Timo Meiendresch

Quantile Regression Analysis

Methodology and Applications

Module: Seminar in Statistics and Econometrics
Lecturers: Dr. Jörg Breitung, Dr. Dominik Wied

Summer Semester 2018
Cologne

Contents

1 Introduction **3**

2 Literature **5**

3 Methodology **6**

3.1 Quantiles and the Quantile Function 6

3.2 Sample Quantiles and the Conditional Quantile Function 7

3.3 The Quantile Regression Model 9

3.4 Test Procedures 11

4 Applications **14**

4.1 Fama-French Three-Factor Model 14

4.2 Happiness Survey 17

4.3 Wage Regression 19

5 Summary **21**

6 References **23**

1 Introduction

Quantile regression (QR) has become a widely used statistical method to analyse the relationship between dependent and independent variables. Whereas ordinary least squares (OLS) relates covariates to the conditional expectation function,

$$E(y|x) = x'\beta \tag{1}$$

QR links to the conditional quantile function:

$$Q_\tau(y|x) = x'\beta_\tau \tag{2}$$

More precisely, OLS relates a set of covariates x to the expectation of the dependent variable y, leading to a model with a single parameter vector β. This vector contains the estimated average effect of the dependent variable on the response variable y. Thus, least squares regression characterises the relation using the mean as parameter.

In the case of equal sample variation, which is assumed in least squares regression, this may be a sufficient outcome as the rate of change is constant across quantiles. However, in situations of unequal variation, there may exist different rates of change across quantiles, implying different quantile-specific slopes. It can be argued that the single slope in least squares regression delivers an incomplete picture in those kind of situations. In addition, there may exist effects beyond the mean which are not detected or misrepresented using the OLS approach here.

As OLS relates to the conditional mean, QR relates covariates x to the conditional quantile function $Q_\tau(y|x)$, where τ indicates the quantile of the dependent variable y. A usual approach is to estimate a set of multiple parameter vectors β_τ for different quantiles. Thus, relationships beyond the mean can be described, giving a more complete view of the relations between dependent variable and covariates. In the setting of unequal variances this leads to multiple slopes for different conditional quantile functions. This approach has been proven useful in situations where "extreme" observations are of interest, i.e. whenever the effect of x on the tails of y is relevant. Further advantages include the robustness of this approach and that no distributional assumption for the error term is required.

Illustration

To illustrate the utility of such a model, consider the example of figure (1), where the relation between the dependent variable "Sale Price" and independent variable "Living Area" is illustrated. Whereas the conditional mean does not capture the heteroscedastic

structure in the data, multiple QR functions (at the bottom of figure 1) seem to deliver a more detailed picture of the functional relationship between errors and covariates.

Figure 2 presents a visual summary of the parameters. It can be seen that the slope of the QR functions for the variable "Living Area" increases from the lower to the upper tail of the distribution y as τ increases. The slope parameter varies across quantiles and differs significantly from the OLS estimate. A general observation is an increase in sampling variation leading to the observed pattern of our conditional quantiles. In particular, the effects on the tails of the distribution differ from those in the center of the distribution. Those effects beyond the mean are commonly of particular interest QR analysis.

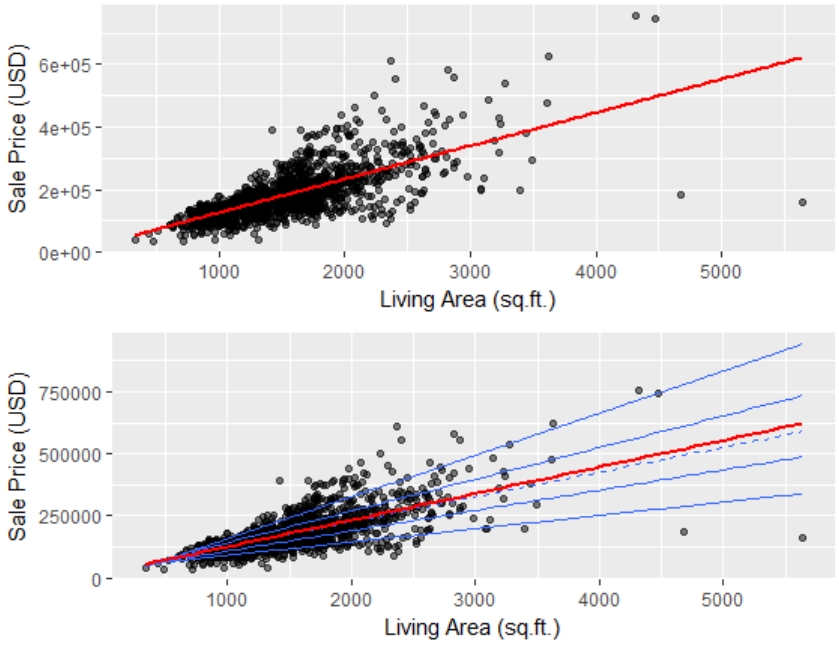

Figure 1: Illustration of conditional mean (top and bottom) in red and conditional quantiles (bottom) for $\tau \in \{0.1, 0.25, 0.75, 0.9\}$ in blue. The median regression line ($\tau = 0.5$) is indicated by the dashed blue line.

The outline of this paper is as follows. Section 2 gives an overview of theoretical and applied QR in literature before section 3 outlines the methodological framework. Section 4 features some applications using the framework that I have presented. Results of the analyses will be presented, together with some remarks on their interpretation before concluding in section 5.

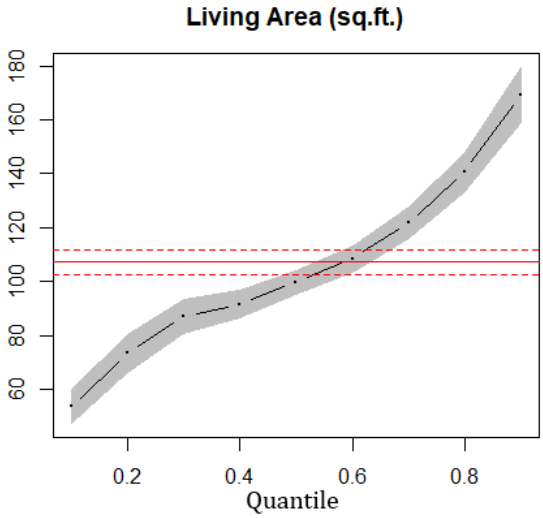

Figure 2: Visual summary of QR coefficients. Quantiles τ are shown on the x-axis with respective values on y-axis. Dependent variable is "Sale Price".

2 Literature

Literature on Methodology

Modern QR was introduced in Koenker and Bassett (1978, 1982), presenting a framework to estimate conditional quantiles using asymmetrically weighted residuals. Since then, it has sequentially been extended adding primarily inferential procedures to the framework. In section 3 I will focus on some techniques which have been shown to add particular value in applied analysis rather than focusing on the theoretical details.

It should be noted that even some basic inferential procedures are highly complex and can barely be covered in a single paper or even a single book. The focus of this paper is to deliver a basic understanding of the main ideas and how regression quantiles can be used in an applied setting. For details of the presented techniques I refer to the respective paper whereas in general, I recommend to refer to the books Koenker (2005) and Koenker et al. (2017), which cover all the details and additional extensions.

Literature on Applications

Modeling conditional quantile functions in empirical work has been widely used in situations where "extreme" observations are of interest and the mean may not be a good representative for the lower part of the distribution. Accordingly, a vast majority of the empirical literature focuses on the interactions between a set of independent variables and their effect on the tails of the variable of interest. Some examples are given below.

Focusing on the lower part of the wage distribution, Buchinsky (1994) claims in a widely acknowledged study that conditional mean methods may misrepresent the effects of education and experience on the lower part of the wage distribution. Those effects appear to be different from what has been previously been presented using OLS. Subsequent studies (e.g. Buchinsky, 1998) extended this approach to wage and gender inequalities indicating heterogenous effects not detected before.

A paper by Allen, Singh and Powell (2011) applied QR on Stock market returns using the the Fama-French three-factor model based on the highly cited work of Fama and French (1993). Emphasising the importance of extreme events to risk analysis their results indicate significant differences among quantiles for the effect of the three factors on different parts of the return distribution. They argue that QR seems to be more efficient than OLS in analysing those extreme events.

In the area of Development Economics Abrevaya (2001) uses QR to complement his analysis of low birthweights. In particular, his results indicate high disparities between children born to black and white mothers at the lower tail of the distribution, which are significantly larger than for the conditional mean. Moreover, his findings show that the clear weight differences between boys and girls almost vanish for the lower tail of the distribution among other findings.

3 Methodology

The following section introduces the framework of QR and covers some inferential procedures. In a first step, I will define the required concepts such as quantiles and the quantile function before introducing the reformulation of the optimisation problem which is the crucial point to estimate conditional quantiles.

3.1 Quantiles and the Quantile Function

The quantile q_τ, with $\tau \in (0, 1)$, is the value that splits a probability distribution or sample into a lower and upper part with the property that a fraction τ is below or equal to the value, whereas a fraction of $1 - \tau$ is above this value. In the one-dimensional case the sample quantile can be obtained by arranging the N-values in ascending order, where the quantile q_τ is the $(N \cdot \tau)$th smallest value. This can be formalized in the following way with Y being a random variable with distribution function F. The quantile q_τ is the value for which the following equations hold:

$$F(y) = P(y \leq q_\tau) \geq \tau \qquad and \qquad F(y) = P(y \geq q_\tau) \geq 1 - \tau, \tag{3}$$

In the case of the standard normal distribution with $Y \sim N(0,1)$, we receive $F(0) = P(y \leq 0) \geq 0.5$, $F(1.645) = P(y \leq 1.645) \geq 0.95$ and $F(1.96) = P(y \leq 1.96) \geq 0.975$, corresponding to the quantiles $q_{0.5} = 0$, $q_{0.95} = 1.65$ and $q_{0.975} = 1.96$.

Some commonly used quantiles are the median, $\tau = 0.5$ as measure of central tendency as well as the lower and upper quartiles, $\tau = 0.25$ and $\tau = 0.75$. A main feature of quantiles, which also applies to regression quantiles, are their robustness and invariance to outliers.

It can already be seen that there is some resemblance in the definition of equation (3) and the cumulative distribution function given by

$$F(y) := P(Y \leq y), \tag{4}$$

Using the definition for the CDF we can define the quantile function, which is closely, as the function $Q : (0,1) \to \mathbb{R}$ given by

$$Q(\tau) = inf\{y \mid F(y) \geq \tau\}, \qquad \tau \in (0,1), \tag{5}$$

Hence, the quantile function maps a probability τ to the minimum value of the realization y, where the CDF is greater or equal to the probability τ. It can be seen that there is an inverse relation between CDF and the quantile function which can be expressed as

$$Q(\tau) = F^{-1}(\tau), \tag{6}$$

assuming a strictly increasing and continous CDF. The quantile function maps the probability τ to its respective quantile and the cumulative distribution $F(q_\tau) = \tau$ maps the quantile to its respective probability.

3.2 Sample Quantiles and the Conditional Quantile Function

Conventionally, the sample quantile can be obtained by ordering the respective values. In multivariate cases this is rather inconventient. As presented by Koenker and Bassett (1978) it is possible to equivalently express the sample quantile q_τ as an optimisation problem. Accordingly, the result for the quantile of the sample $y_1, y_2, ..., y_n$ can be formalized as the solution to the optimisation problem given by

$$\hat{Q}_y(\tau) = arg \min_Q \sum_{i=1}^{n} \rho_\tau(y_i - Q), \tag{7}$$

where ρ_τ is the weight function, also known as "check-function" (see figure 3) given by

$$\rho_\tau(u) = u(\tau - \mathbf{1}_{(u<0)}) \tag{8}$$

with $\mathbf{1}_{(u<0)}$ as indicator function. In this setting, the τ-th quantile becomes a new interpretation as the minimizer of the loss function ρ_τ, which asymmetrically weighs our sample error ($u = y_i - Q$). For the special case $\tau = 0.5$ we receive the median regression or least absolute deviation (LAD), where the optimisation problem simplifies to

$$\hat{Q}_y(\tau = 0.5) = min \sum_{i=1}^{n} |y_i - Q| \tag{9}$$

The further away the quantile deviates from the median the higher is the asymmetry of the loss function which can be seen in figure 3. In a last step we need to specify

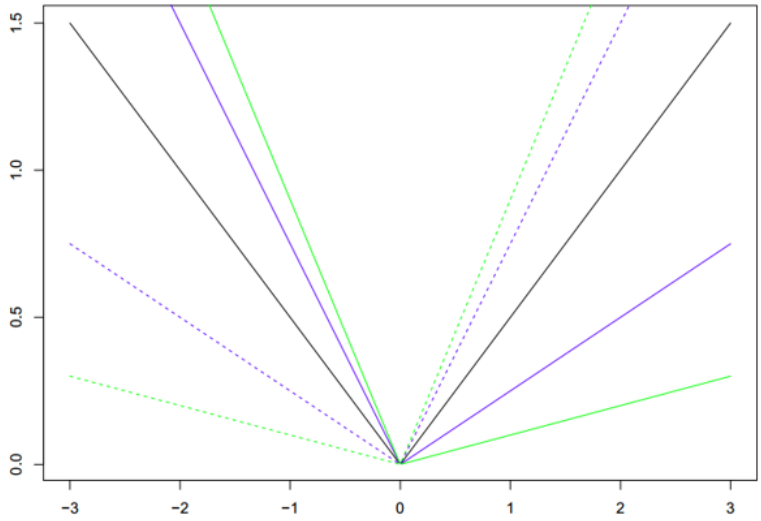

Figure 3: "Check function" illustrating the loss for different τ.

our quantile regression model. Assuming that the errors are stochastically independent and the weighted average of the errors is zero, (i.e. $F_u(0) = \tau$), we can express the linear quantile model for a given sample as

$$Q_y(\tau|x) = x_i'\beta_\tau \tag{10}$$

Replacing Q with this term in equation (9), the τth conditional quantile can be estimated using the optimisation problem given by

$$\hat{\beta}_\tau = arg \min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta_\tau), \tag{11}$$

where ρ_τ is again the quantile loss function of equation (8) with $u_i = y_i - x_i'\beta_\tau$. This optimisation problem is a nondifferentiable function and, unlike least squares regression, does not have a closed form solution. However, linear programming algorithms (e.g. simplex method) can be used to determine the QR functions efficiently.

3.3 The Quantile Regression Model

The estimation of the parameter vector yields

$$Q_y(\tau|x) = x_i'\hat{\beta}_\tau, \tag{12}$$

$$= \hat{\beta}_{0,\tau}x_0 + \hat{\beta}_{1,\tau}x_1 + ... + \hat{\beta}_{k,\tau}x_k \tag{13}$$

indicating the quantile-specific parameter vectors of x. Interpretation of our parameters remains the same as those of other linear models with the difference that they now indicate the rate of change for a conditional quantile.

QR can be motivated based on the form of the underlying model. As stated in the Introduction, OLS implicitly assumes equal variations leading to a constant slope across quantiles as the effect of x on y is equal for all parts of the distribution. This is also known as a location shift model (see top left of figure (4)), where the errors are assumed to be independent of x. In this type of model only the location of our conditional distribution y shifts with differing quantiles, whereas the shape remains the same. QR then yields a set of parallel hyperplanes as illustrated in figure 4. The slope parameter is equal for all quantiles and only the intercept term varies along the distribution. This is in contrast to a location scale shift model, where unequal variation leads to more than one slope for different quantile regressions. Figure 5 contains simulated data of a log-normal distribution with an additional interaction between covariates and errors leading to a fanning out of the distribution. The key aspect is that the errors are now an unknown function of the covariates caused by non-symmetric distributions, outliers, or other complex interactions between errors and covariates.

A common approach dealing with heteroscedastic errors or non-normal distributions include among others the use of robust standard errors, Generalized Linear Models or Weighted Least Squares. These models try to capture the underlying functional relationships between errors and covariates using distributional assumptions or weighting schemes. These approaches lead to a robust estimate of the conditional mean. QR on the other hand does not require any distributional assumptions and can be considered as a "robust" alternative. In practice, however, this property is rarely the primary reason to employ QR as there are other, more efficient, ways to deal with those situations. The main use in empirical literature has been for applications beyond the mean, where

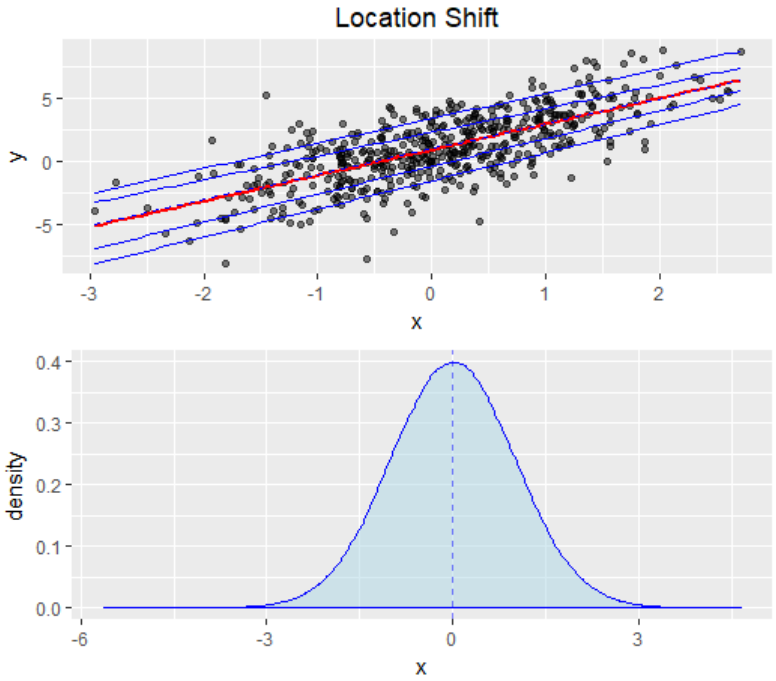

Figure 4: Conditional quantiles (blue) and conditional mean (red) for a Location shift model with simulated data. Kernel density is estimated using a gaussian kernel.

the effect of x on the tails of the distribution of y is of interest. Thus, relationships between conditional quantile and independent variables may be uncovered which were either too weak or differed significantly from the conditioanal mean counterpart.

Properties and Standard Errors

Since the quantile regression does not have a closed form solution like least squares, the asymptotic distribution of $\hat{\beta}_\tau$ is a little bit more complicated. Following Koenker (2005) the QR is asymptotically normal taking the following form in non-iid settings:

$$\sqrt{N}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N(\mathbf{0}, \tau(1-\tau)H_n^{-1}J_nH_n^{-1}), \tag{14}$$

where

$$J_n(\tau) = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \tag{15}$$

$$\tag{16}$$

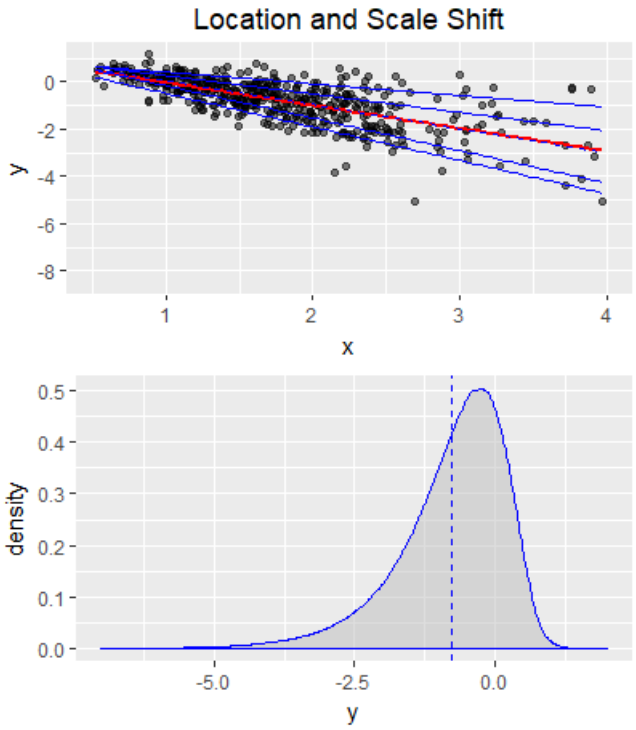

Figure 5: Conditional quantiles (blue) and conditional mean (red) for a location scale shift model with simulated data. Kernel density estimate is based on a gaussian kernel.

and

$$H_n(\tau) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i \, x_i' \, f_i(Q_y(\tau|x)). \tag{17}$$

Here, $f_i(Q_y(\tau|x))$ indicates the conditional density of y which is rather complicated to estimate. Two methods for estimating the matrix H_n and receiving standard errors include the so-called Sandwich estimator (Hendricks and Koenker, 1991) and the approach described by Powell (1991) based on kernel density estimation. A probably easier approach is the use of bootstrapped standard errors for which also different variations exist. Another frequently used method is to compute intervals based on the so-called rank inversion method outlined for example in section 3.5.5. of Koenker (2005).

3.4 Test Procedures

As QR has emerged as a powerful complement to least squares regression, inferential procedures and assessment of the model have become an integral part of research in this area. Unfortunately, inference for quantile regression often involves elements of non-

parametric density estimation making the subject more complicated than comparable counterparts in OLS.

Using the standard errors of the previous section gives a first impression of the precision and to assess the significance of our covariates. One of the next questions that naturally arises is whether the estimated slope parameters are equal across quantiles and if they conform with the implicit assumption of equal variation. Note that this may also be considered as a test of location shift hypothesis. A conventional way to answer this question is the Wald test approach for QR as outlined in Koenker and Bassett (1982b). The Wald test is based on the asymptotic normality outlined in section 3.3 and can be used to test whether the slopes of a set of conditional quantile functions are equal. Thus, for a general linear hypothesis consider the vector of

$$\xi = (\beta(\tau_1)', ..., \beta(\tau_m)')', \tag{18}$$

where our hypothesis takes the form

$$H_0 : R\xi = r \qquad vs. \qquad H_1 : R\xi \neq r. \tag{19}$$

A test statistic is then given by

$$T_n = n(R\hat{\xi} - r)' \left[RV^{-1}R'\right]^{-1}(R\hat{\xi} - r), \tag{20}$$

with V_n as [mp x mp] matrix

$$V_n(\tau_i, \tau_j) = [t_i \wedge \tau_j - \tau_i\tau_j]H_n(\tau_i)^{-1}J_n(\tau_i, \tau_j)H_n(\tau_j)^{-1} \tag{21}$$

Under the null hypothesis the test statistic is asymptotically χ^2_q-distributed where q is the rank of our test matrix R. Note that the test statistic requires the estimation of the matrix H_n similar to equation (17). As mentioned before, this can be done using the approach of Hendricks and Koenker (1991). This test can be used in several variations. Besides testing the equality of slopes, joint equality tests can be implemented as well as tests of whether a nested model with different covariate specification is preferred to the less restricted one. Moreover, nonlinear hypotheses on the vector of ξ can be tested in a slightly adjusted framework.

To test for a location scale model a somewhat more complex test is required, namely the Khmaladze test presented in Koenker and Xiao (2002). This test is used for the hypothesis that a linear model is of the location scale shift form, where the appearance of unknown nuisance parameters adds to the complexity of this approach. The test is based on the Doob-Meyer Martingale transformation which has been proposed by

Khmaladze (1981) and adapted to QR.

For more details and additional well-developed theory of asymptotic inference I refer to Koenker (2005). In addition, resampling methods and corresponding test procedures are extensively covered in Koenker et al. (2017).

Goodness of Fit

Another helpful concept is the of goodness of fit as outlined in Koenker and Machado (1999). Whereas R^2 provides a global measure for the fit of a least squares model, the Koenker-Machado measure is a local quantile-specific goodness of fit. The central idea of this measure is to compare the weighted absolute sum of residuals of an intercept-only model and a fully specified model. First, consider our conditional quantile function as the partition of

$$Q_{y_i}(\tau|x) = x'_{i,0}\beta_{\tau,0} + x'_{i,1}\beta_{\tau,1} + ... + x'_{i,k}\beta_{\tau,k} \tag{22}$$

where $x'_{i,0} = 1$ indicates the intercept of our model. In a next step, estimate the restricted (intercept only) as well as the unrestricted model model as

$$\hat{\beta}_\tau = \min_\beta \sum \rho_\tau(y_i - x'_i\beta) \tag{23}$$

and

$$\tilde{\beta}_\tau = \min_{\beta_1} \sum \rho_\tau(y_i - x'_{i,0}\beta_0). \tag{24}$$

Denoting the weighted sum of residuals of the restricted and unrestricted form of the model by $\tilde{V}_\tau(\tilde{\beta})$ and $\hat{V}_\tau(\hat{\beta})$, the goodness-of-fit criterion is defined as

$$R^{KM} = 1 - \hat{V}_\tau(\hat{\beta})/\tilde{V}_\tau(\tilde{\beta}), \tag{25}$$

with $\hat{V}_\tau \leq \tilde{V}_\tau$, ensuring that the measure is between 0 and 1. The R^{KM} is a measure for the relative success of the quantile regression models representing a local measure for the fit of our intended model. This measure is helpful in assessing the quality of the fit of a conditinoal quantile on different parts of the distribution and is also the basis for various test procedures as proposed in Koenker and Machado (1999).

4 Applications

In the following chapter I will demonstrate previously introduced techniques based on three different datasets. The main application is to use QR for estimating the Fama-French Three-Factor model and investigate abnormal returns. Additional applications are included to illustrate the QR approach.

4.1 Fama-French Three-Factor Model

One of the most commonly cited work in finance is the paper of Fama and French (1993), where the authors present what has become known as the Fama-French Three-Factor model. In short, this model extends the traditional capital asset pricing model (CAPM) by two factors to explain asset returns. These monthly factors are available through the Kenneth R. French data library and are updated on a regular basis. The dependent variable is the monthly excess return of the Dow Jones Industrial Index which also used in Allen et al. (2011) whose results I have introduced in section 2.

Theory

The basic CAPM model uses the risk factor β of an asset to explain its return. Assuming a linear relationship this can be expressed as:

$$r_i = r_f + (r_m - r_f)\,\beta, \tag{26}$$

where r_i is the expected return of an asset i and $r_m - r_f$ the market return over a risk-free asset (usually a government bond with long maturity). The Fama-French model adds two factors to account for the size of assets and book-to-market ratios yielding

$$r_i = r_f + (r_m - r_f)\,\beta + SMB \cdot \beta_s + HML \cdot \beta_h, \tag{27}$$

where SMB (Small Minus Big) represents the average return of small versus big companies and HML (Hig Minus Low) the average return of high value companies measured by their book-to-market ratio versus those with low book-to-market ratios. This model can be reformulated as

$$r_i - r_f = \alpha + (r_m - r_f)\,\beta_\tau + SMB \cdot \beta_{s,\tau} + HML \cdot \beta_{h,\tau} + \epsilon_{i,\tau}, \tag{28}$$

where $r_i - r_f$ is the excess return for asset i.

Results

Table 1 presents the QR and OLS results. The QR estimates are highly significant and lie around the OLS estimates. Figure 6 presents a visual summary of the QR coefficients, where the solid line represents the point estimates of the coefficient for τ. The shaded grey area depicts a 90 percent confidence interval and the solid red line corresponds to the OLS estimate with 90 percent confidence interval (red dotted line).

	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
(Intercept)	-0.0167***	-0.0091***	-0.0013***	0.0069***	0.0129***	-0.0012***
	(0.0001)	(0.0001)	(0.0008)	(0.0000)	(0.0013)	(0.0007)
Mkt – RF	0.9789***	0.9873***	0.9684***	0.9373***	0.9146***	0.9583***
	(.0206)	(.0212)	(.017)	(.0199)	(.0276)	(.0156)
SMB	-0.2218***	-0.229***	-0.256***	-0.2647***	-0.2702***	-0.2246***
	(.0309)	(.0264)	(.0233)	(.0197)	(.042)	(.0222)
HML	0.0863**	0.1123***	0.0815***	0.1074***	0.132**	0.1324***
	(.0329)	(.0245)	(.026)	(.028)	(.0451)	(.0239)
R^2						0.9055
Koenker-Machado R	0.7261	0.7043	0.6851	0.6813	0.6691	

Standard errors based on Hendricks and Koenker (1991) in parentheses.

Koenker-Machado R is based on Koenker and Machado (1999).

Significant codes: '*': significant at 5% level (2-sided); '**' 1% level; '***': 0.1 % level

Table 1: Summary Statistics of QR and OLS. Dependent variable "Excess Return".

It seems that the coefficients for the market risk factor (Mkt - RF) are decreasing from the lower to the upper tail as can be seen on the second panel of figure 6. A similar pattern appears to be present for our factor SMB, whereas the QR coefficients for HML seem to increase. It may be of interest that these observations are consistent with Allen et al. (2011). The R^{KM} (Koenker-Machado goodness-of-fit) is slightly decreasing with higher values for τ, indicating a slightly decreased quality of the fit for upper quantiles.

However, based on the Wald test, as shown in table 2, we cannot confirm the initial observations. The results of this test indicate that there is no significant inequality of slopes for our specified model.

To test the location scale hypothesis, I have used the Khmaladze test as an alternative to check for a location shift and, in addition to check the location scale shift hypothesis. The results are shown in figure 7. The results seem to confirm the Wald test for the location shift hypothesis. In addition, the joint location scale shift hypothesis cannot be rejected (critical value of 16.00 at 0.01 level). Note however that the factor HML is significant at the 0.05 level indicating a location-scale effect for this variable.

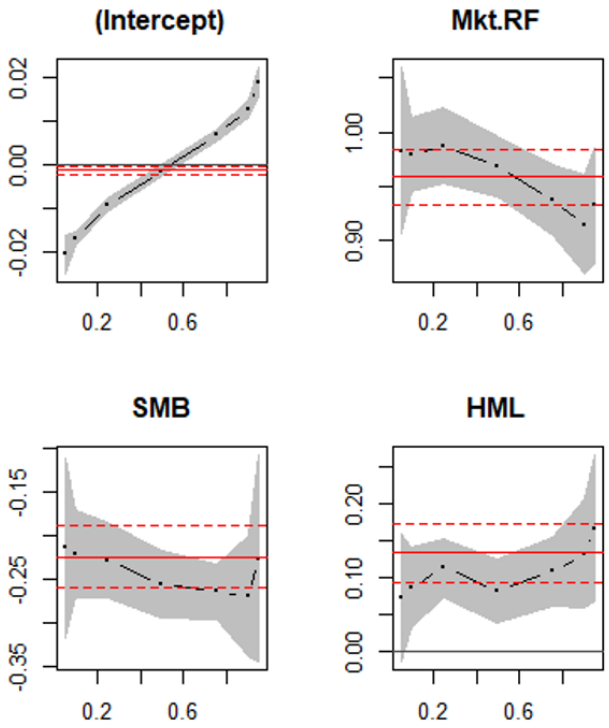

Figure 6: Visual summary of the quantile regression results for the Fama-French three-factor model.

Wald test ($\tau = \{0.1,0.5,0.9\}$)		
	F value	P-value
Mkt-RF	2.2503	0.1058
SMB	0.6758	0.5089
HML	0.6597	0.5172

Table 2: Wald Test for the Fama-French three-factor model

Conclusion

The results could not confirm clear patterns of heterogeneous quantile-effects, neither for the Dow Jones Index, for the S and P 500 (results not included), nor for the NASDAQ (results not included). Allen et al. (2011) found weak heterogeneous quantile-effects at the lower tail of the distribution. Whereas, those results could be replicated for their specific period of time (January 2002 - May 2009), I I did not find any indication that could be transferred for other time periods.

Khmaladze Test		
	Location Shift	Location Scale Shift
Mkt-RF	1.60	0.83
SMB	1.51	1.05
HML	0.82	2.18
Joint Effect	3.80	3.69

Figure 7: Khmaladze test for the Fama-French three-factor model. Critical values for individual variables are 1.923 at 0.05, and 2.42 at 0.01 level. The joint test at 0.01 level has a critical value of 16.00.

4.2 Happiness Survey

The following analysis is based on the data of the German Socio-Economic Panel (SOEP) and relates happiness to a set of socioeconomic factors. Of particular interest of this study is the lower tail of the happiness distribution. This can be motivated by the steady rise of research regarding people suffering from unhappiness and depressive states. The study of how these factors affect the lower tail of the happiness distribution may therefore deliver valuable insights.

Results

Table 3 shows the results of a regression model consisting of 10 socioeconomic factors and the dependent variable of happiness. Figure 8 contains a visual summary of six selected factors from this regression. Note that the quantile regression coefficients for "health", "educ", "unemployed", "married" and "migback" lie outside the confidence intervals of the OLS estimator for the lower tail of the distribution. This is a first indication of heterogeneous quantile effects for the mentioned variable. In the first panel for "health" we can see that the coefficient decreases with increasing quantile, indicating that the relation between happiness and health is stronger for the lower tail of the distribution. Similar observations apply to the variables "educ" and "married", where the coefficients for the lower quantile differs significantly from the OLS results. Moreover, for the variable "unemployed" the effect seem to exert a strong negative effect on the lower quantiles whereas the effect on the upper quantiles almost disappears as can be inferred from the fourth panel in figure 8.

The results of the Wald test (table 4), which jointly rejects the location shift hypothesis and indicates heterogeneous slope parameters confirms these observations for the aforementioned variables, except for the variable "female".

	Quantile Regressions					
	10%	25%	50%	75%	90%	OLS
(Intercept)	-1.4442**	0.7732	4.3314***	6.6503***	8.3984***	3.5466***
	(0.674)	(0.4972)	(0.2972)	(0.391)	0.4080)	(0.2996)
health	0.534***	0.5044***	-0.3866***	0.2894***	0.2349***	0.3785***
	(0.0133)	(0.0105)	(0.0068)	(0.0073)	(0.009)	(0.0063)
educ	0.4858***	0.3085***	0.0502	-0.0219	-0.1556***	0.1676***
	(0.0979)	(0.0701)	(0.041)	(0.0482)	(0.0564)	(0.0425)
$educ^2$	-0.0164***	-0.0102***	-0.0017	0.0003	0.0046**	-0.0059***
	(0.0036)	(0.0025)	(0.0015)	(0.0017)	(0.002)	(0.0016)
hinc	0.001***	0.001***	0.001***	0.0001***	0.0001***	0.001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
unemployed	-1.3753***	-0.7730***	-0.5962***	-0.3060***	-0.0860	-0.5719***
	(0.2087)	(0.1209)	(0.0831	(0.0722)	(0.1)	(0.0617)
age	-0.0364***	-0.0232***	-0.0212***	-0.0302***	-0.0231***	-0.029***
	(0.008)	(0.0061)	(0.0831)	(0.0048)	(0.0058)	(0.0047)
age^2	0.0004***	0.003***	0.003***	0.0004***	0.0003***	0.0004***
	(0.001)	(0.001)	(0.00)	(0.00)	(0.0001)	(0.0000)
female	0.1025**	0.0989**	0.0983***	0.1386***	0.1314***	0.1167***
	(0.0555)	(0.0387)	(0.0215)	(0.0264	(0.0289)	(0.0264)
married	0.4643***	0.3702***	0.279***	0.2475***	0.1670***	0.306***
	(0.0618)	(0.0449)	(0.0288)	(0.0307)	(0.0350)	(0.0299)
migback	0.0504	0.0533	0.1404***	0.2677***	0.3976***	0.209***
	(0.0646)	(0.0464)	(0.027)	(0.0333)	(0.0368)	(0.03)
R^2						0.2817
Koenker-Machado R	0.277	0.2972	0.2309	0.1853	0.1547	

Standard errors based on Hendricks and Koenker (1991) in parentheses.

Koenker-Machado R is based on Koenker and Machado (1999).

Significant codes: '*': significant at 5% level (2-sided); '**' 1% level; '***': 0.1 % level

Table 3: Summary of QR and OLS results. Dependent variable is "happiness".

Conclusion

Applying QR to analyse the lower tail of the happiness distribution seem to add additional insights. Whereas the general pattern of results remains comparable to OLS, the estimated conditional quantile functions show substantial disparity for the effects of the variables on the lower part of the distribution. In particular, variables "health", "educ" and "married" appear to exert a stronger absolute effect and seem to be underestimated using OLS. In addition, "unemployed" appears to have a stronger negative effect on the lower tail leading to the impression that OLS underestimates the negative effect of unemployment on unhappy people. An additional finding indicates that the effect of high income ("hinc") and gender ("female") seem to be consistent along quantiles.

In brief, it seems as if OLS in this analysis does not capture the absolute size of the effects appropriately. Whereas the general pattern is similar for QR and OLS, the latter seems to over- or underestate the effects of certain variables on the lower part of the happiness distribution. Hence, QR seems to deliver a more nunanced picture of

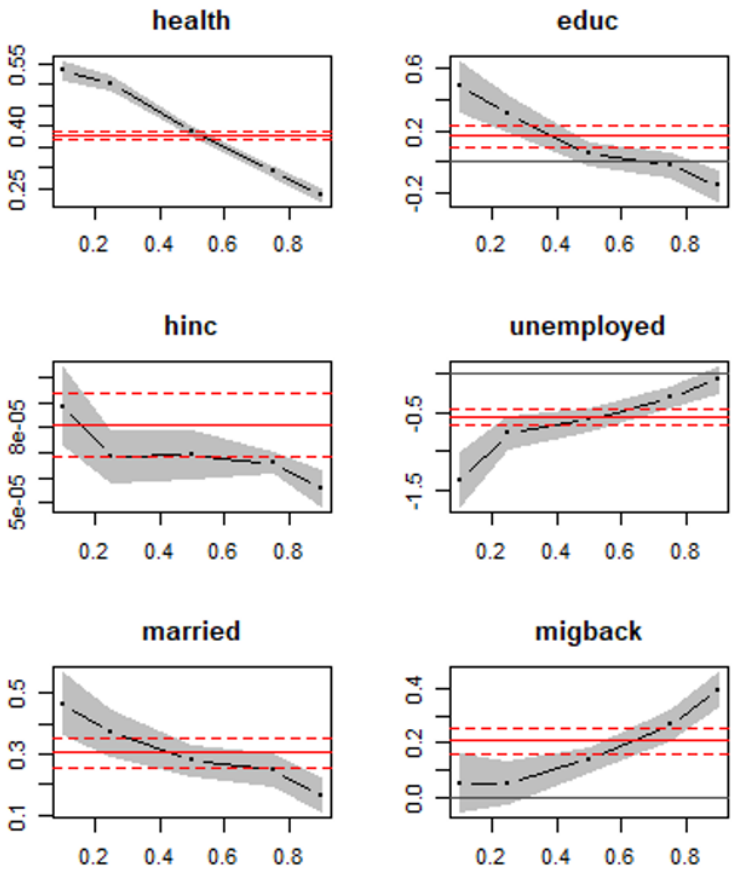

Figure 8: Visual summary of the QR and OLS results.

the relationship between dependent and independent variables.

4.3 Wage Regression

The last example covers a wage regression. The following analysis is based on survey data and examines the relationship between wage and three socioeconomic variables. The main focus of this analysis are whether heterogeneous effects on the wage distribution are present, misrepresented or not detected using OLS.

Results

Results of the QR and OLS results are summarized in table 5 and figure 9. It can be seen that coefficients for schooling and the "male" dummy variable increases steeply with increasing quantiles, whereas the coefficients for experience seem to decrease. In addition, the Wald test indicates that the location shift interpretation seems to be implausible in this case. Accordingly, it can be concluded that the slope parameters

Wald Test		
	F-value	P-value
health	142.01	0.0000***
educ	10.55	0.0000***
educ^2	8.64	0.0000***
hinc	3.63	0.0058**
unemployed	10.43	0.0000***
age	2.49	0.041*
age^2	2.99	0.018*
female	0.81	0.5179
married	5.47	0.0002***
migback	13.26	0.0000***
Significant codes: '*': significant at 5% level (2-sided); '**' 1% level; '***': 0.1 % level		

Table 4: Results of the Wald test.

for schooling and the dummy-variable are not equal, leading to quantile-specific effects on the dependent variable. This does not hold for the experience variable.

	Quantile regressions					
	10%	25%	50%	75%	90%	OLS
(Intercept)	-3.2522***	-3.3691***	-3.1815***	-3.0379***	-2.128**	-3.38***
	(0.4531)	(0.381)	(0.4394)	(0.5729)	(1.0351)	(0.465)
school	0.3686***	0.4598***	0.564***	0.7263***	0.8196***	0.6388***
	(0.0312)	(0.0264)	(0.0314)	(0.039)	(0.0758)	(0.0328)
exper	0.1621***	0.1673***	0.1596***	0.1187***	0.0943**	0.1248***
	(0.0227)	(0.019)	(0.0225)	(0.0287)	(0.265)	(0.0238)
male	0.603***	0.8953***	1.1654***	1.605***	2.0043***	1.344***
	(0.1076)	(0.0932)	(0.1081)	(0.143)	(0.265)	(0.1077)
Standard errors based on Hendricks and Koenker (1991) in parentheses.						
Significant codes: '*': significant at 5% level (2-sided); '**' 1% level; '***': 0.1 % level						

Table 5: Summary of QR and OLS results. Dependent variable is "log wage".

Conclusion

The results seem to suggest unequal variation along different quantiles, leading to quantile-specific effects of the variables. The effect of schooling on the lower part of the distribution seems to be smaller for the lower part of the distribution. Specifically,

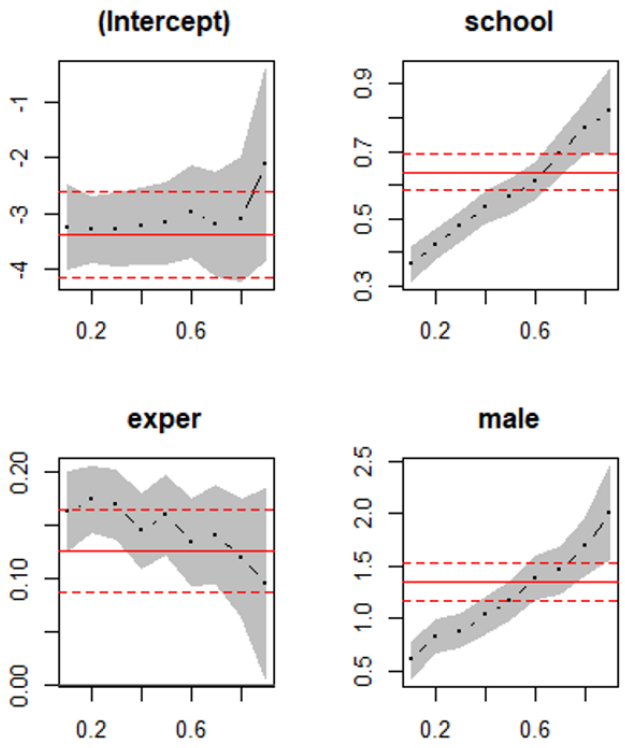

Figure 9: Visual summary of the QR and OLS results. Dependent variable is "log wage".

the effect of schooling and being male on the lower tail appears to be much weaker than OLS suggests. This indicates that for higher incomes the absolute difference of being male is also higher than for lower incomes and that the effect of schooling on income increases with higher income as well. It should be noted however, that the results illustrate an obvious functional relationship between higher sampling variation at the upper quantiles of the wage distribution. Consequently, this analysis should be taken with caution as other statistical methods or appropriate transformations appear to be preferred to the QR approach here.

5 Summary

This paper introduced the non-parametric QR approach as a method to analyze the relationship between dependent and independent variables using asymmetrically weighted errors. QR can be motivated based on the concepts of a location shift and location-scale shift model, corresponding to models with equal and unequal variation. Unequal variation due to a functional relationship between error and covariates leads to quantile-specific effects exerting different effects on different parts of the underlying distribution.

Wald Test		
	F value	P-value
School	16.23	0.0000
Exper	0.88	0.4723
male	10.33	0.0000
Joint Test	9.3589	0.0000

Table 6: Wald test results for QR.

To test these hypotheses I have introduced two test procedures, a version of the Wald test and the Khmaladze test. Whereas the Wald test can be used to test for the inequality of slopes of different quantiles, which is equivalent of testing the location shift hypothesis, the Khmaladze test uses a different procedure which enables to additionally test the location-scale shift hypothesis. For evaluating the fit of the model a local goodness of fit measure has been presented.

In empirical analysis, QR has emerged as a useful complement to the least squares method, particularly if the tails of the distribution are of primary interest for the investigation. By estimating conditional quantiles, effects at the tails may be detected which may be different or not detected using OLS. Note however that other techniques which incorporate the structure of the error terms or non-normal distributions are frequently more efficient if a robust measure of central tendency is desired. In case the functional relationship between error term and covariates can be clearly described, other methods should be considered.

6 References

[1] Abrevaya, J. 2002. "The effects of demographics and maternal behavior on the distribution of birth outcomes". In Economic applications of quantile regression, 247-257. Physica, Heidelberg.

[2] Allen, D. E., and S.R. Powell (2011). "Asset Pricing, the Fama-French Factor Model and the Implications of Quantile Regression Analysis. Financial Econometrics Modeling: Market Microstructure, Factor Models and Financial Risk Measures, 176-193. Palgrave Macmillan, London.

[3] Buchinsky, M. (1994). "Changes in the US wage structure 1963-1987: Application of quantile regression". Econometrics: Journal of the Econometric Society, 404-458.

[4] Buchinsky, M. (1998). "Recent advances in quantile regression models: a practical guideline for empirical research". Journal of human resources, 88-126.

[5] Cameron, A. C. and Trivedi, P. K. (2005). "Microeconomoetrics: methods and applications". Cambridge university press.

[6] Gutenbrunner, C. and Jurecková, J. (1992). "Regression rank scores and regression quantiles". The Annals of Statistics, 305-330.

[7] Hendricks, W. and R. Koenker (1991), "Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity". Annals of Mathematical Statistics 21, 309-310.

[8] Khmaladze, È. V. (1981). "Martingale approach in the theory of goodness-of-fit tests". Teoriya Veroatnostei i ee Primeneniya, 26(2), 246-265.

[9] Koenker, R. and K.F. Hallock (2001). "Quantile Regression." The Journal of Economic Perspectives 15(4), 143-156.

[10] Koenker, R. and G. Bassett Jr. (1978). "Regression quantiles". Econometrica: Journal of the Econometric Society, 33-50.

[11] Koenker, R. and G. Bassett (1982). "Robust Tests for Heteroscedasticity Based on Regression Quantiles". Econometrica, 50(1), 43-61.

[12] Koenker, R., Chernozhukov V., He, X., and Peng, L. (2017). "Handbook of Quantile Regression. CRC Press.

[13] Koenker, R. and J.A. Machado (1999). "Goodness of fit and related inference processes for quantile regression". Journal of the american statistical association, 94(448), 1296-1310.

[14] Koenker, R. and Xiao, Z. (2002). "Inference on the quantile regression process". Econometrica, 70(4), 1583-1612.

[15] Fahrmeir, L., T. Kneib, S. Lang and B. Marx (2007). "Regression". Springer-Verlag Berlin Heidelberg

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht.

Köln, den 31. Juli 2018