

Quantile Regression on a system of ODEs

Adam Macbale

May, 2019

Introduction

Abstract

It is very common to use Ordinary Least Squares (OLS) regression and the conditional mean function to make our predictions even when in reality we should rarely expect our data to be so well behaved as to lend themselves to the assumptions of OLS. In this paper, we briefly introduce Quantile Regression (QR) and then apply it to a system of ordinary differential equations describing the population dynamics between Canadian lynx and snowshoe hares during the period 1845-1935. We use the `quantreg` package in `R` to show how QR can reveal nuances that OLS will completely miss and how QR can assess and detect heterogeneity in our data.

Motivation

In almost any inferential analysis we come across, we will see reports of the mean response of some dependent variable relative to a unit change in some set of independent variables. In all of our studies and training at the university regarding inference, we primarily and somewhat exclusively deal with the average, with the conditional or unconditional mean function. Most of the time, for our purposes, this is sufficient given what we aim to do and given what we do or don't know about our data sets. However, I myself can recall several occasions in my early coursework where I couldn't help but wonder if there might be some other ways to speak about our data. I recall wondering how and why we had quietly agreed that the mean response was sufficient to making sound or complete analyses regarding our questions. When I could muster the courage to vocalize this concern and risk displaying this ignorance of mine, after a slight pause I would be encouraged to remember that we have also the median response, if we cared to look, and thus that our coverage was compact and complete enough. But even with the median response, it feels as though we stand on two small islands of solid ground in a great sea of statistics. What of all the distance in between? And beyond? What else is going on in our data? Relying on the mean response is most useful when our explanatory variables are at their mean values to begin with. Moreover, assuming the mean response for all values of our explanatory variables assumes that their effects remain constant even as their own values vary, an assumption which often is not true. What if, as is often the case, our variables act differently at different values? Or if we have substantial outliers? Or perhaps we have a long-tailed distribution? This is the question of heterogeneity.

Regression Quantiles

Perhaps the first to notably address this problem of heterogeneity seriously was Francis Galton in his book *Natural Inheritance* of 1890 where he took issue with his colleagues' sole focus on the average. Quite famously, he quipped "It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once." To make for robust analyses, he suggested the quantile function. If we have a random variable Y , we define the population quantile as the value $Q_Y(\tau)$ which satisfies, for $0 \leq \tau \leq 1$,

$$P(Y \leq Q_Y(\tau)) = \tau$$

That is, if $F_Y(y)$ is the c.d.f of Y ,

$$Q_Y(\tau) = F_Y^{-1}(\tau)$$

Initially, the notion of regression quantiles seems to indicate that we might need to sort and order our sample if we aimed to employ quantiles of any kind. While this poses little problem with a simple bivariate model where we have a response dependent on a single covariate, the problem becomes real when we deal with more complex models with many and perhaps even multivariate covariates. Surprisingly, the problem can be cast as a linear minimization problem, circumventing any need to order or sort our sample observations (Koenker and Basset, 1978). The problem is analogous to the well known one of minimizing the sum of squared residuals as would be done in a conventional OLS analysis. In the seminal work on the subject, Roger Koenker and Gilbert Basset coin the term “regression quantiles” and detail the same issue discussed above of frequently blind and hopeful assumption of a Gaussian error distribution even when such parametric models do not hold up to scrutiny or even a priori expectations.

Before the method advanced by Koenker and Basset, if robust estimation was desired, it was common to employ linear combinations of order statistics, as suggested by Bickel (1973), or simply to justify the removal of outliers so that a closer adherence to normality was observed. In other cases, the sample mean would be modified to give less weight to outlying observations. These precursory methods are all discussed in Koenker and Basset (1978), fueling the motivation to concoct the generalization of conditional regression quantiles from the method of least absolute deviation which preceded it.

Putting it extremely simply, the line of thought can be outlined as follows. Given a sample (Y_i, x_i) with x_i p -dimensional, we identify the conditional mean regression, $x_i'\hat{b}$, as the solution, over $b \in \mathcal{R}^p$,

$$\hat{b} = \arg \min_b \sum_i (Y_i - x_i'b)^2$$

Modifying this problem slightly, we can identify the conditional median as the solution

$$\hat{b} = \arg \min_b \sum_i |Y_i - x_i'b|$$

This is a generalization of the fact that the least absolute deviations estimator for a single sample is the sample median, $Q_Y(.50)$ (early work by Boscovich about 1755). As shown in detail in Koenker (2005), the solution to the absolute deviations problem can be found using linear programming, which leads to finite sample and asymptotic results for the regression median that are direct generalizations of properties of the one-sample median. Precisely because this problem can be recast as a linear programming problem, we are gifted with the fact that any regression quantile solution must be one which fits at least p observations exactly, which is a product of the subgradient conditions which dictate the solution (Tableman, Kim, and Portnoy 2004). Also note that linear programming algorithms allow regression quantiles to be computed in times comparable to those for least squares solutions: generally a bit slower, but at times even faster (Koenker and Portnoy, 1997). It is then natural to suppose that perhaps an asymmetric weighting of the residuals might yield the other quantiles. Following this thought, let $\rho_\tau(u)$ be the “check” function, that is,

$$\rho_\tau(u) = u(\tau - 1_{(u < 0)}) = \tau u^+ + (1 - \tau)u^-$$

where “+” and “-” indicate the positive and negative parts of u , taken positively. So if $u > 0$, $\rho_\tau(u) = \tau u$ and if $u < 0$, $\rho_\tau(u) = (1 - \tau)|u|$. If $\tau = .5$, then $\rho_{.5}(u)$ weights each residual equally and is exactly the absolute value function (see Figure 1) that will yield the median quantile just as above.

A different τ will weight the residuals differently, and thus produce a different regression quantile. For example, setting $\tau = .25$, $\rho_{.25}(u)$ produces a weighting function like that seen in Figure 2.

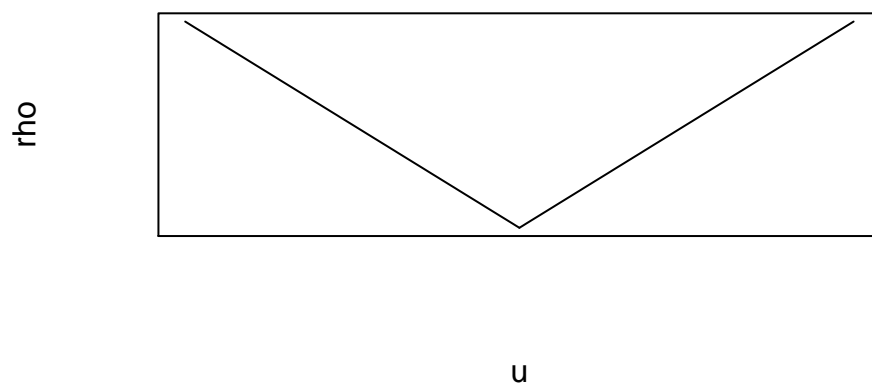


Figure 1: Weighting function at median

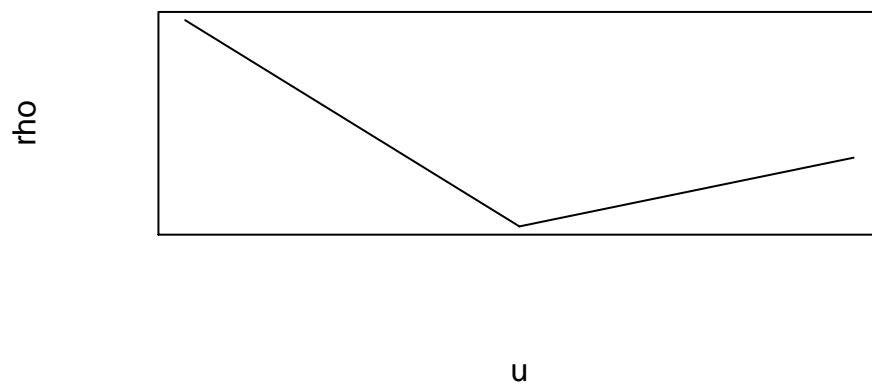


Figure 2: Weighting function at 25th quantile

Instead of using squared error loss or least absolute deviation, we use the check function above, transforming the problem into the minimization

$$\hat{b}(\tau) = \arg \min_b \sum_i \rho_\tau(Y_i - x_i' b)$$

The solution to this problem will yield any desired regression quantile. The important aspects of regression quantiles to note are that they provide a way to naturally order a set of structured data, that they truly generalize the notion of quantiles from the population quantiles, they allow us to account for heterogeneity by allowing our parameter coefficients, $b(\tau)$, to vary, and we've made no assumptions anywhere about the underlying distribution of the data. Moreover, if we weren't sure about whether or not heterogeneity actually exists in our data, regression quantiles will not only allow us to assess it but they will also give us the tools to detect it because any observed significant variation in $b(\tau)$ necessarily means that the structure of the linear relationship is changing as we vary τ . If the data were homogeneous, this would be reflected by coefficients that remain constant across the sample population.

There are some important and useful statistical properties of regression quantiles, too, formulated in Koenker and Basset (1978). Regression quantiles also enjoy the usual equivariance and asymptotic results that we depend on. That is, we can rescale or reparametrize the results of a quantile regression freely without fear of losing the integrity of our predictions, and the finite sample theory of regression quantiles ensures that we can compute meaningful confidence intervals around their estimates. More specifically, if an iid error distribution is justified (or if we go ahead and assume it regardless as we do here), then we have the limiting distribution result that

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \rightarrow \mathcal{N}(0, \omega^2(\tau)(X'X)^{-1})$$

where $\omega^2(\tau) = \tau(1 - \tau)/f^2(F^{-1}(\tau))$ and $f(F^{-1})$ is the density of the error distribution evaluated at the τ^{th} quantile (Koenker 2005). In perhaps more complicated, and realistic, cases, a bootstrap method and some variations of it have also been used to compute confidence intervals (Koenker 2005). These methods involve sampling pairs of observations (x_i, y_i) $i = 1, 2, \dots, n$ with replacement and estimating $\hat{\beta}_n^*(\tau)$ each time. Repeating this process some fixed number of times yields a sample covariance matrix which is a valid estimator of the covariance matrix of the original estimator (Koenker 2005).

A digression for an application

The data

This project started as a venture into model selection of ODEs which is quite a difficult problem in statistics due to the rarity of an analytic solution for any system of ODEs. Even when initial conditions are known or assumed, numerical methods must be used with parameter values that first have to be estimated from measurements of the system in the presence of noise. Zhang et al. (2015) looked to address this problem by suggesting a variation of a least squares approach to select from a set of potential ODE models which find common use in ecology and biology. In particular, they were seeking a way to select the most appropriate model of a predator-prey dynamic between Canadian lynx and snowshoe hares.

The data were collected by the Hudson Bay Company over the period 1845 to 1935, shown in Figure 3. The data are the recorded numbers of pelts, in thousands, collected by the company each year, meaning the data are actually proxy data for the populations of hares and lynx.

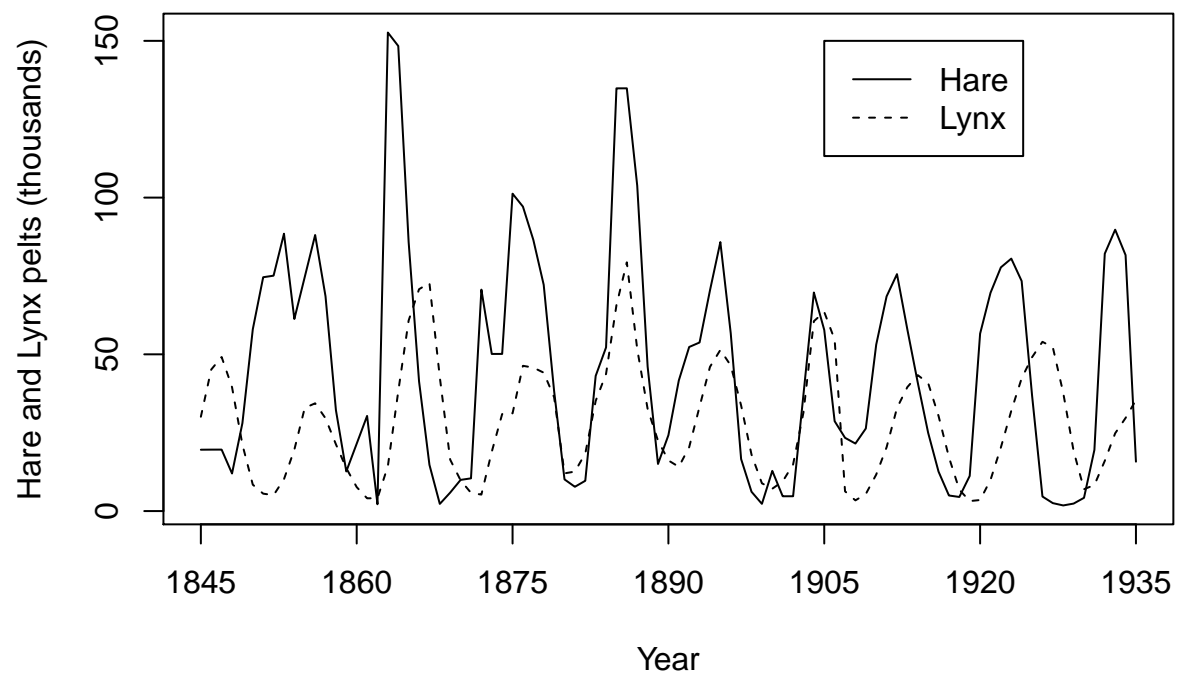


Figure 3: Canadian lynx and snowshoe hares, Hudson Bay Company

The method

Zhang et al. were tasked with choosing one model from a set of four where each model was actually a special case of a more general model. The four models and the more general one are listed below, with H representing the hare population and P representing the lynx population. Their method consisted of a two-stage selection process where they first use the parameter cascading method (Ramsey et al. 2007) on the full, general, model to obtain estimates of all the parameters. The parameter cascade method uses a linear combination of basis functions (cubic B-spline basis functions in this case) to estimate the dynamical process. The second stage is to reduce the full model to one of the special cases shown below by minimizing a least squares type criterion with an adaptive LASSO type penalty whereby some of the parameter estimates may be zero, thus reducing the model to one of the four special cases. Finally, because of the lack of the oracle property of these estimates after the least squares type minimization, they use another parameter cascade on the reduced model to obtain their final estimates.

From this point, instead of pursuing model selection, we looked at the system of ODEs and the complex method used by Zhang et al. and wondered how a quantile regression would fare in the way of estimating the parameter values of the chosen model.

The models

The Lotka-Volterra model

$$\begin{aligned}\frac{dH}{dt} &= rH - aHP \\ \frac{dP}{dt} &= eHP - vP\end{aligned}$$

The logistic prey model

$$\begin{aligned}\frac{dH}{dt} &= rH(1 - H/g) - aHP \\ \frac{dP}{dt} &= eHP - vP\end{aligned}$$

The density-dependent predator death model

$$\begin{aligned}\frac{dH}{dt} &= rH - aHP \\ \frac{dP}{dt} &= eHP - vP - bP^2\end{aligned}$$

The predator-dependent functional response model

$$\begin{aligned}\frac{dH}{dt} &= rH - aHP/(1 + zP) \\ \frac{dP}{dt} &= eHP/(1 + zP) - vP\end{aligned}$$

And finally, the more general model,

$$\begin{aligned}\frac{dH}{dt} &= rH - aHP - kH^2 + az\frac{P}{1+zP}HP \\ \frac{dP}{dt} &= eHP - vP - bP^2 - ez\frac{P}{1+zP}HP\end{aligned}$$

It can be seen that each model is a special case of the final model listed if certain of the parameters are set to 0. In the analysis of Zhang et al., the *Lotka-Volterra* model was finally selected, meaning the parameters $\theta = (r, a, e, v)$ were found to be significantly nonzero. This gave us something to compare to with our regression quantiles approach. However because we were not pursuing a model selection method, we decided to use the full model in our quantile regression.

To model the derivatives, we elected simply to take the first differences at each point in the time-series. Thus, modeling the hare population, for example, looks like

$$H_{i+1} - H_i \approx \frac{dH_i}{dt} = rH_i - aH_iP_i - kH_i^2 + az\frac{P_i}{1+zP_i}H_iP_i$$

There are a few things to note here. First, the population we are modeling is no longer simply the population counts, but rather the differences in hare and lynx population each year. Since the unconditional distribution of differences is highly correlated, we are considering a sample from a conditional distribution of differences given (H_i, P_i) . On that note, since we are dealing with a system of ODEs, our data is actually bivariate and for this to work we are implicitly assuming that the model errors for H_i and P_i are independent. Secondly, the system is non-linear in z . To address this, we first made a call to the `nlrq()` function in the `r` package `quantreg` (Koenker, R. (2018). `quantreg: Quantile Regression`. R package <http://www.r-project.org>) but later decided to drop the non-linear term from the model and run a normal `rq()` command since the z term was not expected to be significant anyway given the results from Zhang et al.

Analysis

Plotting the estimated regression quantiles at $\tau = .5$, we see right away that median regression fits the data quite well; in Figure 4, the black lines are the data, and the dashed lines are the 50th regression quantile fits. Juxtaposing each graph of estimates with their `rq` and `nlrq` outputs, we see the fit is just about the same.

In Figures 5 and 6, we show residual plots, for each population, from an OLS regression, side by side with quantile regressions at the three inner quartiles performed with the `rq()` function (since performing an OLS with the nonlinear version of the model would not be possible without fixing the z term). Looking first at the hare population (Figure 5) we see that the conditional mean and median regressions are nearly identical in their predictive power. Interestingly however, we see that the shape of the residuals seems to morph slightly as we increase from $\tau = .25$ up to $\tau = .75$. The same is true for the lynx population (Figure 6). A more pronounced change of shape would indicate that the parameters may affect the population differently as they vary. Note that it is important, if we are inclined to compare our analysis to an OLS analysis, that we compare the conditional mean with more than just the conditional median. Indeed, comparing those two alone might convince us that the populations really are homogeneous, but the key with regression quantiles is to look along the distribution and this is why we include the $\tau = .25$ and $\tau = .75$ quantiles.

We can plot the change in the estimated coefficients over τ (Figures 7 and 8) and if we find that the estimates are changing, then we have evidence to believe that the effects of the parameters are actually changing as we move about the distribution. These plots are the most important part of any quantile regression analysis. Figure 7 shows the coefficients vs. τ plots for the full model including the nonlinear z term, and Figure 8 is with the z term removed. Note that there seem to be two z terms, zH and zP . This is only because there is no way, using the `quantreg` package anyway, to input a system of equations into the model statement for `rq` or `nlrq`, so the program is run as if they are separate parameters when really they represent the same quantity in the model. For the sake of avoiding errors in the code and output, I resolved to distinguishing the parameters. Also, the lack of confidence intervals around the `nlrq` plots in Figure 7 is a product of the difficulty of extracting the standard errors for each regression quantile in the `nlrq` output. Attempting to extract the standard errors produced values that did not match the summary output of the `nlrq` object, so I surrendered and resolved to forego these confidence intervals.

Another look at Figures 7 and 8 will raise concerns about how very different some of the estimates look from their counterparts in their opposite plots. For example, the v coefficient runs negative in the `nlrq`

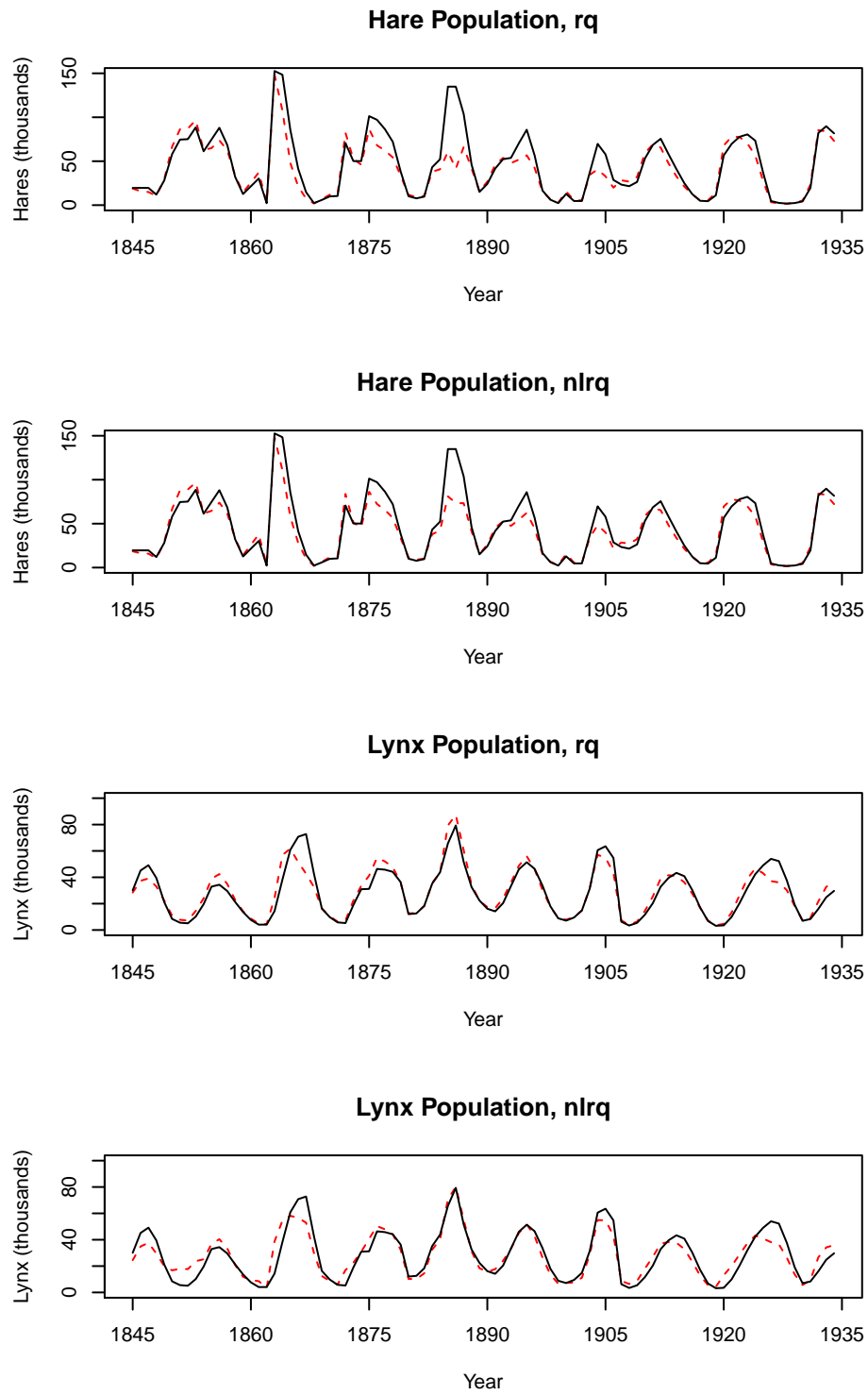


Figure 4: Median regression quantile estimates overlayed onto the data

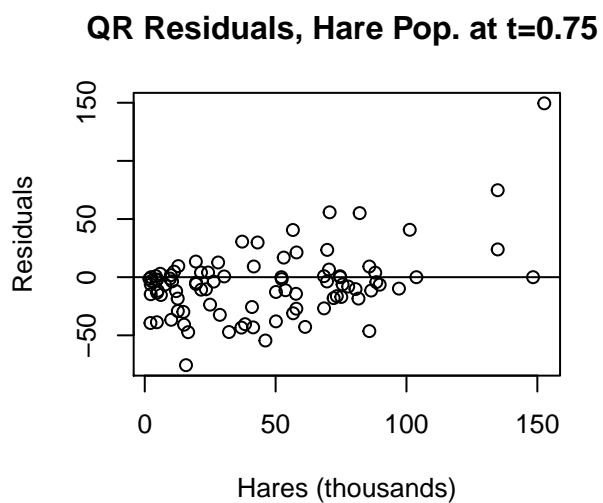
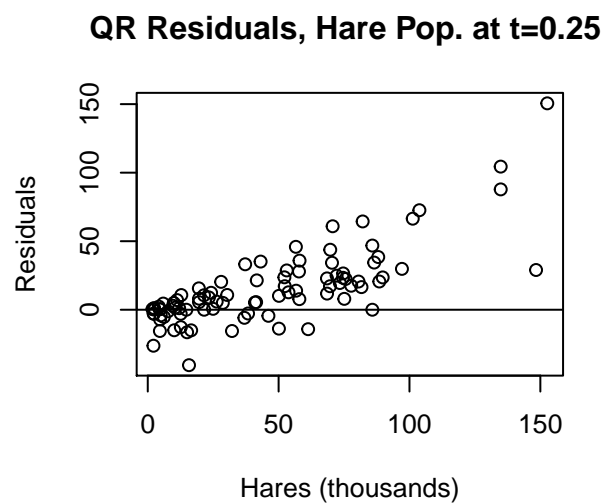
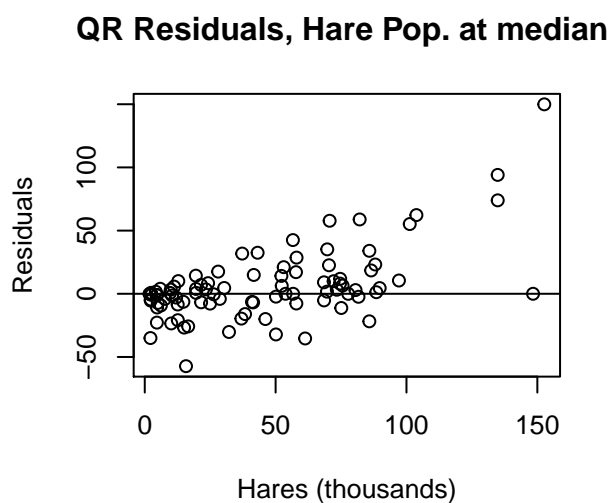
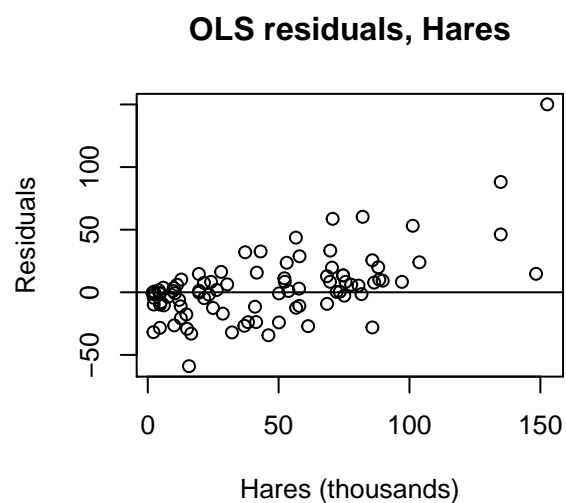


Figure 5: QR residuals vs OLS residuals for hare population differences

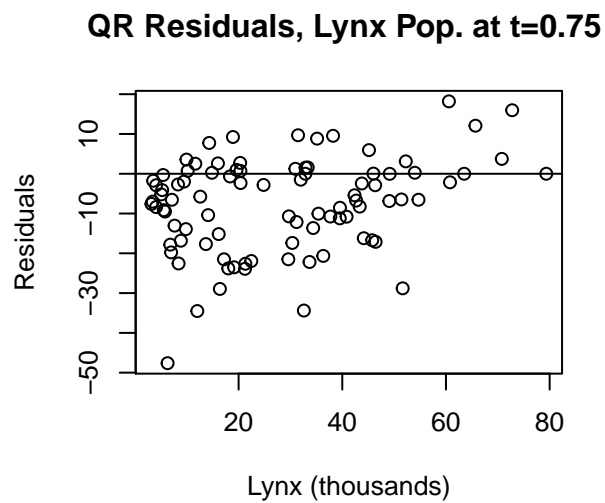
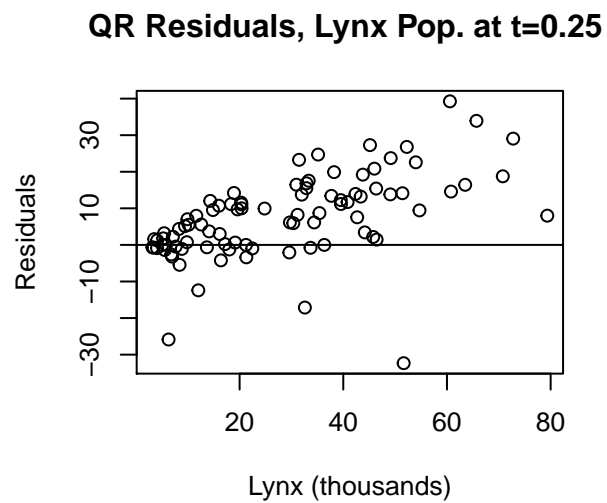
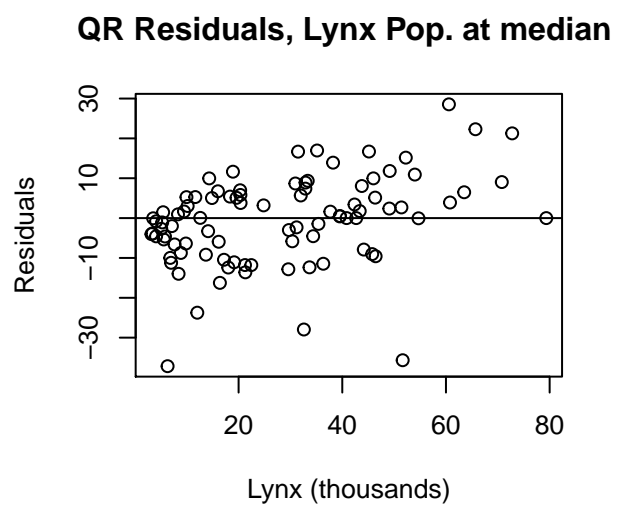
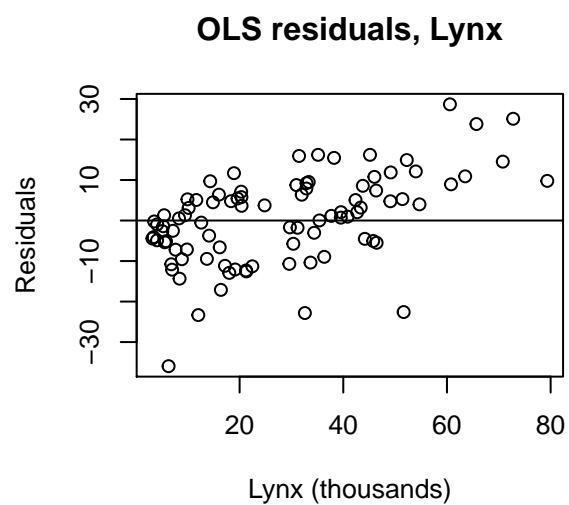


Figure 6: QR residuals vs OLS residuals for lynx population differences

version, while it turns positive in the `rq` plot. The same curiosity occurs for other estimates, namely k , and b . This is simply due to the way `nlrq` requires the user to specify the model. The call statements for `nlrq` require explicit specification of the model including negative signs, where `rq` uses the usual formula statements that one might use in an `lm()` call. Once this is accounted for, the signs of the estimates do not disagree as it appears they do by the τ plots in these figures.

Having accounted for the discrepancy between the `nlrq` and `rq` estimates for parameters v , k , and b , the second most notable difference regards the coefficients of the product terms, a and e . In the linear version of the model (Figure 8), these coefficients move wildly throughout the distribution, while a call to `nlrq` settles down a quite substantially. This is not repeated with e , where it seems we have run into an issue of identifiability with the z parameter. The values of both e and z shoot off astronomically at the same time and for the same range, and yet we know, or at least expect, from Zhang et al. that the product term should be important in the derivative of the lynx population. Partly for this reason, and partly for the ensuing simplicity that doing so would offer, we chose to remove the z term and continue with the linear model.

We've overlaid the OLS estimates of each coefficient (in the `rq` plots in Figure 8) onto the τ plots so that it is easily seen where the two methods disagree. The solid, straight, red lines are the conditional mean estimates, surrounded by their corresponding 95% confidence intervals, the solid black curves are the quantile regression estimates as a function of τ with 95% confidence intervals, and the dashed black lines (Figure 7 and 8) are 0. `rq()` computes confidence intervals by estimating a sandwich form of the covariance matrix which involves estimating the conditional density at each of the sample observations (see `rq()` documentation). For reference, our model is

$$\begin{aligned}\frac{dH}{dt} &= rH - aHP - kH^2 + az\frac{P}{1+zP}HP \\ \frac{dP}{dt} &= eHP - vP - bP^2 - ez\frac{P}{1+zP}HP\end{aligned}$$

where the linear version can be obtained by setting $z = 0$.

From the τ plots in Figure 8, we can see that, for example, the conditional mean and conditional quantile estimates of the r coefficient, which represents the H parameter in our model, are in good agreement throughout most of the distribution. When the hare population reaches values in the upper quantiles of its distribution, the conditional mean turns out to underestimate the future hare population. This makes sense. The hare population is not likely to stay too high for too long since a large population puts a great deal of stress on ecological resources and so the difference in the population between time steps should be large, yet the conditional mean estimate would not predict this. Indeed, looking back at the graph of the data, we see that large peaks in the hare population are quickly followed by returns to near its minimum. The important part is that the effect of the H parameter, that is the value of its coefficient, changes as the value of H itself changes. A conventional OLS analysis would assume the estimated coefficient is constant, as is clearly seen in the τ plots.

Additionally, we can look at the overlap of the confidence intervals of the OLS and QR estimates to see if they are statistically different from one another. For the r and v coefficients, we see that there is statistical difference in the lower and upper tails of the distribution, and this is in fact where we are most likely to employ QR in the first place. Often times we are interested not simply in what happens at the mean of a distribution, but rather which variables are causing the extreme events to occur. This is precisely what QR is useful for. In this case, if we estimated via pelt counting an unusually high number of hares or lynx, we might be better able to estimate the next year's population using the results from a quantile regression instead.

In our case here, we must remember that we have a bivariate system so that considering the hare or lynx population by themselves is insufficient. In fact, if we did, we would conclude that a large hare population is enough to predict a future large hare population. While more hares do in fact beget more hares, more hares also provide more food to the lynx which is why the derivative of the hare population decreases with the product of the hare and lynx population. An upper quantile level of hares mixed with an upper quantile level of lynx should produce a staggering change in the number of hares in the next year. We see this play out

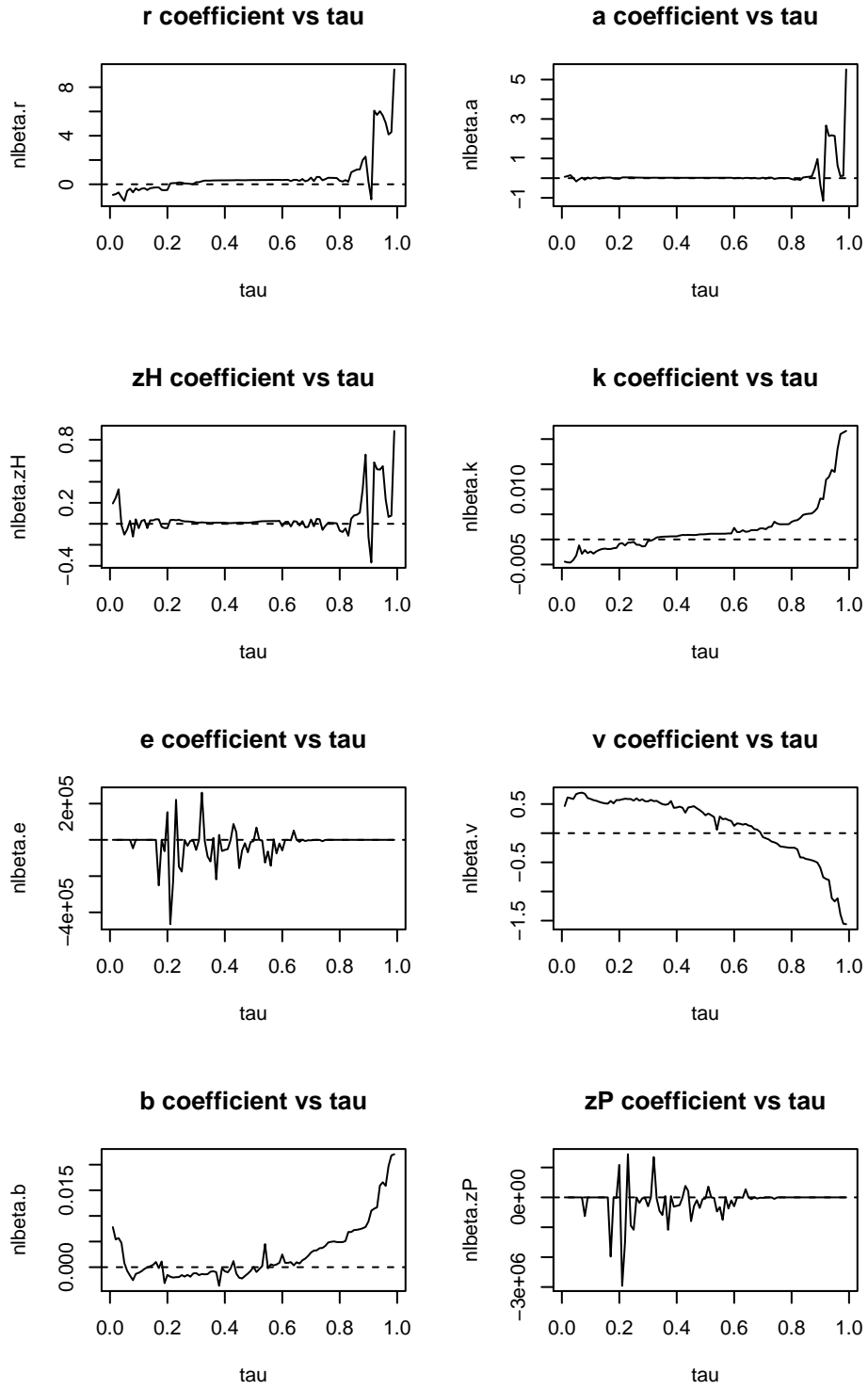


Figure 7: Nonlinear RQ tau plots, full model

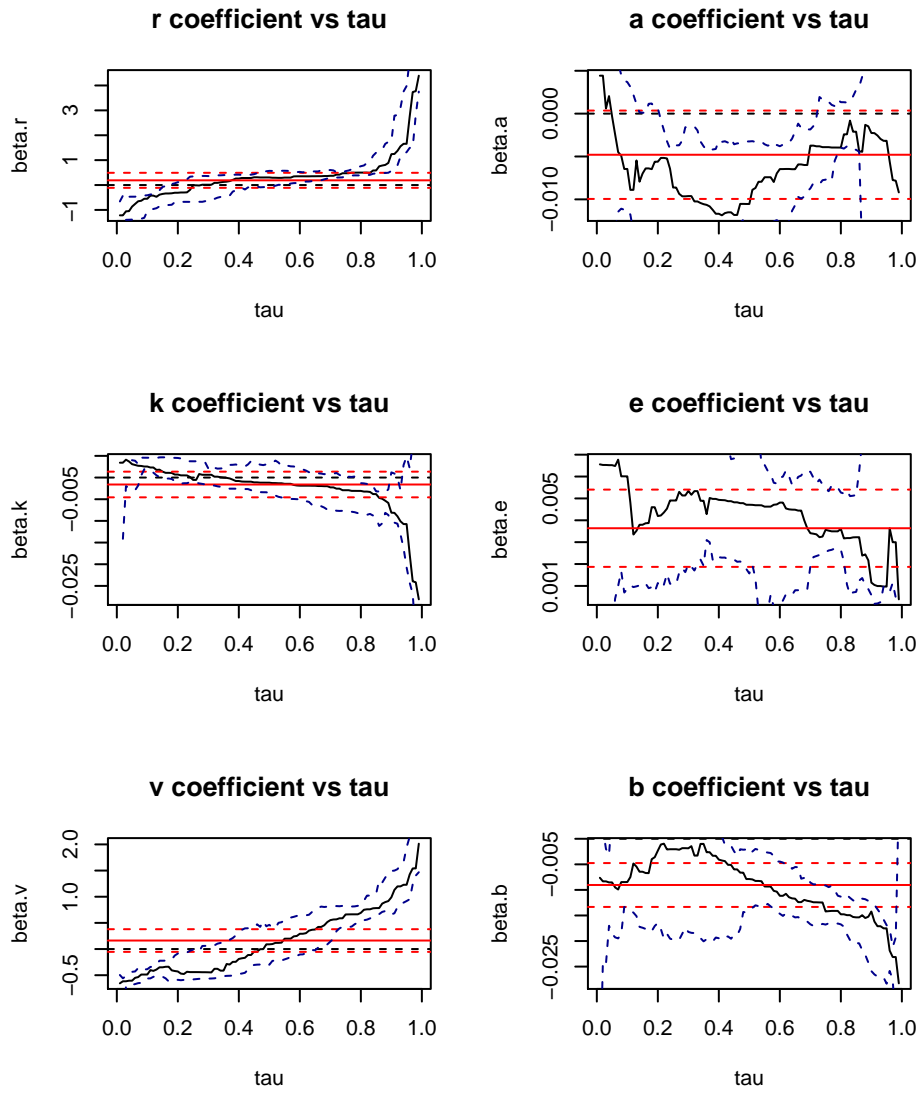


Figure 8: Linear RQ tau plots, full model

in the graph of the data, and we find it in the significance of the a and e coefficients. While the estimates appear small and near 0 due to the scaling factor of the original data, we still find significance throughout most of the distribution. We see that a plays a role in the inner quantiles, indicating that the interplay of the hare and lynx populations is a strong regulator of the population dynamics for normal levels of the populations. Similarly for e , we see significance for most of the inner distribution. When we get to the tails, that is, when the populations of either the hares or lynx become unusually small or large, the effects of r and v become stronger. In Tables 1-4 we report the estimated regression quantile coefficients for various quantiles, alongside the estimates of the values produced by Zhang et al. for comparison with the parameter cascade method. For the sake of closer comparison, we omit the squared terms from our tables, since Zhang et al. only produce estimates for parameters (r, a, e, v) .

Table 1: \hat{r} RQ estimates and confidence intervals

	\hat{r}	95% CI
Zhang et al.	0.0094748	(0.02464468, 0.16485132)
.25 RQ	-0.0112569	(-0.6738953, 0.3451345)
.50 RQ	0.3045929	(-0.0621178, 0.569568)
.75 RQ	0.4743864	(0.3648049, 0.5590242)
.90 RQ	1.2417735	(0.5639014, 1.7485412)

Table 2: \hat{a} RQ estimates and confidence intervals

	\hat{a}	95% CI
Zhang et al.	0.004135	(0.00107544, 0.00719456)
.25 RQ	-0.0086189	(-0.0163257, -0.0039612)
.50 RQ	-0.0105393	(-0.0127738, -0.0032422)
.75 RQ	-0.0039113	(-0.0077023, 0.0013747)
.90 RQ	-0.0022609	(-0.0165994, 0.0224526)

Table 3: \hat{e} RQ estimates and confidence intervals

	\hat{e}	95% CI
Zhang et al.	0.001331	(-0.00001356, 0.00267556)
.25 RQ	0.0049043	(0.0010819, 0.0082428)
.50 RQ	0.00471	(0.0015683, 0.0078608)
.75 RQ	0.0035648	(0.002548, 0.0054381)
.90 RQ	0.0011492	$(5.852477 \times 10^{-4}, 0.0079783)$

Table 4: \hat{v} RQ estimates and confidence intervals

	\hat{v}	95% CI
Zhang et al.	0.050430	(-0.00050256, 0.10136256)
.25 RQ	-0.440831	(-0.5805314, -0.0111248)
.50 RQ	0.0902075	(-0.30017, 0.4628192)
.75 RQ	0.6608239	(0.4196643, 0.8319801)
.90 RQ	0.9879832	(0.6618794, 1.4927957)

It must be remembered that `rq()` is reporting its estimates with regards to how the model was specified in the function call which is not exactly the same as the derivative models above due to sign changes. Therefore, the negative estimates seen for a and (most of) v contribute as they should when they are input into the actual derivatives specified above. Furthermore, since the data are originally in units of thousands, we took the liberty of dividing the estimates and confidence intervals produced by Zhang et al. by 1000 to more easily compare them with our own.

If we had only the estimates from the more complex parameter cascade method exercised by Zhang et al., we would be left to conclude that the parameters e and v are not significant, which would render meaningless the entire second equation in the selected ODE model. If we open our arms to possible heterogeneity with a QR, we would find many new insights into the population dynamics. We would see, for example that the coefficients corresponding to the product terms in the model, a and e , are most responsible for the variation around the middle parts of the distribution, which we see by noting the difference from 0 for these coefficients throughout the $\tau = .25$ to $\tau = .75$ quantiles. We would also be able to distinguish the effects of the linear population terms, that is H and P whose coefficients are r and v , from the product terms and see that those coefficients are most effective at the tails of the distributions.

Conclusion

A median regression on the hare-lynx data was able to fit the data quite well, and QR proved to detect valuable insights even though we had a very limited data set to work with. The regression quantiles were able to detect some heterogeneity in the data towards the upper and lower quantiles of the distribution, to which an OLS analysis would have been blind. Moreover, the QR method did at least as well as the more complex method introduced by Zhang et al., and arguably performed better in that it was able to identify regions of significance for certain parameters, whereas the least squares type method used by Zhang et al. would have had us essentially discard two of the four parameters which the second stage of their own method had deemed to be significantly nonzero.

Quantile regression is a robust regression method offering researchers a useful way to detect and assess heterogeneity in their data when the assumptions of OLS are not met. As seen in this case, regression quantiles are more suited to making predictions regarding extreme circumstances than a conventional OLS analysis or even a more complicated adaptation like that suggested by Zhang et al., since an OLS regression can only treat these extreme cases as average ones. Moreover, QR can adequately treat average or median cases as well, making QR a robust method that should be conducted along with any OLS analysis.

References

- Koenker, Roger. Quantile Regression. Cambridge University Press, 2005.
- Koenker, R and Basset, G. Jr., (1978). Regression Quantiles. *Econometrica*, Vol. 46, No. 1, pp. 33-50.
- Koenker, R. and Hallock, K., (2001). Quantile Regression. *Journal of Economic Perspectives*, vol. 15, no. 4, 143-156.
- Koenker, R. and Portnoy, S. (1997). The Gaussian Hare and the Laplacian Tortoise: computability of squared error vs. absolute-error estimators, *Stat. Science*, 12, 279-300
- Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B* 69, 741-796.
- Tableman, M., Kim, J. S., Portnoy, S. *Survival Analysis Using S: Analysis of time-to-event data, with a contribution from Stephen Portnoy*. Chapman & Hall/CRC, 2004.
- Zhang, X., Cao, J., and Carroll, R. J. (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, 71, 131-138.