Quantile Regression

Anik Chakraborty
Pratyusha Mukhopadhyay
Swarnajit Podder

Supervisor: **Prof. Tathagata Bandyopadhyay**

July 7, 2022

Methodology

Consider the simple bivariate regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where, u_i is independently and identically distributed errors and (β_0, β_1) are the parameters of the model. So, the quantile functions of y_i are

$$Q(\tau|x) = \beta_0 + \beta_1 x + F_u^{-1}(\tau) \tag{2}$$

where, F_u denotes the common distribution function of the errors. So, by quantile regression for τ-th quantile we estimate the parameters $(\beta_0 + F_u^{-1}(\tau), \beta_1)$.

Assumptions

1. Here, we consider different error distributions like, Normal(0,4), Cauchy(0,2), Laplace(0,2) and t_2. Samples of size 100 are drawn from them as the error (u) values.
2. Samples of size 100 are drawn from $U(0,1)$ distribution and they are treated as the explanatory variable (x).

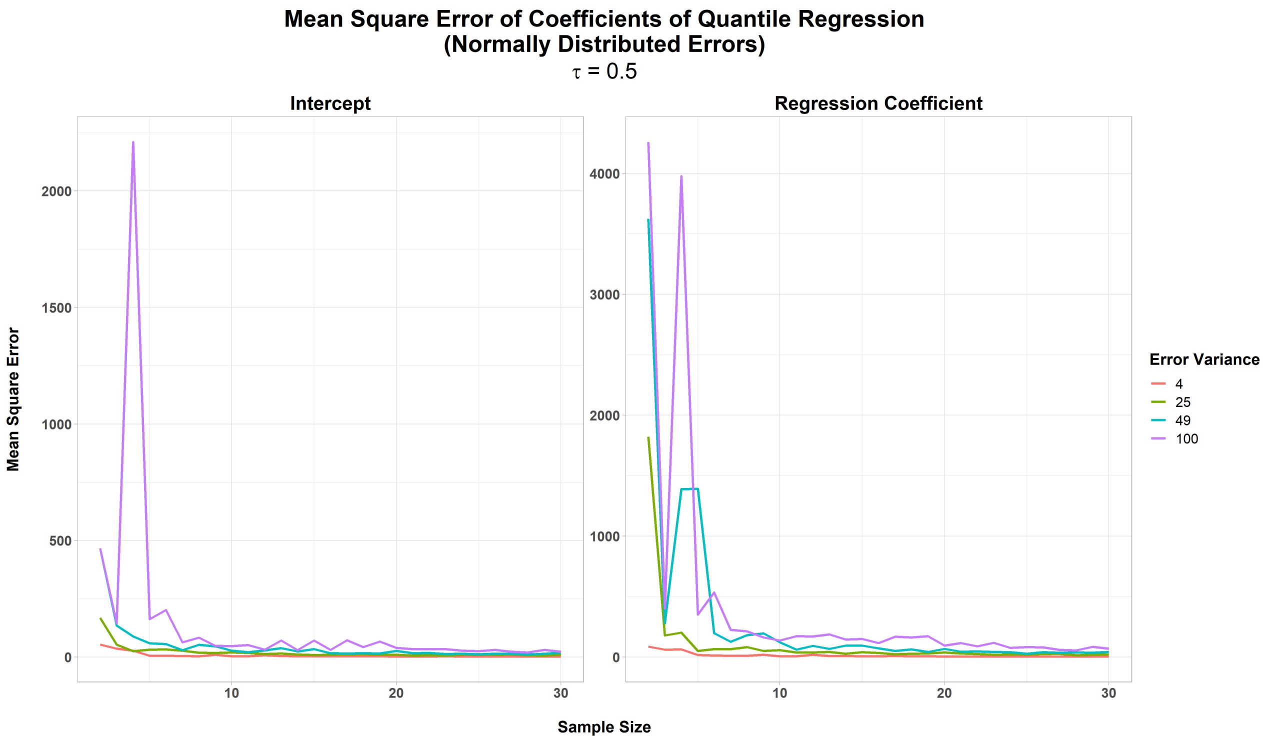
Based on these observations, we estimate the parameters of our Quantile Regression models through simulation. The results obtained are shown in the following table:

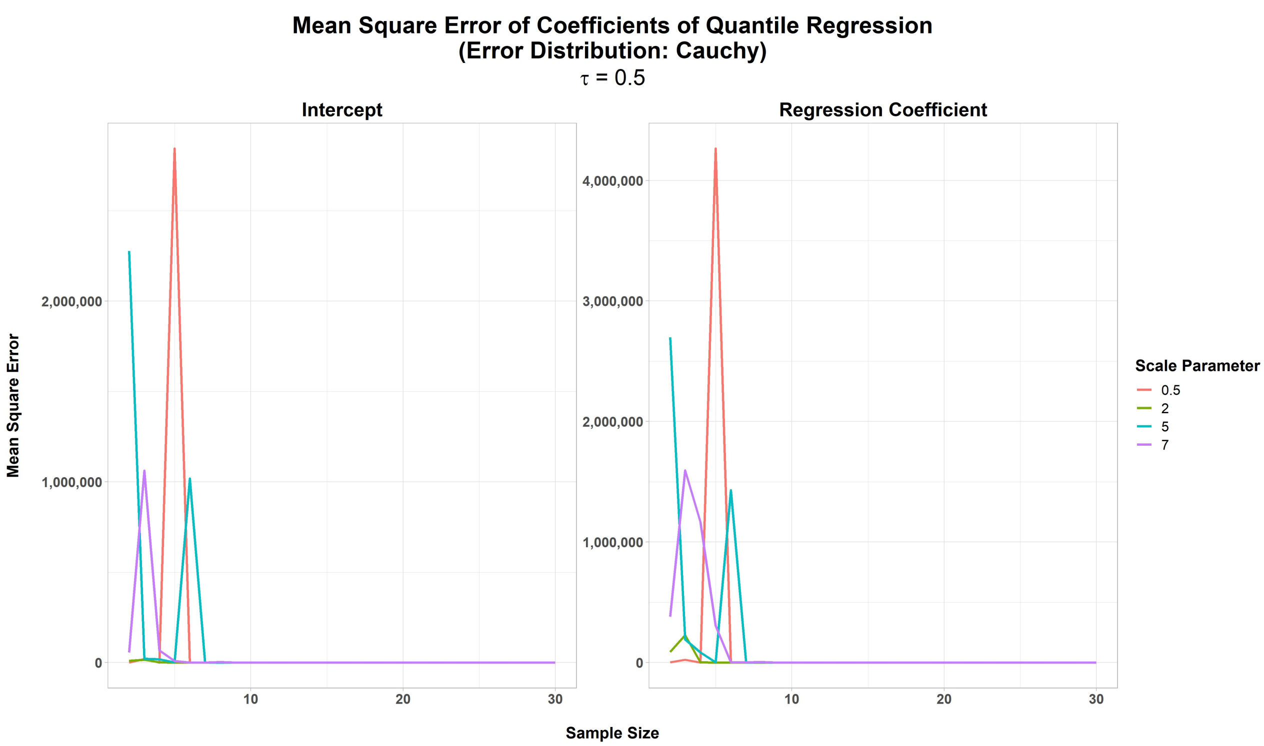
Table 1: Results

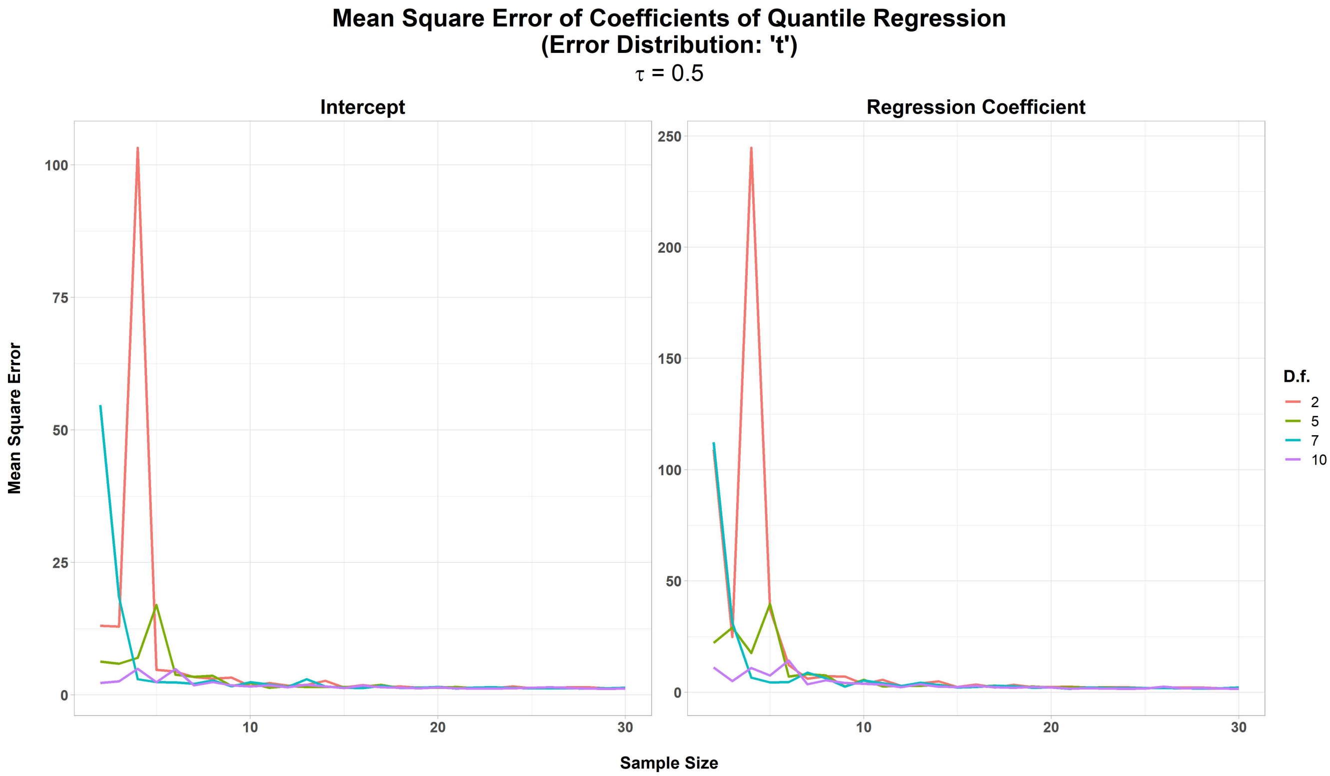
Model: $Y = \beta_0 + \beta_1 x$						
Sample Size			**100**			
Distribution of x (Covariate)			$U(0,1)$			
Distribution of u (Error)			$\mathcal{N}(0,4)$	$\mathcal{DE}(0,2)$	$\mathcal{C}(0,2)$	t_2
$\tau = 0.05$	β_0	**Actual**	-1.29	-2.61	-10.63	-0.92
		Estimate	-1.26	-2.60	-12.91	-1.18
		MSE	0.86	3.49	293.39	3.87
	β_1	**Actual**	5.00			
		Estimate	5.01	4.93	3.65	5.13
		MSE	2.57	9.90	728.67	8.82
$\tau = 0.5$	β_0	**Actual**	2.00			
		Estimate	1.99	2.01	2.01	2.00
		MSE	0.37	0.21	0.36	0.08
	β_1	**Actual**	5.00			
		Estimate	4.99	4.99	4.99	4.99
		MSE	0.95	0.60	1.39	0.27
$\tau = 0.95$	β_0	**Actual**	5.29	6.61	14.63	4.92
		Estimate	5.25	6.54	17.25	5.11
		MSE	0.86	2.46	247.14	3.59
	β_1	**Actual**	5.00			
		Estimate	5.04	5.27	5.30	4.97
		MSE	2.33	10.96	670.26	8.58

MSE and Sample Size for Different Error Distributions

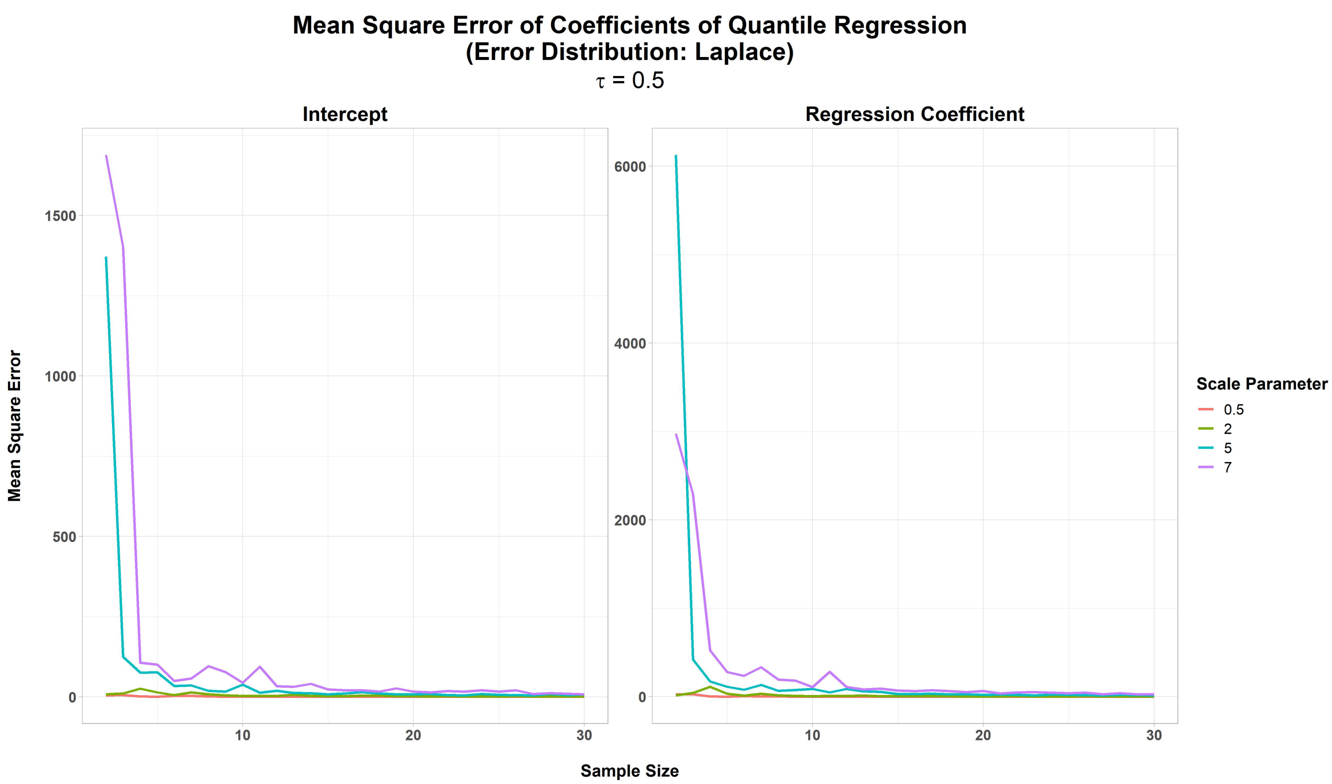
We have plotted the values of Mean Square Error against sample sizes for both the 'intercept' term and the 'regression coefficient' of median-regression taking four choices of error variance into account. The diagrams are shown below:







Mean Square Error of Coefficients of Quantile Regression
(Error Distribution: 't')
τ = 0.5
Intercept
Regression Coefficient
Mean Square Error
Sample Size
D.f.
2
5
7
10



Mean Square Error of Coefficients of Quantile Regression
(Error Distribution: Laplace)
τ = 0.5
Intercept
Regression Coefficient
Mean Square Error
Sample Size
Scale Parameter
0.5
2
5
7

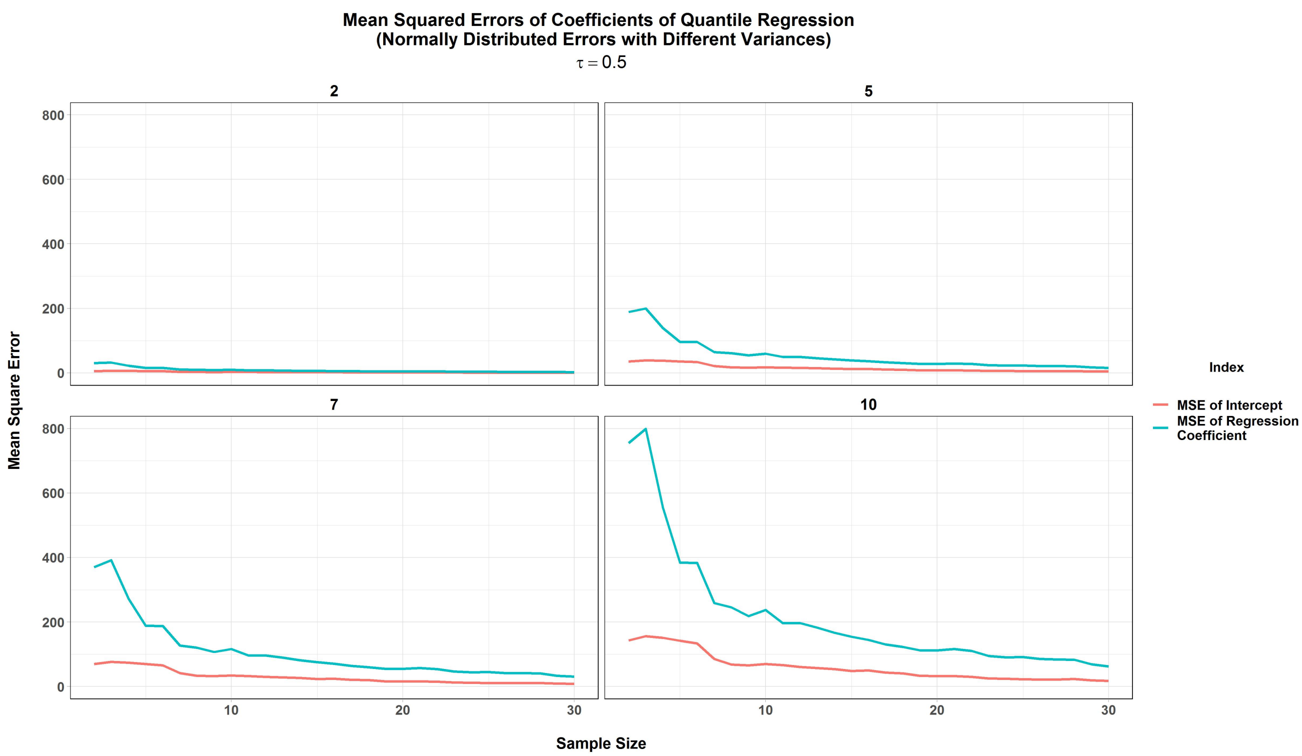
Observations

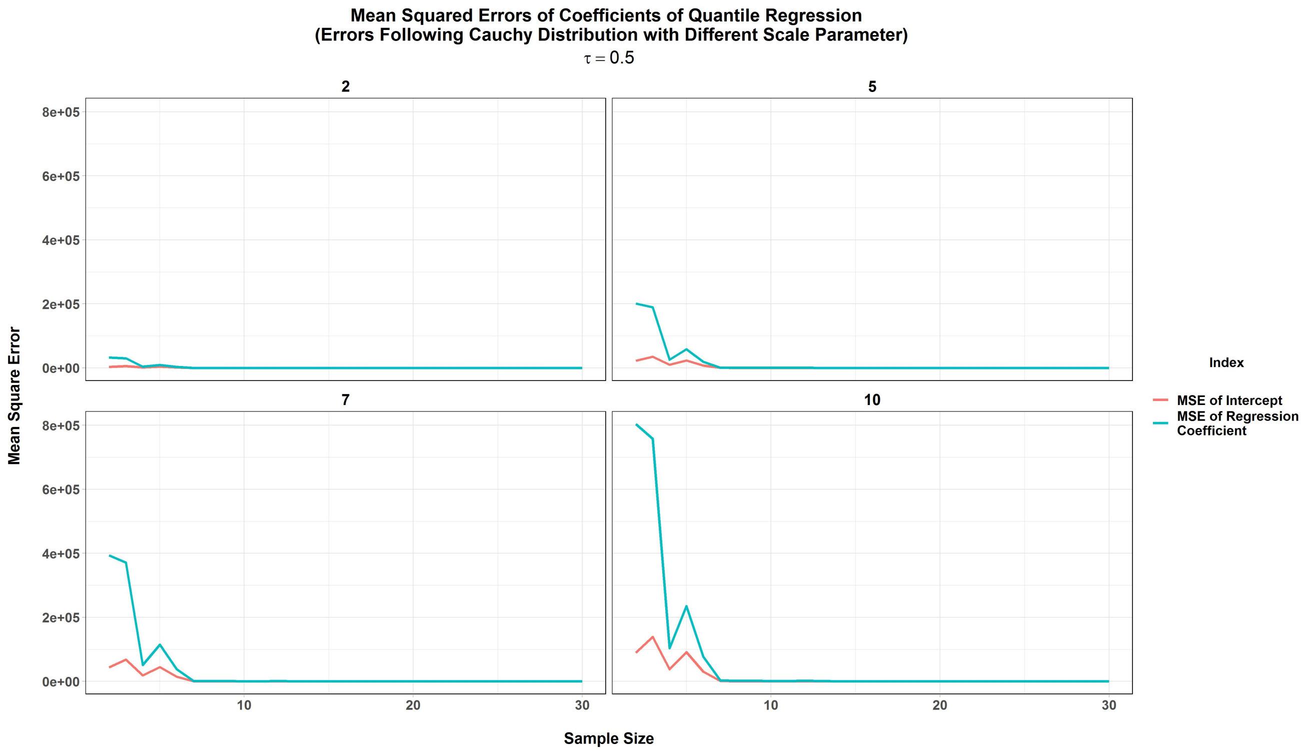
From the above four graphs we observe that -

- Mean Square Error (MSE) gradually decreases with increase in the size of the sample.
- If the errors are i.i.d. Cauchy random variables then the magnitude of MSE is very high.
- In case of the errors following Cauchy distribution the fluctuation in the values of MSE for small sample sizes is very high but that vanishes as we going on increasing the sample size.

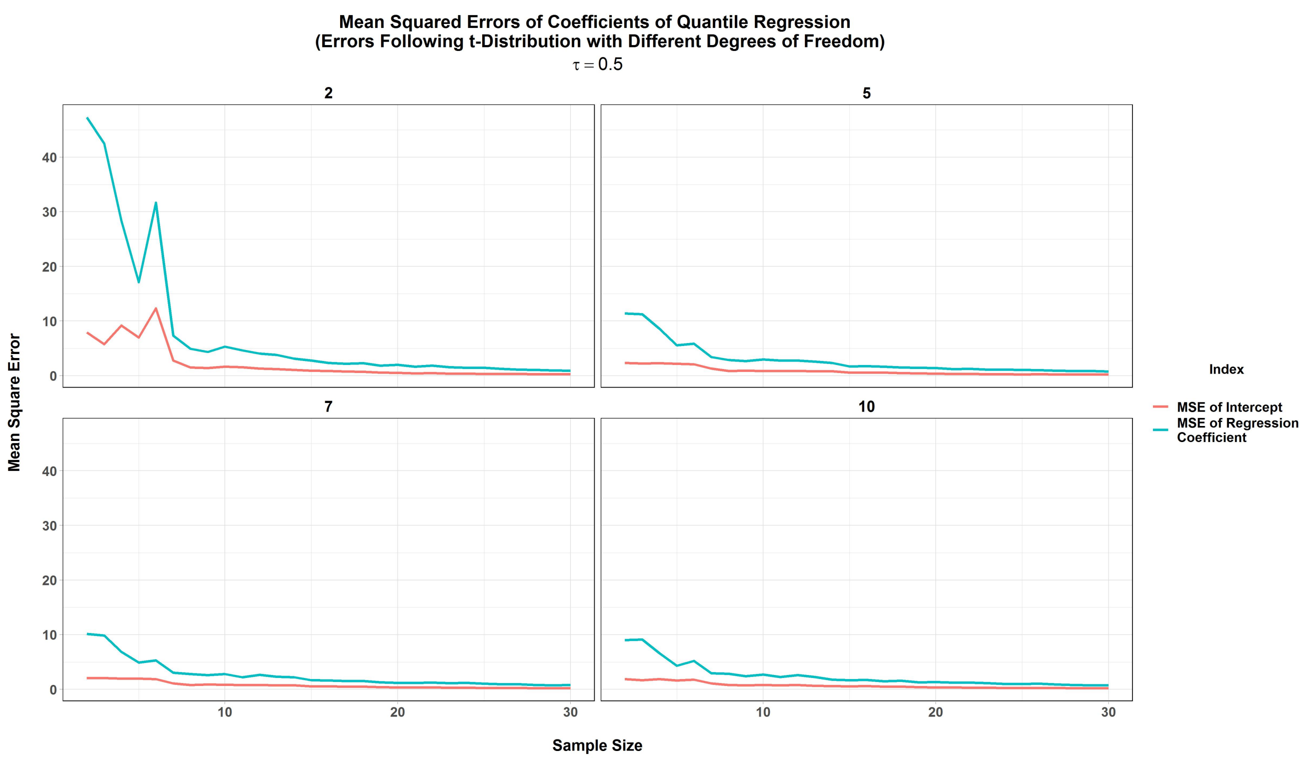
Comparison between Intercept and Regression Coefficient

Next, we have plotted the same but now to compare the MSEs of a term and the regression coefficient in a panel. The four panels are for four choices of variability of the error distribution. The diagrams are shown below:

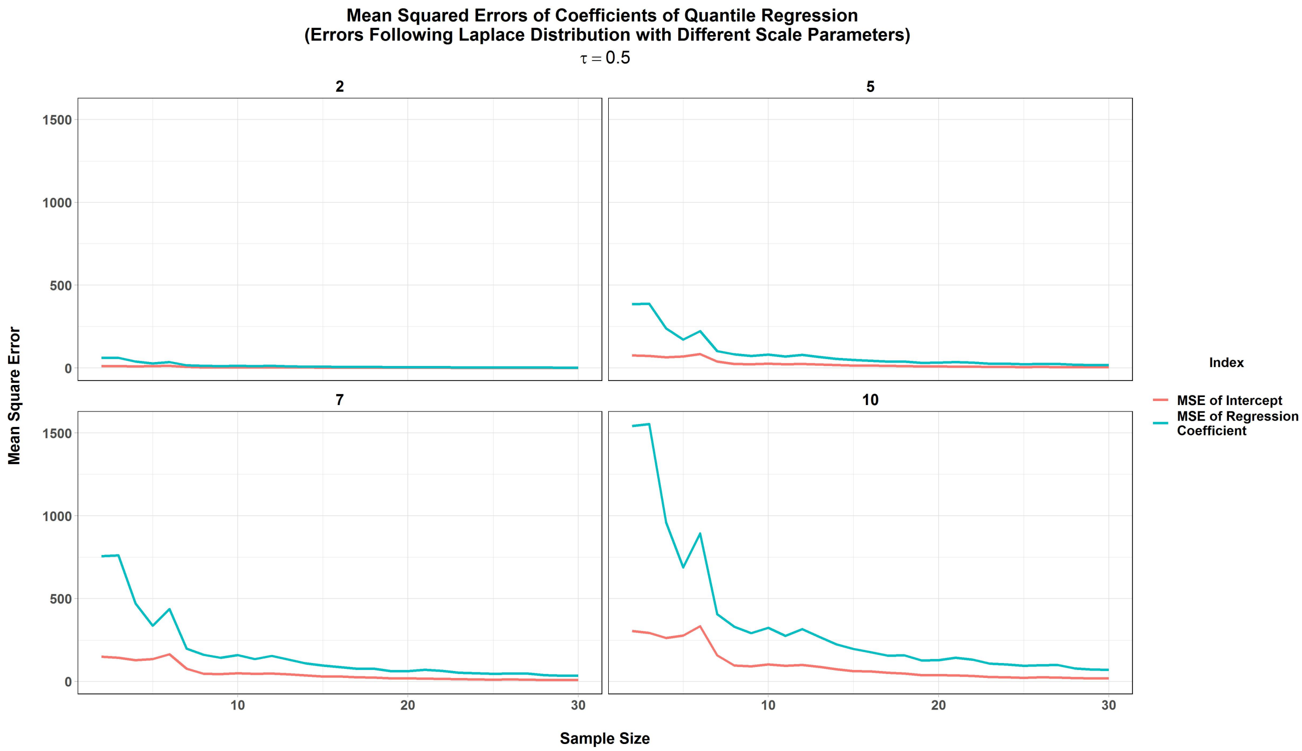




Mean Squared Errors of Coefficients of Quantile Regression
(Errors Following Cauchy Distribution with Different Scale Parameter)
τ = 0.5
2
5
7
10
8e+05
6e+05
4e+05
2e+05
0e+00
Mean Square Error
10
20
30
Sample Size
Index
MSE of Intercept
MSE of Regression Coefficient



Mean Squared Errors of Coefficients of Quantile Regression
(Errors Following t-Distribution with Different Degrees of Freedom)
τ = 0.5
2
5
7
10
40
30
20
10
0
Mean Square Error
10
20
30
Sample Size
Index
MSE of Intercept
MSE of Regression Coefficient



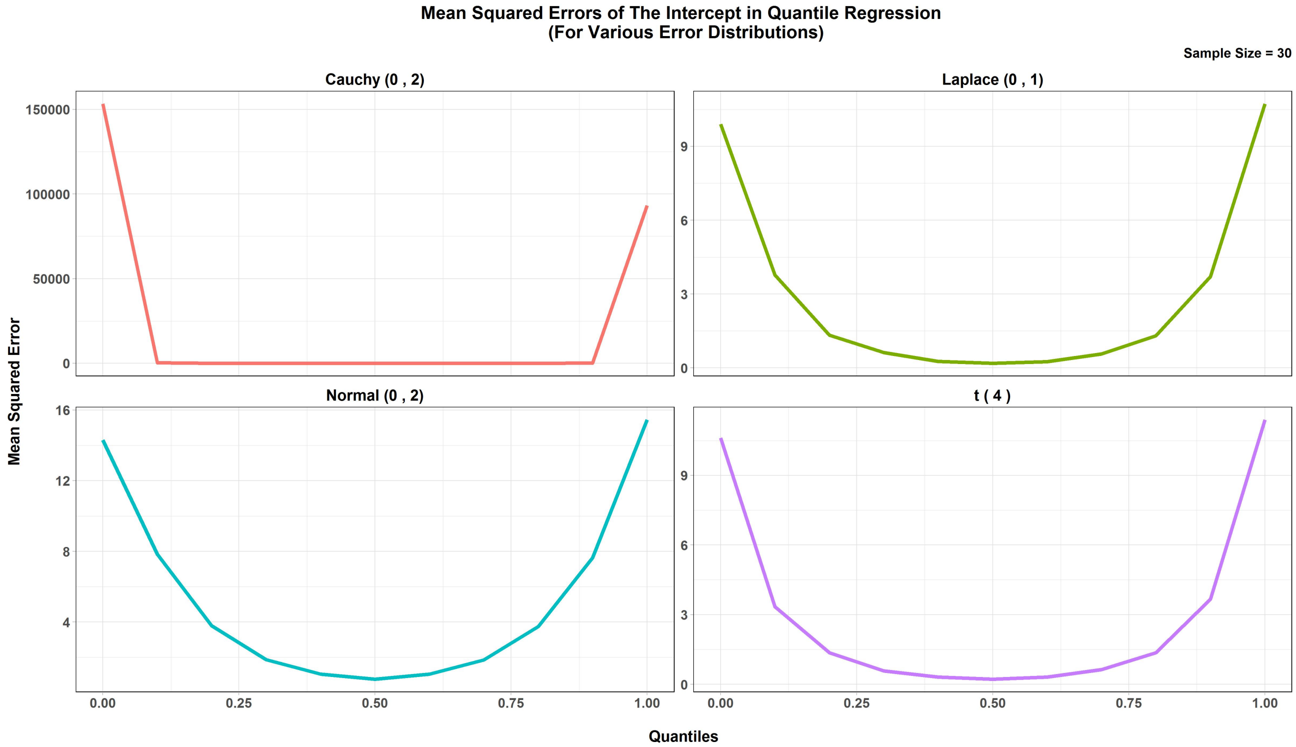
Observations

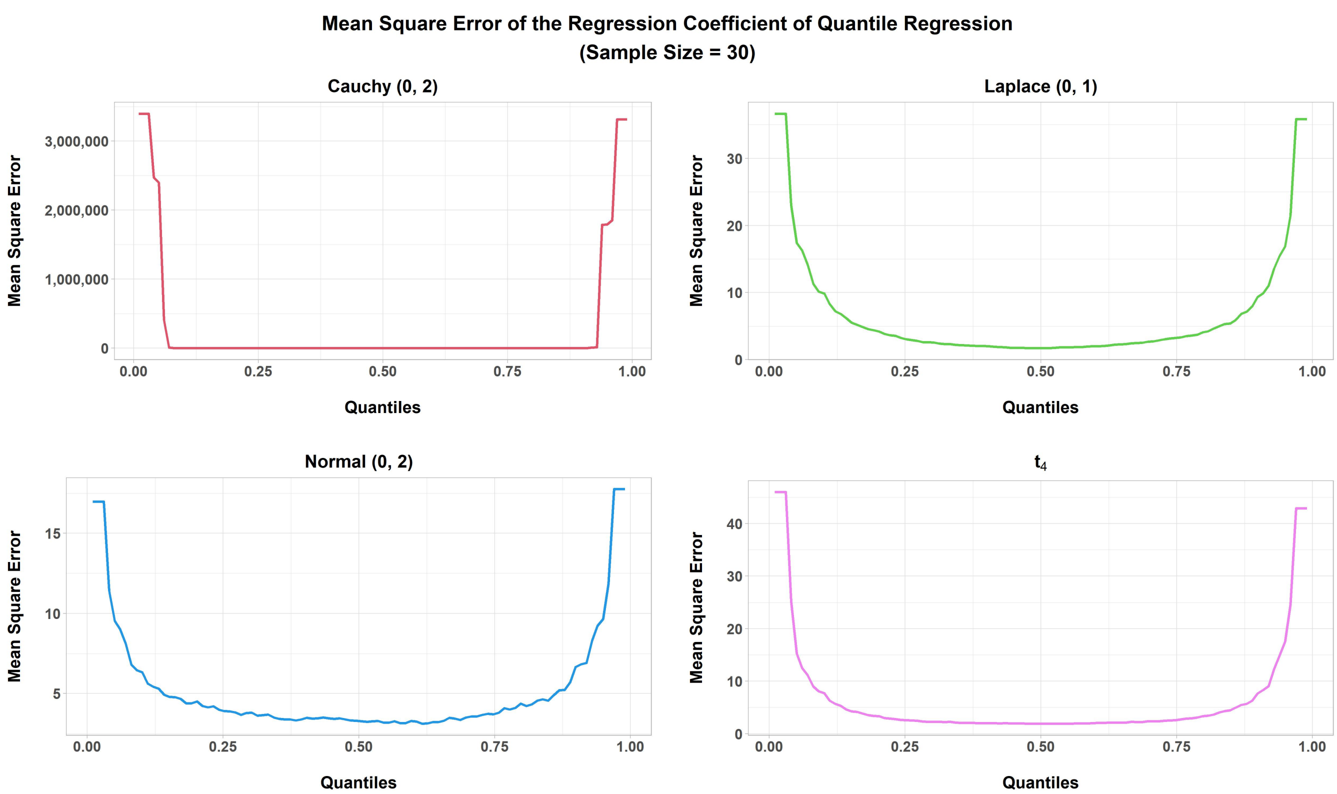
From the above four graphs we observe that -

- MSE of the regression coefficient is far more rapidly decreasing than the MSE of the a term with increase in the sample size.
- With the increase in the variability of the error distribution the MSE also increases.
- In case of errors following Laplace distribution there is an initial decrease in the MSEs up to sample size 5 and again a sudden up-rise at sample size 6, but, fluctuations smoothed out as sample size increases.
- In case of t-distribution, fluctuations in the curves seem to be more for lower degrees of freedom.

MSE and Quantiles

Here, we have plotted MSE values against quantiles, with respect to different error distributions. The sample size is kept fixed at 30 and the scale parameter is 2 for Cauchy and rest of the distributions with variance 2. The predictor is drawn from Beta(3,2) distribution. The diagrams are shown below:





Observations

From the above two graphs we observe that -

- For both a term and regression coefficient, MSE increases as it departs from the central value of the quantile in either side for all distributions.
- For Normal, t and Laplace the increase of MSE in both sides is symmetric whereas for Cauchy it seems to be asymmetric.
- The slope of the curves increases rapidly in case of Normal in either side and the order of the distributions in this aspect is Normal > Laplace > t > Cauchy.

We can draw one general conclusion from the two graphs that, quantile regression provides a 'good' result if the distribution of the errors have a long and thick tail (as in case of Cauchy, the estimates can be said to be 'good' for almost in the full range).

Measurement Error in the Response Variable

Now, suppose there is some error in measurement of the response variable 'y'. To observe the effect of this error, we consider a noise term 'ϵ' which follows some probability distribution independent of the model error 'u'. Now, our model becomes -

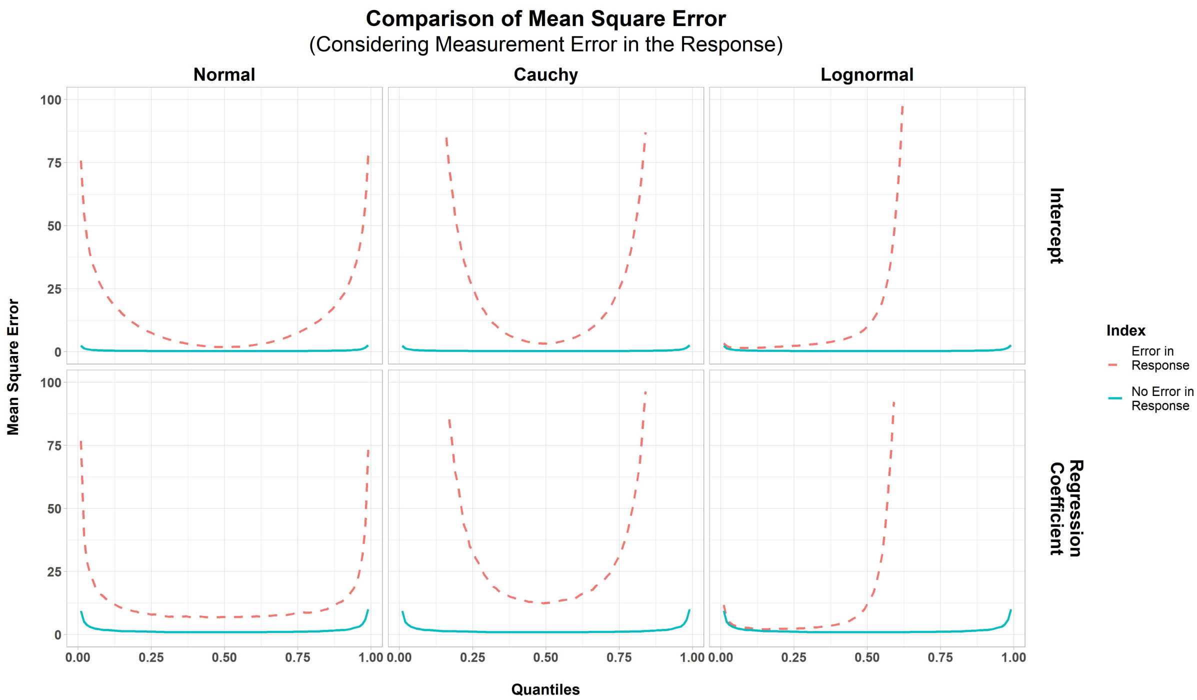
$$y_i + \epsilon_i = \beta_0 + \beta_1 x_i + u_i \tag{3}$$

Here, we have considered the distribution of 'u' as $\mathcal{N}(0,4)$ and the distribution of 'ϵ' as $\mathcal{N}(0,25)$, $\mathcal{C}(0,5)$ and $\Lambda(0,25)$ respectively.

Following the same simulation technique as followed before, we want to compare between the estimates obtained with and without any noise added with the values of 'y'. The results obtained are in the following table:

Table 2: Results

<table>
<tr><th colspan="7">Model: $Y = \beta_0 + \beta_1 x$</th></tr>
<tr><th colspan="4">Sample Size</th><th colspan="3">100</th></tr>
<tr><th colspan="4">Distribution of x (Covariate)</th><th colspan="3">$U(0,1)$</th></tr>
<tr><th colspan="4">Distribution of u (Model Error)</th><th colspan="3">$\mathcal{N}(0,4)$</th></tr>
<tr><th colspan="4">Distribution of ϵ (Noise)</th><th>$\mathcal{N}(0,25)$</th><th>$\mathcal{C}(0,5)$</th><th>$\Lambda(0,25)$</th></tr>
<tr><td rowspan="10">$\tau = 0.05$</td><td rowspan="5">β_0</td><td colspan="2">Actual</td><td colspan="3">-1.29</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>-1.23</td><td>-1.25</td><td>-1.21</td></tr>
<tr><td>MSE</td><td>0.78</td><td>0.71</td><td>0.72</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>-6.79</td><td>-36.62</td><td>-0.39</td></tr>
<tr><td>MSE</td><td>35.57</td><td>3064</td><td>1.71</td></tr>
<tr><td rowspan="5">β_1</td><td colspan="2">Actual</td><td colspan="3">5.00</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>4.97</td><td>4.97</td><td>4.91</td></tr>
<tr><td>MSE</td><td>2.08</td><td>2.16</td><td>2.13</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>5.11</td><td>5.60</td><td>4.94</td></tr>
<tr><td>MSE</td><td>14.44</td><td>4590</td><td>2.51</td></tr>
<tr><td rowspan="10">$\tau = 0.5$</td><td rowspan="5">β_0</td><td colspan="2">Actual</td><td colspan="3">2.00</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>1.98</td><td>2.00</td><td>2.01</td></tr>
<tr><td>MSE</td><td>0.23</td><td>0.22</td><td>0.29</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>1.96</td><td>2.05</td><td>4.68</td></tr>
<tr><td>MSE</td><td>1.58</td><td>3.05</td><td>11.27</td></tr>
<tr><td rowspan="5">β_1</td><td colspan="2">Actual</td><td colspan="3">5.00</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>5.04</td><td>4.99</td><td>4.98</td></tr>
<tr><td>MSE</td><td>0.69</td><td>0.66</td><td>0.88</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>5.05</td><td>4.86</td><td>5.06</td></tr>
<tr><td>MSE</td><td>5.04</td><td>8.87</td><td>10.80</td></tr>
<tr><td rowspan="10">$\tau = 0.95$</td><td rowspan="5">β_0</td><td colspan="2">Actual</td><td colspan="3">5.29</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>5.25</td><td>5.28</td><td>5.32</td></tr>
<tr><td>MSE</td><td>0.80</td><td>0.72</td><td>0.65</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>10.77</td><td>42.41</td><td>9343</td></tr>
<tr><td>MSE</td><td>35.42</td><td>3199</td><td>753016877</td></tr>
<tr><td rowspan="5">β_1</td><td colspan="2">Actual</td><td colspan="3">5.00</td></tr>
<tr><td rowspan="2">No Noise</td><td>Estimate</td><td>5.08</td><td>4.93</td><td>4.93</td></tr>
<tr><td>MSE</td><td>2.48</td><td>2.33</td><td>2.05</td></tr>
<tr><td rowspan="2">Noise</td><td>Estimate</td><td>4.94</td><td>4.53</td><td>212</td></tr>
<tr><td>MSE</td><td>16.30</td><td>4688</td><td>1201727436</td></tr>
</table>



Observations

When 'ϵ' follows Normal Distribution

1. When $\tau = 0.5$, there is not much effect on the estimates of the model parameters and the MSE also does not increase much.

2. When $\tau = 0.05$ or 0.95, the estimates of the model parameters shifts either way and the shift is almost symmetric. Also, the MSE increases to some extent.

When 'ϵ' follows Cauchy Distribution

1. When $\tau = 0.5$, there is not much effect on the estimates of the model parameters and the MSE also does not increase much.

2. When $\tau = 0.05$ or 0.95, there is a huge shift in the estimates of the model parameters in either direction and the shift is almost symmetric. Also, the MSE increases drastically.

When 'ϵ' follows Lognormal Distribution

1. When $\tau = 0.05$, there is not much effect on the estimates of the model parameters and the MSE also does not increase much.

2. When $\tau = 0.5$, there is some increase in the estimates of the model parameters. Also, the MSE increases to some extent.

3. When $\tau = 0.95$, there is a huge increase in the estimates of the model parameters and also, the MSE increases drastically.
   This is happening as the noise-term follows Lognormal distribution which is highly positively-skewed distribution.