Quantile Regression Tutorial and Empirical Analysis

Chen Sizhou, Zhao Xuru

Introduction

*Quantile regression was first proposed by Koenker and Bassett(1978).*When we want to figure out the relationship between two items(Let's call them X and Y in the following discussion.), we do regression. Regression is always applied by 3 steps:

1. Suppose an function of X and Y
2. Determine all unknown parameters by letting function approach training data(the data we know).
3. Measure how near the data is approached by function by MSE.

Therefore, we always get a expectation of y given x. That is to say, when we want to study how much J spends on his coffee everyday for work, after this kind of regression, we know how much we expect him to spend on his coffee.

Since it is always expectation told, we can't answer questions like **"How much would J spend on top coffee?"**, or **"How much would J consume with low workload that day?"**. These are questions about the data itself rather than regression, when we want to know how his spending density changes, we sometimes turn to quantile for help.

Theory Explanation

Now we know what is regression, but before we come to quantile regression, we have to understand **what is quantile**. Suppose we have a continuous distribution of X, and a proportion τ. Then we could find a point with τ of the area of the distribution is left to it. That is so say, this point item is better than τ of its companions.(And in our example, coffee that is bought cheaper than this cup!)

$$p(X \leq X_\tau) = \tau$$

We often take τ as a percentage rate for simplicity(also called percentile).

Then we do quantile regression. Different from regression we talked above, we focus on the best fitting a certain quantile of Y with given X. For example, how education year affects people with top 10 percent income in population. Therefore, the loss function we use changes as well.

How to develop the loss function? Let's first consider a special case of quantile–median. Then we know the loss function is L1 loss:

$$\text{Loss function} = |e|$$

Then, to make it more complex, we want to know about the lowest 10 percent people. We have to weight them more, and weight the other 90 percent less.

Let τ be the quantile percentage, and overestimate or underestimate each part below or under it. When $\tau < .5$, we would overestimate y_i which is less than τ level estimated $\hat{y}$.

$$\text{Loss function} = \sum w_\tau \left| y_i - \widehat{y} \right| = \sum_{i:y_i<\widehat{y}}^{T} (1-\tau) \left| y_i - \widehat{y} \right| + \sum_{t:y_i \geq \widehat{y}}^{T} \tau \left| y_i - \widehat{y} \right|$$
$$= \sum_{i:y_i<\widehat{y}}^{T} (\tau - 1) \left(y_i - \widehat{y} \right) + \sum_{t:y_i \geq \widehat{y}}^{T} \tau \left(y_i - \widehat{y} \right)$$

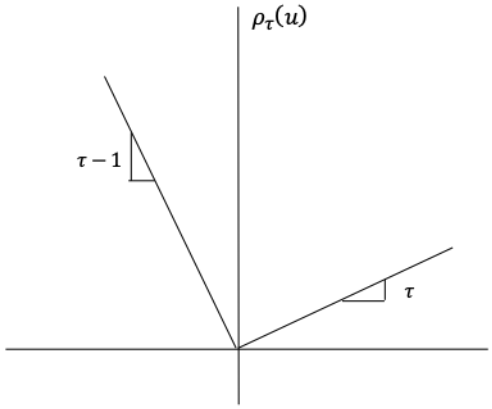


Figure 1: quantile regression

And as the figure shown, at exactly right quantile τ, there's no descent or rising gradient by loss function. Therefore, we could determine the predicted value as follows:

$$\widehat{Q}_Y(\tau) = \text{argmin}_{\xi_\tau \in R} \left(\sum_{i:Y_i \geq \xi_\tau} \tau \left| Y_i - \xi_r \right| + \sum_{i:Y_i<\xi_\tau} (1-\tau) \left| Y_i - \xi_r \right| \right)$$

This is a linear function of parameters.

By giving τ different value, we could derive different curves representing various regression standard with τ of the observed data points of y.

For estimated β, at the τth quantile, per unit change in x would result in β unit change of y.

Since quantile loss fisrt derivative is not continuous at zero point, which may cause unstable calculation. People develop Quantile Huber Loss to smooth it.

$$\text{Loss function} = \mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \leq \kappa \\ \kappa \left(|u| - \frac{1}{2}\kappa \right), & \text{otherwise} \end{cases}$$

Properties and Application

Theoretically, there's no accurate description of any random distribution, therefore, we consider these descriptions as a spacial projection. Quantile regression ensures that the result would be a contraction under Wassestein's distance. That is to say, we could know every transfer of probability density.

When we do quantile regression ,we take advantages of it like:

1. **More robust results.** Large outliers cannot effect the prediction very much.

2. **No assumption on distributions**, which means no parameters that would lead to error term on distribution are assumed. (Fewer hypothesis made before regression)

3. **Shows heterogeneity of the studied group.** Rather than predict the expected value, it gives a description of different segment of the group.

This method is widely adopted when doing policy analysis valuation. For example, it helps the research on birthweight of babies for giving a more complete picture of the low tail that may be ignored by mean square estimation. A more significant covariate effect has been shown.(Abrevaya,2001).

(Space is left for prupose)

Empirical Analysis

Abstract In this case, the subject of our study is the second-hand houses in Xiamen. Through statistical analysis, we will explore the relation between house prices and factors like, floor area, floor, exposition of house, etc. The mainly focused part is the **quantile regression of house price on the floor area** and conclusions actually show that the floor area acts differently with different quantile of house prices. Additionally, we will also show some descriptive analysis on distributions of house prices under other independent variables respectively.

Background and Introduction

House market in Xiamen has been talked over time these years and it is somehow the representative house market in China. In recent years, more and more people from other provinces come to Xiamen then work and live here, and they actually simulate the house market there, especially the second-hand house market. However, the majority of owners often emphasizes too hard on one or some factors on the impact of prices, causing miscalculation of their own housing prices, and results in the large differences between expected prices and real prices.

To solve the problem above, people tend to use linear regression with all possible factors put into the model, **regardless of the varying working mechanism of independent variables in different data sectors**, which is said as *quantile* in this report. Trying to get the information lost back by linear regression model, **linear quantile regression model with a single variable** is used in this report, giving a simplified illustration of how unit price changes with floor area in different data units.

Data Sources and Descriptions

The data we use in this case includes 3203 observations, and is from *Gouxionghui*. It has already been pre-processed. This dataset is collected within 5 years for sure, but the specific time is not clear yet. Variables below can be divided into dependent variable and independent variables. The dependent variable is *Unit.Price*, representing the price per square meter (unit: 10,000 RMB). Among independent variables, only *Area* is a numerical variable, representing the floor area of a house (unit: square meter), and other independent variables are all factorial variables.

```
summary(house)
```

```
##    Unit.Price          Area          Floor       Hall       Room      School
##  Min.   :0.8828   Min.   : 32.0   High  :1156   <2 :1073   <2 :3123   0:1997
##  1st Qu.:3.2594   1st Qu.: 82.0   Low   : 899   >=2:2130   >=2:  80   1:1206
##  Median :4.1947   Median :101.0   Middle:1148
##  Mean   :4.4135   Mean   :111.8
##  3rd Qu.:5.4567   3rd Qu.:130.0
##  Max.   :8.3203   Max.   :375.0
```

```
##        Exposition            Time      Subway      District        City
##  north-south:2469   after 2007 :1544   0:3203   Huli   : 571   Xiamen:3203
##  others     : 443   before 2007:1659            Outside:1645
##  south      : 291                               Siming : 987
##
##
##
```

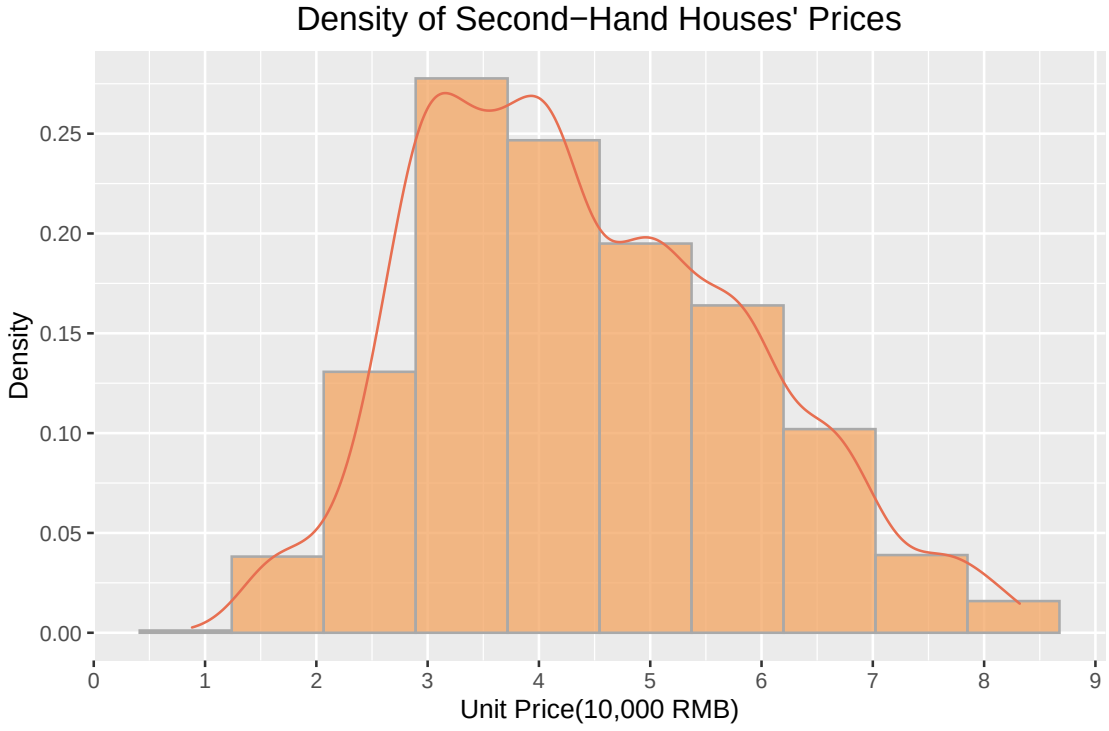
The information of dataset showen above actually involve some useless variables like *Subway*, *Time* and *City*. These variables can not contribute to our analysis so we will ignore them for sure in the part below. Additionally, the **1** under variable *School* represents that the house is close to schools and **Outside** under variable *District* means that the house locates somewhere outside the island, like *Xiangan*.

Descriptive Analysis

Since we only focus on the effect of **Area** on **Unit.price** in the part of regression analysis, we will show some descriptive analysis about unit price under other independent variables in this part and draw some useful conclusions for the subsequent study.

Dependent Variable: Unit Price

In this case, the minimum of the unit price is $8,828 yuan/m^2$ according to the one locates outside the island with $34m^2$, and the maximum of the unit price is $83,203 yuan/m^2$ according to the one locates Huli with $248m^2$. From the histogram(Graph.1) below, the distribution of unit price is **right-skewed**. Specifically, the **average** of unit price is $44,135 yuan/m^2$ and the **median** is $41,947 yuan/m^2$. These figures accord with our cognition to the house market that a few of houses with extremely high prices pull up the average level of house prices. **Generally speaking, there are huge differences among second-hand house prices in Xiamen and more than half of these houses have entered into the epoch of 40,000 yuan.**

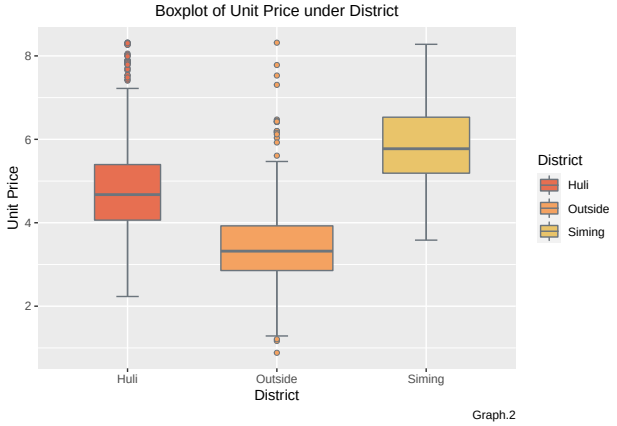


Graph.1

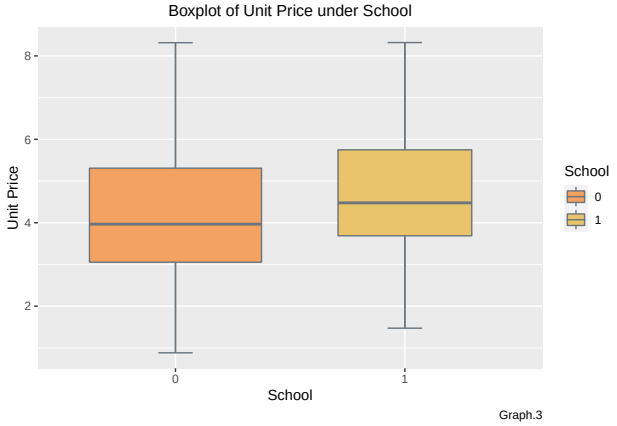
Independent Variable: Exogenous Factors

From Garph.2, it is obvious that differences of unit prices among districts are huge. Price level in Siming district is much more higher than the other two, and price level outside the island is the lowest. These differences can be explained by the distribution of schools and transportations. Combining the width of each box, the second-hand house market outside the island takes the biggest share of the whole market and this is probably because that resident population mainly live in Siming and Huli.

From Graph.3, we can see that houses close to school tend to have a higher unit price level but the upper bounds of them don't differ much. What's more, houses that are close to school take less share of the market. In general, schools and the resource allocation between regions do have an effect on house's unit price level, and this is accord with our expectations.



Graph.2

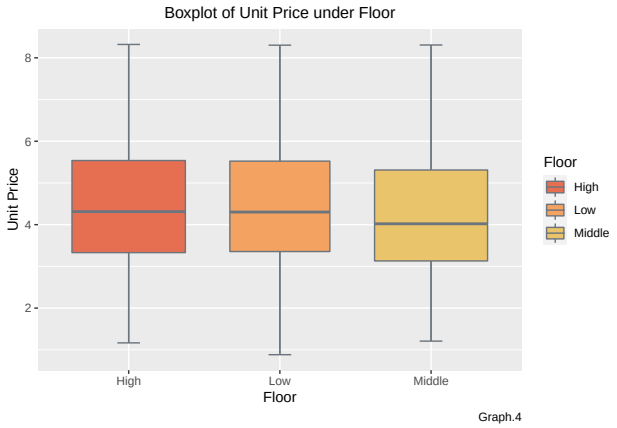


Graph.3

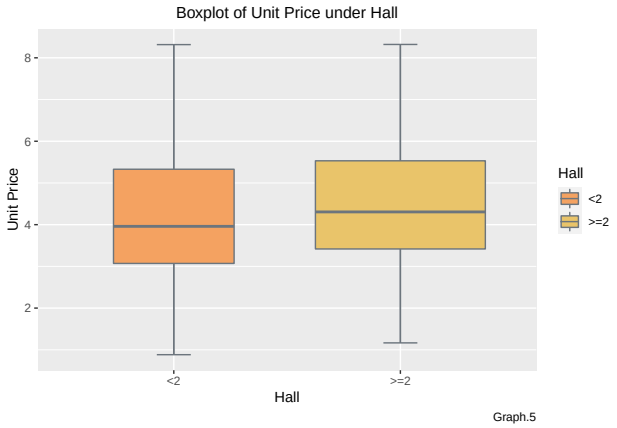
Independent Variable: Endogenous Factors

Now it's turn to see factors belong to houses themselves. We will **first** describe Graph.4 and Graph.6, **then** Graph.5 and Graph.7 and **finally** Graph.8. Graph.4 actually tells that most people don't really care about the floor and it may be caused by the spread of elevators, and somehow, middle floors have a lower unit price level. For market share, houses at low floor take the least market share, and the other two are similar with each other. For Graph.6, it seems that houses with south exposition take the least market share in second-hand house market and have the lowest unit price level. In addition, most second-hand houses have south-north exposition and houses with other expositions have a higher unit price level.

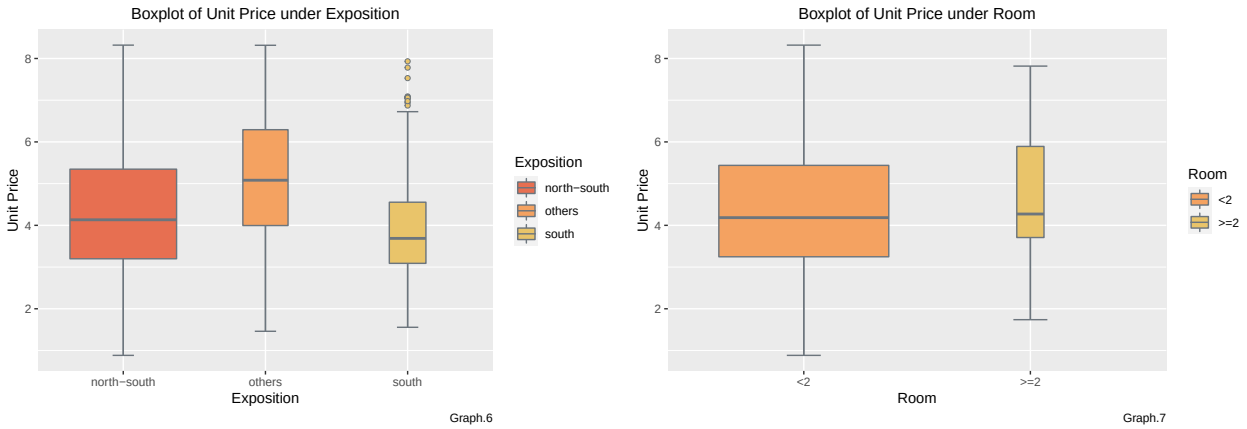
Graph.5 and Graph.7 talk about unit prices under halls and rooms. As we can see, the number of halls and rooms don't cause that much on unit prices compared to other factors. On one hand, people may really don't care about the number of halls and rooms; on the other hand, owners of houses haven't realized that the number of halls and rooms could be an attractive advertisement for the renter. For market share, majority of houses have halls greater than or equal to 2 and rooms less than 2.



Graph.4

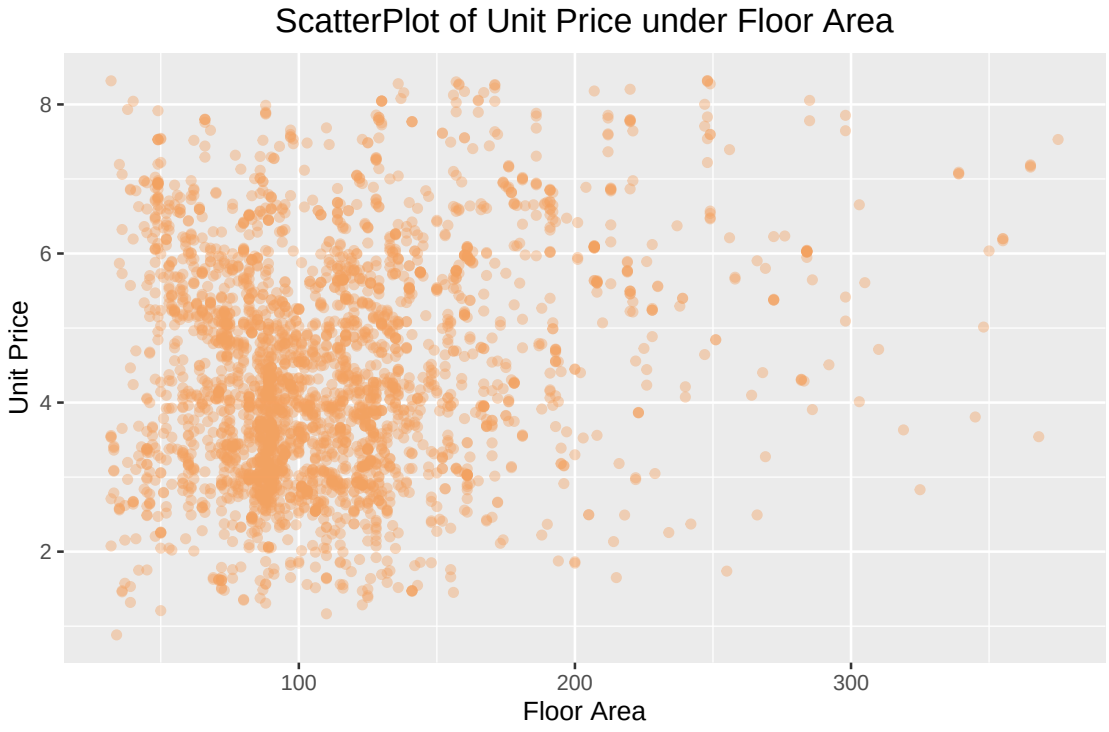


Graph.5



Graph.6 Graph.7

The scatter plot in Graph.8 describes the relationship between floor area(x axis) and unit prices(y axis). Easily to see that the unit prices of houses with the same floor area fluctuate a lot, especially around $100m^2$ and if we only use linear regression model, there must be information loss. **Above all, descriptive analysis shows that exogenous factors can cause much effect on the second-house's unit prices and to explore the effect of floor area, we are going to set up a quantile model in the next part**.



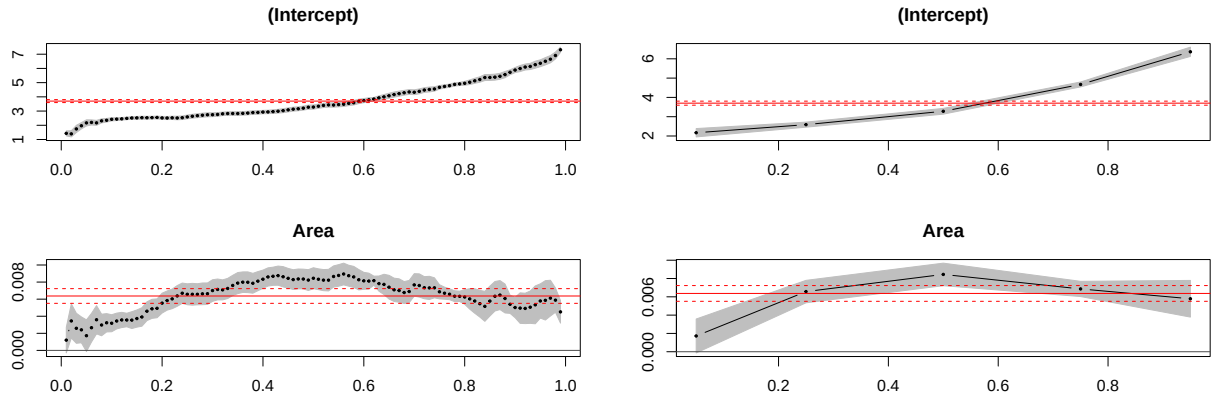
Graph.8

Model Building

To figure out the simplified mechanism of floor area on unit price, here we only focus on one independent variable and use the simplest linear quantile regression. Our discussion below will be divided into 4 parts. **First**, we will compare estimated coefficients at different quantiles; the **second** part is the comparison of fitted curves at different quantiles; the **third** one is about the comparison of unit price's distributions on houses with small floor area and large area.The **last** part will do model comparisons.

Comparison of Coefficient Estimates at Different Quantiles

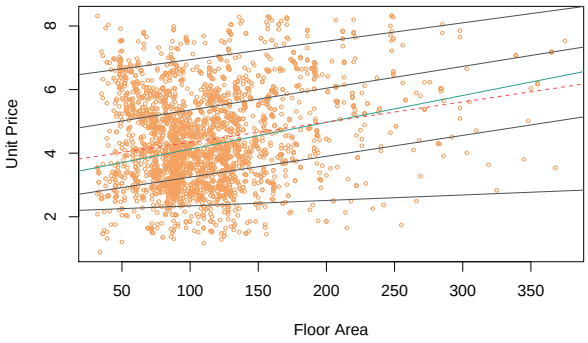
```
fit1 = summary( rq(Unit.Price ~ Area, tau = 1:99/100,data=house) )
fit2 = summary( rq(Unit.Price ~ Area, tau = c(0.05,0.25,0.5,0.75,0.95),data =house) )
plot(fit1)
plot(fit2)
```



Figures above conclude two bunches of quantile regressions at different quantile. In the first bunch, we just simply draw quantiles from 1% to 99% and in the second bunch we pick 5 quantiles to capture the trend of changed effect of floor area on unit price. Combining the two figures, the shape of intercepts is more like a convex curve and the shape of area's coefficients is more like a concave curve, which tells that intercepts grow faster and faster and keep increasing with the increasing quantile and coefficients grow more and more slowly around 0.5 quantile and begin to drop faster and faster when quantile is above 0.5.

More in detail, intercepts and coefficients are all positive. Scale of intercepts is around 1~7 and doesn't have economic meaning in this case. Scale of area's coefficient is around 0.000~0.008, which means that keeping other factors equal, $1m^2$ bigger the house become, 0~80yuan will the unit price increase according to each quantile. **Generally speaking, floor area has the greatest impact on unit price of second-hand houses in Xiamen when unit price is around 41,947yuan and shows a weaker impact when unit price become far from 41,947yuan.**

Comparison of Fitted Curves at Different Quantiles

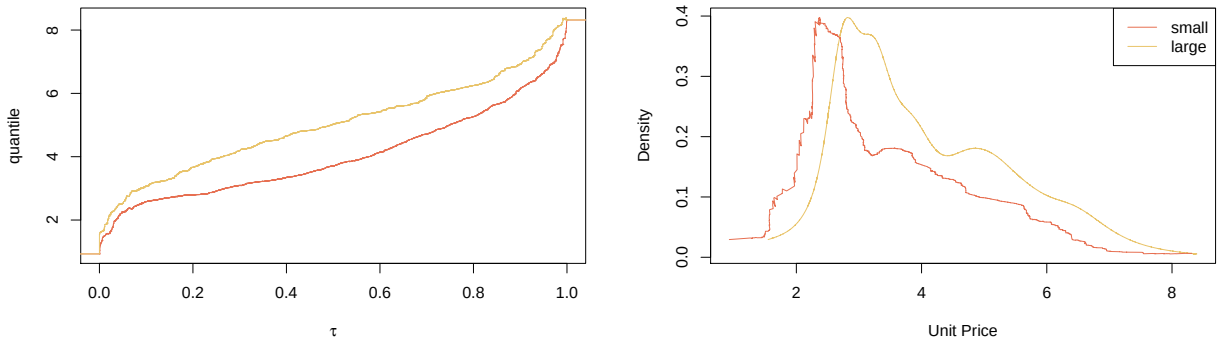


Just as the first part showes, here we draw fitted curves in black at the same quantile as above(0.05,0.25,0.75,0.95), the green curve here is quantile regression at 0.5 and the red curve here is the linear regression model. This figure actually proves that if we use linear regression model to fit the underlying data, we are more likely to lose information. From this figure, we can see that intercepts do grow faster and faster and coefficients increase first and then decrease.

Comparison of Unit price's Distributions on Houses with Small and Large Floor Area

These two figures below show the comparison of unit prices of floor area(Area) at 0.05 quantile and 0.95 quantile respectively. Here we simply define floor area at 0.05 quantile as small house and 0.95 as large house.

Form the left-hand side graph, the gaps between maximum and minimum unit price between small house and large house are similar to each other, but the divergency of unit price between small house and large house is still large at almost each quantile. From the right-hand side graph, unit prices of small and large house are both right-skewed, and small house is more right-skewed compared to large house. **This is to say, no matter the house is small or large, there is always some extremely expensive houses pulling up the average level. And more generally, large houses tend to have higher unit prices in Xiamen second-hand house market.**



Model Comparisions

```
attach(house)
fit_1 = rq(Unit.Price ~ Area, tau = 0.05)
fit_2 = rq(Unit.Price ~ Area, tau = 0.25)
fit_3 = rq(Unit.Price ~ Area, tau = 0.5)
fit_4 = rq(Unit.Price ~ Area, tau = 0.75)
fit_5 = rq(Unit.Price ~ Area, tau = 0.95)
anova(fit_1,fit_2,fit_3,fit_4,fit_5)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: Unit.Price ~ Area
## Joint Test of Equality of Slopes: tau in {  0.05 0.25 0.5 0.75 0.95  }
##
##   Df Resid Df F value   Pr(>F)
## 1  4    16011  8.2644 1.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(house)
```

Purpose in this part is to explore whether the working mechanism of independent variable changes over different quantiles and whether those variations are significant. One more time we choose quantiles showen in code above and the `anova()` here tells us that the variation of floor area's working mechanism on unit price is significant, and we can say that **floor area of second-hand house does have different effect at different quantile of unit price**.

Conclusions and Development

In this section, we will conclude our analysis and show both advantages and disadvantages of our **model and data processing** .

Case Conclusion

Actually, the idea of this report comes from the linear model analysis done by others and we observe that the dataset can give us more information by using quantile regression model. According to analysis above, it is true that linear model fitted will lose much information and we did see the strength of quantile regression model in this case. The result above can give second-hand house sellers a new thought to evaluate their asset in Xiamen and somehow trade more efficiently.

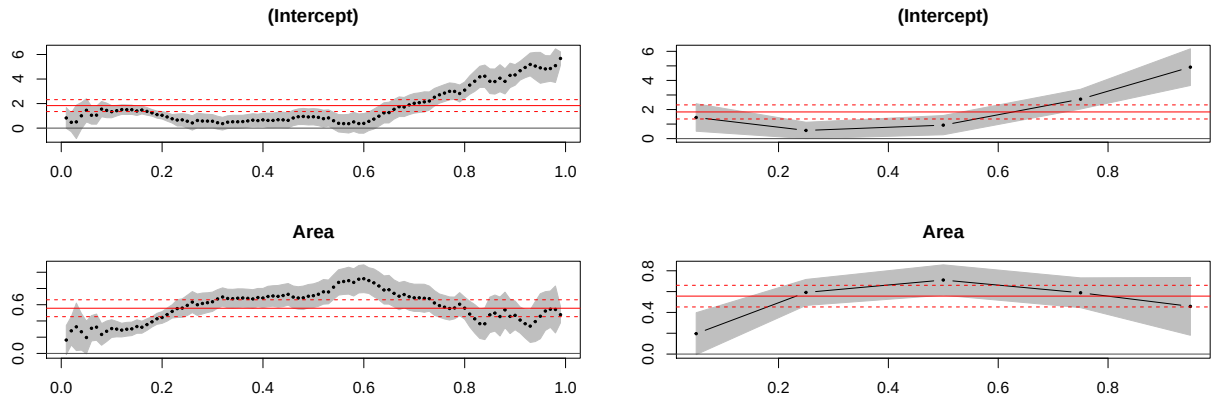
Model Evaluation

Model used in this report is an extremely poor one actually. It seems that we almost ignore all the other independent variables. In fact, we do regressions putting all the independent variables into linear quantile regression model. Results show that *Area* here is the only variable worthy talking and it can be an good example for us to present. Moreover, the existence of other independent variables actually doesn't affect that much on the working mechanism of *Area*. For further development, maybe we can use nonlinear quantile regression model next time since we still don't figure out its theory clearly.

Data Processing Evaluation

The coefficient of *Area* is small in this report and that is because the unit of our variables. The unit of unit price here is 10,000RMB but the unit of floor area is square meter. The reason we didn't take a `log()` or `scale()` is that we want to reserve the basic economic meaning of data and coefficient. Now we will give it a try to see if we take a log of our independent variable. **In fact,unlike OLS regression, when doing quantile regression, the predicted value won't deviate due to pretreatment like log or scale. And we wake advantages of this property.**

```
house$Area <- log(house$Area)
fit3 <- summary( rq(Unit.Price ~ Area, tau = 1:99/100,data=house) )
fit4 <- summary( rq(Unit.Price ~ Area, tau = c(0.05,0.25,0.5,0.75,0.95),data =house) )
plot(fit3)
plot(fit4)
```



Trends and results don't change much actually and only value itself become larger.

References

1. Nguyen, Quantile Regression
2. Koenker, R. And K. Hallock. (2001) "Quantile Regression," Journal of Economic Perspectives, 15(4).
3. Koenker, R. (2017) "Quantile Regression 40 Years On," Annual Review of Economics, 9.
4. https://www.thoughtco.com/what-is-a-quantile-3126239
5. https://zhuanlan.zhihu.com/p/60912847