for values 1.5 and 2.9

```
%let covarq=n1.5 n2.9;                    *** numeric covariate value;
```

(iii) filepath: This is simply a valid directory path depending on user's operating system

```
%let filepath=H:\myfolder1\myfolder2\myfolder3;
```

(iv) filename: This is a valid file name without file extension(such as rtf or pdf).

```
%let filename=myfilename;    *should be valid name without extension (e.g .rtf or .pdf);
```

When listing more than one quantile and covariate quantile (global macro variables *reg_quantile* and *covariq*) a space must be included to separate the quantiles and each quantile must be specified as p2 p5 etc.

3.3 MACRO INVOCATION

The macro invocation call is as follows:

```
%final_quant(data =,yvar =,xvar =, model =,log =); * model 1;
```

Macro call input parameters:

1. data: Input data set, as described above
2. yvar:response outcome variable
3. xvar: covariate variable of interest for adjustment
4. *model: A* simple numeric sequence of the model (the first macro call should have model=1, the subsequent macro calls could take value greater than 1)
5. *log: It* is the indicator in case outcome yvar be log transformed (log=1 if log transformed ,otherwise 0).

In this example, we use two response variables (LK_AREA, RK_AREA) and their corresponding covariates (BSA, AGE) . Quantile regression (2^{nd}, 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}) is obtained for males and females (gender as grouping variable) separately. In addition, we are requesting the program to calculate the reference values of the response at the 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} of the covariate values.

Below we show 3 macro calls to build 3 models:

```
%final_quant(data=data,yvar=LK_AREA,xvar=BSA, model=1,log=0); * model 1;
 %final_quant(data=data,yvar= RK_AREA,xvar=BSA, model=2,log=0); * model 2;
 %final_quant(data=data,yvar= LK_AREA,xvar=AGE, model=3,log=0); * model 3;
```

The *final_quant call* macro results in 3 output datasets

(i) *all_nocovar2*: A table that contains covariate unadjusted estimates of percentiles for overall cohort and for males and females for *xvar* listed in macro calls

(ii) *all_eq :A table that* contains the estimates of the quantile regressions and equations for each of the percentiles for overall cohort and for males and females for models used in each macro call

(iii) *all_pred*: A table that contains the covariate adjusted percentiles and CI for the quantiles specified for the covariate.

3.4 RTF OUTPUT FILES AND PLOTS

To create an RTF files and regression plots, user needs following two calls. It is simply done as follows

```
% first_quant_output ();
%do_plot();
```

The first_quant_output() call produces an rtf file with 3 tables.

Table 1: Reference Percentile Estimates with 95% CI: Overall and by Gender
Table 2: Covariate Adjusted Reference Percentile Estimates with 95% CI: Overall and by Gender
Table 3: Covariate Adjusted Regression Coefficients and Regression Equations: Overall and by Gender

The %*do_plot*() call gives a plot of predictive percentile for response variable with grouping variable(males, females) overlaid with the raw data.

3.5 MACRO MAIN PROGRAM

We give below sequence of code that user needs to follow:

```
%let reg_quantile=p2 p5 p10 p25 p50 p75 p95; **** Outcome quantiles ;
%let covarq=p5 p25 p50 p75 p95;*** covariate percentiles;

%let filepath=H:\myfolder1\myfolder2\myfolder3;      *should be valid path;
%let filename=myfilename;    *should be valid name without extension(e.g .rtf or
```

```
.pdf);

%include "quantile_reg_macro.sas";  **** macro functions bundle ;

%final_quant(data=,yvar= ,xvar=, model=1,log=0); * model 1;
%final_quant(data=,yvar=,xvar=, model=2,log=1); * model 2;
%final_quant(data=,yvar=,xvar=, model=3,log=0); * model 3;

/**** input parameters for output tables ********************/
%first_quant_print();

/**** output plot ********************/
%do_plot();
```

3.6 SAS PROCEDURES USED IN MACROS

The QUANTREG [[ref] procedure from SAS/STAT is used to estimate reference quantiles and confidence intervals with and without covariate in quantile regression. The slope and intercept from quantileregression were used to calculate each percentile equations and confidence intervals. Means procedure is used to calculate covariate quantiles specified by global variable *covarq*. The covariate adjusted percentiles and confidence intervals for outcome were then calculated from estimates of quantile regression. Furthermore, several data steps and PROC SQL calls are made during the calculations. SAS Graphics Template Language (GTL) is used for the plots.

4. EXAMPLE

4.1 REFERENCE VALUES FOR KIDNEY SIZE

We demonstrate the use of our SAS codes in estimating reference values for kidney size adjusting for Body Surface Area (BSA) in normal population. Our data for renal length, width and area is determined from ^{99m}Tc- MAG3 renal scintigraphy. Our study sample consisted of 44 men (mean age 41 years) and 62 women (mean age 39.9 years). These 106 subjects were determined to be normal based on the fact that all underwent screening for possible kidney donation. For each participant, right and left kidney size (area in cm^2 and width and height in cm) and the BSA are obtained

We are interested in assessing the reference values for extreme percentile (p5,p10,p90,p95) for kidney area BSA as a covariate for males and females. We are developing four models for kidney area and length for left and right kidney respectively. We are interested in obtaining the reference values at five percentile values (p5,p25,p50,p75,p95) of BSA. Use the following to invoke the macro:

```
%let reg_quantile=p5 p10 p90 p95;
%let covarq=p5 p25 p50 p75 p95;
%let filepath=H:\Kidney\Paper;
%let filename=quant_paper_output_v5;

%include "quantile_reg_macro_v7.sas";

%final_quant(data=roi2,yvar=LK_Area,xvar=BSA, model=1,log=0);
%final_quant(data=roi2,yvar=RK_Area,xvar=BSA, model=2,log=0);
%final_quant(data=roi2,yvar=LK_longaxis,xvar=BSA, model=3,log=0);
/**** get RTF output ********************/
%first_quant_print();

/**** output plot ********************/
%do_plot();
```

The macro call *%first_quant_print()* outputs three tables as rtf file with filename provided in macro variable *filename* in location specified in the *filepath* variable. The macro call *%do_plot()* gives a rtf file for side-by-side percentile plots for males and females.

The Table 1 provides covariate unadjusted percentiles and 95% confidence intervals for four independent variables by overall and by gender. The Table 2 provides covariate adjusted reference percentile estimates with 95% CI for overall dataset and by gender. The Table 3 provides covariate adjusted regression coefficients and regression equations for overall dataset and by Gender. Due to space limitation, we have included output from model 1 (LK_Area with BSA) only. In Fig. 1, the reference values from regression equation LK_Area male and females are plotted against BSA. Tables 1,2,3 and Fig 1 are given at the end of the paper.

REFERENCES

1. Taylor AT, Shenvi NV, Folks RD, Manatunga A. 2012 Reference values for renal size obtained from MAG3 Scintography. Clin Nucl Med 2012 Accepted

2. Campbell WW, Robinson LR. 1993 Deriving reference values in electrodiagnostic medicine. Muscle Nerve 1993;16:424-428.

3. Koenker R, Bassett G. 1978 Regression Quantiles. Econometrica 1978;46:33-50.

4. Limin P, Joanne Wuu, Benatar M. 2009 Developing reference data for nerve conduction studies: An application of quantile regression. Muscle Nerve. 2009 Nov;40(5):763-71.

5. Limin P, Joanne Wuu, Benatar M. 2009 Reference data for commonly used sensory and motor nerve conduction studies. Muscle Nerve. 2009 Nov;40(5): 772-94.

ACKNOWLEGEMENTS

This research was funded by a United States National Institutes of Health (NIH) grant RO1-EB008838; specifically, the research was supported by the National Institute of Biomedical Imaging and Bioengineering and the National Institute of Diabetes and Digestive and Kidney Diseases.

SAS PROGRAM AVAILABILITY

The SAS program was written in v9.2. The macro was run in SAS v9.3 in the PC environment on a desktop computer with an Intel® Pentium® 4 processor of 1.80GHz speed and 512MB of RAM. The code is available by requesting the author.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Neeta V. Shenvi
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta, GA 30322
E-mail: nshenvi@emory.edu

Table 1. Reference Percentile Estimates with 95% CI: Overall and by Gender

		==========percentile [Lower CI, Upper CI] ==========				
	Gender	p5	p10	p85	p90	p95
LK_Area	overall	51.1 CI[48.5, 53.7]	52.4 CI[50.7, 54.1]	71.0 CI[68.0, 74.0]	72.6 CI[70.3, 74.9]	74.6 CI[69.9, 79.2]
	Females	49.7 CI[46.7, 52.6]	52.4 CI[49.8, 55.0]	69.6 CI[64.6, 74.7]	71.5 CI[68.0, 74.9]	73.2 CI[70.6, 75.8]
	Males	51.7 CI[48.2, 55.2]	53.6 CI[50.1, 57.0]	72.6 CI[66.7, 78.5]	74.5 CI[67.1, 81.8]	79.9 CI[74.0, 85.8]

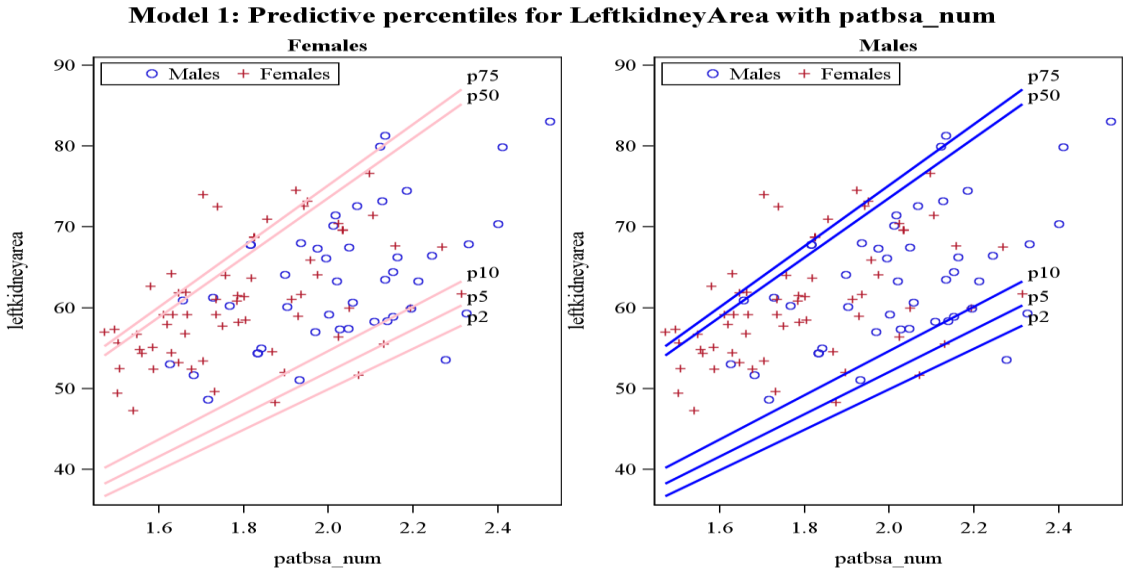
Table 2. Covariate Adjusted Reference Percentile Estimates with 95% CI: Overall and by Gender

					====percentile [Lower CI , Upper CI]==========				
Y	X	Gender	Covar Q	Covar Q value	p5	p10	p85	p90	p95
LK_Area	BSA	overall	p5	1.5	40.1 [37.8, 42.3]	42.1 [40.6, 43.5]	56.6 [54.8, 58.5]	57.8 [56.5, 59.1]	59.0 [56.5, 61.5]
			p25	1.7	44.4 [41.9, 46.9]	46.6 [45.0, 48.2]	62.7 [60.7, 64.8]	64.1 [62.6, 65.5]	65.4 [62.6, 68.1]
			p75	2.1	53.9 [50.9, 57.0]	56.6 [54.7, 58.6]	76.2 [73.8, 78.7]	77.9 [76.1, 79.6]	79.5 [76.1, 82.8]
			p95	2.3	60.3 [56.9, 63.7]	63.3 [61.1, 65.5]	85.2 [82.4, 88.0]	87.0 [85.0, 89.0]	88.8 [85.1, 92.5]
		Females	p5	1.5	40.1 [37.7, 42.4]	42.2 [39.5, 44.9]	57.8 [55.8, 59.8]	59.0 [56.6, 61.5]	60.7 [56.0, 65.3]
			p25	1.7	44.4 [41.7, 47.0]	46.8 [43.8, 49.7]	64.1 [61.8, 66.3]	65.4 [62.7, 68.1]	67.2 [62.0, 72.4]
			p75	2.1	53.9 [50.7, 57.2]	56.8 [53.2, 60.5]	77.9 [75.2, 80.6]	79.5 [76.2, 82.8]	81.7 [75.4, 88.0]
			p95	2.3	60.3 [56.7, 63.9]	63.5 [59.4, 67.6]	87.0 [84.0, 90.0]	88.8 [85.1, 92.5]	91.3 [84.2, 98.3]

					====percentile [Lower CI , Upper CI]=============				
Y	X	Gender	Covar Q	Covar Q value	p5	p10	p85	p90	p95
		Males	p5	1.5	39.2 [34.7, 43.8]	42.0 [38.9, 45.1]	54.0 [51.5, 56.5]	54.6 [51.3, 57.8]	57.4 [54.5, 60.4]
			p25	1.7	43.5 [38.4, 48.5]	46.5 [43.1, 49.9]	59.8 [57.1, 62.6]	60.4 [56.8, 64.1]	63.6 [60.4, 66.9]
			p75	2.1	52.8 [46.7, 58.9]	56.5 [52.4, 60.7]	72.7 [69.3, 76.1]	73.5 [69.1, 77.9]	77.4 [73.4, 81.3]
			p95	2.3	59.0 [52.2, 65.9]	63.2 [58.6, 67.8]	81.3 [77.5, 85.1]	82.1 [77.2, 87.0]	86.5 [82.1, 90.8]

Table 3. Covariate Adjusted Regression Coefficients and Regression Equations: Overall and by Gender

				Regression Coefficient Estimate ± Std Err and Reression Equation				
Y	X	Gender		p5	p10	p85	p90	p95
LK_Area	BSA	overall	Intercept	32.5 ± 10.2	31.6 ± 10.3	22.1 ± 8.0	23.8 ± 11.9	12.6 ± 18.4
			Slope	9.6 ± 5.5	11.9 ± 5.5	24.1 ± 4.2	24.6 ± 6.4	31.7 ± 9.5
			equation	Y= 32.5 + (9.617) * X	Y= 31.6 + (11.93) * X	Y= 22.1 + (24.15) * X	Y= 23.8 + (24.62) * X	Y= 12.6 + (31.73) * X
		Females	Intercept	39.3 ± 20.7	35.1 ± 13.7	11.1 ± 18.7	19.5 ± 28.0	28.0 ± 75.0
			Slope	6.0 ± 12.2	9.6 ± 7.7	31.3 ± 10.2	27.3 ± 14.9	23.2 ± 45.1
			equation	Y= 39.3 + (5.997) * X	Y= 35.1 + (9.592) * X	Y= 11.1 + (31.26) * X	Y= 19.5 + (27.35) * X	Y= 28.0 + (23.21) * X
		Males	Intercept	29.1 ± 102.4	27.2 ± 15.9	18.0 ± 12.7	25.2 ± 18.2	-3.9 ± 77.6
			Slope	11.4 ± 51.8	14.5 ± 8.1	25.9 ± 6.5	22.9 ± 9.2	39.5 ± 42.8
			equation	Y= 29.1 + (11.37) * X	Y= 27.2 + (14.55) * X	Y= 18.0 + (25.91) * X	Y= 25.2 + (22.91) * X	Y= -3.9 + (39.50) * X


Model 1: Predictive percentiles for LeftkidneyArea with patbsa_num
Females
Males
Females
p75
p50
p10
p5
p2
leftkidneyarea
patbsa_num
90
80
70
60
50
40
1.6
1.8
2.0
2.2
2.4
Males
Females
Males
p75
p50
p10
p5
p2
leftkidneyarea
patbsa_num