Chapter 12. Special topics

12.1. Nonparametric regression

12.1.1 Bootstrapping

Appropriate for handling problems of sample dependence, non-normal distribution, sampling error in x (calibration analysis).

Given model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Step 1: create a new dataset by $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, where e_i is resampled with replacement form the observed residuals.

Step 2: Do a regression on the new data

Step 3: Repeat the process for R times.

We need library "boot" for implementation. The R program attached for doing bootstrapping for data: **hl.dat**, is **boot.lm.r**. To run **boot.lm.r**:

>boot.lm.r(hl.dat).

Compare the outputs with those of **lm**.

12.1.2 Monotonic regression (based on rank)

If the regression relationship is believed not a straight line but Y is monotonically increasing with X, this method is useful.

Data: X: $x_1, x_2, \ldots, x_n$.

Y: $y_1, y_2, \ldots, y_n$.

Steps:

(1) Obtain the ranks $R(X)$ and $R(Y)$ for X and Y, respectively. Use average ranks in case of ties.

(2) Fit a regression line to the ranks:

$$R_y = \hat{\beta}_0 + \hat{\beta}_1 R_x,$$

where $\hat{\beta}_0 = \dfrac{(1-\hat{\beta}_1)(n+1)}{2}$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} R_{x_i} R_{y_i} - n(n+1)^2/4}{\sum_{i=1}^{n} R_{x_i}^2 - n(n+1)^2/4}$

(3) To predict y at x_0, we need to obtain a rank $R(x_0)$:

(a) If x_0 equals one of the observed X's, let $R(x_0)$ equal the rank of that x_i.

(b) If x_0 lies between two adjacent values x_i and x_j where $x_i < x_0 < x_j$, interpolate between their respective ranks to get $R(x_0)$:

$R(x_0) = R(x_i) + \dfrac{x_0 - x_i}{x_j - x_i}\left[R(x_j) - R(x_i)\right]$. This "rank" may not be an integer.

(c) If x_0 is less than the smallest observed X or greater than the largest X, do not attempt to extrapolate. Information on the regression of Y on X is available only within the observed range of X.

(4) Substitute $R(x_0)$ for x to get an estimated rank to $R(y_0)$ using the rank linear model:

$$R_y = \hat{\beta}_0 + \hat{\beta}_1 R_x$$

(5) Use $R(y_0)$ to estimate original $\hat{y}$:

(a) If $R(y_0)$ equals the rank of one of the observation y_i, let the estimate $\hat{y}$ equal that observation y_i.

(b) If $R(y_0)$ lies between the ranks of two adjacent values y_i and y_j where $y_i < y_j$, so that $R(y_i) < R(y_0) < R(y_j)$, interpolate between y_i and y_j:

$$\hat{y}_0 = y_i + \frac{R(y_0) - R(y_i)}{R(x_j) - R(x_i)}\left[y_j - y_i\right].$$

(c) If $R(y_0)$ is greater than the largest observed rank of Y, let $\hat{y}$ equal the largest observed Y. If $R(y_0)$ is less than the smallest observed rank of Y, let $\hat{y}$ equal the smallest observed Y.

(6) Since no assumptions are being involved in the rank regression, there are no confidence intervals, hypothesis tests, prediction intervals, etc. The assessment of goodness-of-fit is based on R2 between the observation Y and estimated $\hat{Y}$: $r^2_{Y\hat{Y}}$.

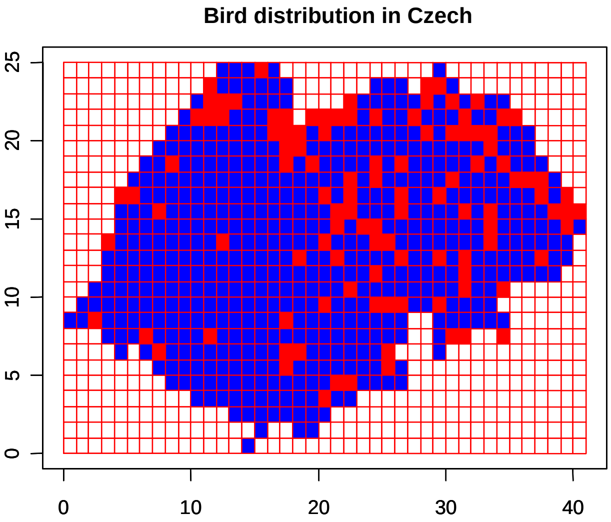
12.2. Local regression (lowess/loess) is a useful technique for EDA to detect patterns

R-code for local regression:

```
>plot(hl$dbh, hl$htt)
>lines(lowess(hl$dbh,hl$htt),col=2)
```

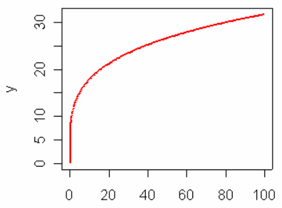
Try the distribution of Czech birds (occupancy versus no of clusters):

```
>plot(czech.bird.dat$occup,czech.bird.dat$nclst)
> bird.lo=loess(nclst~occup,data=czech.bird.dat)
>id=order(czech.bird.dat$occup)
> lines(czech.bird.dat$occup[id],bird.lo$fit[id],col="red")
>lines(czech.bird.dat$occup[id],bird.lo$fit[id]+1.96*predict(bird.lo,se=T)$se.fit[id],col="blue",lty=8)
>lines(czech.bird.dat$occup[id],bird.lo$fit[id]-1.96*predict(bird.lo,se=T)$se.fit[id],col="blue",lty=8)
```



12.4. Nonlinear regression

Linear regression models provide a broad and rich framework that suits many applications. However, linear regression cannot be adequate for all problems. Another important type of models is nonlinear models.



There are two ways to incorporate error terms in nonlinear models:

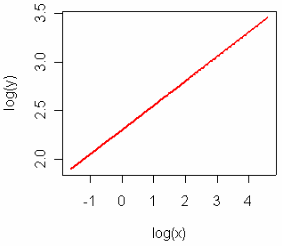
(1). Multiplicative error:

$$y = \beta_0 x^{\beta_1} \varepsilon, \quad \text{where } \varepsilon \text{ follows a lognormal distribution}$$

We can log-transform this nonlinear model into a linear model

$$\log(y) = \log(\beta_0) + \beta_1 \log(x) + \log(\varepsilon)$$

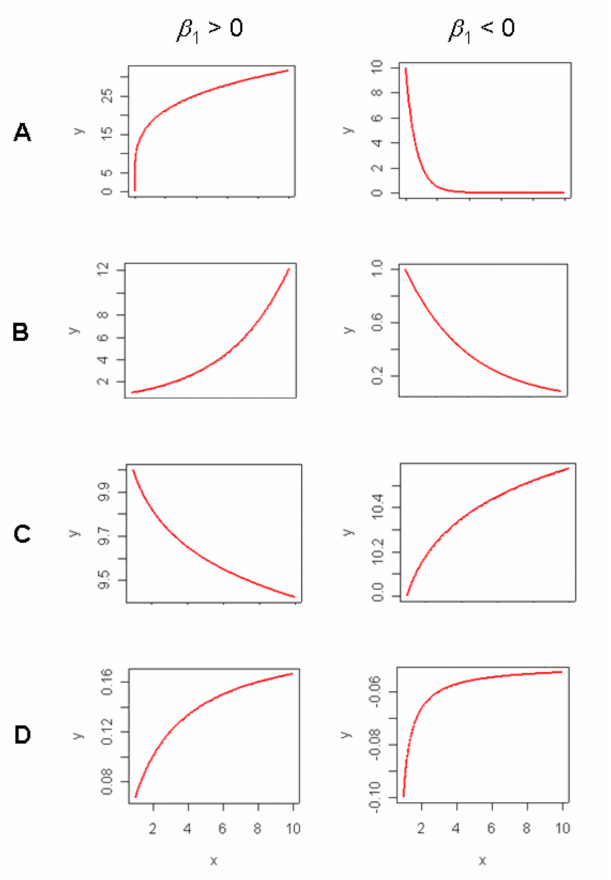
Or write as $y' = \beta_0' + \beta_1 x' + \varepsilon'$, where $\varepsilon' \sim N(0, \sigma^2)$



If you believe your model has multiplicative error, you should use the linear regression methods as we have learned. The way to do it is to linearize the nonlinear models.

Here are some models you can linearize them by transformation:

Figure	Linearizable functions	Transformation	Linear form
A	$y = \beta_0 x^{\beta_1}$	$y' = \log(y)$, $x' = \log(x)$	$y' = \log(\beta_0) + \beta_1 x'$
B	$y = \beta_0 e^{\beta_1 x}$	$y' = \log(y)$	$y' = \log(\beta_0) + \beta_1 x$
C	$y = \beta_0 + \beta_1 \log(x)$	$x' = \log(x)$	$y = \beta_0 + \beta_1 x'$
D	$y = \dfrac{x}{\beta_0 + \beta_1 x}$	$y' = \dfrac{1}{y}$, $x' = \dfrac{1}{x}$	$y' = \beta_1 + \beta_0 x'$



The shapes of the models given in the previous table.

(2). Additive error:

$$y = \beta_0 x^{\beta_1} + \varepsilon, \qquad \text{where } \varepsilon \sim N(0, \sigma^2)$$

Nonlinear regression is based on this model. In general, we can write a nonlinear regression model as:

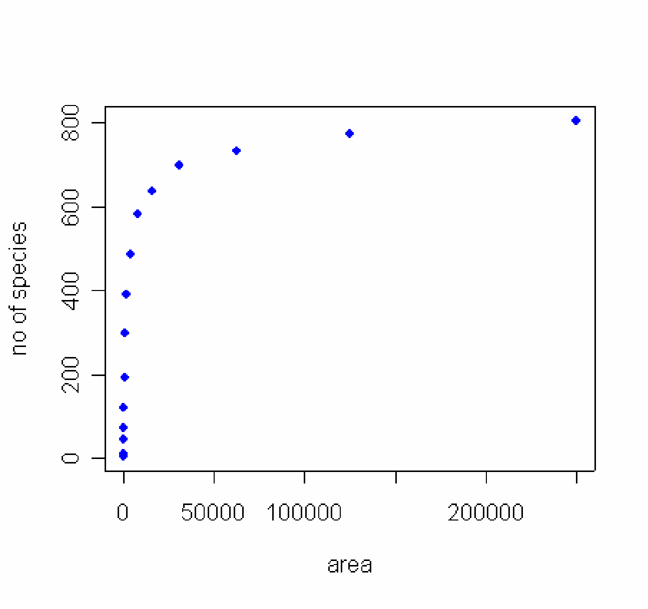
$$y_i = f(x_i, \theta) + \varepsilon, \qquad i = 1, 2, \ldots, n.$$

Given a set of observation data, our purpose is to find a suitable model and fit it to the data.

> sparea.dat:

area	nsp (no of species)

250000	806
125000	775
62500	734
31250	699
15625	636
7812	581
3906.25	486
1953.12	390
976.56	300
488.28	193
244.14	119
122.07	74
61.035	46
30.518	12
15.259	11
7.6294	6
3.8147	4
3.8147	3



Steps:

1. In order to model the species-area data, we have to find a model which can capture the shape of the curve. There are several models which may be useful, including:

Power model: $y = \beta_0 x^{\beta_1}$

Logarithmic model: $y = \beta_0 + \beta_1 \log(x)$

Michaelis-Menten model: $y = \dfrac{x}{\beta_0 + \beta_1 x}$

2. Parameter estimation – nonlinear least squares:

$$S(\beta) = \sum_{i=1}^{n} \left[y_i - \beta_0 x_i^{\beta_1} \right]^2 \rightarrow \text{minimum.}$$

3. R-code:

```
> sparea.nls=nls(nsp~beta0*area^beta1,start=c(beta0=10,beta1=0.2),
     data=sparea.dat)
```

```
> summary(sparea.nls)

Formula: nsp ~ beta0 * area^beta1

Parameters:
      Estimate Std. Error t value Pr(>|t|)
beta0 57.50154   16.12696   3.566  0.00258 **
beta1  0.22634    0.02599   8.709 1.81e-07 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 99.47 on 16 degrees of freedom

Correlation of Parameter Estimates:
        beta0
beta1 -0.9813
```

4. Measurement of goodness-of-fit:

   In general, R^2 is not appropriate for nonlinear regression because the regression and residual sum of squares do not necessarily add to the total sum of squares. That is why the **R** output does not provide R^2. However, R^2 may still be valid if it is calculated using following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

R-code:

```
>SSE=sum((sparea.dat$nsp-predict(sparea.nls))^2)
>SST=sum((sparea.dat$nsp-mean(sparea.dat$nsp))^2)
>R2=1-SSE/SST
```

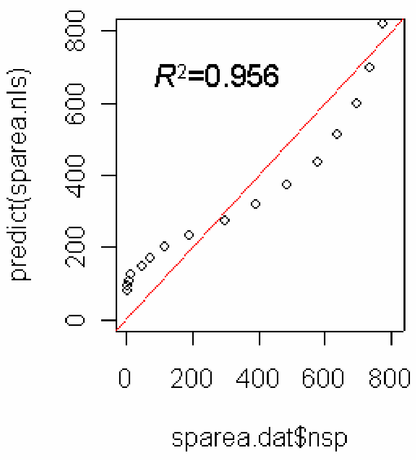
We get $R^2 = 0.9034577$

Another useful measure is the correlation between the observation and estimated values:

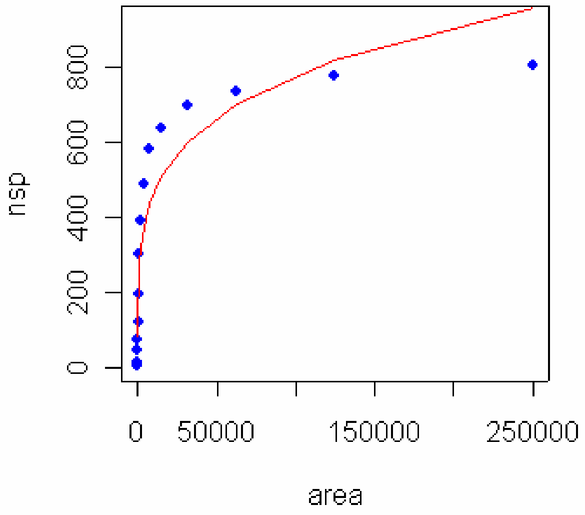
$$R^2 = R^2_{y\hat{y}}.$$

R-code:

```
> cor(sparea.dat$nsp,predict(sparea.nls))
> plot(sparea.dat$nsp,predict(sparea.nls))
```



5. Plot observations and prediction:



Although the R^2 seems reasonably high, the power model:

$$y = 57.501x^{0.226}$$

does not appear to be a good model. It underestimates nsp at small area, but overestimates nsp at large area.

Exercise: Try to fit other two models:

Logarithmic model: $y = \beta_0 + \beta_1 \log(x)$

Michaelis-Menten model: $y = \dfrac{x}{\beta_0 + \beta_1 x}$

6. You may also want to fit the power model using the linearized form:

$$\log(y) = \log(\beta_0) + \beta_1 \log(x)$$

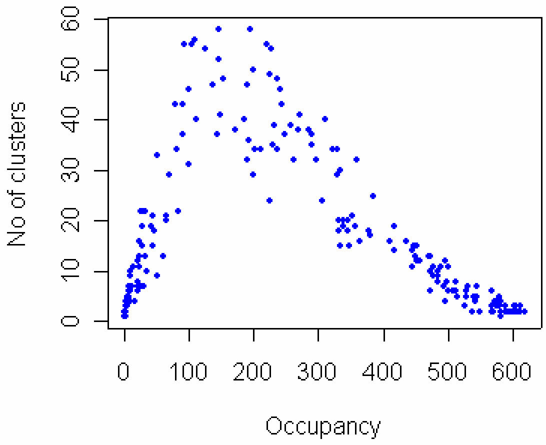
R-code:

```
>sparea.lm=lm(log(nsp)~log(area),data=sparea.dat)
```

You will find the output from the **lm** is very different from the **nls**. Which one we should use? This is a difficult question without a simple answer. We may choose one with highest R^2. It is also useful to look at residuals.

7. **Exercise:** To model the relationship between occup and nclst for the Czech bird data, using model:

$$y = x^{\alpha} \exp(-\beta x^{\gamma})$$



12.5. Quantile regression

(http://en.wikibooks.org/wiki/Statistics:Numerical_Methods/Quantile_Regression)

Quantile regression as introduced by Koenker and Bassett (1978) seeks to complement classical linear regression analysis. The primary goal of the OLS is to determine the conditional mean of random variable Y, given some explanatory variable x_i, reaching the expected value $E[Y \mid x_i]$. Quantile regression goes beyond this and enables one to pose such a question at any quantile of the conditional distribution function.

OLS is the best, linear, and unbiased (BLUE) estimator, if following four assumptions hold: (1) The explanatory variable x_i is non-stochastic, (2) The expectations of the error term ε_i are zero, i.e. $E[\varepsilon_i] = 0$, (3) Homoscedasticity - the variance of the error terms ε_i is constant, i.e. $var(\varepsilon_i) = \sigma^2$, and (4) No autocorrelation, i.e. $cov(\varepsilon_i,\varepsilon_j) = 0$, for $i \neq j$.

However, frequently one or more of these assumptions are violated, resulting in that OLS is not anymore the BLUE estimator. In constrast, quantile regression can tackle following issues:

(i) The error terms are not necessarily constant across a distribution (heteroscedasticity),

(ii) QR is robust to outliers,

(iii) QR can consider the entire spectrum of distribution. This is particularly useful in ecological application as ecologists are often interested in limiting factors which locate in the tails of a distribution. (By focusing on the mean as a measure of location, information about the tails of a distribution are lost in OLS.)

1. What are quantiles?

Quantile is defined in terms of cumulative distribution (or distribution function). A quantile is simply the value that corresponds to a specified proportion of an (ordered) sample of a population. For instance a very commonly used quantile is the median M, which is equal to a proportion of 0.5 of the ordered data. This corresponds to a quantile with a probability of 0.5 of occurrence.

More formally stated, let Y be a continuous random variable with a distribution function $F_Y(y)$ such that

$$F_Y(y) = P(Y \leq y) = \tau$$

which states that for the distribution function $F_Y(y)$ one can determine for a given value y the probability τ of occurrence. Now if one is dealing with quantiles, one wants to do the opposite: to determine for a given probability τ of the sample data set the corresponding value y. A τ^{th} quantile refers in a sample data to the probability τ for a value y:

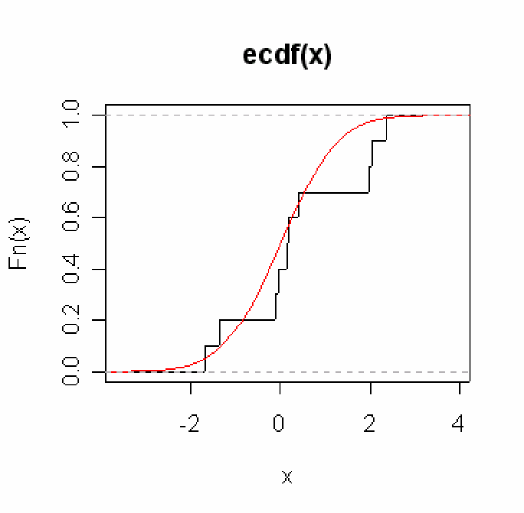
$y_\tau = F_Y^{-1}(\tau)$ (*)

That is y_τ is equal to the inverse of the function $F_Y(\tau)$ for a probability τ.

Quantile: y_p is called the p^{th} quantile of the random variable Y, if $p(X<y_p) \le p$ and

$p(Y>y_p) \le 1-p$. **[R: >quantile(y,p) produces y_p]**

However, a problem that frequently occurs is that an empirical distribution function is a step function. A solution to this problem is to smooth the empirical distribution function through replacing it a with continuous linear function $\hat{F}(y)$. There are several algorithms available which are well described in Handl (2000) and more in detail with an evaluation of the different algorithms and their efficiency in computer packages in Hyndman and Fan (1996). Only with smoothed cumulative function can one apply any division into quantiles of the data set as suitable for the purpose of the analysis.



Smooth the cumulative function

2. Quantile regression

QR essentially transforms a conditional distribution function into a conditional quantile function by slicing it into segments. These segments describe the cumulative distribution of a conditional dependent variable *Y* given the explanatory variable x_i with the use of quantiles as defined in eq. (*).

For a dependent variable *Y* given the explanatory variable $X = x$ and fixed τ, $0 < \tau < 1$, the conditional quantile function is defined as the τ^{th} quantile $Q_{Y|X}(\tau \mid x)$ of the conditional distribution function $F_{Y|X}(y \mid x)$.

One can nicely illustrate QR when comparing it with OLS. In OLS, modeling a conditional distribution function of a random sample $(y_1,..., y_n)$ with a parametric function $\mu(x_i,\beta)$ where x_i represents the independent variables, β the corresponding estimates and μ the conditional mean, OLS is formulated as:

$$\sum_{i=1}^{n}\left(y_i - \mu(x_i,\beta)\right)^2 \rightarrow \text{min}$$

QR can be formulated in a similar fashion. Central feature thereby becomes ρ_τ, which serves as an indicator function:

$$\rho_\tau = \begin{cases} \tau \bullet x & if \;\; x > 0 \\ \\ (\tau - 1) \bullet x & if \;\; x < 0 \end{cases}$$

This indicator-function ensures that: (1) all ρ_τ are positive, (2) the scale is according to the probability τ. In QR we minimize now following function:

$$\sum_{i=1}^{n}\rho_\tau\left(y_i - \xi(x_i,\beta)\right) \rightarrow \text{min}$$

Here, as opposed to OLS, the minimization is done for each subsection defined by ρ_τ, where the estimate of the τ^{th} quantile function is achieved with the parametric function $\xi(x_i,\beta)$.

Features that characterize QR and differentiate it from other regression methods are following:

(1) The entire conditional distribution of the dependent variable Y can be characterized through different values of τ.

(2) Heteroscedasticity can be detected. If the data is heteroscedastic, median regression estimators can be more efficient than mean regression estimators.

(3) The minimization problem as illustrated in the above equation can be solved efficiently by linear programming methods, making estimation easy.

(4) Quantile functions are also equivariant to monotone transformations. That is $Q_{h(Y|X)}(x_\tau) = h(Q_{(Y\,|\,X)}(x_\tau))$, for any function.

(5) Quantiles are robust in regards to outliers.

A graphical illustration of Quantile Regression

Consider Fig. 1. For a given explanatory value of x_i the density for a conditional dependent variable Y is indicated by the size of the balloon. The bigger the balloon, the higher is the density, with the mode, i.e. where the density is the highest, for a given x_i being the biggest balloon. QR essentially connects the equally sized balloons, i.e. probabilities, across the different values of x_i, thereby allowing one to focus on the interrelationship between the explanatory variable x_i and the dependent variable Y for the different quantiles, as can be seen in Fig. 2. These subsets, marked by the quantile lines, reflect the probability density of the dependent variable Y given x_i.

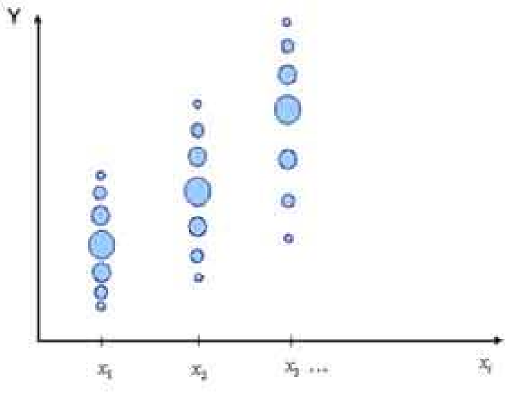


Fig. 1. Probabilities of occurrence for individual explanatory variables

Fig. 2 shows the relationship between household income and household food expenditure, across a spectrum of quantiles: $\tau \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$, indicated by the thin blue lines that separate the different color sections, are superimposed on the data points. The conditional median (τ = 0.5) is indicated by a thick dark blue line, the conditional mean by a light white line. The color sections thereby represent the subsections of the data as generated by the quantiles.

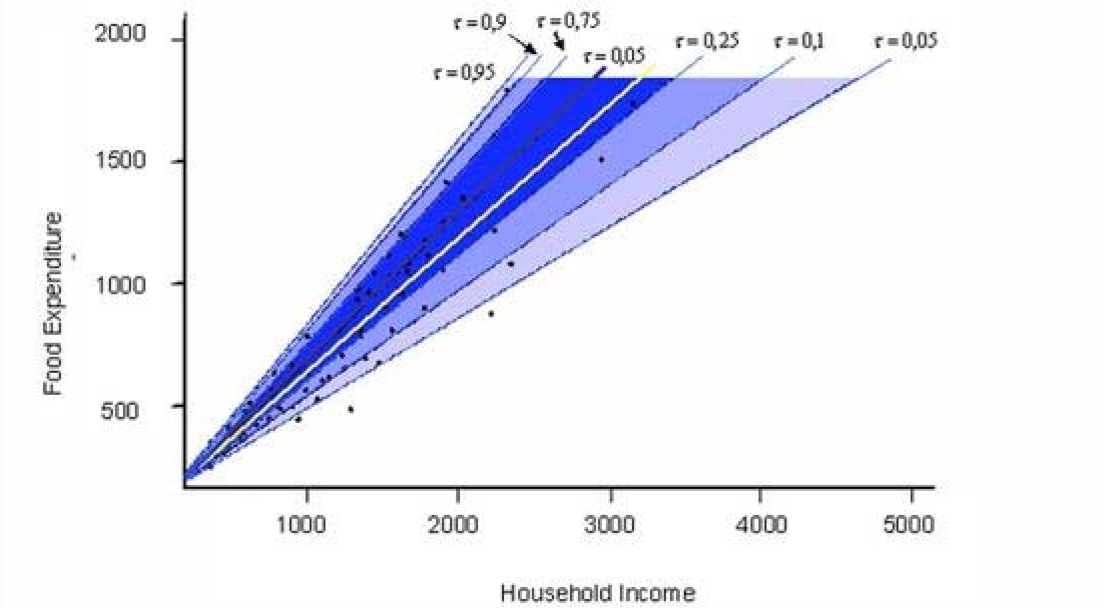


Figure 2 can be understood as a contour plot representing a 3-D graph, with food expenditure and income on the respective *y* and *x* axis. The third dimension arises from the probability density of the respective values. The density of a value is thereby indicated by the darkness of the shade of blue, the darker the color, the higher is the probability of occurrence. For instance, on the outer bounds, where the blue is very light, the probability density for the given data set is relatively low, as they are marked by the quantiles 0.05 to 0.1 and 0.9 to 0.95. It is important to notice that Fig. 2 represents for each subsections the individual probability of occurrence, however, quantiles utilize the cumulative probability of a conditional function. For example, τ of 0.05 means that 5% of observations are expected to fall below this line, a τ of 0.25 for instance means that 25% of the observations are expected to fall below this and the 0.1 line.

The graph in Fig. 2 suggests that the error variance is not constant across the distribution. The dispersion of food expenditure increases as household income goes up. Also the data is skewed to the left, indicated by the spacing of the quantile lines that decreases above the median and also by the relative position of the median which lies above the mean. This suggests that homoscedasticity is violated.

R implementation: rq in package **quantreg.** Install it and run the example !

3. A QR application

The Boston Housing example, first analyzed by Belsley et al. (1980). The original data comprised 506 observations for 14 variables stemming from the census of the Boston metropolitan area.

This analysis utilizes as the dependent variable the median value of owner occupied homes (a metric variable, abbreviated with H) and investigates the effects of 4 independent variables as shown in table 1. These variables were selected as they best illustrate the difference between OLS and QR. A simple multiple linear regression model is assumed.

Table1: The explanatory variables

Name	Short	What it is	type
NonrTail	T	Proportion of non-retail business acres	metric
NoorOoms	O	Average number of rooms per dwelling	metric
Age	A	Proportion of owner-built dwellings prior to 1940	metric
PupilTeacher	P	Pupil-teacher ratio	metric

In the following an OLS model was first estimated:

$E[H_i \mid T_i,O_i,A_i,P_i] = \alpha + \beta T_i + \delta O_i + \gamma A_i + \lambda P_i$

The results are:

Table2: OLS estimates

$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\gamma}$	$\hat{\lambda}$

36.459	0.021	38.010	0.001	-0.953

Analyzing this data set via QR, utilizing the τ^{th} quantiles: $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the model is characterized as follows:

$Q_H[\tau \mid T_i, O_i, A_i, P_i] = \alpha_\tau + \beta_\tau T_i + \delta_\tau O_i + \gamma_\tau A_i + \lambda_\tau P_i$

As an illustrative example, the equation for the 0.1^{th} quantile is given below:

$$[\rho_{0.1}(y_1 - \beta x_1) + \rho_{0.1}(y_2 - \beta x_2) + .. + \rho_{0.1}(y_n - \beta x_n)] \rightarrow \min$$

$$\text{with } \rho_{0.1}(y_i - \beta x_i) = \begin{cases} 0.1 \bullet (y_i - \beta x_i) & \text{if } (y_i - \beta x_i) > 0 \\ -0.9 \bullet (y_i - \beta x_i) & \text{if } (y_i - \beta x_i) < 0 \end{cases}$$

Table3: Quantile Regression estimates

τ	$\hat{\alpha}_\tau$	$\hat{\beta}_\tau$	$\hat{\delta}_\tau$	$\hat{\gamma}_\tau$	$\hat{\lambda}_\tau$
0.1	23.442	0.087	29.606	-0.022	-0.443
0.3	15.7130	-0.001	45.281	-0.037	-0.617
0.5	14.8500	0.022	53.252	-0.031	-0.737
0.7	20.7910	-0.021	50.999	-0.003	-0.925
0.9	34.0310	-0.067	51.353	0.004	-1.257

Comparing Tables 1 and 2, we can find that QR method can make much more subtle inferences of the effect of the explanatory variables on the dependent variable. Of particular interest are quantile estimates that are relatively different as compared to other quantiles for the same estimate.

Probably the most interesting result and most illustrative in regards to an understanding of the functioning of QR and pointing to the differences with OLS are the results for the independent variable of the proportion of non-retail business acres (T_i). OLS indicates that this variable has a positive influence on the dependent variable, the value of homes, with an estimate of $\hat{\beta} = 0.021$, i.e. the value of houses increases as the proportion of non-retail business acres (T_i) increases.

From Table 2 we find a more differentiated picture. For the 0.1 quantile, we find an estimate of $\hat{\beta} = 0.087$ which would suggest that for this low quantile the effect seems to be even stronger than is suggested by OLS. Here house prices go up when the proportion of non-retail businesses (T_i) goes up, too. However, considering the other quantiles, this effect is not quite as strong anymore, for the 0.7th and 0.9th quantile this effect seems to be even reversed indicated by the parameter $\hat{\beta} = -0.021$ and $\hat{\beta} = -0.062$. These values indicate that in these quantiles the house price is negatively influenced by an increase of non-retail business acres (T_i). The influence of non-retail business acres (T_i) seems to be obviously very ambiguous on the dependent variable of housing price, depending on which quantile one is looking at. The general recommendation from OLS that if the proportion of non-retail business acres (T_i) increases, the house prices would increase can obviously not be generalized. A policy recommendation on the OLS estimate could therefore be grossly misleading.

One would intuitively find the statement that the average number of rooms of a property (O_i) positively influences the value of a house, to be true. This is also suggested by OLS with an estimate of $\hat{\delta} = 38.099$. Now QR also confirms this statement, however, it also allows for much subtler conclusions.

This analysis makes clear, that QR allows one to make much more differentiated statements than OLS. Sometimes OLS estimates can even be misleading what the true relationship between an explanatory and a dependent variable is as the effects can be very different for different subsection of the sample.

References

Belsley, D. A., Kuh, E. and Welsch, R.E.. 1980 Applied Multivariate Statistical Analysis. Regression Diagnostics, Wiley.

Koenker, R. and Bassett, G.W. 1978. Regression Quantiles. Econometrica, 46, 33–50.

Koenker, R. and Hallock, K. F. 2000. Quantile Regression an Introduction, available at http://www.econ.uiuc.edu/~roger/research/intro/intro.html

Cade, B.S. and Noon, B.R. 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1:412-420.

Cade, B.S., Terrell, J.W. and Schroeder, R.L. 1999. Estimating effects of limiting factors with regression quantiles. Ecology 80:311-323.

Dunham, J.B. and Cade, B.S. and Terrell, J.W. 2002. Influence of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. Transactions of the American Fisheries Society 131:86-98.