Quantile Regression

**Quantile regression: what is it?**

Let y \in \mathbb{R} be some response variable of interest, and let x \in \mathbb{R}^p be a vector of features or predictors that we want to use to model the response. In linear regression, we are trying to estimate the conditional mean function, \mathbb{E}[y \mid x], by a linear combination of the features.

While the conditional mean function is often what we want to model, sometimes we may want to model something else. On a [recent episode of the Linear Digressions podcast](http://lineardigressions.com/episodes/2019/1/12/quantile-regression), Katie and Ben talked about a [situation in Uber](https://eng.uber.com/analyzing-experiment-outcomes/) where it might make sense to model a **conditional quantile function**.

Let’s make this more concrete. Say Uber came up with a new algorithm for dispatching drivers and we are interested in how this algorithm fares in terms of wait times for consumers. A simple (linear regression) model for this is

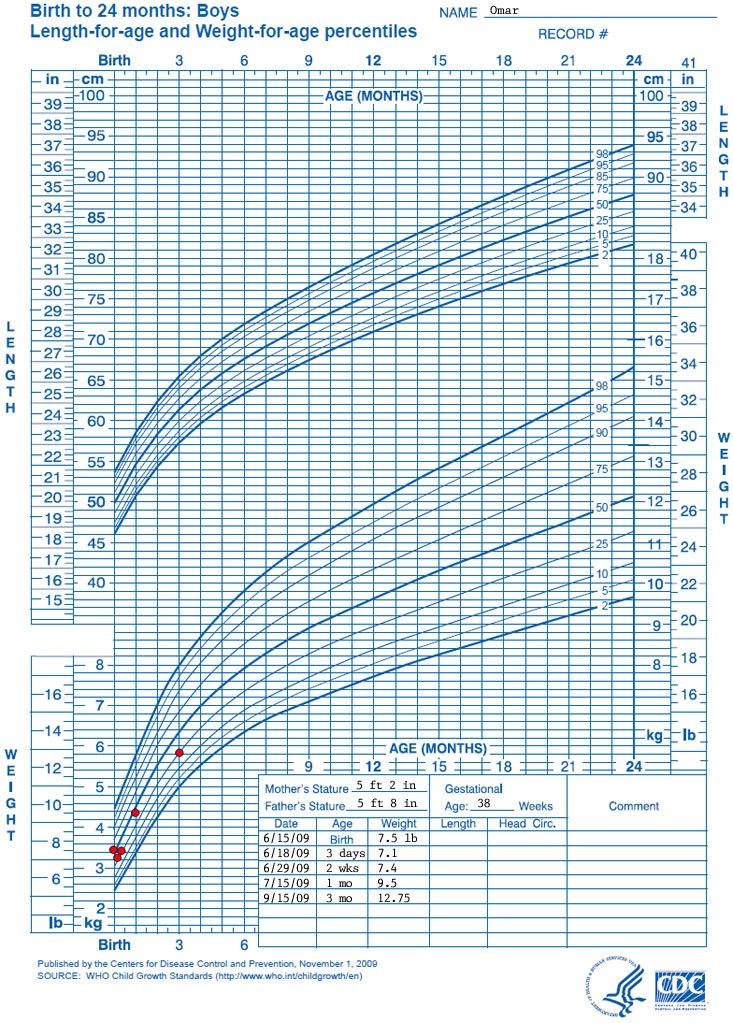
\mathbb{E}[wait\_time \mid algorithm] = a + b \cdot algorithm,

where algorithm = 1 if the new algorithm was used to dispatch a driver, and algorithm = 0 if the previous algorithm was used. From this model, we can say that under the old algorithm, mean wait time was a, but under the new algorithm, mean wait time is a + b. So if b < 0, I would infer that my new algorithm is "doing better".

But is it really? What if the new algorithm improves wait times for 90% of customers by 1 min, but makes the wait times for the remaining 10% longer by 5 min? Overall, I would see a decrease in mean wait time, but things got significantly worse for a segment of my population. What if that 10% whose wait times became 5 minutes longer were already having the longest wait times to begin with? That seems like a bad situation to have, but our earlier model would not pick it up.

One way to pick up such situations is to model ***conditional quantile functions*** instead. That is, trying to estimate the mean of y given the features x, let’s trying to estimate a quantile of y given the features x. In our example above, instead of trying to estimate the mean wait time, we could estimate the 95th quantile wait time to catch anything going wrong out in the tails of the distribution.

Another example where estimating conditional quantiles is useful is in growth charts. All of you have probably seen one of these charts below in a doctor’s office before. Each line in the growth chart represents some quantile for length/weight given the person’s age. We track our children’s length and weight on this chart to see if they are growing normally or not.

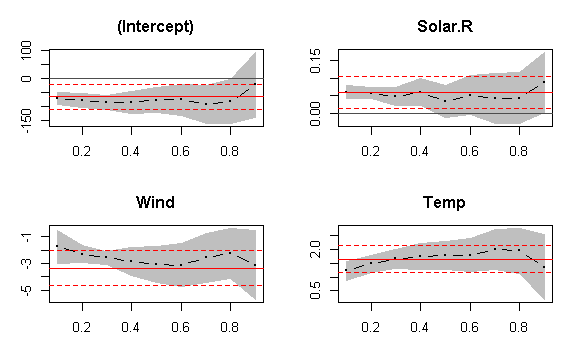


**Quantile regression** is a regression method for estimating these conditional quantile functions. Just as linear regression estimates the conditional mean function as a linear combination of the predictors, quantile regression estimates the conditional quantile function as a linear combination of the predictors.

**Quantile regression in R**

We can perform quantile regression in R easily with the [quantreg](https://cran.r-project.org/package=quantreg" \t "_blank) package. I will demonstrate how to use it on the mtcars dataset. (For more details on the quantreg package, you can read the package’s vignette [here](https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf).)

Let’s load our packages and data:



The dotted red lines are the 95% confidence interval for linear regression and the shaded grey area is the 95% confidence interval for each of the quantreg estimates.

In the plot below, we can visualize that the quantreg estimates are within the bounds of the linear regression estimates which suggests that there may not be a statistically significant difference.