Sample size and power calculation for univariate case in quantile regression

IOP Conf. Series: Journal of Physics: Conf. Series **948** (2018) 012072 doi:10.1088/1742-6596/948/1/012072

Sample size and power calculation for univariate case in quantile regression

Ferra Yanuar[1]

[1]Mathematics Department, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, 25163, Padang – Indonesia

E-mail: ferrayanuar@yahoo.co.id

Abstract. The purpose of this study is to calculate the statistical power and sample size in simple linear regression model based on quantile approach. The statistical theoretical framework isthen implemented to generate data using R. For any given covariate and regression coefficient, we generate a random variable and error. There are two conditions for error distributions here; normal and nonnormal distribution. This study resulted that for normal error term, sample size is large if the effect size is small. Meanwhile, the level of statistical power is also affected by effect size, the more effect size the more level of power. For nonnormal error terms, it isn't recommended using small effect size, moderate effect size unless sample size more than 320 and large effect size unless sample size more than 160 because it resulted in lower statistical power.

1. Introduction

The least square estimator has many limitations. This estimator estimates the relationships between the conditional mean of the covariates on the response, or it only measures the mean of the response at given values of covariates [1]. The leastsquare estimator can be applied if the classical assumptions are fulfilled unless the conditional mean is not appropriate anymore. Another limitation of the leastsquare estimator is the response variable must follow normal distribution. If not, the conditional mean cannot appropriate to be applied. Due to these limitations, numerous estimator methods have been proposed. One of more popular estimator methods than others is quantile regression approach. Quantile regression estimates the relationship between covariates at different percentile points of the response variable. Quantile regression also has no assumption about the distribution of error. This estimator is widely used in many disciplines, such as in social science research, health research, and environmental sciences.

In many disciplines of research, the study of statistical power and sample size calculation are important factors to be considered in pilot study which then required for future implementation [2]. The accurate statistical power and reasonable sample size must be allowed in order to achieve the purposes of the study.

The sample size is an estimate of how many objects to be involved in a study. In determining the sample size, it needs predetermined parameters of the corresponding probability distribution, such as mean and standard deviations. Sample size can be determined based on these parameters. There are two criteria that can be used to determine sample size, precision analysis, and power analysis.

In calculating the reasonable sample size, it uses the concept of type I error, denoted by α. Type I error or α is the probability of reaching incorrect conclusion. Sample size calculation in precision analysis is usually referred to as the maximum error of an estimate of the unknown parameter. It allows the maximum half width of the $(1-\alpha)100\%$ confidence interval for this error.

Let $X_1, X_2, \dots, X_n$ are identically independent distribution normal random variables with mean μ and variance σ^2. The $(1-\alpha)100\%$ confidence interval for μ if σ^2 is known, is given by following :

$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$,

Then error or ε is defined as :

$$\varepsilon = |\bar{X} - \mu| = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, the sample size is calculated by :

$$n = \frac{z_{1-\alpha/2}^2\sigma^2}{\varepsilon^2} \quad (1)$$

Using Chebyshev's inequality to obtain the following nonparametric approach :

$$P(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2},$$

then set $1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - p$, with $p = P\left(|Z| \geq z_{1-\alpha/2}\right)$. Thus, the sample size can be calculated by:

$$n = \frac{\sigma^2}{p\varepsilon^2}. \quad (2)$$

Besides controlling the Type I error, or α, the researcher is also suggested to minimize Type II error, β, in hypothesis testing while maintaining Type I error at a certain pre-specifed level. The statistical power here is represented by $1-\beta$, the complement of type II error βwhich is the probability of rejecting the impact or change as it occurs. Meanwhile the $(1-\beta)\ 100\%$ is the power of the test represents the probability that a significant result will be detected when the alternative hypothesis is true. Sample size calculation under assumptions of this power of the test is called power analysis.

For two-sample test with equal sample size and having normal distribution when the variances for both samples are known, the sample size can be calculated by[3]:

$$n = \frac{(\sigma_1^2 + \sigma_2^2)\left(z_{1-\beta}+z_{1-\alpha/2}\right)^2}{\delta^2}, \quad (3)$$

where δ is the difference between two population means which are observed. The problems of this study are how to determine the reasonable power and sample size in simple linear regression based on quantile estimation approach.

2. Data and Methods

In this study, we give brief explanation related to quantile regression approach. This study aims to determine the statistical power and calculate the reasonable sample size for univariate case in the quantile regression. In this research, we use the generated data which the method is then implemented and tested.

2.1. Quantile Regression Approach

Let $\boldsymbol{y}$ is the response variable, $\boldsymbol{x}$ is a $p \times 1$ vector of p indicator variables for the ith observation. Let F_ε be the distribution function of error term which is ε with associated probability density function f_ε. The F_ε is supposed to be strictly increasing and absolutely continuous for every $\varepsilon \in \mathcal{J}$for some interval$\mathcal{J} \subset \Re$. Then let $Q_\tau(\boldsymbol{x})$ denote the τ-th ($0 < \tau < 1$) quantile regression function of y given $\boldsymbol{x}$.

The quantile regression model is[4] :

$$Q_\tau(\boldsymbol{x}) = \boldsymbol{x}'\beta(\tau) + F_\varepsilon^{-1}(\tau) \quad (4)$$

Normally, we assume $F_\varepsilon^{-1}(\tau)$ is 0 and τ-th quantile of dependent variable can be expresssed as a linear function of indicator variables. For $i = 1,2,\dots,n$ observations, the quantile regression estimation for $\hat{\beta}(\tau)$ are given by :

$$\hat{\beta}(\tau) = min \sum_i \rho_\tau \left(y_i - \boldsymbol{x}_i^T \beta(\tau)\right), \quad (5)$$

where $\rho_\tau(\varepsilon)$ is the loss function defined by:

$$\rho_\tau(\varepsilon) = \varepsilon(\tau - I(\varepsilon < 0)) \quad (6)$$

We also may write (3) as:

$$\rho_\tau(\varepsilon) = \varepsilon(\tau I(\varepsilon > 0) - (1-t)I(\varepsilon < 0)) \text{or } \rho_\tau(\varepsilon) = \frac{|\varepsilon| + (2\tau - 1)\varepsilon}{2}$$

For independent and identic distribution (iid) errors $\varepsilon_i, i = 1,2,\dots,n$, the asymptotic distribution for $\hat{\beta}$ in quantile regression has the following distribution [5][6]:

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \sim N\left(0, \frac{\tau(1-\tau)}{f\left(F^{-1}(\tau)\right)^2} E(X'X)^{-1}\right) \quad (7)$$

where $f\left(F^{-1}(\tau)\right) > 0,$ and suppose $n^{-1}X'X \equiv Q_n$ converges to a positive definite matrix Q_0, i.e $E(X'X)$.

The errors can be assumed non iid, Koenker [6] showed that the asymptotic distribution of the estimator $\hat{\beta}$ has the following form :

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \sim N(0, \tau(1-\tau)H^{-1}J_n H^{-1}) \quad (8)$$

where

$J_n = n^{-1}X'X$ and $H = \lim_{n\to\infty} n^{-1}X'X\, f_\varepsilon(\tau)$.

2.2. Statistical Framework for Sample size and Power Calculation

In univariate case, we have $x_i,\ i = 1,2,\dots,n_1$ and $y_i,\ i = 1,2,\dots,n_2$. Both groups of observations are independent and normally distribution with means μ_x and μ_y and variances σ_x^2 and σ_y^2, respectively. We then have the following hypothesis to be tested [7][3] :

H_o: $\mu_x = \mu_y$ vs H_a: $\mu_x \neq \mu_y$

The Type I error of the test is :

$$\alpha = P(Z > c|H_0) \quad (9)$$

Where Z is a test statistic and c is the critical value of the rejecting the null hypothesis H_o.

The Type II error is defined as following:

$$\beta = 1 - P(Z > c|H_a) \quad (10)$$

Thus, we have :

power = 1- $\beta = P(Z > c|H_a)$.

For condition σ_x^2 and σ_y^2 are known, and $n_x = n_y = n,$ we can implement Z test:

$$Z = \frac{\mu_x - \mu_y}{\sqrt{\frac{\sigma_x^2 + \sigma_y^2}{n}}} \quad (11)$$

The null hypothesis or H_o is rejected if the absolute value of Z is larger or equal $Z_{1-\alpha/2}$, represented by $|Z| \geq Z_{1-\alpha/2}$.

Under alternative hypothesis we have $\delta = \mu_x - \mu_y$ with $\delta > 0$and $Z_a \sim N(\mu^*, 1)$with $\mu^* = \frac{\delta}{\sqrt{\frac{\sigma_x^2+\sigma_y^2}{n}}} > 0$. Parameter δ is known as predefined constant and as effect size, it plays an important role in the determination of statistical power and reasonable sample size [3].

Therefore, based on equation (10), we have corresponding power here given by :

$$\beta = 1 - P\left(|Z| \geq Z_{1-\frac{\alpha}{2}}|H_a\right) \approx P\left(Z > Z_{1-\frac{\alpha}{2}}\right)$$
$$= P\left(Z - \mu^* > Z_{1-\frac{\alpha}{2}} - \mu^*\right)$$
$$= P\left(Z^* > Z_{1-\frac{\alpha}{2}} - \mu^*\right)$$

With Z^* has standard normal distribution. To achieve desired power of (1-β)100%, set $-z_\beta = z_{\frac{\alpha}{2}} - \mu^*$.

The formula for calculate reasonable sample size then can be obtained by using the same logic. Let β_1 is the coefficient of indicator variable *x*, following is hypothesis test for β_1 :

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Under H_a, $\delta = \beta_1 + 0 = \beta_1$, with $\mu^* = \frac{\delta}{\sqrt{\frac{\sigma_{\beta_1}^2}{n}}} > 0$.

By modifying the equation (3), the formula to determine sample size is given by:

$$n = \frac{\sigma_{\beta_1}^2 \left(z_{1-\beta} + z_{1-\alpha/2}\right)^2}{\delta^2} \quad (12)$$

The value of δ is assummed to be known and the unknown variable $\sigma_{\beta_1}^2$ is calculated using (7) for iid errors :

$$\sigma_{\beta_1}^2 = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2} E(X'X)^{-1}$$

Test statistics under H_0 to test β_1 is given by :

$$T = \frac{\widehat{\beta_1} - 0}{\sqrt{\frac{s_{\beta_1}^2}{n}}} = \frac{\widehat{\beta_1}}{\sqrt{\frac{s_{\beta_1}^2}{n}}}.$$

Under H_a, statistics T has non-central *t* distribution with $\mu^* = \frac{\delta_1}{\sqrt{\frac{s_{\beta_1}^2}{n}}}$. Sample size is calculated using following approximation:

$$P\left(|T| > t_{n-p,1-\alpha/2}|H_a\right) \approx P\left(T > t_{n-p,1-\alpha/2}|H_a\right) \quad (13)$$

We need *n* large enough to achieve desired power of $(1-\beta)$ x 100%. Since sample size *n*which is involved in degrees of freedom and non-central parameter, it is quite hard to find the closed form of*n*. It can be solved using numerical method.

2.3. Generated Data for Error Distribution

We apply generated data in this present study to implement the theoretical framework. For any given covariate $\boldsymbol{x_i}$ and regression coefficient β, we generate a random variable $\boldsymbol{Y_i}$ and error ε_i. There are two conditions for error disributions here; normal and non normal distribution.

Assuming random variable $\boldsymbol{Y_i}$has normal distribution with mean parameter μ_i and variance σ^2, meanwhileerror term has the same distibution. We set the regresssion coefficient $\beta_0 = 1$, $\beta_1 = 1.5$ and $\tau = 0.1, 0.2, \ldots, 0.90$.

3. Result

The simulation study is done with the design as represented above. Following tables show the result of simulation study. Table 1 and table 2 present the estimated sample size and estimated power respectively for normal error term.

Table 1. Estimated Sample Size for Normal Distribution of Error.

Quantil (τ)	Effect size (δ)		
	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.4$
0.1	22	5	3
0.2	63	11	5
0.3	104	17	7
0.4	147	24	10
0.5	193	32	13
0.6	246	40	16
0.7	310	50	20
0.8	393	64	26
0.9	526	85	34

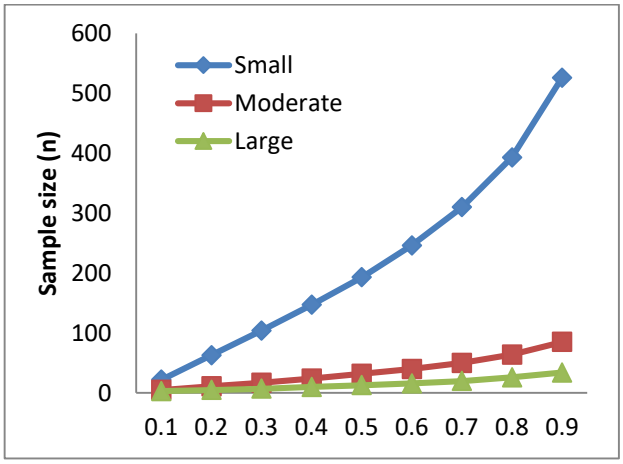

Figure 1. Estimated Sample Size for Normal Distribution of Error.

Table 1 and figure 1 show the estimated sample size under selected points of quantil and three conditions of effect size ($\boldsymbol{\delta}$)**.** The sample size range from n = 3 to n = 526 for the setting that we considered. The table informs us that if effect size is small, it needs large sample size, the opposite sample size is small when effect size is large.

Table 2. Estimated Power for Normal Distribution of Error.

Sample size (n)	Effect size (δ)		
	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.4$
20	0.09	0.19	0.34
40	0.11	0.30	0.55
80	0.16	0.47	0.80
160	0.23	0.72	0.97
320	0.35	0.93	0.99
640	0.56	0.99	1.00

Table 2 shows the empirical power under selected sample size and three conditions of effect sizes, small
($\delta = 0.1$), moderate ($\delta = 0.25$) and large ($\delta = 0.4$). This simulation study observed that the empirical power ranges from 9% to 100%. When effect size is small, the empirical power is far away from 80%. When effect size is moderate, the empirical power is more than 70% for sample size more

than 160. Meanwhile in, large effect size, the empirical power more than 80% when sample size is 80 or more.

The result of simulation study for estimated sample size and estimated power respectively for nonnormal error term are presented in this following Figure 2 and Table 3.

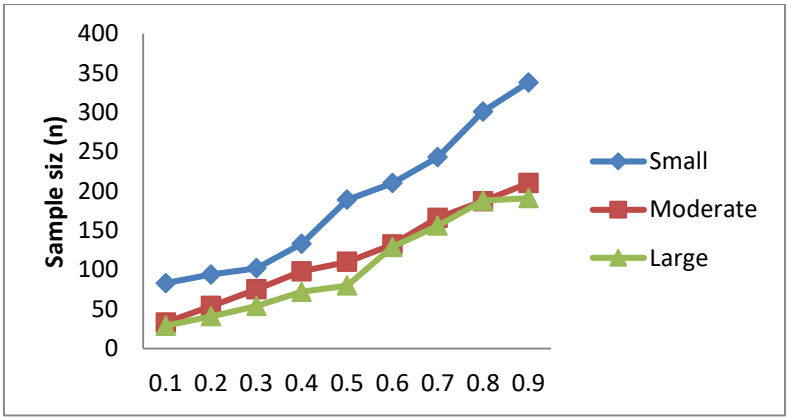

Figure 2. Estimated Sample Size for Non-Normal Distribution of Error.

Figure 2 informs us that sample size is small when effect size is large. Conversely, it needs large sample size if effect size is small. Meanwhile, Table 3 inform us the empirical power for nonnormal error term for selected sample size for three conditions of effect sizes.

Table 3. Estimated Power for Non-Normal Distribution of Error.

Sample size (n)	Effect size (δ)		
	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.4$
20	-	-	0.27
40	-	0.22	0.43
80	0.10	0.36	0.65
160	0.15	0.65	0.82
320	0.29	0.72	0.89
640	0.52	0.92	0.97

4. Summary

In this present study, we reviewed the statistical power and sample size calculation for univariate case in quantile regression model. The statistical theoretical framework was then implemented to generate data using R software. There are two assumptions for the error term, normal distribution, and nonnormal distribution. This study resulted that for normal error term, sample size is large if the effect size is small. Meanwhile, the level of statistical power is affected by effect size as well, the more effect size the more level of power. For nonnormal error terms, it isn't recommended to use small effect size, moderate effect size unless sample size more than 320 and large effect size unless sample size more than 160.

This study described in how to calculate the statistical power and sample size in univariate case only based on quantile approach. For future work, we will explore the statistical power and sample size calculation for multivariate case then.

References

[1] F Yanuar 2014 The Estimation Process in Bayesian Structural Equation Modeling Approach *J. Phys. Conf. Ser.*

[2] G Shan, S Moonie, and J Shen 2014 Sample size calculation based on efficient unconditional tests for clinical trials with historical controls *J. Biopharm. Stat.*
[3] Q Shao and Y Wang 2009 Statistical power calculation and sample size determination for environmental studies with data below detection limits *Water Resour. Res*. **45** 9
[4] G Tarr 2012 Small sample performance of quantile regression confidence intervals *J. Stat. Comput. Simul.* **82** 1 pp 81-94
[5] R Koenker and G Bassett 1978 Regression Quantiles *Econometrica* **46** 1 pp 33-50
[6] R Koenker 2005 Quantile Regression *J. Stat. Comput. Simul.*
[7] Z Gong 2016 Estimation of Sample Size and Power For Quantile Regression