Quantile Regression

Mason Lowery, Katie Byrne, Alyssa Selvey

08/07/2022

1 Introduction

The primary objective of this paper is to explain the method of quantile regression and when it is to be appropriately implemented. At its foundation, quantile regression follows a similar structure to that of linear regression analysis, where we are interested in modeling the relationship between a continuous dependent variable and one or more predictor variables. A Regression Model, in general, gives two valuable takeaways about a data set: the ability to understand the relationship between the explanatory variables and dependent variables and the ability to predict the dependent variable given values of the explanatory variables, where inferences can be made. However, the overarching difference between normal linear regression and quantile regression, is that linear regression most commonly uses the method of ordinary least squares to predict the conditional mean of the response variable, whereas quantile regression estimates the relationship between the predictors and certain quantiles of the response variable, usually the conditional median. Quantile regression was first introduced by Roger Koenker and Gilbert Bassett in 1978, but only recently has it become more practical for large data. [1]

1.1 History of Quantile Regression

The idea of quantile regression was first introduced by Ruder Josip Bošković during his ambassadorship to London in 1760 [2]. The original proposal was the idea of trying to estimate the regression slope for the median. His proposal stemmed from his interest in the ellipticity of the earth. Bošković believed that the estimation of the ellipticity of the earth can be solved through the following formula

$$min \sum [y_i - \alpha - \beta sin^2 \theta_i] \tag{1}$$

using the five observations at various latitudes such as Paris and Rome [3]. The previous formula was also being constrained from the zero equaling mean residual,

$$n^{-1} \sum (y_i - \alpha - \beta sin^2 \theta_i) \tag{2}$$

Bošković preceded the next advancements for this regression that was further expanded by Pierre-Simon Laplace and later Francis Edgeworth. Laplace further expanded on Bošković's solution that the problem could rather be solved utilizing a computed weighted mean. He furthered his belief through his complete theory for the asympotic behavior in a scalar case for the weighted median [3]. Furthermore, acting as a revival of Bošković's 1760 proposal, Edgeworth created his theory of 'plural method'. Edgeworth was in direct rival of the popular least-squares method that was founded in the early 1800s by Adrien-Marie Legendre and Carl Friedrich Gauss [3]. Agreeing with the results of Laplace, Edgeworth believed that by dropping the mean constraint for the residuals that it would result in a more accurate plural median rather than the least-squares estimator when there was differing observations [3]. Each of the previous mathematicians, paved the way for Roger Koenker and Gilbert Bassett Jr's 1978 study of quantile regrssion. Koenker and Bassett's study has been noted as the most important and modern study on quantile regression. Their quantile regression study is an extension of the least-squares method for the conditional mean to illustrate the models of conditional quantile functions [3]. The 1978 study has been the basis of later applications and studies using quantile regression.

2 Methods

2.1 Method of Ordinary Least Squares Regression

Given that quantile regression is technically similar and formulated from linear regression, we must first explain the process of normal linear regression. Linear regression is a linear modeling method that attempts to explain the relationship between one continuous dependent variable and one or more predictor variables. [4]

The most common method of regression analysis is the Ordinary Least Squares method (OLS). A simple linear regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i \tag{3}$$

Y_i is the dependent variable (continuous)
X_1 is the independent predictor variable (categorical or continuous)
β_0 is the intercept
β_1 is the coefficient of corresponding to the independent variable.
ϵ_i is the error term.
where $\epsilon\ N(0, \sigma^2)$

We can find the line of best fit by minimizing the mean square error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 X_{i1}))^2 \tag{4}$$

Equations for solving for β_0 and β_1 we have:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} yi}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \tag{5}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{6}$$

After fitting a model, we want to know if our independent variable is a significant predictor of the response variable. We do this through hypothesis testing.

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Where we reject the null hypothesis if $p < \alpha$, and conclude that β_1 is not equal to zero. Similarly, we can run another hypothesis test if β_0 is not equal to zero, this tells us if our model itself is statistically significant.

2.1.1 Multiple Linear Regression

Multiple linear regression is an extension to simple linear regression, where we aim to increase the number of predictor variables in our model. This is important to note, because the application later seen in this paper will contain multiple independent variable. A multiple linear regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_k x_{ik} + \epsilon_i \tag{7}$$

We can estimate the vector β of the coefficients by the formula below:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{8}$$

2.2 Method of Quantile Regression

To start, the Conditional Quantile Function (CQF) is defined as the inverse of its conditional distribution function. The CQF of Y_i given X_i fully explains the relationship between Y and X. [5]

$$Q_\tau(Y_i|X_i) = F_Y^{-1}(\tau|X_i) = inf\{y : F_y(y|X_i) \geq \tau\} \tag{9}$$

where $F_Y(y|Xi)$ is the distribution function for Y_i conditional on X_i.

Contrasting from multiple linear regression where the β coefficients are constant, in quantile regression the beta coefficients are functions dependent upon τ which is whatever τ^{th} quantile we are interested in.

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)X_1 + ... + \beta_k(\tau)X_{ik} + \epsilon_i \tag{10}$$

In order to determine the β coefficients for each quantile, which is a very similar process as seen in OLS, but for quantile regression we must reduce the median absolute deviation.

$$MAD = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1(\tau)X_{i1} + ... + \beta_k X_{ik}(\tau)) \tag{11}$$

where ρ_τ is referred to as the "check function".

The check function, often referred to as the quantile loss function or the criterion function, is defined as follows [6]:

$$\rho_\tau(u) = \left\{ \begin{array}{ll} \tau u, & \text{if } u > 0 \\ (\tau - 1)u, & \text{if } u \leq 0 \end{array} = u(\tau - I)(u < 0)) \right\} \tag{12}$$

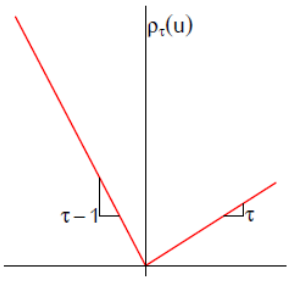

Figure 1: Graph of Check Function

Simply enough, the check function gets its name due to its usual resemblance to the a check mark. [7] When $\rho_\tau\ (0) = 0$, it increases linearly with slope τ as u moves positively to the right and it increases linearly with slope $\tau - 1$ as u moves negatively to the left. [7] This can be seen in the graph above, where the red line shows $\tau = 0.25$, if we are interested in the median of the 25th percentile, we are interested in having 75% of the

error be positive and 25% to be negative. In order to effectively minimize the median absolute deviation for $\tau = 0.25$, we have to penalize extreme negative errors more heavily than large positive errors, more precisely a penalty of 0.75 is placed on the negative errors and 0.25 on the positive errors. [8] $\tau = 0.50$ is a special case, corresponding to median regression, resulting in us having symmetric weighting of observations with positive and negative residuals, resembling the absolute value function. To summarize, OLS prediction intervals are constructed assuming that residuals have equal variance and are normally distributed, when these assumptions are violated the quantile check function provides logical prediction intervals for these residuals.

2.2.1 ANOVA Test for Comparing Slopes

When performing quantile regression, our primary interest is in two things: do the quantile coefficients differ significantly from the OLS model (shown later) and is there a significant difference in the coefficients between the different quantiles. The ANOVA test can be utilized to to determine if there is such a difference in slopes. The ANOVA test uses the following table to compare different slopes, a significant difference in slopes between quantiles justifies our use of quantile regression [9].

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Value
Between Groups	$SSB = \sum n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = \frac{SSB}{k-1}$	$f = MSB/MSE$
Error	$SSB = \sum\sum(\bar{X}_j - \bar{X})^2$	$df_2 = N - k$	$MSE = \frac{SSE}{N-k}$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

Table 1: ANOVA Test Table

2.2.2 Quantile Regression or Linear Regression

Primary Differences Between Modeling Methods

Linear Regression

- Estimates the conditional mean
- Normal distribution of response/error
- Equal variance
- Sensitive to outliers
- Computationally inexpensive

Quantile Regression

- Estimates the conditional quantiles
- Makes no distribution assumptions
- No assumption of Equal Variance
- Robust against outlier
- Computationally intensive

3 Illustrative examples with R

3.1 Quantile and Linear Regression Comparison

For this comparison, a data set imported from Kaggle containing 28 observations from a classroom was utilized. The data set contained the hours a student put into studying for a test and the score that they received on it [10]. A simple linear regression model is fitted to to predict how many hours of studying would give a student a certain score. We will estimate the conditional quantiles, using quantile regression to compare our β coefficients differ from the OLS model.

Consider the following dataset:

Hours	Scores	Hours	Scores
2.5	21	5.1	47
3.2	27	8.5	75
3.5	30	1.5	20
9.2	88	5.5	60
8.3	81	2.7	25
7.7	85	5.9	62
4.5	41	3.3	42
1.1	17	8.9	95
6.1	67	7.4	69
2.7	30	4.8	54
3.8	35	6.9	76
7.8	86	1.3	19
2.6	30	6.1	63

Table 2: Data Table for Hours and Scores

3.1.1 Linear Regression Model

In R Studio, the summary output for the linear model yields the equation shown below. Fitting this equation gives a good idea of how β coefficients of the OLS model compare to the coefficients of the different quantiles for τ.

R Code:

```
reg <- lm(Scores = Hours, data=data)
summary(reg)
```

$$\hat{Scores} = 3.2017 + 9.6774 * Hours \tag{13}$$

3.1.2 Quantile Regression Model

The focus for this model will be on the τ values (0.25, 0.50, 0.75 and 0.90). Using R Studio, we will see how the coefficient estimates differ for each of the τ^{th} quantile [11].

R Code:

Model + Summary

```
qreg25 <- rq(Scores = Hours, data=data, tau=0.25)
summary(qreg25)

qreg50 <- rq(Scores = Hours, data=data, tau=0.50)
summary(qreg50)

qreg75 <- rq(Scores = Hours, data=data, tau=0.75)
summary(qreg75)

qreg90 <- rq(Scores = Hours, data=data, tau=0.90)
summary(qreg90)
```

Graph Code:

```
ggplot(data, aes(Hours, Scores)) +
geom_point() + geom_quantile(quantiles=0.25, aes(color = "25th
Quantile")) +
geom_quantile(quantiles=0.50,aes(color="MAD")) +
geom_quantile(quantiles=0.75,aes(color = "75th Quantile")) +
geom_quantile(quantiles=0.90,aes(color = "90th Quantile")) +
geom_smooth(method='lm',aes(color="OLS"),lty=2,size=0.5,se=F)
+
scale_color_manual(name='Regression Lines', breaks = c("25th
Quantile","MAD","75th Quantile","90th Quantile","OLS"),
values = c("25th Quantile"="red", "MAD" = "blue", "75th
Quantile"="orange", "90th Quantile"="purple", "OLS"=
"black"))
```

$$\tau = 0.25 > \hat{Q_{0.25}(Scores)} = -0.88889 + 9.44444 * Hours$$

$$\tau = 0.50 > \hat{Q_{0.50}(Scores)} = 5.91826 + 9.34783 * Hours$$

$$\tau = 0.75 > \hat{Q_{0.75}(Scores)} = 6.00000 + 10.00000 * Hours$$

$$\tau = 0.90 > Q_{0.90}(\hat{Scores}) = 5.674164 + 10.29851 * Hours$$

For example, the 0.50 quantile model predicts that for every hour increase in study time there is a 5.91826 increase in the score received. This is differing from the 0.90 quantile where for every hour increase in study time there is a 10.29851 increase in the score received.

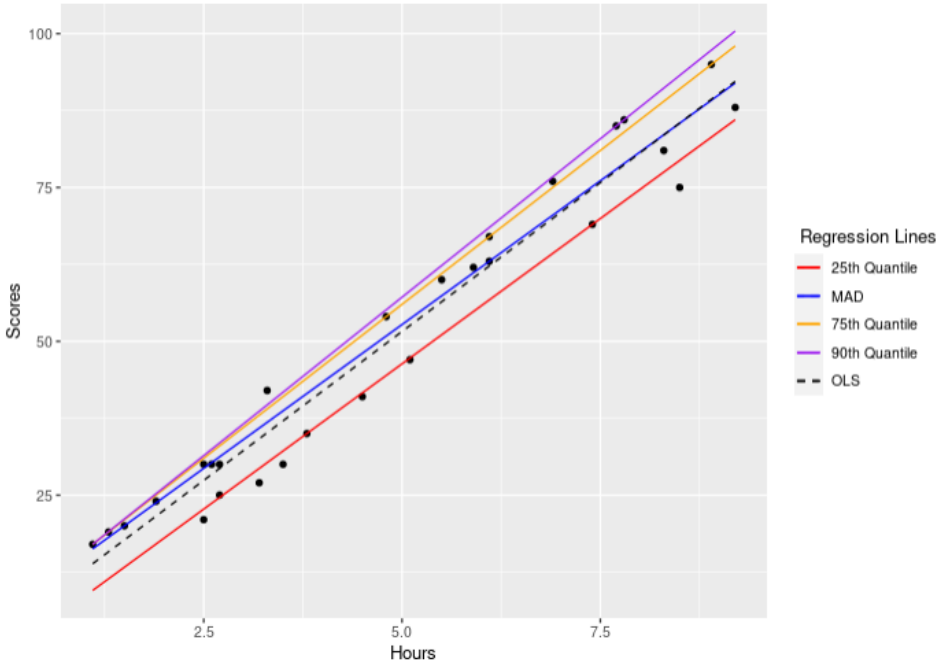

Figure 2: Regression Lines for all Quantiles and OLS line

The differences in the quantile regression lines and OLS line can be visualized in the plot displayed above. We see from the graph that there may not be a significant difference in the slopes of the different quantiles, we can perform an ANOVA to verify this.

3.1.3 Do the Quantile Regression Coefficients Differ Across Quantiles?

R Code:

```
anova(qreg25, qreg50, qreg75, qreg90)
```

DF	Resid DF	Test Statistic	p-value	Quantiles
3	109	0.7916	0.5011	25th, 50th, 75th, 90th

Table 3: Anova Output from R: Equality of Slopes between 25th, 50th, 75th, 90th Quantiles

From the results above, we see that there is not a significant difference in β coefficients across the different quantiles, $p > \alpha$. Based on this, we may suggest looking into other methods to analyze this dataset, ie. quantile regression does not tell us an interesting story for this data.

4 Illustrative examples with R

4.1 Quantile Regression Application 2 (Multiple Predictors)

Consider the following dataset shown below.

Obs	Sex	BMI	Children	Smoker	Expenses
1	Female	27.9	0	yes	16884.92
2	Male	33.8	1	no	1725.55
3	Male	33.0	1	no	4449.46
4	Male	22.7	0	no	21984.47
5	Male	28.9	0	no	3866.86
6	Female	25.7	0	no	3756.62
7	Female	33.4	1	no	8240.59
8	Female	27.7	1	no	7281.51
9	Male	29.8	1	no	6406.41
...	...	...	...	...	...
1338	Female	29.1	0	yes	29141.36

Table 4: Medical Insurance Data

The table above shows a subset of 10 observations from a Medical Insurance Cost dataset imported into R from Kaggle. [12] The data includes the following information, sex (male = 1, female=0), body mass index (BMI), children (0 or 1, no kids or has kids), smoker (yes or no), expenses (cost of medical insurance premiums). For our application of quantile regression, sex, BMI, children, smoker, will be used to predict the cost of medical insurance premium per person (expenses).

The quantile regression line will be as follows:

$$\hat{Expenses} = \beta_0(\tau) + \beta_1(\tau)Sex + \beta_2(\tau)Children + \beta_3(\tau)Smoker + \beta_4(\tau)BMI$$

R Code: hist(data1$expenses, xlab="Expenses", ylab="Frequency", main = "Histogram of Response")

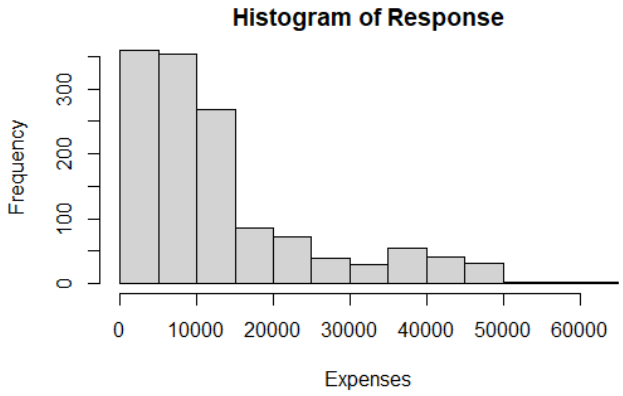

Figure 3: Histogram of Response Variable

From the figure above, we can see that our response variable is not normally distributed. Unlike linear regression, quantile regression is not affected by this, therefore we can conclude that our use of quantile regression could be of appropriate use to analyze this dataset.

To confirm our suspicions that linear regression assumptions are violated. We can also check for heteroscedasticity using the Breusch-Pagan test.

R Code:

```
bptest(reg)
```

Test Statistic	**DF**	***p*-value**
83.023	4	<0.0001

Table 5: Breusch-Pagan Test for Heteroscedasticity

From the results above we can conclude that there is heteroscedasticity and that linear regression assumptions are violated. Again, while quantile regression is not affected by the violation of any of these assumptions, it's important to justify our use of it for this data.

In R we will display the output summary of the "lm" function and the "qr" function for τ=0.10, τ=0.50, and τ=0.90. This will give us an exact estimation how much the β coefficients change across quantiles.

R Code:

```
reg < − lm(expenses   bmi + smoker + sex + children, data=data1)
summary(reg)

quantreg10 < − rq(expenses   bmi + smoker + sex + children , data=data1, tau=0.10)
summary(quantreg10)

quantreg50 < − rq(expenses   bmi + smoker + sex + children , data=data1, tau=0.50)
summary(quantreg50)

quantreg90 < − rq(expenses   bmi + smoker + sex + children , data=data1, tau=0.90)
summary(quantreg90)
```

Parameter	**τ =0.10**	**τ =0.50**	**τ =0.90**	**OLS**
Intercept	644.1800*	-4017.0730*	3840.9533*	-4054.200*
Sex	433.18403*	-335.1294	-853.9010	-303.4
Children	1897.3275*	1449.9323*	1076.6288	1329.600*
Smoker	14884.7322*	27348.7223*	29509.1998*	23605.800*
BMI	42.293*	341.3453*	358.7325*	387.600*

Table 6: β Coefficients for τ = 0.10, 0.50, 0.90

note: asterisk indicates $p < \alpha$ where $alpha = 0.05$

From the tables above, we see that only two variables were not significant predictors of expenses across all three specified quantiles, sex (male=1, female=0), and children (1=has children, 0=no children). Sex was also found to not be a significant predictor of expenses in the OLS model.

4.1.1 Are the β Coefficients from Quantile Regression Significantly Different from OLS?

From the previous page, in the summary output above the Smoker variable is highlighted. For this variable we will analyze if the β values obtained from quantile regression are significantly different from the OLS model. We can do this by creating a 95% confidence interval for three quantiles listed above for their respective β values

for Smoker, if the β coefficient for Smoker in the OLS model lies within the confidence interval, they are not significantly different. We construct the confidence interval using the formula below:

$$CI_{95} = (\beta_i - 1.96 * SE, \beta_i + 1.96 * SE) \tag{14}$$

note: $\beta_{Smoker} = 23605.800$ in OLS output

$$\tau = 0.10 -> CI_{95} = (14884.7322 - 1.96 * 366.6492, 14884.7322 + 1.96 * 366.6492) = (14166.1002, 15603.3642)$$

Here we can see the β estimate from the normal linear regression output falls well outside the confidence interval for the 10th quantile. Therefore our β estimate for Smoker in the 10th quantile is significantly different from the OLS model.

$$\tau = 0.50 -> CI_{95} = (27348.7223 - 1.96 * 1186.3560, 1186.3560 + 1.96 * 1186.3560) = (25023.4645, 29673.9801)$$

In this case, the β for Smoker in the OLS model also falls outside the confidence interval for the 50th quantile. Therefore our β estimate for Smoker in the 50th quantile is significantly different from the OLS model.

$$\tau = 0.90 -> CI_{95} = (29.509.1998 - 1.96 * 637.46160, 29.509.1998 + 1.96 * 637.46160) = (28259.78, 30758.62)$$

Lastly, we see that the β for Smoker falls outside of the confidence interval generated for the 90th quantile, so we can conclude that our β for Smoker is significantly different than that of the β produced from the OLS model.

Our interpretation for the Smoker variable would be as follows: individuals who classify as smokers spend an additional $\$14,844.73$ for those who with lower medical insurance premiums (10th quantile), smokers with higher insurance premiums spend $\$29,509.20$ (90th quantile). Simply put, for individuals who smoke the cost of insurance premiums increase as quantiles increase.

R Code:

quantregall< − rq(expenses bmi + smoker + sex + children, tau=seq(0.10, 0.90, by=0.05), data=data1)

quantregplot< − summary(quantregall) plot(quantregplot)

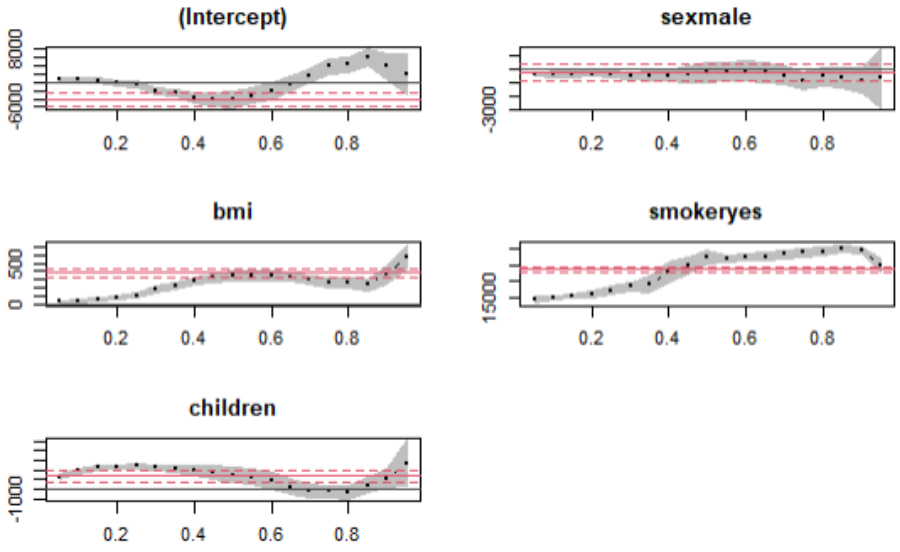

Figure 4: Quantile Regression Plots from R

From the quantile regression plots above in R, each black dot is the slope coefficient for the quantile indicated on the x axis. The solid red lines are the OLS coefficients and the dotted red lines are its respective confidence interval. [13]. Similarly to our calculations from the previous page, plotting all of our predictor variables across every quantile can also gives a general understanding of if the β coefficients generated from quantile regression are significantly different from the OLS model. From Figure 3, we can see that sex maintains within the upper and lower bounds of the OLS regression line, meaning there is no significant difference in β coefficients from the OLS model in any quantile for this variable. Children and BMI typically fall outside or just inside the OLS bounds. We can also see that the plots correspond to our analysis for the Smoker variable, the β coefficients were significantly different from the OLS model across the 10th, 50th, and 90th quantiles, and from the plot above it seems to cross somewhere between the 40th-45th quantile.

4.1.2 Do the Quantile Regression Coefficients Differ Across Quantiles?

R Code:

```
anova(quantreg10, quantreg50, quantreg90)
```

DF	Resid DF	Test Statistic	p-value
8	4006	67.74	<0.0001

Table 7: Anova Output from R: Equality of Slopes between 10th, 50th, 90th Quantile

Further justifying our use of quantile regression, would be to test for a significant difference between the coefficients of different quantiles. Shown by the Anova table above, we can test for a significant difference between the slopes for the 10th, 50th, and 90th quantile. From the table we see that we have $p < 0.0001$, meaning we would reject the null hypothesis that our slopes are the same and conclude that at least one is different.

5 Conclusion

This paper discussed some of core mathematical concepts of both the linear and quantile regression models. The ultimate difference between the two types of statistical analysis is that quantile regression gives a more in depth view of the relationship between the predictor variables and response by allowing us to view the conditional quantiles of the dependent variable, as opposed to only analyzing the conditional mean. We also saw that quantile regression can be a useful alternative to linear regression when core assumptions are violated, such as non-normal distribution of the response/error and constant variance of residuals. The key findings from our first application is that the test scores received variable differed significantly from the OLS estimates across most quantile levels. This was shown through calculating the corresponding model equations and interpreting the coefficients. In our second application, we found that the Smoker variable also differed significantly from the OLS estimates across most quantiles, this was shown through calculating confidence intervals from our quantiles of interest and through our plots of all our independent variables across the conditional quantiles.

References

[1] R. N. Rodriguez and Y. Yao, "Five Things You Should Know About Quantile Regression." SAS Institute Inc., May 2017.

[2] J. J. O'Connor and E. F. Robertson, "Ruggero giuseppe boscovich," Aug 2006.

[3] R. Koenker, "Galton, edgeworth, frisch, and prospects for quantile regression in econometrics," *Journal of Econometrics*, vol. 95, pp. 347–374, 2000.

[4] N.A., "Statistics for Research Projects." Massachusetts Institute of Technology, N.D.

[5] Z. Xiao, "Time Series Analysis: Methods and Applications," *Handbook of Statistics*, 2012.

[6] R. Xi, "Quantile Regression Lecture 1." Peking University, ND.

[7] A. B. Owen, "Lecture 18 Quantile Regression." Standard University, December 2019.

[8] R. Susmel, "Lecture 10 Robust and Quantile Regression." C.T. Bauer College of Business, Spring 2014.

[9] Cuemath, "Anova test,"

[10] "Student study hour v2," 2021.

[11] RStudio Team, *RStudio: Integrated Development Environment for R.* RStudio, PBC., Boston, MA, 2020.

[12] N.A., "Medical Cost Personal Datasets." Kaggle, 2018.

[13] C. Ford, "Getting Started With Quantile Regression." University of Virginia Library, September 2015.