Journal of Data Science,18(1). P. 148 – 160,2019
DOI:10.6339/JDS.202001_18(1).0008

QUANTIFYING DISEASE SEVERITY OF CYSTIC FIBROSIS USING QUANTILE REGRESSION METHODS

Kameryn Denaro[1], Barbara A. Bailey[2], Douglas J. Conrad[3]

[1.] Teaching and Learning Research Center, University of California, Irvine
[2.] Department of Mathematics and Statistics, San Diego State University [3.] Department of Medicine, University of California, San Diego

Abstract

This article presents a classification of disease severity for patients with cystic fibrosis (CF). CF is a genetic disease that dramatically decreases life expectancy and quality. The disease is characterized by polymicrobial infections which lead to lung remodeling and airway mucus plugging. In order to quantify disease severity of CF patients and compute a continuous severity index measure, quantile regression, rank scores, and corresponding normalized ranks are calculated for CF patients. Based on the rank scores calculated from the set of quantile regression models, a continuous severity index is computed for each CF patient and can be considered a robust estimate of CF disease severity.

Key words: Cystic Fibrosis, FEV1, quantile regression, robust regression, rank scores, severity index.

1. Introduction

Cystic fibrosis (CF) is a genetic disease that greatly decreases life expectancy and quality. CF is characterized by decreased mucociliary clearance, chronic polymicrobial infections which lead to airway wall remodeling. The hallmark of CF is chronic, progressive obstructive lung disease (Kulich et al., 2005). The pulmonary disease progresses over time and causes significant disability, and eventually respiratory failure and death. Among people with CF, the median predicted life expectancy for people born in the 2013-2017 US birth cohort is 44 years of age (CF National Registry, 2017).

Lung function is a primary indicator of health for people with CF whether measured as Forced Expiratory Volume (FEV1), forced vital capacity (FVC), or FEV1 percent predicted. FEV1–measured using a spirometer–is the volume of air that can be forced out of the lungs in the first second after a full inspiration and is a common marker of disease severity for CF patients (American Thoracic Society, 2005; Davis et al., 1997; Ramsey and Boat, 1994). Two other measures of lung function are; (1) FVC, the volume of air that a patient is able to expire after full inspiration (2) FEV1 percent predicted, the percent of a nonsmoking population normalized for gender, height, age and ethnicity.

The classification of CF patients into different levels of severity allows researchers and clinicians to establish the associations between different severity of lung function and clinical progress. The severity of the CF disease based on longitudinal lung function data has been classified by grouping patients based on their age and lung function (Schluchter et al., 2006; Szczesniak et al., 2017; Wagner et al., 2011). In their work, Schluchter et al. (2006) determined severity of CF patients in the following way; patients younger than thirty-four and in top quartile of FEV1 percent predicted were classified as mild, CF patients younger than thirty-four and in the lowest quartile were classified as severe, and CF patients thirty-four and older were all classified as mild. This classification of CF patients was an extension of the research by Wagner et al. (2011), in which the t-test, Wilcoxon test, and the two-part test statistic was applied to sequence data in order to identify taxa that differ between two groups.

Conrad and Bailey (2015) used the modern statistical learning method of random forests to cluster CF patients using common clinical variables. Phenotypes were identified using a proximity matrix generated by the unsupervised Random Forests algorithm and subsequent clustering by the Partitioning around Medoids (PAM) algorithm. This research provided a description of the different classes or groups of patients. However, only the patients' best FEV1% and FVC% predicted for the prior 12 months was included as a measure of lung function.

Recently, Szczesniak et al. (2017) applied sparse functional principal components analysis to classify patients into distinct phenotypes using longitudinal FEV1 trajectories. In their research, the classes were determined by comparing a patient's single score to the first and third quartiles of scores from the first functional principal component. Previous studies have been designed to compare severity of CF and dichotomizing CF into distinct groups (i.e. mild versus severe CF), looking at one measure of lung function at a particular age, or using a percentile of lung function to determine severity.

The primary goal of our research is to use a range of quantile regression models to gain a

better understanding of disease severity classification and the factors that affect disease prognosis for CF patients. This paper presents a new approach to describing disease severity of CF patients by using quantile regression rank scores and the corresponding normalized ranks. In this approach, we propose using the ranks calculated from a series or range of quantile regression models. The series of quantile regression models use FEV1 as the response and age, BMI, and gender as predictors. The main advantage of using quantile regression is that we can obtain a range of models and we are able to see how the effects of covariates can change across the quantiles. Based on the rank scores calculated from the set of quantile regression models, a continuous severity index is computed for each CF patient and can be considered a robust estimate of CF disease severity.

2. Methods

2.1 Data

Data was collected from CF patients (age 18 and over) from the Adult Cystic Fibrosis Clinic at the University of California, San Diego. Forty-four patients were randomly selected from the patients evaluated at least once during the 2012 calendar year. Patients who received a lung transplant were excluded from the study. The patients were ranked by an experienced CF clinician for their overall health outlook using sequential pairwise comparisons. These patients had 2,501 outpatient encounters during which lung function was assessed using spirometry; 50% of the CF patients had forced expository volume (FEV1) measurements that decreased by 0.46 liters over the length of their care. 43.18% of the CF patients are female. The summary statistics for FEV1, BMI, and age at entry are given in Table 1. In addition to gender, BMI at entry, age, and FEV1 measurements, we have an independent secondary source of information on the patients' CF disease severity.

Table 1: Summary statistics for the CF clinic patients.

	Mean (Standard Deviation)	Median (IQR)
FEV1 at entry (liters)	2.46 (0.85)	2.45 (1.22)
BMI at entry (kg/m^2)	22.24 (3.35)	21.03 (4.96)
Age at entry (years)	23.83 (6.82)	20.83 (9.54)
Length of follow-up (years)	11.75 (6.61)	11.88 (8.43)

Multi-threaded computing (parallel processing) for analyses were performed in the Microsoft R Application Network (Microsoft Corporation, 2016); an enhanced distribution of the R language and environment for statistical computing. This research was implemented using the R package *quantreg* (Koenker, 2015).

2.2 Quantile Regression

Quantile regression methods have been explored in numerous research areas (Casady and Cryer, 1976; Daouia et al., 2011; Eide and Showalter, 1998; He and Shi, 1998; Ma and He, 2014; Møller et al., 2008; Portnoy and Koenker, 1997; Xiong and M., 2019; Zhang et al., 2017;

Zhou and Portnoy, 1998). Linear quantile regression and other quantile methods are discussed in Koenker (2000), in which he developed quantile regression methods in the linear and the nonlinear case.

Let (y_i, x_i) be a sample from some population, where x_i is a $p \times 1$ vector of regressors. Let τ be the quantile of interest in the interval (0, 1). Let $\beta_\tau = \left(\beta_{\tau 1}, \ldots, \beta_{\tau p}\right)^T$ be the slopes associated with the respective covariates associated with a particular quantile of interest, and let the random error be ϵ_i. Therefore, the quantile regression model for the τth quantile is given by:

$$y_i = x_i^T \beta_\tau + \epsilon_i \qquad (1)$$

The τth conditional quantile of y_i given x_i is defined to be

$$Q_\tau\left(y_i \mid x_i\right) = x_i^T \beta_\tau \qquad (2)$$

The errors are defined to be $\epsilon_i = y_i - x_i^T \beta_\tau$. For this model, the assumption is that the ϵ_i's are independent and identically distributed asymmetric Laplace random variables where $Q_\tau(\epsilon_i | x_i) = 0$. The parameter estimates are chosen so that the loss function is minimized:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau\left(y_i - x_i^T \beta_\tau\right) \qquad (3)$$

where the loss function is $\rho_\tau(u) = u(\tau - I(u < 0))$ and I(·) is the indicator function. In order to solve the minimization problem presented in (3), we solve:

$$\arg\min_{\beta_\tau} \sum_{i=1}^{n} w_i(\tau)\left|y_i - x_i^T \beta_\tau\right| \qquad (4)$$

where

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i - x_i^T \beta_\tau \geq 0 \\ 1 - \tau & \text{if } y_i - x_i^T \beta_\tau < 0 \end{cases}$$

Solving the dual linear programming problem yields the estimates of the quantile specific regression parameters and gives equivalent results to those found by (3); it is available and implemented using the R package quantreg (Koenker, 2015).

2.3 Regression Rank Scores

As an extension to quantile regression, we calculate regression rank scores by using the order statistics to calculate the rank for each observation in the dataset. The rank based method for quantile regression that we propose is a special case of generalized rank tests for the linear

regression model (Koenker, 2005).

Linear rank statistics can be used to obtain the rank for each observation in the dataset while considering the conditional distribution of Y based on X across the quantiles. The quantile regression rank score for each observation can be calculated for each quantile of the distribution as follows. Let $\left(Y_1,\ldots,Y_n\right)$ be the response and let $\left(R_1,\ldots,R_n\right)$ be the associated ranks. The rank generating function (Hajek and Sidak, 1967) is given by:

$$\hat{a}_i(\tau)=\begin{cases}1 & \text{if } \tau < \dfrac{R_i-1}{n}\\ R_i-\tau n & \text{if } \dfrac{R_i-1}{n}\le\tau\le\dfrac{R_i}{n}\\ 0 & \text{if } \tau > \dfrac{R_i}{n}\end{cases} \qquad (5)$$

After fitting the quantile regression lines for a set of $\tau \in (0, 1)$, we can use the rank generating function to calculate quantile regression rank scores for each observation. The regression rank scores (Gutenbrunner and Jureckova, 1992) for each observation, are given by:

$$\hat{b}_i=-\int_0^1 \phi(\tau)d\hat{a}_i(\tau) \qquad (6)$$

where $\phi(\cdot)$ is a function of bounded variation which is constant outside a compact subinterval of (0, 1). The following choices of $\phi(\cdot)$ will result in two important versions of regression rank scores. The Wilcoxon regression rank scores are given by:

$$\hat{b}_i=-\int_0^1\left(\tau-\frac{1}{2}\right)d\hat{a}_i(\tau)=\int_0^1\hat{a}_i(\tau)d\tau-\frac{1}{2} \qquad (7)$$

The normalized regression rank scores are given by:

$$\hat{b}_i=-\int_0^1\Phi^{-1}(\tau)d\hat{a}_i(\tau) \qquad (8)$$

where $\Phi^{-1}(\cdot)$ is the inverse standard normal distribution. Both the Wilcoxon and normalized regression rank scores serve as continuous measures of disease severity.

3. Results

The effect of BMI on lung function varies across the quantiles. For the “most severe” patients (i.e. patients in the lower quantiles), as BMI increases, lung function is still decreasing. However, for the “least severe” patients (i.e. patients in the upper quantiles) increasing BMI is associated with increased lung function. These results are displayed in Figure 1; the scatterplot for FEV1 versus BMI displays a slightly negative correlation in the lower quantiles

and a positive correlation in the upper quantiles.

The estimated slope coefficients versus the quantiles is portrayed in Figure 2, 3, and 4 for τ in (0, 1); the estimated slope coefficients and respective standard errors, test statistics, and p-values for the quantile regression models for the 0.10, 0.50 and 0.90 quantiles are given in Table 2 (Appendix B). Depending on the severity of the patient (being in the lower or upper conditional quantile), the effects of the covariates (and respective significance) varies. This concept yields invaluable information and has the ability to assist practitioners to assess patients on a personalized level based on the severity of their CF disease.

The relationship between FEV1 and age, age^2, BMI, and gender varies across the quantiles; rarely does an individual patient behave like the "average" patient. For the "best" patients, age has a larger effect on FEV1 (in the upper quantiles the relationship between age of spirometry and FEV1 has a larger negative effect). Whereas, for the "worst" patients, age has a smaller effect on FEV1; in the lower quantiles, the relationship between age and FEV1 is close to zero.

The ranks were obtained for every observation across a fine grid of values of τ in the interval (0, 1) and then the normalized rank for each observation was found based on the quantile regression models for FEV1 as the response and age, age^2, BMI, and gender as the predictors. In Figure 5 we see that the FEV1 scores for this particular patient are decreasing, however, when we take into account age, BMI, and gender, the normalized ranks indicate that this patient's lung function is improving. We obtained similar plots of the normalized ranks for all patients and show patients with the lowest severity indices in Figure 6 and the highest severity indices in Figure 7. Overall, the less severe patients have normalized ranks which are positive and increasing with age. Whereas, the more severe patients have normalized ranks which are negative and decreasing with age.

Figure 8 displays the median Wilcoxon ranks and median normalized ranks based on the quantile regression models for FEV1 as the response and age, BMI, and gender as the main effects as well as the FEV1 scores for each of the CF patients. The patients are ordered (based on expert opinion) by severity index with the "least severe" CF patients on the left and the "most severe" patients on the right. The normalized ranks decrease as the severity index increases indicating that the normalized ranks can serve as an indication of what an expert would consider severe CF progression and is more robust than looking at spirometry measurements alone. We have presented the results from both the Wilcoxon ranks and median normalized ranks as continuous measures of severity for comparison. The Wilcoxon ranks consider each point in time as being positive, negative, or zero while the normalized ranks take into account the size of the residual for each observation.

4. Summary and Conclusions

Since CF is a complex disease, FEV1 is serving as one measure of disease severity. The expert in this case is providing a holistic measure of disease severity. Our goal of using the regression rank scores is to provide a continuous measure of disease severity that accounts for the relationship of FEV1 and age, BMI, and gender. This analysis allows the severity of a CF patient's lung function to be calculated based on a range of quantile regression models. This is in contrast to other research that uses a single summary measure of severity. Our

methodology provides more information to describe the severity of a complex disease as well as providing a continuous measure of disease severity.

One limitation of our study is that it was only performed at one site. Thus, we do not know whether or not the severity of CF disease for each patient would change when more patients are added. Additionally, patients were not seen at equally spaced time points. If patients were too sick to come into the clinic we would not have spirometry measurements for them at those times of disease exacerbations and low lung function. Similarly, patients experiencing more problems with their CF disease progression may come to the clinic more often. Since none of the clinic patients were hospitalized during the study period we assume that patients scheduling an appointment was not due to a decrease of lung function.

Although we focus primarily on CF disease progression, these same concepts could be applied to many different problems beyond CF. The methods of this paper can be applied to other diseases to calculate a continuous measure of disease severity. Furthermore, these methods can be used to provide ranks conditional on a set of factors rather than on one summary measure. Extending the methodology to account for longitudinal and clustered data using linear mixed models is an exciting area of ongoing and future research.

By incorporating results from multiple quantile regression models we are able to provide more detailed information on disease progression that takes into account an individuals' patient history and lung function over time. The secondary benefits are that we understand how the conditional distribution of FEV1 changes if the subjects are in the top 10% compared to the bottom 10%, given a particular set of covariates. This fact implies that important variables for prediction can change across a distribution and the effects of covariates may be different for patients with different disease severity levels. As more focus is placed on personalized medicine, there is a need to develop innovative ways to classify disease severity for an individual patient and not just the "typical" or average patient. The use of quantile regression and normalized regression rank scores allow clinicians to understand the severity of CF disease based on the specific history of each patient.

Appendix A

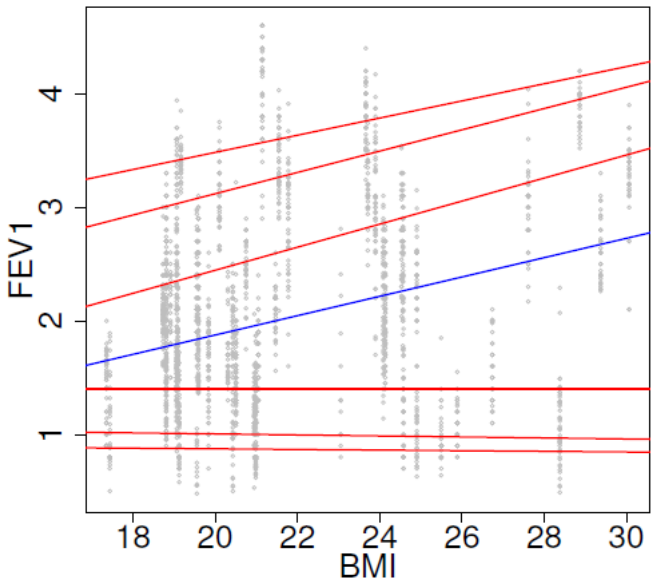

Figure 1: FEV1 modeled by BMI; the median regression line is in blue and the fitted quantile regression lines for $\tau = (0.05, 0.10, 0.25, 0.75, 0.90, 0.95)$ are in red.

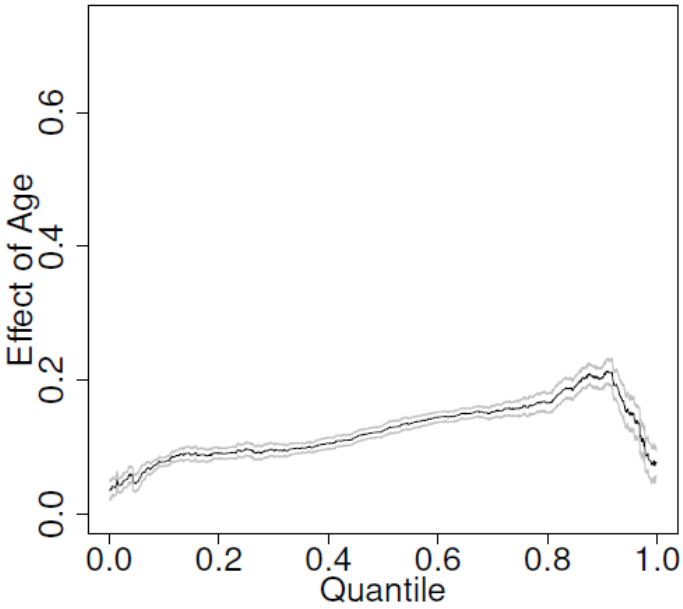

Figure 2: Estimated slope coefficients and standard errors of age across the quantiles.

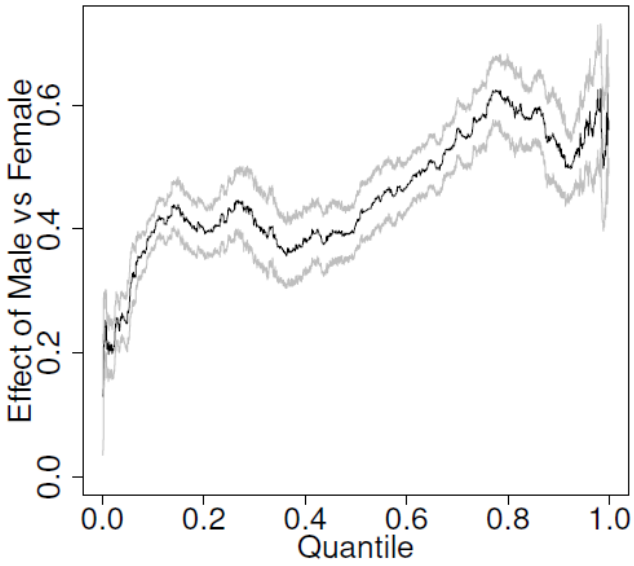

Figure 3: Estimated slope coefficients and standard errors of gender (male versus females) across the quantiles.

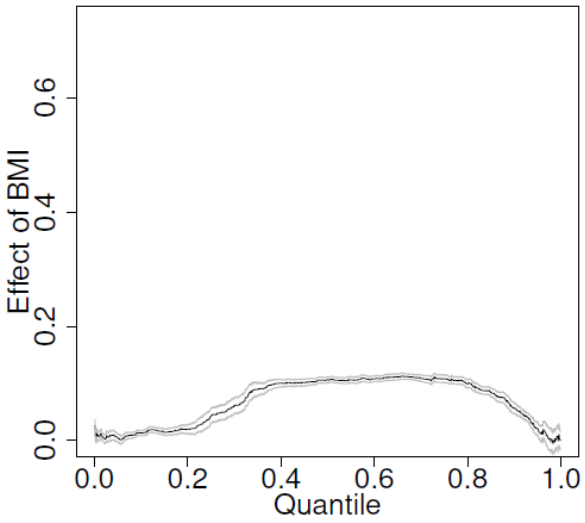

Figure 4: Estimated slope coefficients and standard errors of BMI across the quantiles.

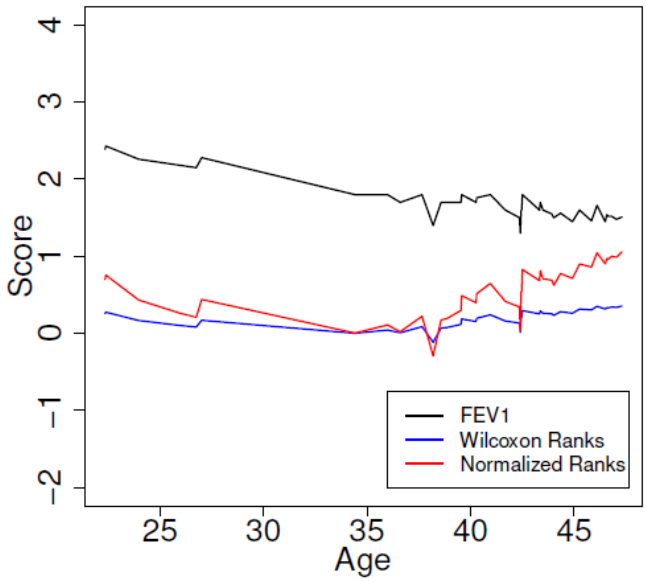

Figure 5: FEV1 and regression rank scores for one patient in our study; regression rank scores are based on a set of quantile regression models with FEV1 as the response and age, BMI, and gender as the main effects.

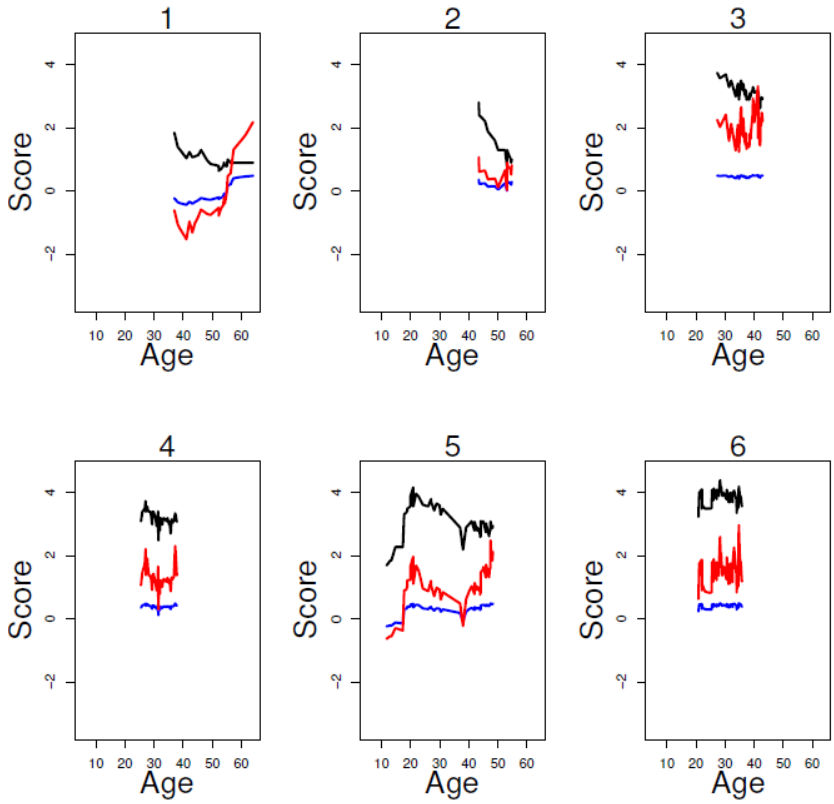

Figure 6: FEV1 (black), Wilcoxon ranks (blue), and normalized ranks (red) for the least severe patients (severity index 1-6) according to independent expert opinion.

Appendix B

Table 2: Quantile regression models for FEV1.

Quantile	Variable	Estimate	Standard Error	t-value	Pr(>|t|)
$\tau = 0.1$	Intercept	-0.2784	0.0882	-3.16	0.0016
	Age	0.0788	0.0041	19.18	0.0000
	Age^2	-0.0015	0.0001	-20.45	0.0000
	Male	0.3950	0.0217	18.19	0.0000
	BMI	0.0132	0.0049	2.67	0.0077
$\tau = 0.5$	Intercept	-1.7605	0.1448	-12.16	0.0000
	Age	0.1249	0.0051	24.54	0.0000
	Age^2	-0.0025	0.0001	-26.96	0.0000
	Male	0.4088	0.0394	10.38	0.0000
	BMI	0.1070	0.0061	17.51	0.0000
$\tau = 0.9$	Intercept	-0.9370	0.2909	-3.22	0.0013
	Age	0.2050	0.0136	15.09	0.0000
	Age^2	-0.0036	0.0002	-15.10	0.0000
	Male	0.5338	0.0526	10.15	0.0000
	BMI	0.0564	0.0060	9.39	0.0000

Bibliography

American Thoracic Society, E. R. S. (2005). Raised volume forced expirations in infants: guidelines for current practice. American Journal of Respiratory and Critical Care Medicine 172, 1463–1471. doi: 10.1164/rccm.200408-1141ST.

Casady, R. J. and Cryer, J. (1976). Monotone percentile regression. The Annals of Statistics 4, 532–541. http://www.jstor.org/stable/2958224.

Conrad, D. and Bailey, B. (2015). Phenotyping of an adult cystic fibrosis patient population. PLoS ONE 10, 1–14. doi: 10.1371/journal.pone.0122705.

Daouia, A., Gardes, L., and Girard, S. (2011). On kernel smoothing for extremal quantile regression.

Bernoulli 19, 2557–2589. doi: 10.3150/12BEJ466.

Davis, P., Byard, P., and M., K. (1997). On kernel smoothing for extremal quantile regression. Pediatric Research 41, 161–165. doi: 10.1203/00006450-199704001-00973.

Eide, E. and Showalter, M. H. (1998). The effect of school quality on student performance: a quantile regression approach. Economics Letters 58, 345–350. doi: 10.1016/S0165-1765(97)00286-3.

Gutenbrunner, C. and Jureckova, J. (1992). Regression rank scores and regression quantiles. The Annals of Statistics 20, 305–330. doi: 10.1214/aos/1176348524.

Hajek, J. and Sidak, Z. (1967). Theory of rank tests. Academic Press. doi: 10.1016/B978-0-12-642350-1.X5017-6.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. Journal of the American Statistical Association 93, 1–17. doi: 10.2307/2670115.

Koenker, R. (2000). Galton, edgeworth, frisch, and prospects for quantile regression in econometrics. Journal of Econometrics 95, 347–374. doi: 10.1016/S0304-4076(99)00043-3.

Koenker, R. (2005). Quantile regression. Econometric Society Monographs. Cambridge University Press. doi: 10.1017/CBO9780511754098.

Koenker, R. (2015). quantreg: Quantile Regression. R package version 5.11.

Kulich, M., Rosenfeld, M., Campbell, J., Kronmal, R., Gibson, R., Goss, C. H., and Ramsey, B. (2005). Disease-specific reference equations for lung function in patients with Cystic Fibrosis. American Journal of Respiratory and Critical Care Medicine 172, 885–891. doi: 10.1164/rccm.2004101335OC.

Ma, X. J. and He, F. X. (2014). Power weighted quantile regression and its application. Journal of Data Science 12, 535–544.

Microsoft Corporation (2016). RevoUtilsMath: Microsoft R Services Math Utilities Package. R package version 8.0.3.

Møller, J. K., Nielsen, H. A., and Madsen, H. (2008). Time-adaptive quantile regression. Computational Statistics & Data Analysis 52, 1292–1303. doi: 10.1016/j.csda.2007.06.027.

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Statistical Science 12, 279–300. http://www.jstor.org/stable/2246217.

Ramsey, B. and Boat, T. (1994). Outcome measures for clinical trials in cystic fibrosis: summary of cystic fibrosis foundation consensus conference. The Journal of Pediatrics 124, 177–192. PMID: 8301420.

Schluchter, M. D., Michael, W. K., Mitchell, L. D., James, R. Y., and Michael, R. K. (2006). Classifying severity of Cystic Fibrosis lung disease using longitudinal pulmonary function data. American Journal of Respiratory Critical Care Medicine 174, 780–786. doi: 10.1164/rccm.2005121919OC.

Szczesniak, R. D., Li, D., Su, W., Brokamp, C., Pestian, J., Seid, M., and Clancy, J. P. (2017). Phenotypes of rapid cystic fibrosis lung disease progression during adolescence and young adulthood. American Journal of Respiratory Critical Care Medicine 196, 471–478. doi: 10.1164/rccm.201612-2574OC.

Wagner, B. D., Roberston, C. E., and Harris, J. K. (2011). Two-part statistics for comparison of sequence variant counts. PLoS ONE 6, 1–8. doi: 10.1371/journal.pone.0020296.

Xiong, W. and M., T. (2019). Weighted quantile regression theory and its application. Journal of Data Science 17, 145–160. doi: 10.6339/JDS.201901 17(1).0007.

Zhang, L., Wang, H., and Zhu, Z. (2017). Composite change point estimation for bent line quantile regression. Annals of the Institute of Statistical Mathematics 69, 145–168. doi: 10.1007/s10463-015-0538-5.

Zhou, K. and Portnoy, S. (1998). Statistical inference on heteroscedastic models based on regression quantiles. Journal of Nonparametric Statistics. 9, 239–260. doi: 10.1080/10485259808832745.