Classical Quantile Regression and One-Dimensional Vector Quantile Regression

Mentor: Alfred Galichon
Xianyang (Eric) Hu

May - Sept. 2023

1 Introduction

A quantile is defined to be the inverse of the cumulative distribution function. It is an important characteristic of the random-variable distribution and provides us with statistical insights into the median, the extremes, etc. Classical quantile regression, like other regression techniques, measures the dependence of a response variable with respect to the explanatory variables by calculating the conditional quantiles of the target variable. Professor Alfred Galichon and his collaborators proposed the idea of "Vector Quantile Regression" (VQR) to extend the one-dimensional case to higher dimensions (i.e. when the response variable can be multi-dimensinal). A vector quantile is defined to be the map that minimizes the average squared distance between an outcome and its preimage. In the one-dimensional case, VQR produces close results relative to the classical case. In this report, I will compare their similar yet different methods of constructing regression, from mathematical deduction to coding with Python.

2 Loss Function – Start with OLS

To discuss regression, let's start with the two most commonly seen: MSE (minimizing squared error) and MAE (minimizing absolute error).

$$\mu = \arg\min_a \mathbb{E}(Y-a)^2$$

$$m = \arg\min_a \mathbb{E}|Y-a|$$

When the penalty functions are different, we will get different results from the regression. The MAE loss function is actually a special case in quantile regression (since median is the 50^{th} quantile). For ordinary least squares, let's first verify that the mean minimizes the expected value $\mathbb{E}(Y-a)^2$:

Proof. To find what minimizes the expected value, we take the derivative with regard to a and set it to 0:

$$\begin{aligned}\mathbb{E}\frac{d}{da}(Y-a)^2 &= \mathbb{E}(-2Y+2a)\\ &= 2a - 2\mathbb{E}(Y)\\ &= 0\end{aligned}$$

This gives $a = \mathbb{E}(Y) = \mu$ and proves the statement $\mu = \arg\min_a \mathbb{E}(Y-a)^2$ □

Next, we may construct the regression in the following way:
Substituting a with dependent variables multiplied by a coefficient β. The loss function is $||y - X\beta||^2$. We need to solve the following problem:

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||^2$$

Proof. If we view $X\hat{\beta}$ as a projection from y to X (i.e. the residual $e = y - X\hat{\beta}$ is orthogonal to the vector space spanned by X), we may have the following simple equation using linear algebra

$$X^T(y - X\hat{\beta}) = 0$$

which gives

$$X^T X\hat{\beta} = X^T y$$

and thus

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

Notice that $X^T X$ should be invertible, which means that the matrix X has full rank and there's no perfect multicollinearity in the variables. □

Let's take this optimization perspective into quantile regression.

3 Classical Quantile Regression

3.1 Definition

3.1.1 Quantile

Given a real-valued random variable X, and its cumulative distribution function F, we will define the τth quantile of X as:

$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x | F(x) \geq \tau\}$$

3.1.2 Quantile Loss Function

Like before, we would like to solve this optimization problem of quantile loss function:

$$\alpha_\tau = \arg\min_\alpha \mathbb{E}[\rho_\tau(Y-\alpha)]$$

where $\rho_\tau(u) = \tau u^+ + (1-\tau)u^-$, $u^+ = \max(u,0)$, $u^- = \max(-u,0)$. It is called the quantile loss function or the check function.
We are going to prove that the unconditional τth quantile is the solution.

Proof. To minimize $\mathbb{E}[\rho_\tau(Y-\alpha)]$, we look for the local minimum point using the same method as above. Set the derivative be 0, $\mathbb{E}[\frac{d}{d\alpha}\rho_\tau(Y-\alpha)] = 0$.

$$\begin{aligned}
\mathbb{E}[\frac{d}{d\alpha}\rho_\tau(Y-\alpha)] &= \mathbb{E}[\tau\frac{d}{d\alpha}(Y-\alpha)^+ + (1-\tau)\frac{d}{d\alpha}(Y-\alpha)^-] \\
&= \mathbb{E}[-\tau\mathbb{1}\{Y>\alpha\} + (1-\tau)\mathbb{1}\{Y<\alpha\}] \\
&= -\tau(1-\mathbb{E}[\mathbb{1}\{Y<\alpha\}]) + (1-\tau)\mathbb{E}[\mathbb{1}\{Y<\alpha\}] \\
&= 0
\end{aligned}$$

Hence,

$$\begin{aligned}
\tau(1-\mathbb{E}[\mathbb{1}\{Y<\alpha\}]) &= (1-\tau)\mathbb{E}[\mathbb{1}\{Y<\alpha\}] \\
\tau(1-F_Y(\alpha)) &= (1-\tau)F_Y(\alpha) \\
F_Y(\alpha) &= \tau \\
\alpha &= Q_Y(\tau)
\end{aligned}$$

□

Similarly, the conditional τth quantile solves

$$\hat{\alpha_\tau}(x) = \arg\min_\alpha \mathbb{E}[\rho_\tau(Y-\alpha)|X=x]$$

To construct regression and compute the coefficients we have

$$\hat{\beta}_\tau(x) = \arg\min_\beta \mathbb{E}[\rho_\tau(Y-X^T\beta)|X=x]$$

$$Q_{Y|X}(\tau|x) = \inf\{y : F(y) \geq \tau | X=x\}$$

3.1.3 Parametric Form

In parametric form we may write the quantile function as a linear formation

$$Q_{Y|X}(\tau|x) = \sum_k \beta_k(\tau)x_k = \beta(\tau)^T x$$

(Parametric assumes a specific functional form for the relationship between the predictor variables and the quantiles of the response variable.)

3.2 Regression Construction: Linear Programming

3.2.1 Definition

To begin with, let's briefly talk about how linear programming works. Linear programming is a mathematical method used for optimizing a linear objective function, subjected to linear equality and linear inequality constraints. A linear programming problem can be represented in its standard form as:

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

x is a vector of variables;
c is a vector of coefficients representing the objective function;
A is a matrix representing the constraints;
b is a vector representing the right-hand side of the constraints

For every linear programming problem (called the primal problem), there exists an associated linear programming problem called its dual problem. The dual is derived from the primal's coefficients, and vice-versa. The process of deriving the dual problem can offer insights into the structure of the original problem and provide bounds on the optimal value of the primal problem.
Let's say the primal problem is:

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

The dual of the above problem is:

$$\begin{aligned} \text{minimize} \quad & b^T y \\ \text{subject to} \quad & A^T y \geq c \\ & y \geq 0 \end{aligned}$$

3.2.2 Primal - Dual Formulation

The goal is to minimize

$$\mathbb{E}[\rho_\tau(Y - X^\top \beta)]$$

First, we would like to formulate the expectation into a linear programming problem. For ease of computation, we consider its sample version.

$$\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta)$$

Let $u_i = (y_i - x_i^\top \beta)^+$ and $v_i = (y_i - x_i^\top \beta)^-$ with $u_i, v_i \geq 0$. $u_i - v_i = y_i - x_i^\top \beta$.

$$\begin{aligned}
\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) &= \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_\tau(u_i - v_i) \\
&= \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \tau u_i + (1-\tau) v_i \\
&= \arg\min_{\beta \in \mathbb{R}^k} [u^\top \mathbf{1}\tau + v^\top \mathbf{1}(1-\tau)]
\end{aligned}$$

Hence, we obtain a primal linear program

$$\arg\min_{\beta \in \mathbb{R}^k} [u^\top \mathbf{1}\tau + v^\top \mathbf{1}(1-\tau)]$$

$$s.t.\ y - X^\top \beta - (u - v) = 0,\ (u, v \geq 0)$$

We can further put this into standard form.
Let $w = (\beta, u, v)$, $c = (0, \mathbf{1}\tau, \mathbf{1}(1-\tau))$, $Z = (X^\top, I_n, -I_n)$.

$$\arg\min_{w \in \mathbb{R}^k} (w^\top c)$$

$$s.t.\ Zw = y$$

Now that we have the primal problem in standard form, it is easy to find its dual.

$$\arg\max_s (y^\top s)$$

$$s.t.\ Z^\top s \leq c$$

Note that the constraint implies that $s \leq \mathbf{1}\tau$ and $-s \leq \mathbf{1}(1-\tau)$, so $\mathbf{1}(\tau - 1) \leq s \leq \mathbf{1}\tau$. Thus this is equivalent to

$$\arg\max_s (y^\top s)$$

$$s.t.\ Xs = 0, s \in [\tau - 1, \tau]^n$$

By change of variables $s = z - (1-\tau)\mathbf{1}$ we have

$$\arg\max_z (y^\top z)$$

$$s.t.\ Xz = (1-\tau)X\mathbf{1}, z \in [0, 1]^n$$

3.2.3 Complementary Slackness

Def. Complementary slackness is a condition in linear programming which states that for every constraint, either the constraint is tight (meaning it holds with equality) or its corresponding dual variable (Lagrange multiplier) is zero. In other words, if a constraint isn't binding, its associated dual variable must be zero, and vice versa. This principle bridges the gap between primal and dual solutions, ensuring that when one has a positive slack (i.e., it's not binding), the other has zero influence on the objective.

In the above program, we can infer that if $y_i < x_i^\top \beta$, (the dual constraint is slack), then $z_i = 0$. If $y_i > x_i^\top \beta$, (the constraint is tight), then $z_i = 1$. Otherwise, $y_i = x_i^\top \beta$. These follow from the KKT conditions of complementary slackness.

In Prof. Galichon's work, he formulated the problem into

$$\mathbb{E}[\tau P + (1-\tau)N]$$

$$s.t.\ \ P - N = Y - X^\top \beta$$

where $P = (Y - X^\top \beta)^+$ and $N = (Y - X^\top \beta)^-$
We may rewrite and eliminate N to get the dual problem

$$\min_{P \geq 0, \beta} \mathbb{E}\left[P + (1-\tau)X^\top \beta\right]$$

$$s.t.\ \ P + X^\top \beta \geq Y$$

Add a slack variable V to the dual problem to represent the constraint

$$\min_{P \geq 0, \beta} \mathbb{E}[P + (1-\tau)X^\top \beta] + \max_{V \geq 0}[V(Y - P - X^\top \beta)]$$

We may rewrite it by combining the min max

$$V_D = \min_{P \geq 0, \beta} \max_{V \geq 0} \mathbb{E}[P + (1-\tau)X^\top \beta + V(Y - P - X^\top \beta)]$$

Using the minimax inequality we have $V_P \geq V_D$ with

$$\begin{aligned} V_P &= \max_{V \geq 0} \min_{P \geq 0, \beta} \mathbb{E}[P + (1-\tau)X^\top \beta + V(Y - P - X^\top \beta)] \\ &= \max_{V \geq 0} \mathbb{E}[VY] + \min_{P \geq 0, \beta} \mathbb{E}[(1-V)P + (1-\tau-V)X^\top \beta + VY] \end{aligned}$$

Then, we have obtained the primal problem

$$\max_{V \geq 0} \mathbb{E}[YV]$$

$$s.t.\ \ V \leq 1\ [P_t \geq 0]$$

$$\mathbb{E}[VX] = (1-\tau)\mathbb{E}[X]$$

The complementary slackness condition gives:

$$V(Y - P - X^\top \beta) = 0$$

Let $V(\tau)$ and $\beta(\tau)$ be solutions to the above program.

$$\begin{cases} Y - X^\top \beta(\tau) < 0 \Longrightarrow V(\tau) = 0 \\ Y - X^\top \beta(\tau) > 0 \Longrightarrow V(\tau) = 1 \end{cases}$$

therefore

$$\mathbb{1}\left\{Y > X^\top \beta(\tau)\right\} \leq V(\tau) \leq \mathbb{1}\left\{Y \geq X^\top \beta(\tau)\right\}.$$

Assume (X, Y) has a continuous distribution. Then for any $\beta, \Pr\left(Y - X^\top \beta = 0\right) = 0$, and therefore one has almost surely $V(\tau) = \mathbb{1}\left\{Y \geq X^\top \beta(\tau)\right\}$.

3.3 Quantile Curve Regression

The previous optimization problem has provided a way to compute β for pointwise values of τ. To compute the whole curve $\tau \to \beta$, we construct quantile curve regression.
Since we know that β_τ solves the primal problem

$$\max_{V_{\tau_i} \in [0,1]} \mathbb{E}[Y V_{\tau_i}]$$

$$s.t.\ \mathbb{E}[V_{\tau_i} X] = (1 - \tau_i)\mathbb{E}[X]\ [\beta]$$

for $\tau_i = \frac{i}{n},\ 0 \leq i \leq n$
Combine them by taking the sum of these problems,

$$\max_{V_{\tau_i} \in [0,1]} \sum_{\tau_i} \mathbb{E}[Y V_{\tau_i}]$$

$$s.t.\ \forall i,\ \mathbb{E}[V_{\tau_i} X] = (1 - \tau_i)\mathbb{E}[X]$$

Taking the limit and we may convert the sum into an integral

$$\max_{V(\cdot) \geq 0} \int_0^1 \mathbb{E}[V_\tau Y] d\tau$$

$$s.t.\ V(\tau) \leq 1$$

$$\mathbb{E}[V(\tau) X] = (1 - \tau)\mathbb{E}[X]$$

Similarly, its dual is constructed as

$$\min_{P \geq 0, \beta} \int_0^1 \mathbb{E}[P(\tau) + (1 - \tau) X^T \beta(\tau)] d\tau$$

$$s.t.\ P(\tau) \geq Y - X^T \beta(\tau)$$

To better compute it with a computer program, we write its sample version. Here, we observe as sample (X_i, Y_i) for $i \in \{1, \ldots, I\}$. We discretize the probability space $[0, 1]$ into T points, $\tau_1 = 0 < \tau_2 < \ldots < \tau_T \leq 1$. K is a matrix representing samples of X. Let $\bar{x}$ be the $1 \times K$ row vector whose k-th entry is $\sum_{1 \leq i \leq I} X_{ij}/I$.

$$\max_{V_{ij} \geq 0} \frac{1}{I} \sum_{i,j} V_{ij} Y_j$$

$$s.t.\ V_{ij} \leq 1$$

$$\frac{1}{I}(VK)_{ik} = (1 - \tau_i)\bar{x}_k$$

In matrix terms, it is

$$\max_{V \geq 0} \frac{1}{I} \mathbf{1}_T^\top VY$$

$$s.t.\ V \leq 1$$

$$\frac{1}{I}(VK) = (1 - \tau)\bar{x}^\top$$

where $\mathbf{1}_T^\top VY$ is a compact way of representing the sum of the products $V_{ij}Y_j$. After vectorization, $v = vec(V)$, it becomes

$$\max_{V \geq 0} \frac{1}{I}(\mathbf{1}_T \otimes V)^\top v$$

$$s.t.\ V \leq 1$$

$$\frac{1}{I}(I_T \otimes X^\top)v = vec((1 - \tau)\bar{x}^\top)$$

This version of the problem is more suitable for optimization solvers that typically operate on vectors.

3.4 Computing

We may implement the quantile regression by solving the LP problem. There are many powerful linear programming solvers, like *linprog* in Matlab. Here, we use Python Gurobi package to compute the linear programming task, with an example dataset from Koenker, Roger and Kevin F. Hallock. "Quantile Regression". Journal of Economic Perspectives, Volume 15, Number 4, Fall 2001, Pages 143–156.

We would like to study the relationship between income and expenditures on food for a sample of working class Belgian households in 1857 (the Engel data). For the use of comparison, we set the quantile as 0.5, least absolute deviation (LAD) model.

```
Y_i_1 = food.reshape((-1,1))
nbt=21
τ_t_1 = np.linspace(0,1,nbt).reshape((-1,1))
A = spr.kron(spr.identity(nbt),X_i_k.T) / nbi
obj = np.kron(np.ones((nbt,1)),Y_i_1).T / nbi
xbar_1_k = X_i_k.mean(axis = 0).reshape((1,-1))
rhs = ((1-τ_t_1) * xbar_1_k).flatten()
qrs_lp=grb.Model()
qrs_lp.setParam( 'OutputFlag', False )
v = qrs_lp.addMVar(shape=nbi*nbt, name="v",lb=0,ub=1)
qrs_lp.setObjective(obj @ v , grb.GRB.MAXIMIZE)
qrs_lp.addConstr(A @ v == rhs)
qrs_lp.optimize()

βqrs_t_k = np.array(qrs_lp.getAttr('pi')).reshape((nbt,nbk))
βqrs_t_k[10,:]
```

```
array([81.48614818,  0.56017469])
```

Figure 1: LP Approach - Quantile Curve Regression

Let's compare the result obtained above with the quantile regression package from statsmodels.

```
In [17]: import statsmodels.api as sm
         import statsmodels.formula.api as smf
         import matplotlib.pyplot as plt
         import scipy.sparse as spr
```

```
In [18]: τ = 0.5
         model = smf.quantreg('food ~ income', engle_data)
         print(model.fit(q=τ).summary())
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                   food   Pseudo R-squared:               0.6205
Model:                       QuantReg   Bandwidth:                      64.51
Method:                 Least Squares   Sparsity:                       209.3
Date:                Thu, 13 Jul 2023   No. Observations:                 235
Time:                        15:40:38   Df Residuals:                     233
                                        Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     81.4862     14.635      5.568      0.000      52.652     110.321
income         0.5602      0.013     42.513      0.000       0.534       0.586
==============================================================================

The condition number is large, 2.38e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2: statsmodels_quantreg Approach

Notice that the coefficients (intercept and income) calculated by the two methods are nearly the same with a high accuracy.

4 One-Dimensional Vector Quantile Regression

4.1 Intuition

We know that in OLS regression, the error and random variable X are not independent since the residual is orthogonal with $X\beta$: $Y = X\beta + \epsilon$. Is there a way to loosen the constraint such that the variable is independent of the error in a certain manner? To form this question in a better way, let X be a random variable in $\mathbb{R}^d$, Y a random variable in $\mathbb{R}$, and X, Y both have continuous distribution (density function exists). Is there a function $Q(X,t)$ such that Q is increasing in t for each X and there exists another random variable U independent from X following a uniform distribution on $[0,1]$ that makes $Y = Q(X,U)$?

First, we need to define the third random variable U. Since we know the quantile map is a nondecreasing map, the function Q could likely relate to a quantile function. We define U such that $Y = Q_{Y|X}(U|X)$. By definition of a quantile function, we have $U = F_{Y|X}(Y|X) = \mu[0,1]$ (the inverse relationship between the quantile function and cumulative distribution function). Thus, we have a random variable U correctly defined, which follows a uniform distribution given certain X.

$$\begin{aligned}\mathbb{P}(U < t \mid X = x) &= \mathbb{P}(F_{Y|X}(Y|x) < t \mid X = x) \\ &= \mathbb{P}(Y < Q_{Y|X}(t|x) \mid X = x) \\ &= t\end{aligned}$$

Thus, we proved that U is uniformly distributed and independent from X.
To find the β such that

$$\arg\min_{\beta} \mathbb{E}[\rho_\tau(Y - \beta^T X)]$$

Still, we look for the local minimum by taking its derivative equal to 0.

$$\mathbb{E}[X\rho'_\tau(Y - \beta^T X)] = \mathbb{E}[X(\mathbb{1}(Y < \beta_\tau^T X) - \tau)] = 0$$

Assume $X_1 = 1$, which means the first component of X 1. This normalization makes it easy for interpretation and the transformation of regressors. Define U such that $Y = \beta(U)^T X$. Then, solve U as a function of Y, X. We plug this into the above equation and use the fact that U is non-decreasing

$$\mathbb{E}[X(\mathbb{1}(\beta(U)^T X < \beta_\tau^T X) - \tau)] = \mathbb{E}[X(\mathbb{1}(U \le \tau) - \tau)] = 0$$

What does this equation suggest about the independence between X and U?
Since the indicator function depends on the size of U, we have conditional expectation from the previous eqaution

$$\mathbb{E}[X|U] = \tau E[X] = 0$$

Thus, $\mathbb{E}[X|U] \sim E[X]$. We have deduced the mean independence between X and U, which is weaker than independence since we only have the expectation.

Like the monotonicity implied by the quantile function, this is also a natural constraint for quantile regression construction. Note that this constraint can be further loosened until full independence. "In both vector and scalar cases, we have that, when the conditional quantile function is linear, the quasi- linear representation coincides with the linear representation and U becomes fully independent of X."

If we define $Q(X, U_\tau)$ as the τ^{th} quantile of the conditional distribution of Y given X, then U_τ follows a uniform distribution on $[0, 1]$. This is also a utilization of the Probability Integral Transform with U as the cumulative density function map.

From the perspective of loss function, instead of minimizing the conditional means as in OLS, quantile regression minimizes the absolute deviation through conditional quantile. Through the formulation in 3.1, the major advantage of quantile regression over ordinary least squares (OLS) regression is its ability to handle non-constant variance (heteroskedasticity) and hence it tends to provide a more complete view of possible causal relationships between variables.

4.2 Definition

4.2.1 Vector Quantile for General VQR

In some fixed nonatomic probability space, $(\Omega, F, \mathbb{P})$, given a random vector Z with values in $\mathbb{R}^k$ defined on this space, we will denote by $\mathscr{L}(Y)$ the law of Z. We fix as a reference measure the uniform measure on the unit cube $[0, 1]^d$

$$\mu_d := \mathcal{U}\left([0, 1]^d\right)$$

Given Y, an integrable $\mathbb{R}^d$-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, Brenier's Theorem states that there exists a unique $U \sim \mu_d$ and a unique convex function defined on $[0, 1]^d$ such that

$$Y = \nabla\varphi(U).$$

The map $\nabla\varphi$ is called the **Brenier's map** between μ_d and $\mathscr{L}(Y)$.
The **vector quantile** of Y is defined to be the Brenier's map between μ_d and $\mathscr{L}(Y)$.
In one-dimensional space, the optimal transport map of Brenier is given by $\nabla\varphi = Q$, where Q is the quantile of Y. Monotonicity persists in both one-dimensional and higher dimensions.

4.2.2 Conditional Vector Quantile

For $m = \mathscr{L}(X)$-a.e. $x \in \mathbb{R}^N$, the vector conditional quantile of Y given $X = x$ is the Brenier's map between $\mu_d := \mathcal{U}\left([0, 1]^d\right)$ and $\nu(. \mid x) := \mathscr{L}(Y \mid X = x)$. We denote this well defined map as $\nabla\varphi_x$ where φ_x is a convex function on $[0, 1]^d$.

We denote by ν^x the conditional probability of Y given $X = x$. Thus, $\nu = m \otimes \nu^x$.

4.3 Linear Programming

Recall the definition of quantile as the inverse of a CDF function. There is a natural constraint of monotonicity imposed on the dual (Koenker and Ng):

$$\begin{aligned}\min_{P\geq 0, N\geq 0, \beta} \int_0^1 \mathbb{E}\left[P(\tau) + (1-\tau)X^\top \beta(\tau)\right] d\tau \\ \text{s.t. } P(\tau) - N(\tau) = Y - X^\top \beta(\tau) \\ X^\top \beta(\tau) \geq X^\top \beta\left(\tau'\right), \tau \geq \tau'\end{aligned}$$

To solve it more easily, we consider its primal formulation (Carlier, Chernozhukov and Galichon). Given that

$$V(\tau) = 1\left\{Y \geq X^\top \beta(\tau)\right\},$$

We have $X^\top \beta(\tau)$ nondecreasing in $\tau \Longrightarrow V(\tau)$ nonincreasing.
Consider the quantile curve regression discussed in the previous section. The primal problem is

$$\begin{aligned}\max_{V(\tau)} \int_0^1 \mathbb{E}[YV(\tau)]d\tau \\ \text{s.t. } 0 \leq V(\tau) \leq 1 \\ \mathbb{E}[V(\tau)X] = (1-\tau)\mathbb{E}[X] \\ V(\tau) \leq V\left(\tau'\right), \tau \geq \tau'\end{aligned}$$

Assume $\tau_1 = 0 < \ldots < \tau_T \leq 1$ and let $\bar{x}$ be a $1 \times K$ row vector whose k-th entry is $\mathbb{E}\left[X_k\right]$. I is the dimension of Y. T is the size of the partition set of τ. The sampled version of the previous primal problem is

$$\begin{aligned}\max_{V_{ti}\geq 0} \frac{1}{I} \sum_{\substack{1\leq i\leq I \\ 1\leq t\leq T}} V_{ti} Y_i \\ s.t.\ V_{ti} \leq 1 \\ \frac{1}{I} \sum_{1\leq i\leq I} V_{ti} X_{ik} = (1-\tau_t)\,\bar{x}_k \\ V_{(t+1)i} \leq V_{ti}\end{aligned}$$

Since $\tau_1 = 0$, we have $V_{1i} = 1$ from the second constraint. The program becomes

$$\begin{aligned}\max_{V_{ti}} \frac{1}{I} \sum_{\substack{1\leq i\leq I \\ 1\leq t\leq T}} V_{ti} Y_i \\ s.t.\ V_{1i} = 1 \\ \frac{1}{I} \sum_{1\leq i\leq I} V_{ti} X_{ik} = (1-\tau_t)\,\bar{x}_k \\ V_{1i} \geq V_{2i} \geq \ldots \geq V_{(T-1)i} \geq V_{Ti} \geq 0\end{aligned}$$

Next, let τ be the $T \times 1$ row matrix with entries τ_k and D be a $T \times T$ matrix defined as

$$D = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \ddots & \vdots & \vdots \\ 0 & -1 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & & 0 & -1 & 1 & 0 \\ 0 & & 0 & 0 & -1 & 1 \end{pmatrix}$$

we have $V^\top D \geq 0$ if and only if

$$V_{1i} \geq V_{2i} \geq \ldots \geq V_{(T-1)i} \geq V_{Ti} \geq 0$$

We can write the sampled version into the following matrix form

$$\begin{aligned} & \frac{1}{I} \max_V 1_T^\top VY \\ s.t. \ & \frac{1}{I} VX = (1_T - \tau)\, \bar{x} \\ & V^\top D 1_T = 1_I \\ & V^\top D \geq 0 \end{aligned}$$

Suppose $\pi = \frac{D^\top V}{I}$and $U = D^{-1}\mathbf{1}_I$, $\mu = D^\top (1_T - \tau)$ and $p = \frac{\mathbf{1}_I}{I}$. It is equivalent to

$$\begin{aligned} & \max_\pi U^\top \pi Y \\ s.t. \ & \pi X = \mu \bar{x} \\ & \pi^\top 1_T = p \\ & \pi \geq 0 \end{aligned}$$

Assume that the first entry of X is one for the ease of computation. If π satisfies the constraints

$$\sum_{i=1}^{I} \pi_{ti} = \mu_t \text{ and } \sum_{t=1}^{T} \pi_{ti} = p_i$$

then π can be thought of as a joint probability on τ and X given the marginal probability constructed above.

$$\begin{aligned} & \max_{\pi \geq 0} \sum_{\substack{1 \leq t \leq T \\ 1 \leq i \leq I}} \pi_{ti} U_t Y_i \\ s.t. \ & \sum_{1 \leq i \leq I} \pi_{ti} X_{ik} = \mu_t \bar{x}_k \\ & \sum_{1 \leq t \leq T} \pi_{ti} = p_i \end{aligned}$$

To compute it with linear programming packages like Gurobi, we need to vectorize the matrices. Given $\text{vec}(A\pi B) = \left(A \otimes B^\top\right) \text{vec}(\pi)$, the constraint becomes

$$\begin{pmatrix} I_T \otimes X^\top \\ 1_T^\top \otimes I_I \end{pmatrix} \text{vec}(\pi) = \begin{pmatrix} \text{vec}(\mu\bar{x}) \\ p \end{pmatrix}$$

There are IT primal variables and $KT + I$ constraints.
β is the vector of Lagrange multipliers of the constraint $\frac{1}{I} VX = (1_T - \tau)\,\bar{x}$ in the former problem.
Let ψ be the vector of Lagrange multipliers of the constraint $\pi X - \mu\bar{x}$ in the latter problem. We have $\beta = D\psi$.

$$(\pi X - \mu\bar{x})^\top \psi = 0$$

thus

$$\left(\frac{1}{I} D^\top VX - D^\top (1_T - \tau)\,\bar{x}\right)^\top \psi = 0$$

and therefore

$$\left(\frac{1}{I} VX - (1_T - \tau)\,\bar{x}\right)^\top D\psi = 0$$

It can be rewritten into the one-dimensional vector quantile regression construction in the continuous case (Carlier, Chernozhukov and Galichon)

$$\begin{aligned} &\max_{\pi} \mathbb{E}_\pi[UY] \\ s.t.\ &U \sim \mu \\ &(X, Y) \sim P \\ &\mathbb{E}[X \mid U] = \mathbb{E}[X] \end{aligned}$$

This is an extension of the optimal transport problem of Monge-Kantorovich. When X is restricted to the constant, it is an optimal transport problem.

If we replace the mean-independence between X and U by independence using conditional probability (scalar VQR, by $Thm.3.3$), we have

$$\begin{aligned} &\max_{\pi} \mathbb{E}_\pi[UY] \\ &\text{s.t.U } \sim \mu \\ &(X, Y) \sim P \\ &X \perp\!\!\!\perp U \end{aligned}$$

The solution to the latter problem is simply $U = F_{Y|X}(Y \mid X)$ and the nonparametric conditional quantile representation is $Y = F^{-1}_{Y|X}(U \mid X)$. This conforms with the classical quantile function we constructed above!
Dual form:

$$\begin{aligned} &\min_{\psi, b} \mathbb{E}_P[\psi(X, Y)] + \bar{x}^\top \mathbb{E}_\mu[b(U)] \\ &\text{s.t. } \psi(x, y) + x^\top b(\tau) \geq \tau y, \forall x, y, \tau \end{aligned}$$

The optimality of (ψ, b) gives us solution

$$\psi(x, y) = \sup_{\tau \in [0,1]} \left\{ \tau y - x^\top b(\tau) \right\}$$

when b is differentiable, the conditional quantile function is linear

$$Y = X^\top \beta(U)$$

where (U, X, Y) are the solutions to the primal problem and $\beta(\tau) = b'(\tau)$.

When the components of Y are independent, we can run the scalar version component by component. The general case will not be discussed here.

4.4 Computation

```
In [21]: D_t_t = spr.diags([1, -1], [ 0, -1], shape=(nbt, nbt))

         U_t_1 = np.linalg.inv(D_t_t.toarray()) @ np.ones( (nbt,1))
         μ_t_1 = D_t_t.T @ (np.ones((nbt,1)) - τ_t_1)

         A1 = spr.kron(spr.identity(nbt),X_i_k.T)
         A2 = spr.kron(np.array(np.repeat(1,nbt)),spr.identity(nbi))
         A = spr.vstack([A1, A2])
         rhs = np.concatenate( [(μ_t_1 * xbar_1_k).flatten(), np.ones(nbi)/nbi])
         obj = np.kron(U_t_1, Y_i_1).T
         vqr_lp=grb.Model()
         pi = vqr_lp.addMVar(shape=nbi*nbt, name="pi")
         vqr_lp.setParam( 'OutputFlag', False )
         vqr_lp.setObjective( obj @ pi, grb.GRB.MAXIMIZE)
         vqr_lp.addConstr(A @ pi == rhs)
         vqr_lp.optimize()

         ϕ_t_k = np.array(vqr_lp.getAttr('pi'))[0:(nbt*nbk)].reshape((nbt,nbk))

         βvqr_t_k = D_t_t.toarray() @ ϕ_t_k

         βvqr_t_k[10,:]

Out[21]: array([81.48614818,  0.56017469])
```

Figure 3: 1-D VQR computation

Using the formulation deduced above, we reached a coefficient result identical to what we computed above. This helps to verify the equivalence between classical quantile regression and one-dimensional vector quantile regression.