```r
library(ggplot2)
#install.packages("lme4")
library(lme4)
#install.packages("DHARMa")
# library(DHARMa)
library(quantreg)
library(dplyr)
library(ggplot2)
library(tinytex)
```

Methods

Design Matrix

The design matrix is defined to be a matrix $\mathbf{X}$ such that $\mathbf{X}_{ij}$ (the j^{th}) column of the i^{th} row of $\mathbf{X}$) represents the value of the j^th variable associated with the i^{th} variable object.

A regression model may be represent via matrix multiplication as

$$y = \mathbf{X}\beta + e$$

where X is the design matrix, β is a vector of the model's coefficient (one for each variable), e is a vector of random errors with a mean zero, and y is the vector outputs for each object.

Ordinary least squares

Ordinary least squares model or OLS, works by creating a line through the data points. Then it calculates the difference between each prediction and observation (residual). And it tries to minimize the squared value of the residuals. The ordinary least squares is defined by:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

The least squares estimates in this case are given by simple formulas

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

How does the minimization of absolute deviations equal the media?

Definition of mean

Assume, without loss of generality, that Y is a continuous random variable. The expected value of the absolute sum of deviations from a given center c can be split into the following two terms:

$$E|Y-c| = \int_{y \in R} |y-c| f(y) dy = \int_{y<c} |y-c| f(y) dy + \int_{y>c} |y-c| f(y) dy$$

If y is less than c, then y-c will always be negative. Therefore, |y-c|=-(c-y). By a similar argument, |y-c| is just (y-c) when y > c.

$$= \int_{y<c} (c-y) f(y) dy + \int_{y>c} (y-c) f(y) dy$$

Since the absolute value is convex, differentiating E|y-c| with respect to c and setting the partial derivatives to zero will lead to the solution of the minimum.

$$\frac{\partial}{\partial c} E|y-c| = 0$$

$$\left\{ (c-y) f(y)|_{-\infty}^{c} + \int_{y<c} \frac{\partial}{\partial c} (c-y) f(y) dy \right\} +$$
$$\left\{ (y-c) f(y)|_{c}^{+\infty} + \int_{y>c} \frac{\partial}{\partial c} (y-c) f(y) dy \right\} = 0$$

The limit of any PDF approaching positive or negative infinity will equal 0, therefore the previous equation simplifies to:

$$\left\{ \int_{y<c} \frac{\partial}{\partial c} (c-y) f(y) dy \right\} +$$
$$\left\{ \int_{y>c} \frac{\partial}{\partial c} (y-c) f(y) dy \right\} = 0$$

Taking the partial, $\frac{\partial}{\partial c}(c-y)f(y)$ = f(y) and $\frac{\partial}{\partial c}(y-c)f(y)$ = -f(y).

$$\left\{\int_{y<c} \theta f(y)dy\right\} +$$
$$\left\{\int_{y>c} -\theta f(y)dy\right\} = 0$$

Using the CDF definition and the notion of reciprocals, the previous equation simplifies to: $F(c) - [1 - F(c)] = 0$ and thus $2F(c) - 1 = 0 \longrightarrow F(c) = \frac{1}{2} \longrightarrow$ c=Me.

Thus the minimization to a weighted least absolute deviation loss function is the value that gives the theta$\widehat{}${th} quantile.

Generalization least absolute deviations

The solution of the minimization problem formulated in Equation (1.2) is thus the median. The above solution does not change by multiplying the two components of $E|Y-c|$ by a constant θ and $(1-\theta)$, respectively. This allows us to formulate the same problem for the generic quantile θ. Namely, using the same strategy for Equation (1.5), we obtain:

$$\frac{\partial}{\partial c}E\left[\rho_\theta(Y-c)\right] = \frac{\partial}{\partial c}\left\{(1-\theta)\int_{-\infty}^{c}|y-c|f(y)dy + \theta\int_{c}^{+\infty}|y-c|f(y)dy\right\}.$$

Repeating the above argument, we easily obtain:

$$\frac{\partial}{\partial c}E\left[\rho_\theta(Y-c)\right] = (1-\theta)F(c) - \theta(1-F(c)) = 0$$

and then q_θ as the solution of the minimization problem:

$$F(c) - \theta F(c) - \theta + \theta F(c) = 0 \Longrightarrow F(c) = \theta \Longrightarrow c = q_\theta.$$

By replacing the sorting with optimization, the above line of reasoning generalizes easily to the regression setting. In fact, interpreting Y as a response variable and $\mathbf{X}$ as a set of predictor variables, the idea of the unconditional mean as the minimizer of Equation (1.1) can be extended to the estimation of the conditional mean function:

$$\widehat{\mu}\left(\mathbf{x}_i,\beta\right) = \underset{\mu}{\operatorname{argmin}} E\left[Y-\mu\left(\mathbf{x}_i,\beta\right)\right]^2,$$

In the case of a linear mean function, $\mu(x_i,\beta) = x_i^T\beta$ so the previous equation becomes:

$$\widehat{\beta} = \underset{\beta}{argmin} E[Y - x_i^T\beta]^2$$

By the same argument,

$$q_\theta = \underset{c}{\operatorname{argmin}} E\left[\rho_\theta(Y-c)\right]$$

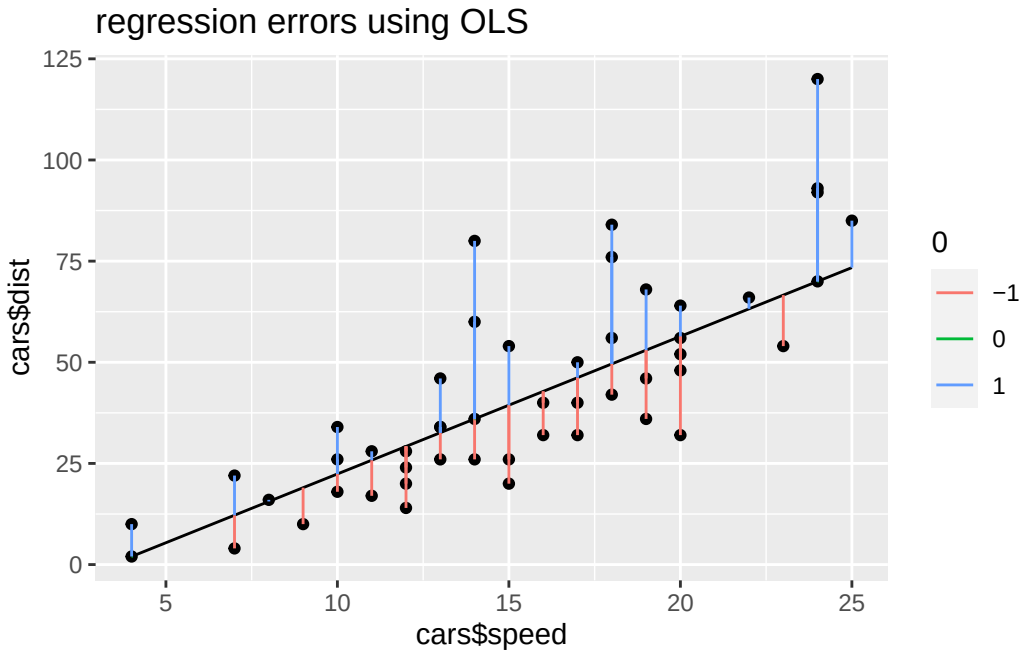
where $\rho_\theta(.)$ denotes the following loss function:

$$\begin{aligned}\rho_\theta(y) &= [\theta - I(y < 0)]y \\ &= [(1-\theta)I(y \leq 0) + \theta I(y > 0)]|y|.\end{aligned}$$

Graphic

```
data(cars)


rq50 <- rq(dist ~ speed, data=cars, tau=0.5)
yhat<-rq50$fitted.values
color = sign(rq50$residuals)
qplot(x=cars$speed, y=cars$dist)+geom_line(y=yhat)+
       geom_segment(aes(x=cars$speed, xend=cars$speed, y=cars$dist, yend=yhat, group=as.fa
       labs(title="regression errors using OLS", color=color)
```

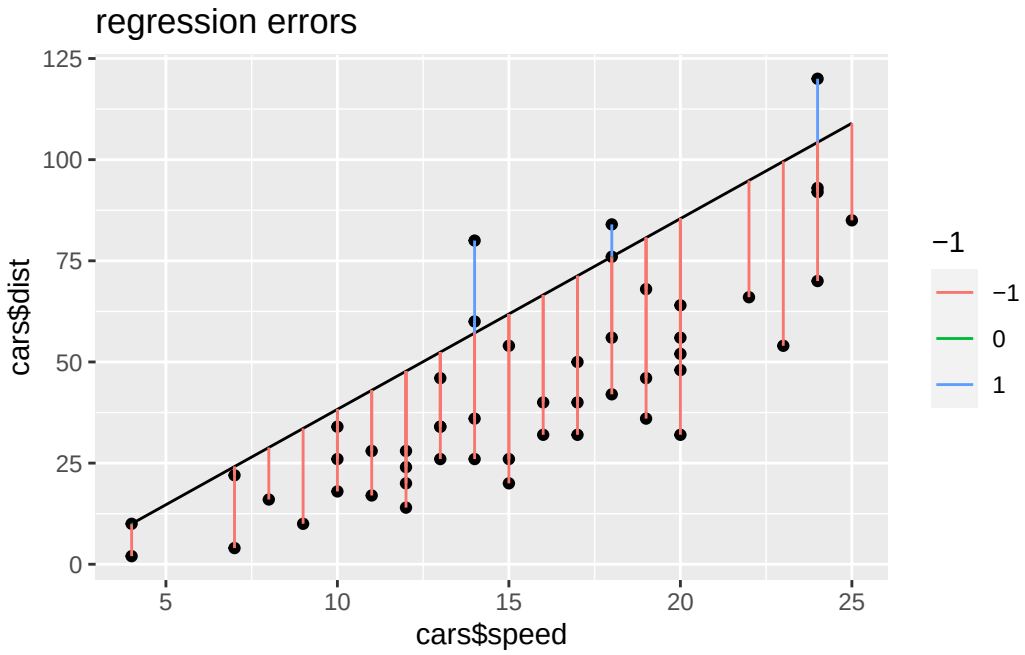



```
table(color)
```

```
color
-1  0  1
24  3 23
```

Notice that approximately half of the distribution of the points are above the QR line and approximately half are above the QR line. Now let's see what happens when we look at the 90th conditional quantile.

```
rq90 <- rq(dist ~ speed, data=cars, tau=0.9)
yhat<-rq90$fitted.values
color = sign(rq90$residuals)
qplot(x=cars$speed, y=cars$dist)+geom_line(y=yhat)+
       geom_segment(aes(x=cars$speed, xend=cars$speed, y=cars$dist, yend=yhat, group=as.fa
       labs(title="regression errors", color=color)
```

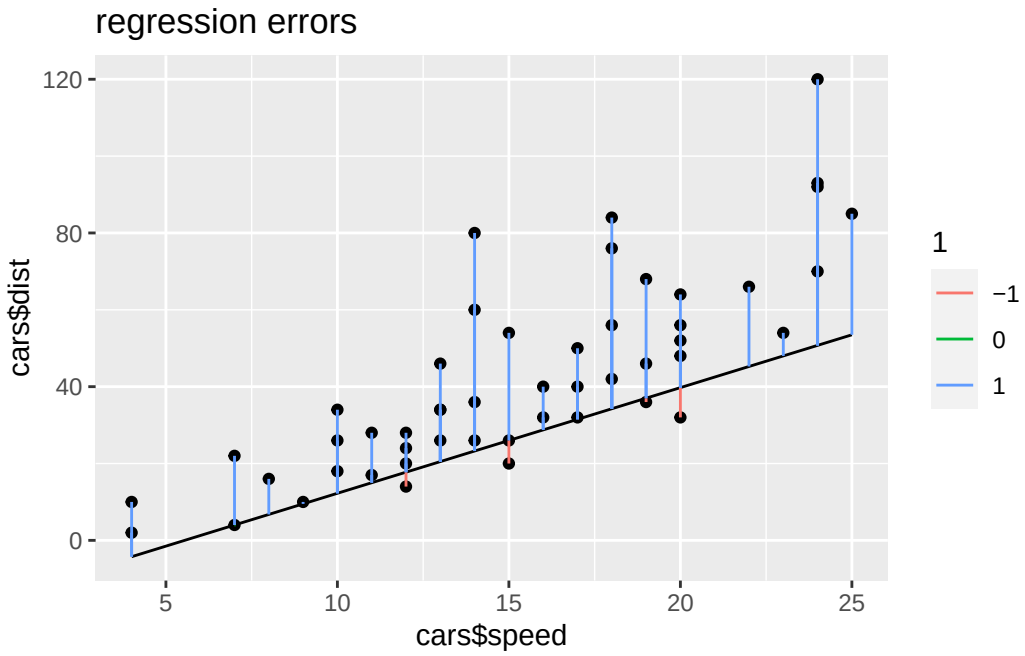



```
table(color)
```

```
color
-1  0  1
44  2  4
```

When we input the .9 for the quantile, we get approximately 90% of the points under the QR line and 10% over the QR line.

```
rq10 <- rq(dist ~ speed, data=cars, tau=0.1)
yhat<-rq10$fitted.values
color = sign(rq10$residuals)
qplot(x=cars$speed, y=cars$dist)+geom_line(y=yhat)+
       geom_segment(aes(x=cars$speed, xend=cars$speed, y=cars$dist, yend=yhat, group=as.fa
       labs(title="regression errors", color=color)
```




```
table(color)
```

```
color
-1  0  1
 5  1 44
```

When we input 0.1 for the 10th percentile, we get approximately 90% of the points above the QR line and 10% below the QR line.

Thus, we can see that QR is not online a robust ## Evaluation metrics

Mean absolute error

The mean absolute error (MAE) is the average magnitude of the errors of the values predicted by the regression and the actual observed values for the response variable. Because it is a simple average, all errors have the same weight, there are no penalties for different magnitude deviations [2]. MAE assumes that the errors are normally distributed, if the error distribution was non-normal, the average may not be a good measure of centrality and can paint a false picture of the goodness-of-fit of the regression curve. MAE also assumes that the errors are unbiased. While the average magnitude of the errors is expected to be non-zero (unless the regression is a perfect fit) the average of the residuals, i.e., the deviation of the predicted value from the actual value, considering underestimation and overestimation. This means on average the regression curve does not over or underestimate.

$$\text{MAE } = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Root mean squared error

It calculates the differences between the predictions and the actual observations (residuals) and then gets their quadratic mean for each. This type of error gives a larger penalty for larger errors [2]. This error also assumes that the errors are unbiased and that they follow a normal distribution. This gives a picture of the size of residuals in comparison to the regression line.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$$

Variance of error

It is a measure of how spread all the errors are from the mean of all errors.

$$\text{Var}(e) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2$$

Min/max error

A measure of the maximum residual for a prediction and the minimum residual.

$$f : X \to \mathbb{R}, \text{ if } (\forall e \in X_{error}) f(e_i) \geq f(e)$$

$$f : X \to \mathbb{R}, \text{ if } (\forall e \in X_{error}) f(e_i) \leq f(e)$$