Bootstrapped Quantile Regression Analysis on Small Samples in Applied Linguistic Research

Trevor Atkins

June 6, 2021

1 Introduction and Motivation

In this report, we are going to use the statistical method of Bootstrapped Quantile Regression (BQR). This method could provide ways to deal with the flaws caused by research data that has small sample sizes. Quantile Regression estimates the conditional median (or other quantiles) of the response variable instead of the mean like in the method of least squares and usually employed when conditions of linear regression are not met.

The dataset used for this experiment has a large sample size of 229 undergraduate students, "Undergraduate students' motivation to learn and attitudes towards English in multilingual Pakistan: A look at shifts in English as a world language (System)", however for the purposes of the report we will randomly sample a small group of those students that do not follow an assumption or condition of standard linear regression/method of least squares. (Nikitina et al., 2019) For this task, we will be recreating a similar analysis on the relationship between students' attitudes toward a target language country and their motivation as done in the paper by Nikitina (Nikitina et al., 2019) with exploratory models. The findings could provide further motivation for research in applied linguistics concerning second-language acquisition. As the smaller experiment had less variables or factors, the challenge in this task will be to find a sample that violates standard linear regression and what the best model is before considering the BQR analysis.

The research question will be: can BQR provide adequate results on small applied linguistic dataset samples that violate standard linear regression constraints?

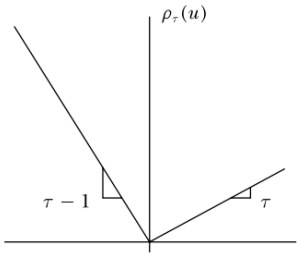
2 Background

2.1 Theory

Quantile Regression (QR) has become more widespread than in the past due to the advancement in computation technology since median regression computations become more intensive with larger datasets compared to the least squares method. QR has often been used in ecology (Cade and Noon, 2003). Bootstrapped quantile regression has even has been used for Covid-19 predictability in the USA (Mavragani and Gkillas, 2020), so it is just as applicable in linguistics and other social sciences. As stated by (Koenker and Hallock, 2001), "in classical linear regression, we also abandon the idea of estimating separate means for grouped data...and we assume that these means fall on a line or some linear surface, and we estimate instead the parameters of this linear model. Least squares estimation provides a convenient method of estimating such conditional mean models. Quantile regression provides an equally convenient method for estimating models for conditional quantile functions."(p.144-145) A quantile is synonymous with percentile, as they are cut points that divide the range of a probability distribution into continuous intervals with equal probabilities or divide observations in a sample in the same way. Koenker explains that quantiles can be defined through a simple alternative expedient as an optimization problem. The median can be defined as the solution to the problem of minimizing a sum of absolute residuals. (Koenker and Hallock, 2001) "The symmetry of the piece-wise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there are the same number of observations above and below the median." (Koenker and Hallock, 2001) In other words, the sample median is taken as an estimator of the population median m,a quantity which splits the distribution into two halves where, if a random variable Y can be measured on the population, then $(P(Y \leq m) = P(Y \geq m) = 0.5$. In particular,for a continuous random variable, m is a solution to the equation $F(m) = 0.5$, where $F(y) = P(Y \leq y)$ is the cumulative distribution function. (Yu, Lu, Stander, 2003)

To generalize the function with a more concrete example like the distribution of growth in children (with upper and lower quartiles for the conditional distribution of heights Y given age X) the quantiles depend on the value of the covariate X and can be found by solving $F(y|x)/ = p$, where $F(y|x) = P(Y \leq y|X = x)$ (Yu, Lu, Stander, 2003) By modelling the relationship between x and the conditional quantiles of y instead of only the conditional mean of y, quantile regression can provide a more comprehensive picture of the effect of independent variables on the dependent variable. The other half of QR being regression - a method to calculate the relationship between a response variable and some covariates. (Yu, Lu, Stander, 2003). Visualizing the relationship between salary and years of a profession for example can be more accurately represented with a quantile regression curve corresponding to a range of values of p because a standard regression fit model only calculates the average relationship with salary and years of profession. (Yu, Lu, Stander, 2003) A significant drawback in QR in general is that parameters are more difficult to estimate than in Gaussian or standard linear regression.(Waldmann, 2018) The inference is hard since the estimators for the coefficients are not available in closed form. (Waldmann, 2018) Closed form means that the mathematical expression has a finite number of standard operations - usually indicating that there is no limit, differentiation, or integration. There are however, various ways to estimate quantile regression parameters such as the conditional median function: $\min_{\beta \in \setminus p} \sum \rho_\tau(y_i - \xi(x_i, \beta))$, where the $\rho_\tau()$ function is the tilted absolute value function appearing in Figure 1 that provides the τth sample quantile as the solution. (Koenker and Hallock, 2001) Another shortcoming of the QR is that if the dependent variable is not continuous or

Figure 1: Example of Quantile Regression ρ function (Koenker and Hallock, K.F., 2001)



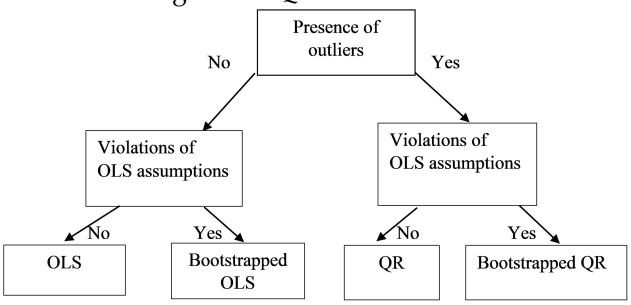
has too many repeated values, it will not be as effective as it relies on sufficient data in the tails of the

distribution and is more computationally intensive than standard linear regression.

2.2 Relevant Literature

Previous work on BQR includes the work done by Nikitina, Paidi, and Furuoka in 2019, where they discuss the methodology of BQR in applied linguistics. They developed a decision tree for when to use BQR with the contents of the data that can be summarized in Figure 2 below, where OLS stands for Ordinary Least Squares and QR is Quantile Regression. (Nikitina, Paidi, Furuoka, 2019) They used this BQR method

Figure 2: BQR Decision Tree



to compare to the OLS method and were able to argue that the BQR can be used to combat the issues of violation of test assumptions and presence of outliers in a dataset of a small sample size. (Nikitina, Paidi, Furuoka, 2019) Their experiment involved determining the relationship between 27 Malaysian public university students' language learning motivation, images or stereotypes about Japan and their attitudes toward the target language country and its people. Another advantage they found with BQR over OLS is that it helps to avoid committing Type I errors - particularly helpful in applied linguistics research "that relies on traditional parametric statistics tends to have a high rate of studies with Type I error..." (Nikitina, Paidi, Furuoka, 2019). Nikitina, Furuoka, and Kamaruddin then examined 19 Korean language learners at a Malaysian university relationship between motivation and attitude with BQR.

3 Materials and Methodology

The dataset used in this report is a random subset of the ”Data for: Undergraduate Students' Motivation to Learn and Attitudes toward English in Multilingual Pakistan: A look at shifts in English as a World Language”. (Rasool and Winke, 2019) The original dataset contains responses made by 229 undergraduate students' from three public universities in the capital of Balochistan province of Pakistan. These responses are collected via a 54-item questionnaire pertaining to learning languages and the motivations or attitudes toward English. Other information included information about the participant such as age, gender, and native language. A random subset has been chosen with a set seed because biases were to be avoided as much as possible. However, out of these random subsets one that violated standard linear regression assumptions and one that did not violate (was power transformed to fix the violation) the assumptions were chosen. The experiment is to test whether there are any reasonable predictors via exploratory models for answering the numerical, ordinal response to ”Item16” - ”I would like to study English even if I were not required.” The numerical, ordinal representation of the responses of the ”Items” that were chosen are from 1 - ”Strongly Disagree” to 7 - ”Strongly Agree”. ”Item16” was chosen since it is a response to an explicit question to the level of motivation and attitude the participant has to learn English voluntarily. In terms of methodology/theory this experiment will test whether there is a difference between using BQR on data that violated or data that followed standard linear regression assumptions as well as why and how to deal with the difference if there is one. The data was preprocessed since the factor of gender was represented by numbers.For the BQR portion of the analysis was produced through the replication of the (Nikitina, Paidi, Furuoka, 2019) paper - the essential R packages being ”lm”, ”boot”, and ”quantreg”. For the purposes of this paper, not all the factors or predictors that were available from the dataset were used. The predictors that were used include: Age, Gender, Mother Tongue, ”Item3” - ”English is the most important language in the world”, ”Item7” - ”Learning English is one of the most important aspects in my life”, ”Item8” - ”I would like to know more about people from English speaking countries”, ”Item48” - ”I want to use English to communicate with people in other countries”, ”Item49” - ”My friends encourage me to learn English”, ”Item52” - ”I wish I could speak English fluently”, ”Item54” - ”I like the people who live in English speaking countries”. The best model was determined by comparing models with increasing complexity via their

Akaike Information Criterion (AIC) - of course, the outcome of the best model differed on the sample. For both of the sample sizes, $N = 36$ and each sample had the gender split as: 21 males and 15 females. The sample that follows the standard linear regression assumptions had the dependent variable power transformed ($\lambda = 1.75$, determined using the BoxCox profile-likelihood procedure) to pass the assumptions of homoscedacity of variance in residuals and residuals being normally distributed. The visualizations of the linear regressions are shown below

Figure 3: Non-violated Model Visual Linear Regression

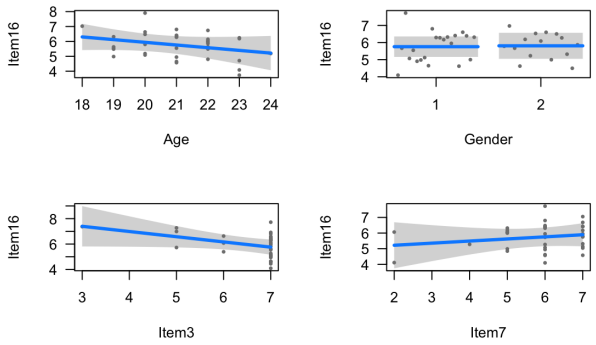


Figure 4: Non-violated Model Visual Linear Regression

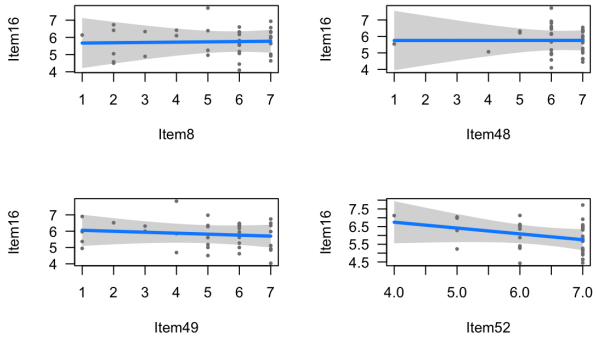


Figure 5: Non-violated Model Visual Linear Regression

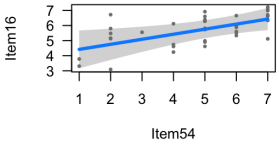


Figure 6: Violated Model Visual Linear Regression

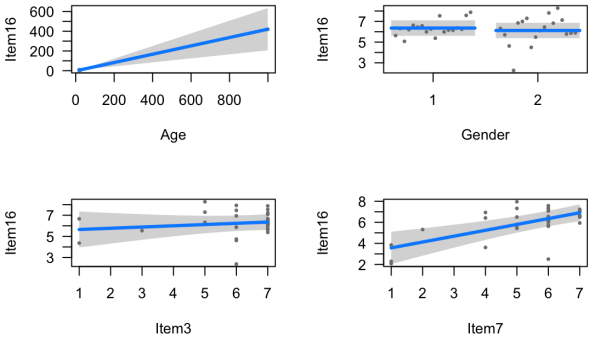
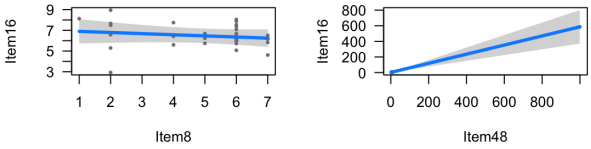


Figure 7: Violated Model Visual Linear Regression



4 Results

The model that was transformed to follow the standard linear regression assumptions produced a result of the OLS summary as shown in Figure 8.

Figure 8: Non-violated Model OLS Summary Result

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.18787   27.89523   2.158   0.0404 *
Age         -1.01708    0.84886  -1.198   0.2417
Gender2      0.05651    2.32303   0.024   0.9808
Item3       -2.46764    1.30539  -1.890   0.0699 .
Item7        0.74976    1.09606   0.684   0.5000
Item8        0.16754    0.82515   0.203   0.8407
Item48       0.22137    0.96657   0.229   0.8206
Item49      -0.34235    0.61106  -0.560   0.5801
Item52      -2.02678    1.41954  -1.428   0.1653
Item54       2.00191    0.77838   2.572   0.0162 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.005 on 26 degrees of freedom
Multiple R-squared:  0.4282,    Adjusted R-squared:  0.2303
F-statistic: 2.163 on 9 and 26 DF,  p-value: 0.06015
```

The summary from Figure 8 shows that the relationship between the motivation and attitudes toward learning English are not significantly influenced with one another since the p-value is greater than 0.05. The Confidence Intervals (CI) have also been calculated, where none of the CI's contained zero, indicating that the null hypothesis of no statistically significant relationship between the predictors and the dependent variable "I would like to study English even if I were not required" could not be rejected.

The result of the model not violating standard linear regression produced a QR summary shown in Figure 9.

Figure 9: Non-violated Model QR Summary Result

```
tau: [1] 0.5

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 31.73732 29.52183    1.07505  0.29223
Age          0.20938  0.89836    0.23307  0.81753
Gender2     -1.97082  2.45849   -0.80164  0.43003
Item3       -1.72891  1.38150   -1.25147  0.22191
Item7        0.22903  1.15998    0.19744  0.84502
Item8       -0.06945  0.87326   -0.07953  0.93722
Item48       1.00049  1.02293    0.97806  0.33706
Item49       0.01728  0.64669    0.02672  0.97889
Item52      -2.66503  1.50232   -1.77395  0.08779
Item54       2.22899  0.82376    2.70586  0.01187
```

Figure 9 indicates that the majority of the predictors have p values > 0.05, except for Item54 - "I like the people who live in English speaking countries". This may be due to the particular random sample that was pulled having more responses skewed toward agreeing to this response or the tendency for the responses to be positive towards learning the English language in the general dataset as found through multiple regression done by (Rasool and Winke, 2019).

With the model that did violate the standard linear regression assumptions produced a result of the OLS summary shown in in Figure 10.

Figure 10: Violated Model OLS Summary Result

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.16964    1.97999  -4.631 7.06e-05 ***
Age          0.41224    0.10288   4.007 0.000393 ***
Gender2     -0.66994    0.40728  -1.645 0.110785
Item3        0.08671    0.13461   0.644 0.524542
Item7        0.55678    0.12776   4.358 0.000150 ***
Item8       -0.11799    0.11103  -1.063 0.296678
Item48       0.59509    0.10141   5.868 2.29e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.147 on 29 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.215e+05 on 6 and 29 DF,  p-value: < 2.2e-16
```

Figure 10 shows that the p value is significant as it is < 0.05. However, it does not express a valid significance since the assumptions are violated. It is also surprising to see that the Adjusted R-Squared to

be at a value of 1, showing that the model fits the data. The result of the model violating standard linear regression produced a QR summary shown in Figure 11.

Figure 11: Violated Model QR Summary Result

```
tau: [1] 0.5

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) -4.81568  1.42306   -3.38404  0.00206
Age          0.28679  0.07394    3.87847  0.00056
Gender2     -0.38895  0.29272   -1.32876  0.19429
Item3       -0.06352  0.09675   -0.65656  0.51664
Item7        0.38113  0.09182    4.15067  0.00027
Item8       -0.24932  0.07980   -3.12443  0.00402
Item48       0.71833  0.07289    9.85532  0.00000
```

Figure 11 shows that there are a lot more predictors indicated as significant than the previous model. The two predictors that do not have significant p values are Gender and Item3 - "English is the most important language in the world". It is surprising to see that Item48 - "I want to use English to communicate with people in other countries" is a zero, indicating that the null hypothesis should be rejected for that predictor.

5 Evaluation and Analysis

There were a lot of limitations to this experiment and comparison. Especially with the model that was power transformed to be able to follow the assumptions of the standard linear regression - the results may have further been misleading due to that transformation. Another limitation was that related experiments used Cronbach's Alpha as a metric to determine the internal consistency and reliability of the sample of responses chosen as well as detecting and removing any outliers or noise in the data. Unfortunately, unlike the research that this paper was trying to replicate the BQR method and BOLS method did not reaffirm each other's findings in both models. Comparing between OLS and BQR with small sample sizes that either violated or transformed to not violate the standard linear regression assumptions have indicated that BQR may be a more accurate representation of small sample populations of a larger dataset. This can be particularly emphasized by the results of the adjusted R-squared value, effect size, and sum of squared

error (SSE) calculations. The OLS model had an adjusted R-squared of 0.23 and the BQR model had an adjusted R-squared of 0.99. This shows that the BQR method can more accurately predict the population according to median distribution than the mean distribution. The partial eta-squared values (effect size) for individual predictors are also higher for more predictors on the BQR model than in the OLS model. The SSE for OLS was 937.4 and the SSE for BQR was 38.2, showing that the BQR model is more of a tight fit to the data than the OLS model. Future work should address the issues of whether OLS or BQR provides a better method to examine relationships with interactions in small sample sizes and compare the results with datasets of reliable sizes.

References

[1] Brian S Cade and Barry R Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003.

[2] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[3] Amaryllis Mavragani and Konstantinos Gkillas. Covid-19 predictability in the united states using google trends time series. *Scientific reports*, 10(1):1–12, 2020.

[4] Larisa Nikitina, Fumitaka Furuoka, and Nurliana Kamaruddin. Language attitudes and l2 motivation of korean language learners in malaysia. *Journal of Language and Education*, 6(2):132–146, 2020.

[5] Larisa Nikitina, Rohayati Paidi, and Fumitaka Furuoka. Using bootstrapped quantile regression analysis for small sample research in applied linguistics: Some methodological considerations. *PloS one*, 14(1):e0210668, 2019.

[6] Ghulam Rasool and Paula Winke. Undergraduate students' motivation to learn and attitudes towards english in multilingual pakistan: A look at shifts in english as a world language. *System*, 82:50–62, 2019.

[7] Elisabeth Waldmann. Quantile regression: a short story on how and why. *Statistical Modelling*, 18(3-4):203–218, 2018.

[8] Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350, 2003.