Linear Programming 1

Exercise 1: Sparse Quantile Regression

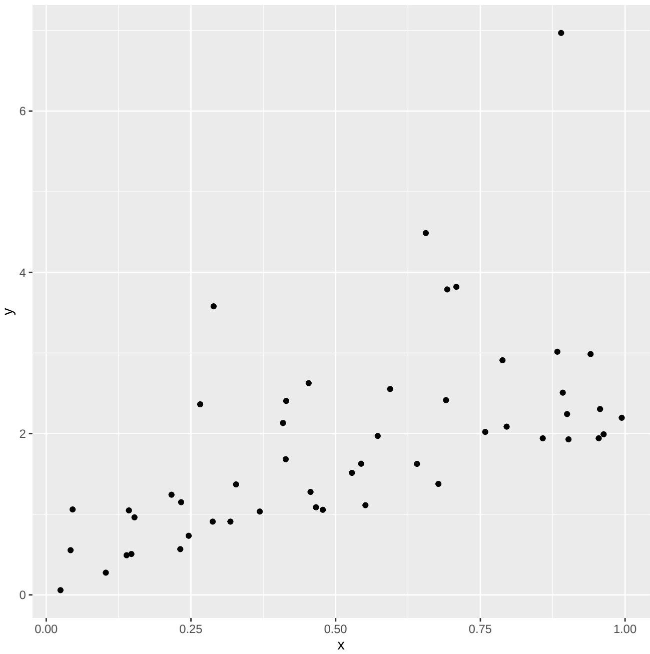
You are given the following data situation:

```
library(ggplot2)

set.seed(123)

# generate 50 numerical observations with skewed error distribution (gamma)
n = 50
x = runif(n)
y = 2 * x + rgamma(n, shape = 1)

ggplot(data.frame(x = x, y = y), aes(x=x, y=y)) +
  geom_point()
```



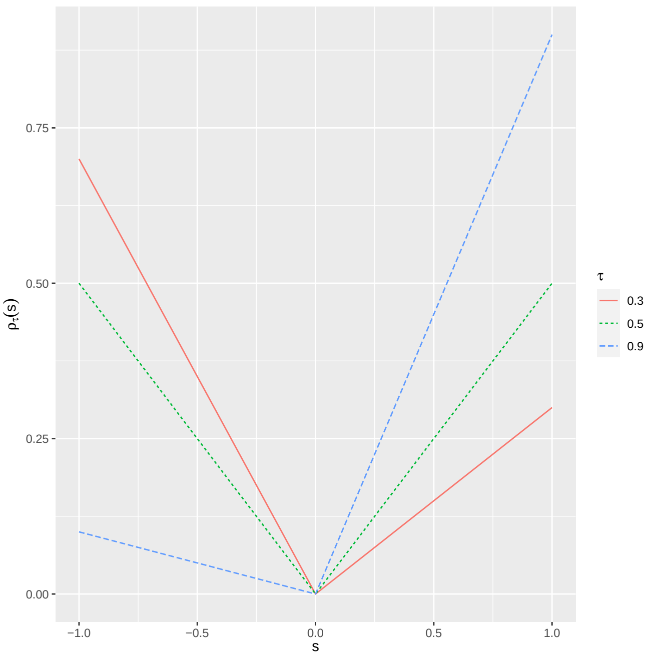
The general multivariate sparse quantile regression can be stated as in the lecture such as

$$\min_{\beta \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y^{(i)} - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}^{(i)})}_{=\mathcal{R}_{\text{emp}}} \text{ s.t. } \|\boldsymbol{\beta}\|_1 \leq t$$

with $\tau \in (0, 1)$

$$\rho_\tau(s) = \begin{cases} \tau \cdot s & \text{for } s > 0 \\ -(1-\tau) \cdot s & \text{for } s \leq 0 \end{cases}$$

The $\rho_\tau(s)$ function is also called pinball loss. For $\tau \neq 0.5$, it asymmetrically assigns weights to the residuals:



In the following we consider only the univariate case:

(a) Find the standard form (as defined in the lecture) of the one-dimensional sparse quantile regression
Hint: If an unconstrained variable x is decomposed into two non-negative variables such that $x = x^+ - x^-$, then the absolute value $|x| = x^+ + x^-$. This works since the optimization algorithm ensures that at most one of the variables x^+, x^- is not zero.

(b) Plot $\mathcal{R}_{\text{emp}}$ for $(\beta_0, \beta_1) \in [-3, 3] \times [-3, 3]$ and $\tau = 0.4$ and mark the feasible region ($t = 1.7$)

(c) Use the `solveLP` command in `R` to solve the sparse 40% quantile regression ($t = 1.7$)

(d) State the corresponding dual formulation of a) (as defined in the lecture)