# QUANTILE REGRESSION

# Quantile Regression

- Standard linear regression techniques summarize the average relationship between a set of regressors and the outcome variable based on the conditional mean function E[Y|X]. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of y.
- There are many cases, such as skewed data, multimodal data, or data with outliers, when the behavior at the conditional mean fails to fully capture the patterns in the data.

# Quantile Regression

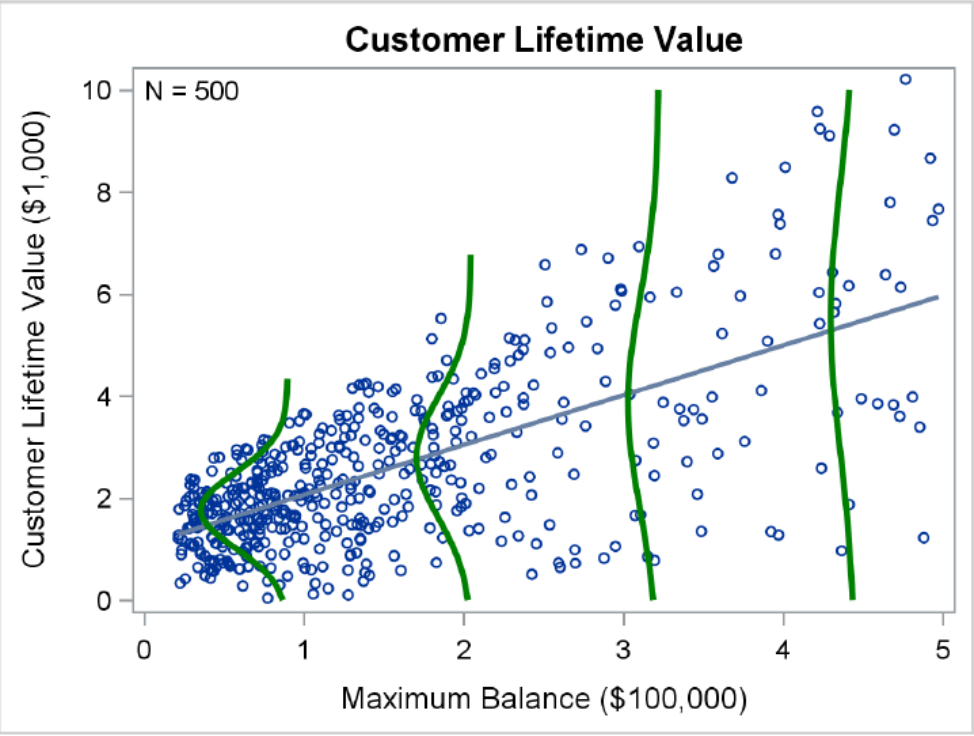
- Quantile regression fits conditional quantiles of the response with a general linear model that assumes no parametric form for the conditional distribution of the response; it gives you information that you would not obtain directly from standard regression methods
- Quantile regression yields valuable insights in applications such as risk management, where answers to important questions lie in modeling the tails of the conditional distribution.

# Quantile Regression?

- By comparison, standard least squares regression models only the conditional mean of the response and is computationally less expensive.
- Quantile regression does not assume a particular parametric distribution for the response, nor does it assume a constant variance for the response, unlike least squares regression.

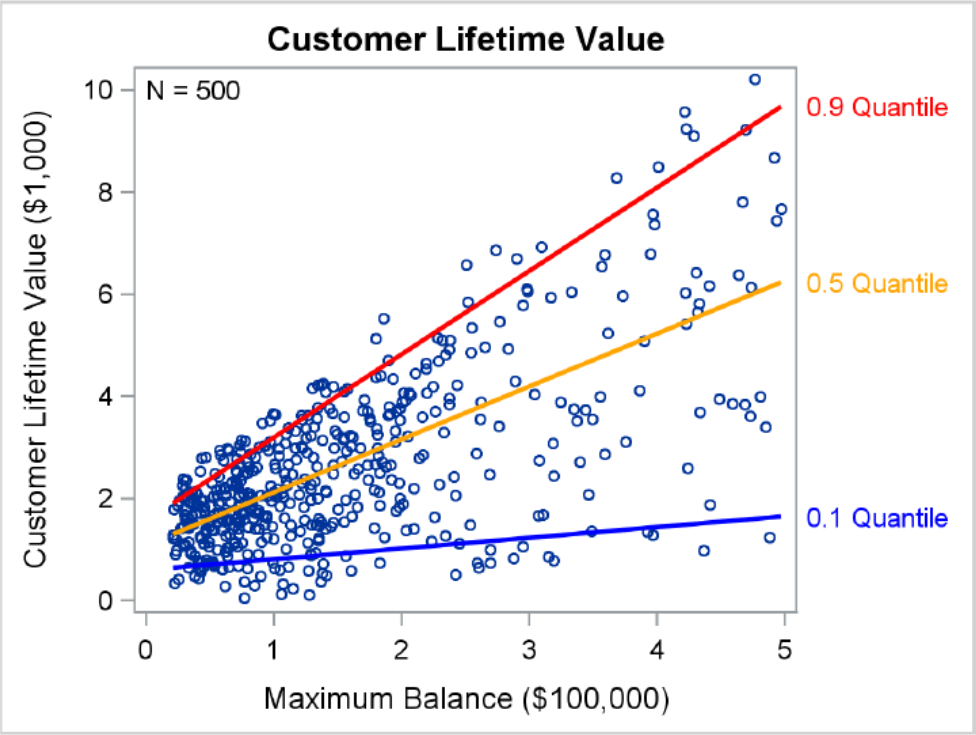
# Quantile Regression

- Least squares regression for a response Y and a predictor X models the conditional mean but it does not capture the conditional variance much less the conditional distribution of Y given



# Quantile Regression

- Figure shows fitted linear regression models for the quantile levels 0.10, 0.50, and 0.90, or equivalently, the 10th, 50th, and 90th percentiles.



# Quantile Regression

| **Linear Regression** | **Quantile Regression** |
|---|---|
| Predicts the conditional mean $E(Y\|X)$ | Predicts conditional quantiles $Q_\tau(Y\|X)$ |
| Applies when $n$ is small | Needs sufficient data |
| Often assumes normality | Is distribution agnostic |
| Does not preserve $E(Y\|X)$ under transformation | Preserves $Q_\tau(Y\|X)$ under transformation |
| Is sensitive to outliers | Is robust to response outliers |
| Is computationally inexpensive | Is computationally intensive |

# Quantile Regression

- Say we quantile regression to analyze Major League Baseball Salary data at the 10%, 25%, 50%, 75%, and 90% quantiles. We will consider the model

$$ln(salary) = \beta_0 + \beta_1 AtBats + \beta_2 Hits + \beta_3 HmRun + \beta_4 Walks + \beta_5 Years + \beta_6 PutOuts$$

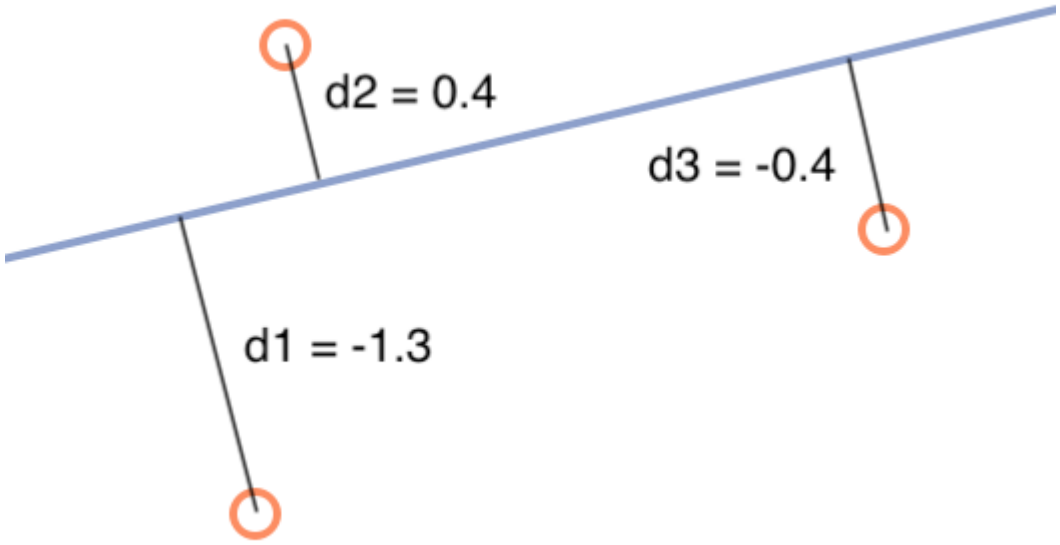
# Quantile Regression

- The least squares estimate minimizes the sum of the squared error terms
- Comparatively, quantile regression minimizes a weighted sum of the positive and negative error terms (where τ is the quantile level):

$$\tau \sum_{y_i > \hat{\beta}_\tau{}' X_i} |y_i - \hat{\beta}_\tau{}' X_i| + (1-\tau) \sum_{y_i < \hat{\beta}_\tau{}' X_i} |y_i - \hat{\beta}_\tau{}' X_i|$$

# Quantile Regression

- *Each orange circle represents an observation while the blue line represents the quantile regression line. The black lines illustrate the distance between the regression line and each observation, which are labelled d1, d2 and d3.*

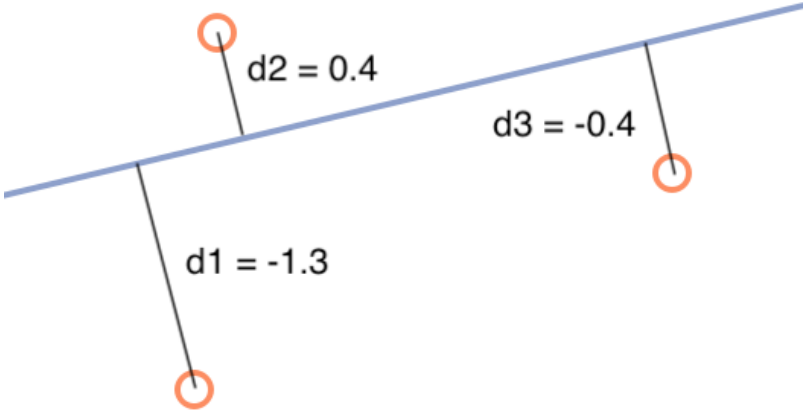


# Quantile Regression

- If τ is equal to 0.9, we can compute the quadratic regression loss for the data in the image, like this:

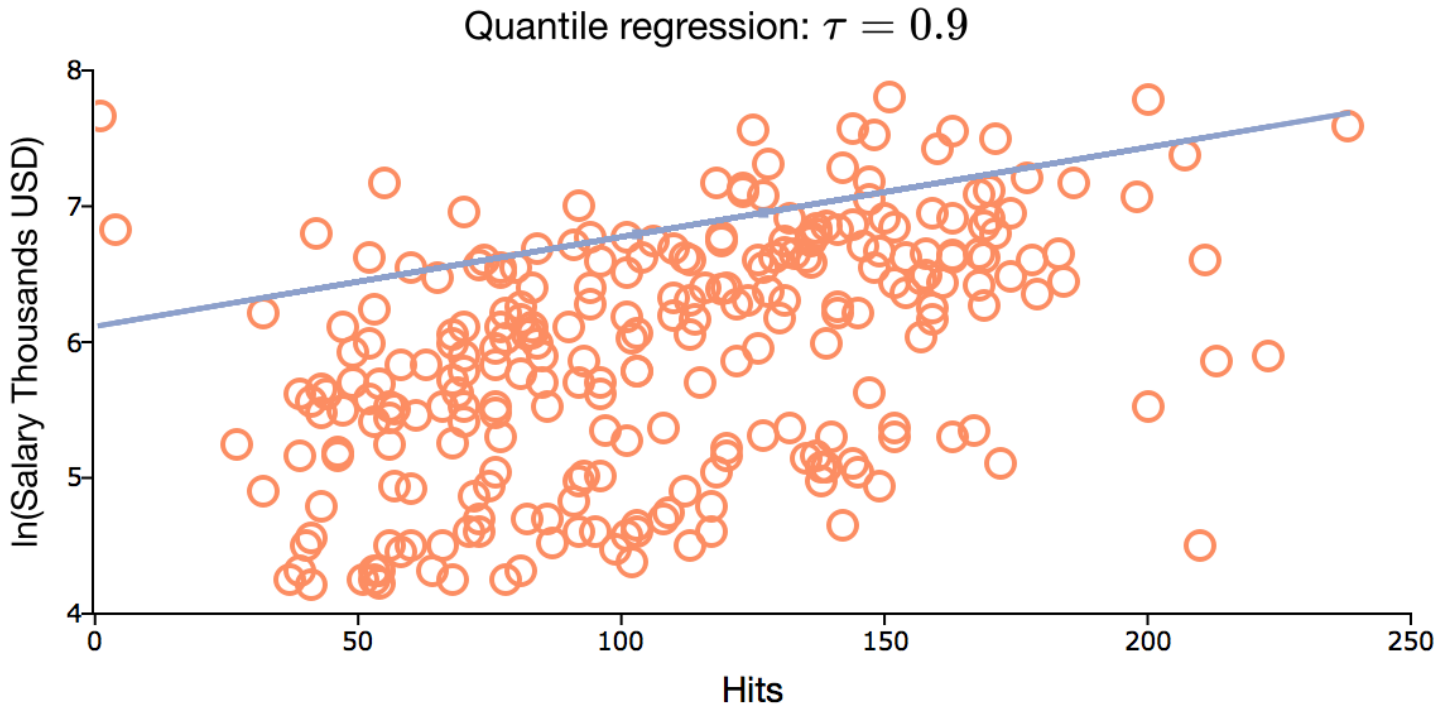
$$\tau(d2) + (1 - \tau)(|d1 + d3|)$$

$$0.9 * 0.4 + 0.1 * (|-1.3 + -0.4|) = 0.53$$



# Quantile Regression

- Optimizing this loss function results in an estimated linear relationship between $y_i$ and $x_i$ where a portion of the data, $\tau$, lies below the line and the remaining portion of the data, $1-\tau$, lies above the line



# Quantile Regression - Applications

- There are at least two motivations for quantile regression: Suppose our dependent variable is bimodal or multimodal that is, it has multiple humps. If we knew what caused the multimodality, we could separate on that variable and do stratified analysis, but if we don't know that, quantile regression might be good. OLS regression will, here, be as misleading as relying on the mean as a measure of centrality for a bimodal distribution.
- If our DV is highly skewed as, for example, income is in many countries we might be interested in what predicts the median (which is the 50th percentile) or some other quantile; just as we usually report median income rather than mean income.

# Quantile Regression - Applications

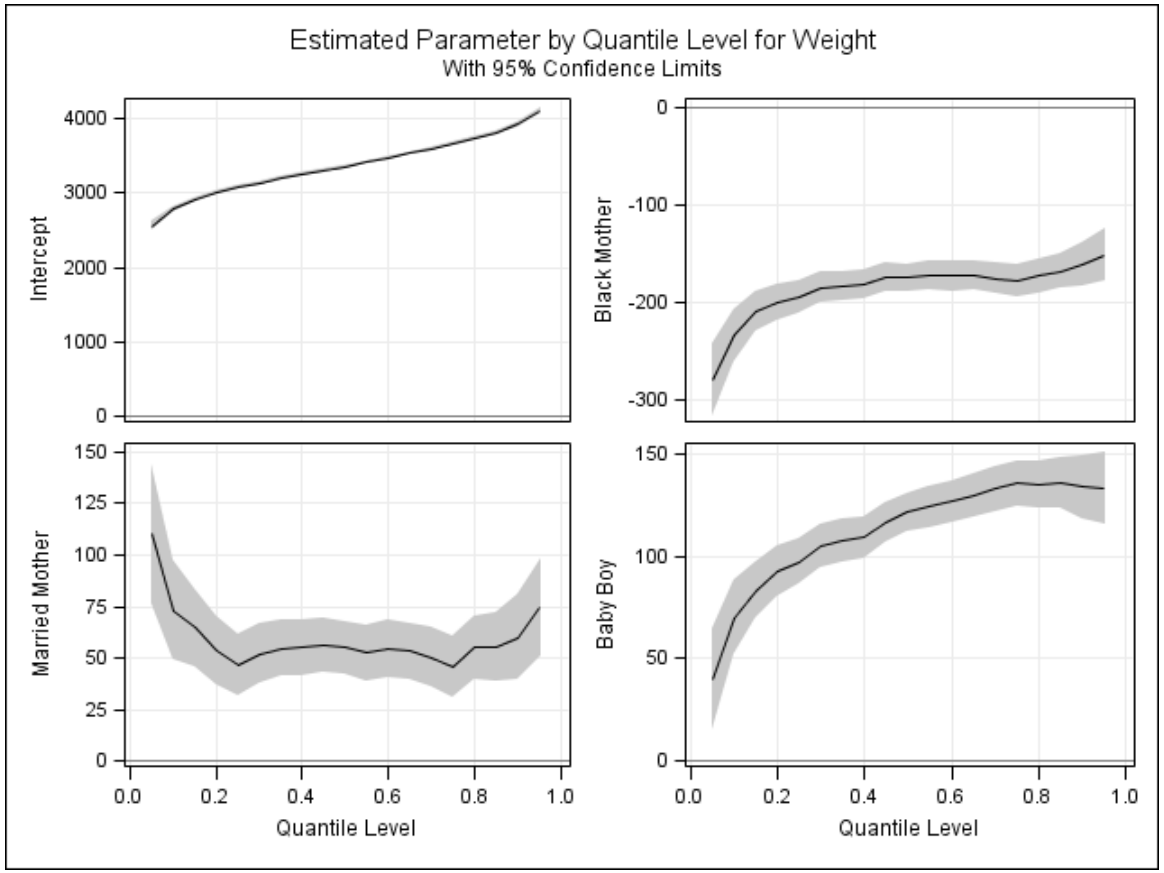
- Predicting low birth weight
- Modeling the mean is inadequate as we are often interested in predicting which mothers are likely to have the lowest weight babies, not the average birth weight of a particular group of mothers.
- We can categorize with respect to birth weight, why don't we just do that then?

# Quantile Regression - Applications

- One model of birth weight includes the child's sex, the mother's marital status, mother's race, the mother's age (as a quadratic), her educational status, whether she had prenatal care, and, if so, in which trimester, whether she smokes, and, if so, how many cigarettes a day, and her weight gain (as a quadratic).

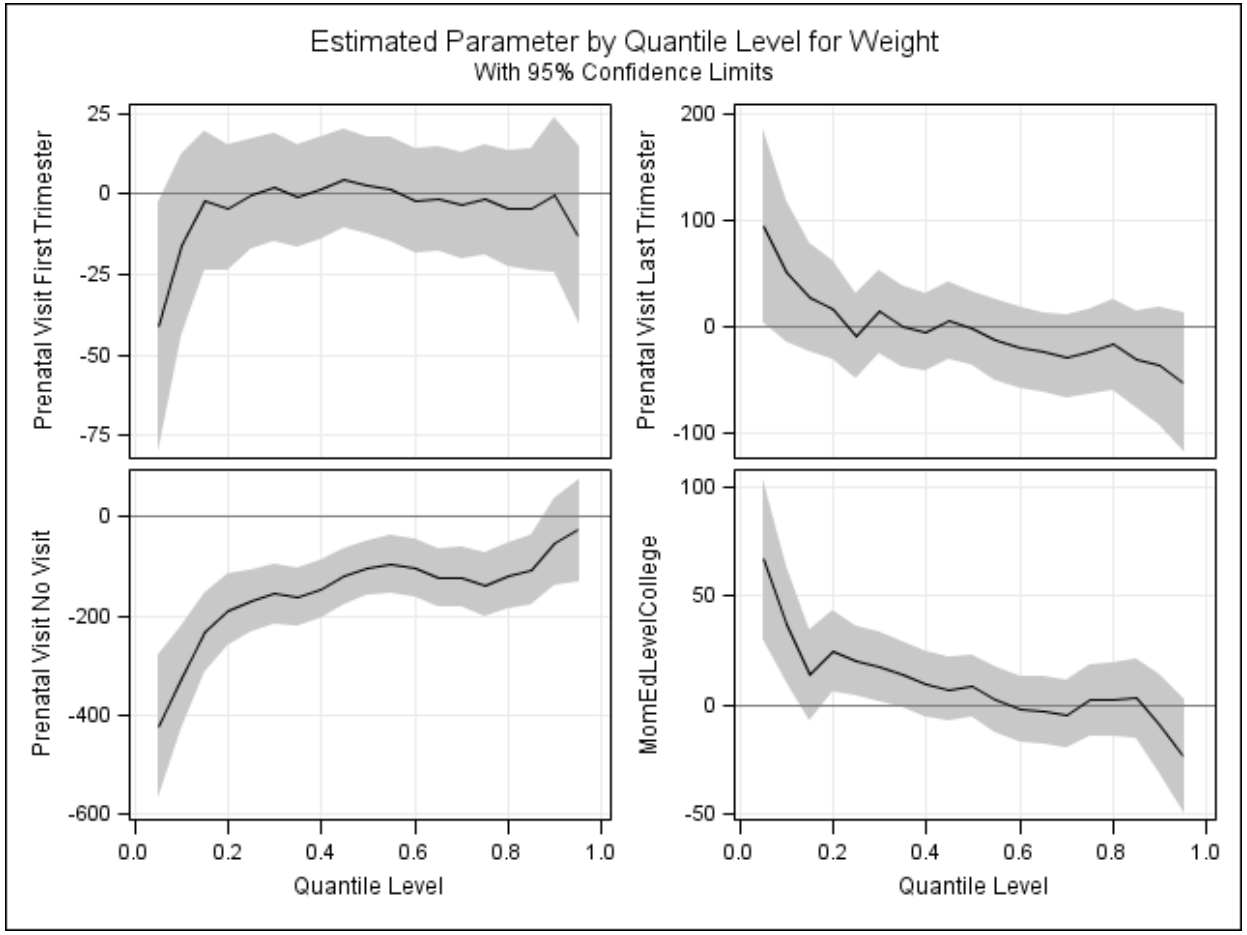
# Quantile Regression - Applications

- If we look at the effects of each of these variables at each quantile.



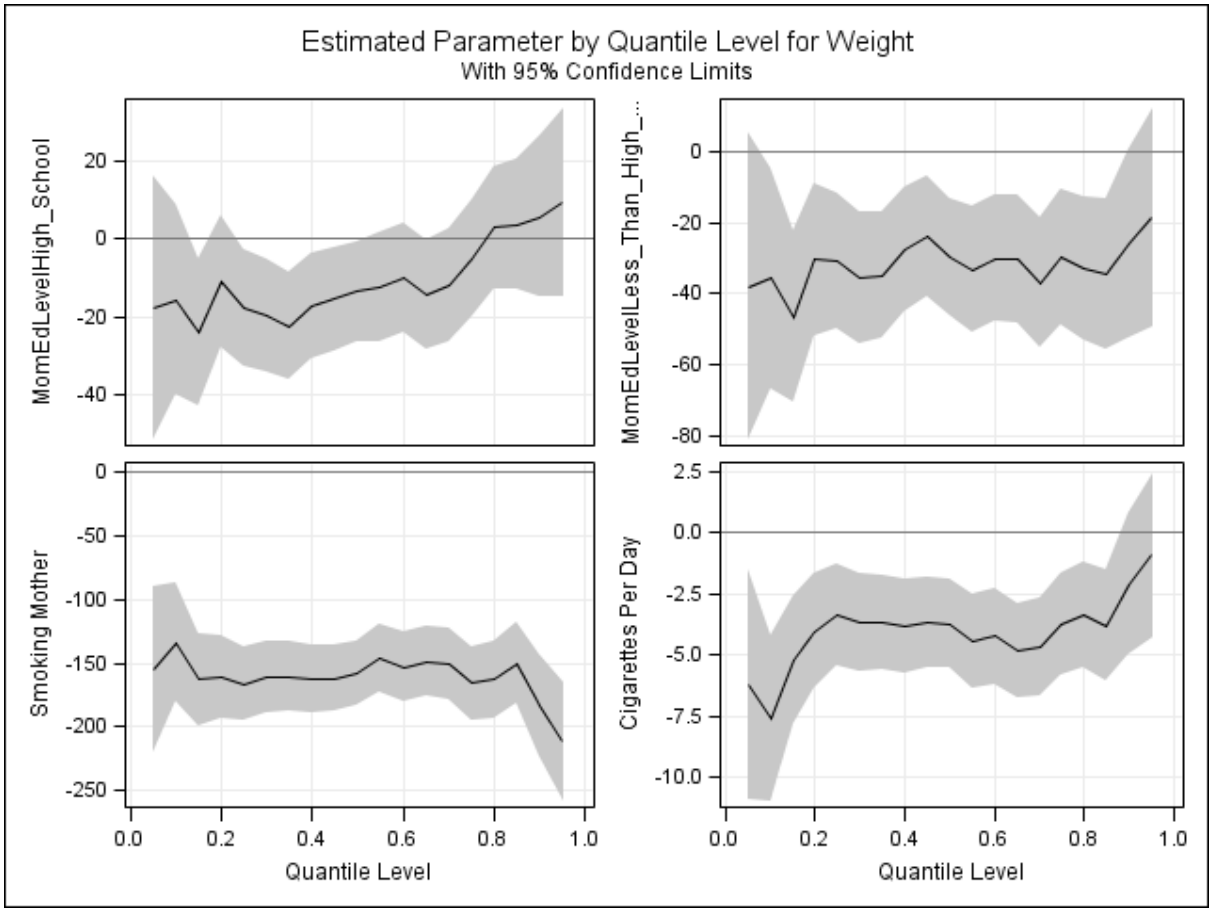
# Quantile Regression - Applications

- If we look at the effects of each of these variables at each quantile.



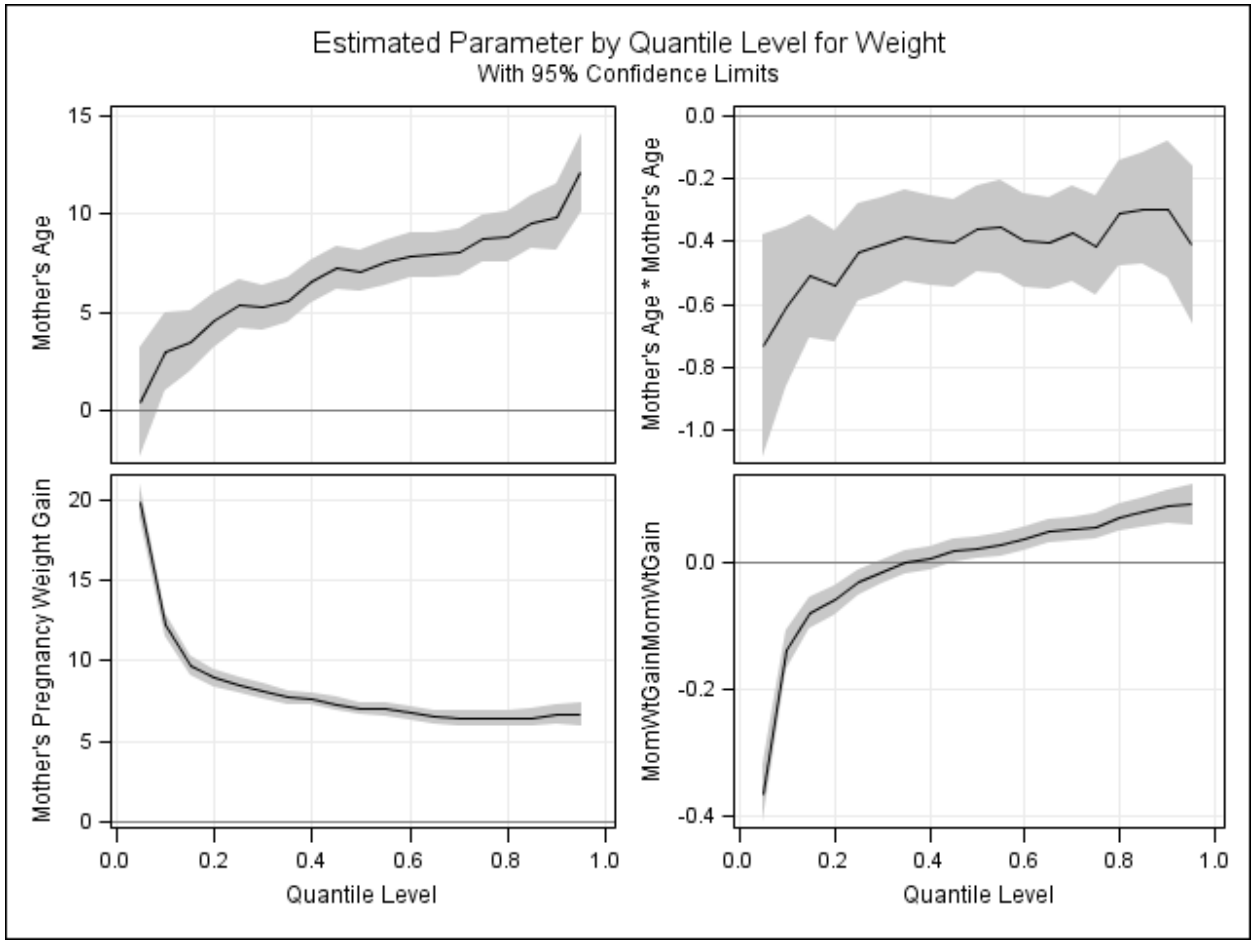
# Quantile Regression - Applications

- If we look at the effects of each of these variables at each quantile.



# Quantile Regression - Applications

- If we look at the effects of each of these variables at each quantile.



# Quantile Regression - Applications

- If we look at the effects of each of these variables at each quantile.