Stats 790 project draft

Gujie Fu 400067957

April 2023

1 Introduction

Quantile regression[KD87] is a technique used to analyze the relationship between a set of predictor variables and specific percentiles (or quantiles) of the response variable. Unlike the usual linear regression, which focuses on predicting the mean of the response variable, quantile regression aims to estimate the conditional quantiles, providing a more comprehensive view of the data distribution. This method is particularly useful when dealing with heterogeneous data, as it allows for capturing different patterns or trends across various quantiles. Quantile regression is robust against outliers and provides insights into the behavior of the response variable at different levels of the distribution, making it a valuable tool for understanding complex relationships in diverse fields such as finance, economics, and environmental sciences. In this project, I will pick up a data set and apply quantile regression to discover some properties and draw conclusions

2 Methodology

Let's have a look at quantile first. The quantile q_τ where $\tau \in (0, 1)$ of a random variable y is defined by

$$P(y \leq q_\tau) \leq \tau or P(y \leq q_\tau) \geq 1 - \tau$$

However, based on the data we observe, assuming there are n observations, we can estimate the empirical quantile by

$$\frac{1}{n}\sum_{i=i}^{n} I(y_i \leq \hat{q}_\tau) \leq \tau or \frac{1}{n}\sum_{i=1}^{n} I(y_i \geq \hat{q}_\tau) \geq 1 - \tau$$

where I is the indicator function. A common loss function to be minimized is defined as

$$\hat{q}_\tau = \underset{q}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i, q) * abs(y_i - q) where$$

$$w_\tau(y_i, q) = \begin{cases} 1-\tau & y_i < \tau \\ 0 & y_i = \tau \\ \tau & y_i > \tau \end{cases}$$

which is a weighted sum of absolute deviations. It is a L_1-norm quantile regression loss function which is called check function of Koenker and Bassett (1978).[LZ12] Instead of one quantile q_τ, we can natually replace the quantile q_τ by β_τ with

$$\hat{\beta}_\tau = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau}(\beta)) * abs(y_i - \eta - i\tau(\beta)) where \eta_{i\tau}(\beta) = \beta^T x_i$$

$$w_\tau(y_i, \eta_{i\tau}(\beta)) = \begin{cases} 1-\tau & y_i < \eta_{i\tau}(\beta) \\ 0 & y_i = \eta_{i\tau}(\beta) \\ \tau & y_i > \eta_{i\tau}(\beta) \end{cases}$$

Since the check function is not differentiable at 0, we can't find a solution similar as in the linear regression problem. Thus we will use a check function which is smooth and differentiable.[Zhe11] The check function is

$$S_{\tau,\alpha}(u) = \tau u + \alpha log(1 + e^{-\frac{u}{\alpha}})$$

The following figure is a comparison of two different check function. The red curve is more smooth than the usual check function around 0.

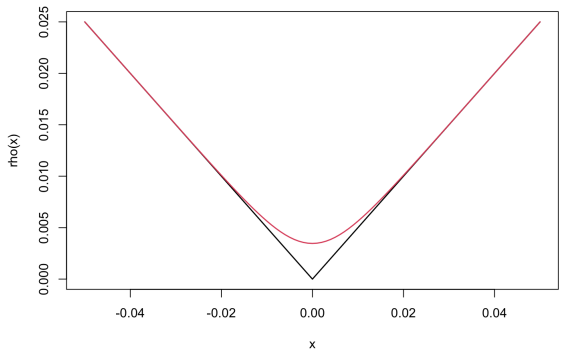

Figure 1: Caption

3 Illustration

The data I use is a build in data in R which is called "age.income" which has 205 pairs observations on Canadian workers from a 1971 Canadian Census Public Use Tape. The following is a scattor plot of the dataset. From the figure we can see that when age is less than 30 years, the income increases very quickly. When age is larger than 30years, it seems the mean of income does not change a

lot but these data points definitely capture more variance, especially when age is larger than 50 years. Hence, we can not assume constant variance and this data shows heteroskedasticity.

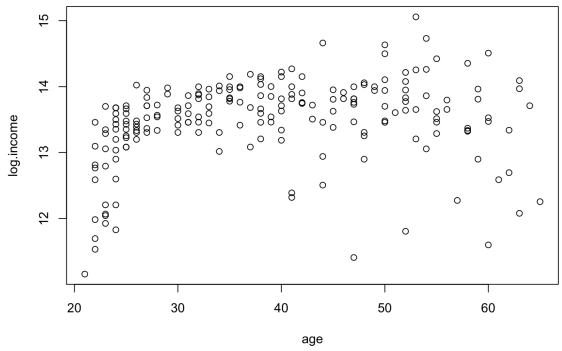

Figure 2: Caption

The following figure contains three quantile regression curve with respect to different quantiles. If we use the most naive quantile regression, then we would expect a straight line instead of a curve, however, in this project I use different degrees of polynomial to fit the response variable. Intuitively, these quantile regression curves are reasonable. The income increases for all three quantiles and for the median which is 50% quantile, income is stable. For the lower 5% and upper 95% quantile, the behavior of income are in opposite direction but eventually both of them decrease.

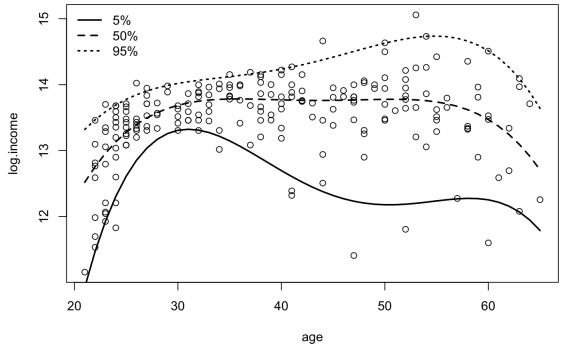

Figure 3: Quantile regression curve of three different quantiles

There is a package in R studio called "quantreg" which can be used to perfrom quantile regression. Thus we can compare the results from my method with the package. The processing time of my method is much longer than the quantile regression package which almost take 6 to seven times longer. My personal guess is that the optimizer takes more iterations to converge. However,

the output of my method is very close to the package. The following two figures are comparison of two methods.

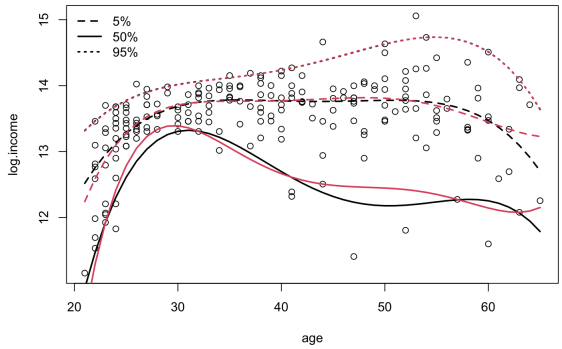

Figure 4: Comparison of two methods (polynomial with degree 4)

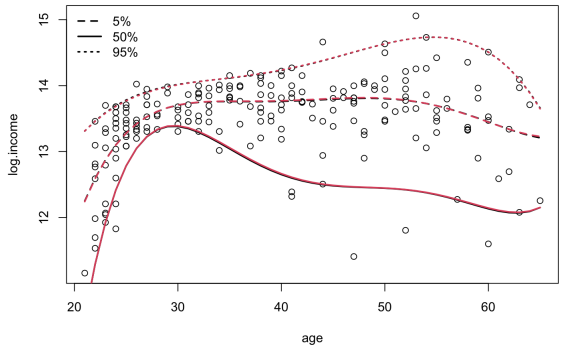

Figure 5: Comparison of two methods (polynomial with degree 4)

From the figure we can see that if we fit the quantile regression model with polynomial with degree 4, then the curve of all three quantiles are very close to the curve from the package. If we use polynomial with degree 5, then there is almost no difference between them. However, the adjusted R^2 of both method is low, which is approximately 32%. The overall performance is not so exciting.

4 Summary

In this project, we introduce quantile regression and use a smooth check function to build it from scratch. The result is graphically satisfying but not numerically both from goodness of fit and processing time. However, it still inspired me to think about data from non normal distribution, or heterogeneous data.

References

[KD87] Roger W Koenker and Valéria D'Orey. "Algorithm AS 229: Computing Regression Quantiles". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3 (1987), pp. 383–393. DOI: 10.2307/2347802.

[Zhe11] Songfeng Zheng. *Gradient descent algorithms for quantile regression with smooth approximation - International Journal of Machine Learning and Cybernetics*. July 2011. DOI: 10.1007/s13042-011-0031-2. URL: https://link.springer.com/article/10.1007/s13042-011-0031-2.

[LZ12] Youjuan Li and Ji Zhu. "L1-Norm Quantile Regression". In: *Journal of Computational and Graphical Statistics* 17 (Jan. 2012). DOI: 10.1198/106186008X289155.