Distributional Regression by Dyadic CART

By Kyle Adams & Sophia Sohail
Mentor: Sabyasachi Chatterjee

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Overview

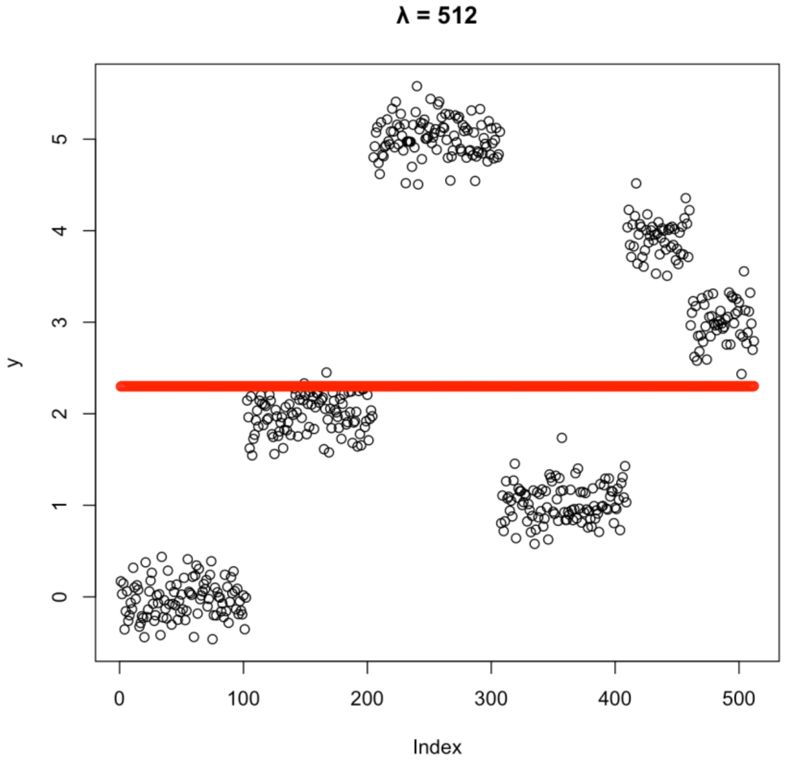
- Goal: Create a regression model to predict the conditional CDFs of various distributions
- Learned about nonparametric curve fitting method: Dyadic CART
- Implemented Cross Validation for Dyadic CART
- Used Dyadic CART to do distributional regression
- We tested our model on both simulated data and a real dataset to determine its accuracy

Dyadic CART

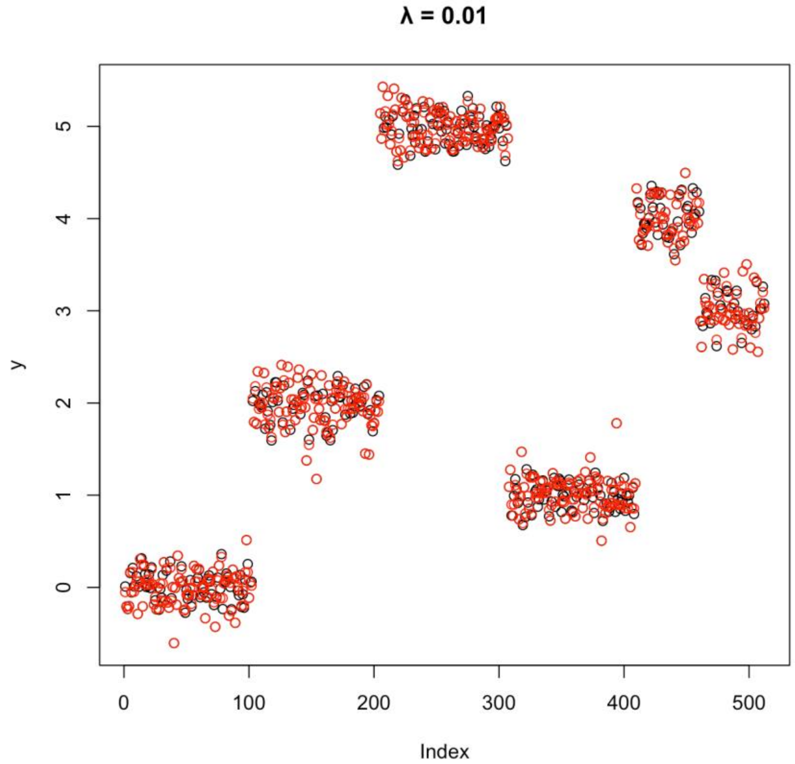
- Dyadic CART: a nonparametric regression method for fitting piecewise functions using a least squares estimate
- Given a vector y, we want to minimize $(y - \pi y)^2 + \lambda|y|$, where π is RDP
- Dyadic CART fits a piecewise constant function over the optimal RDP
- What are the benefits of using Dyadic CART over other regression methods?
  - Can be used for several classes of nonlinear functions!
  - Fast computation – O(n) linear time

Dyadic CART

- Why it's important to pick the correct lambda:

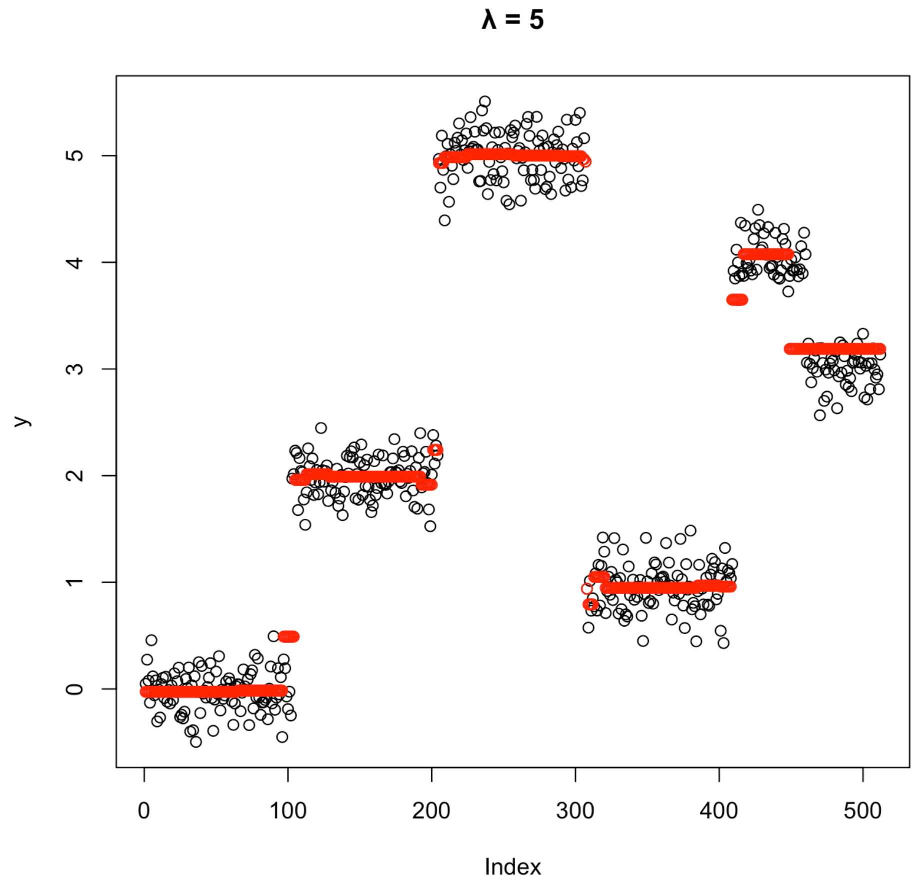


Underfitting



Overfitting

Dyadic CART



Better fit!

Two-Fold Cross Validation

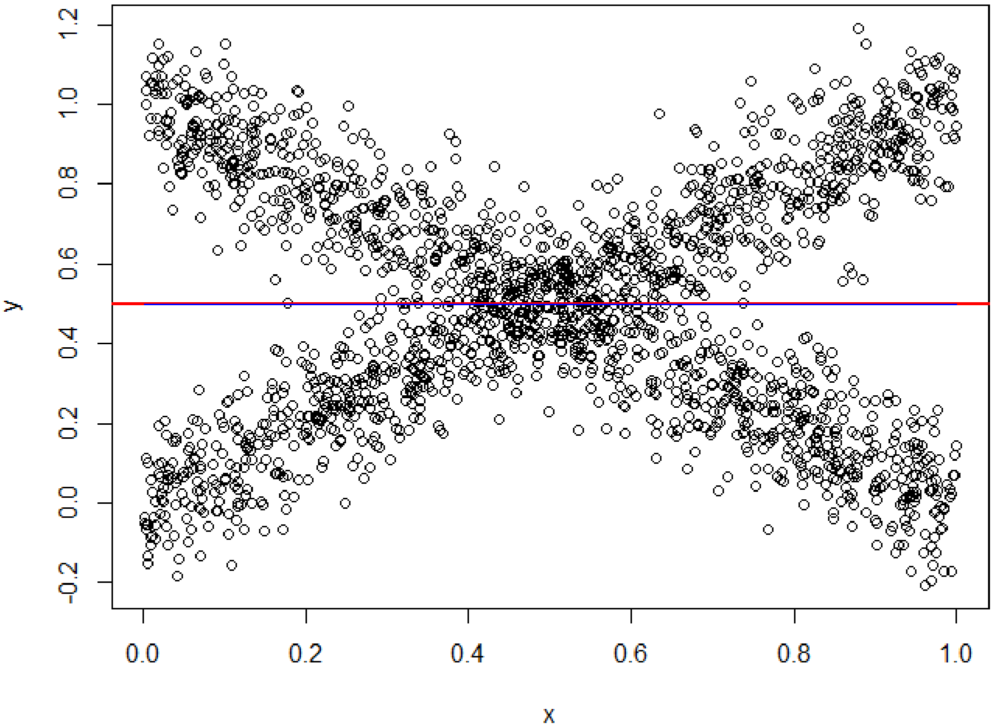
- Two-Fold Cross Validation: A method for dividing data into a training set and a testing set
- Algorithm:
  1. Take a data vector y and divide it into odd and even-indexed observations
  2. Use odd observations as training set and even observations as testing set
  3. Calculate and minimize mean squared error to set optimized tuning parameter – this gives us the best fit for even observations
  4. Repeat steps 2-3 with the roles of even and odd switched – combine even and odd fits to get the final fit
  5. Post processing step to choose a single optimized lambda

Distributional Regression

Motivation of distributional regression: captures information the conditional mean may miss

Consider a case where our data looks like this: $Y|X = x \sim$ **Normal**$(x, 0.1)$ or **Normal**$(1-x, 0.1)$ with probability 0.5

Though our conditional mean estimate is accurate, we miss the pattern of the data



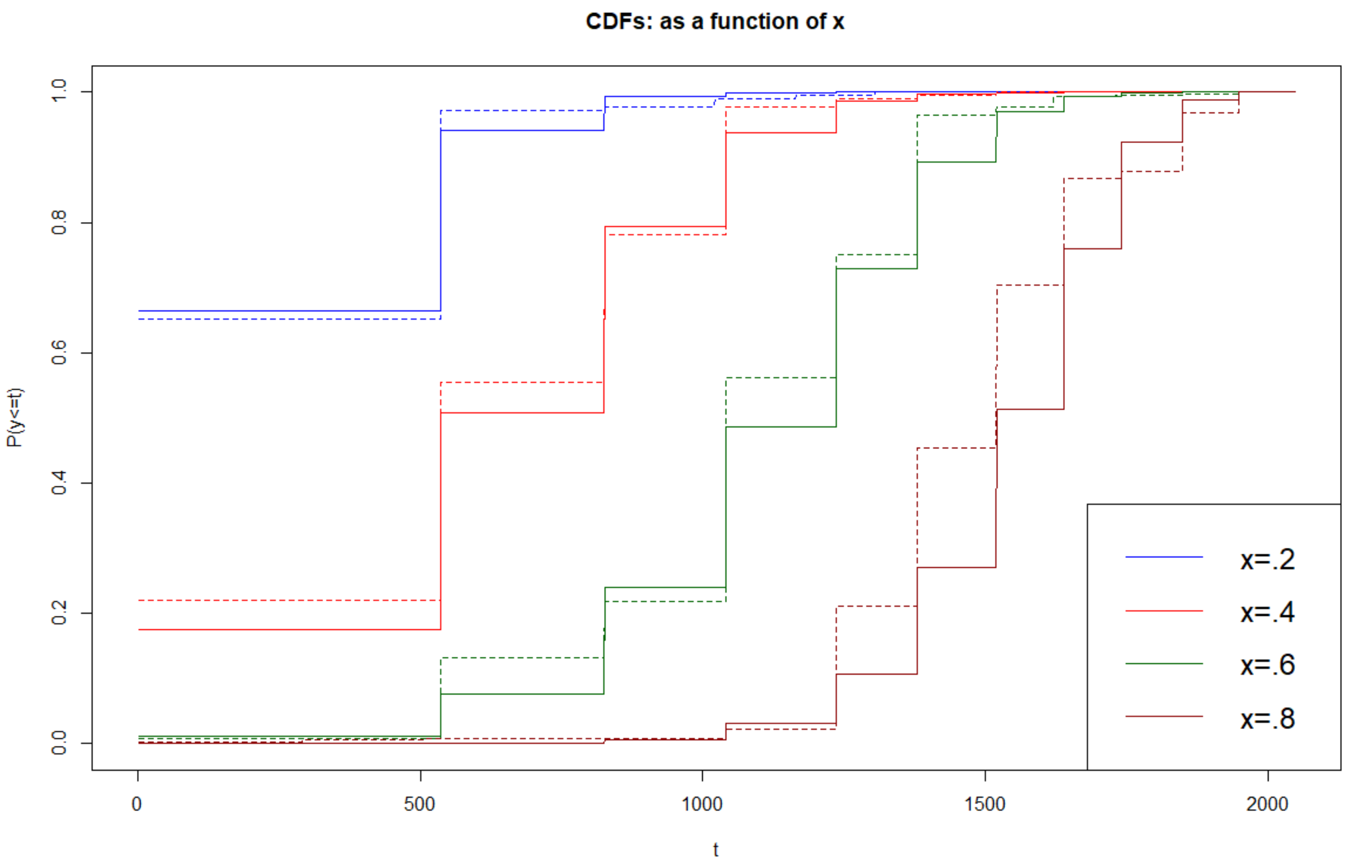
Conditional CDFs

- Conditional CDF: Given a joint distribution of X and Y, the conditional CDF function $F(t, x) = P(Y \le t \mid X = x)$.
- The goal of Distributional Regression is to estimate $F(t, x)$ from data
- We use Dyadic CART + Cross Validation to estimate conditional CDFs

Conditional CDFs

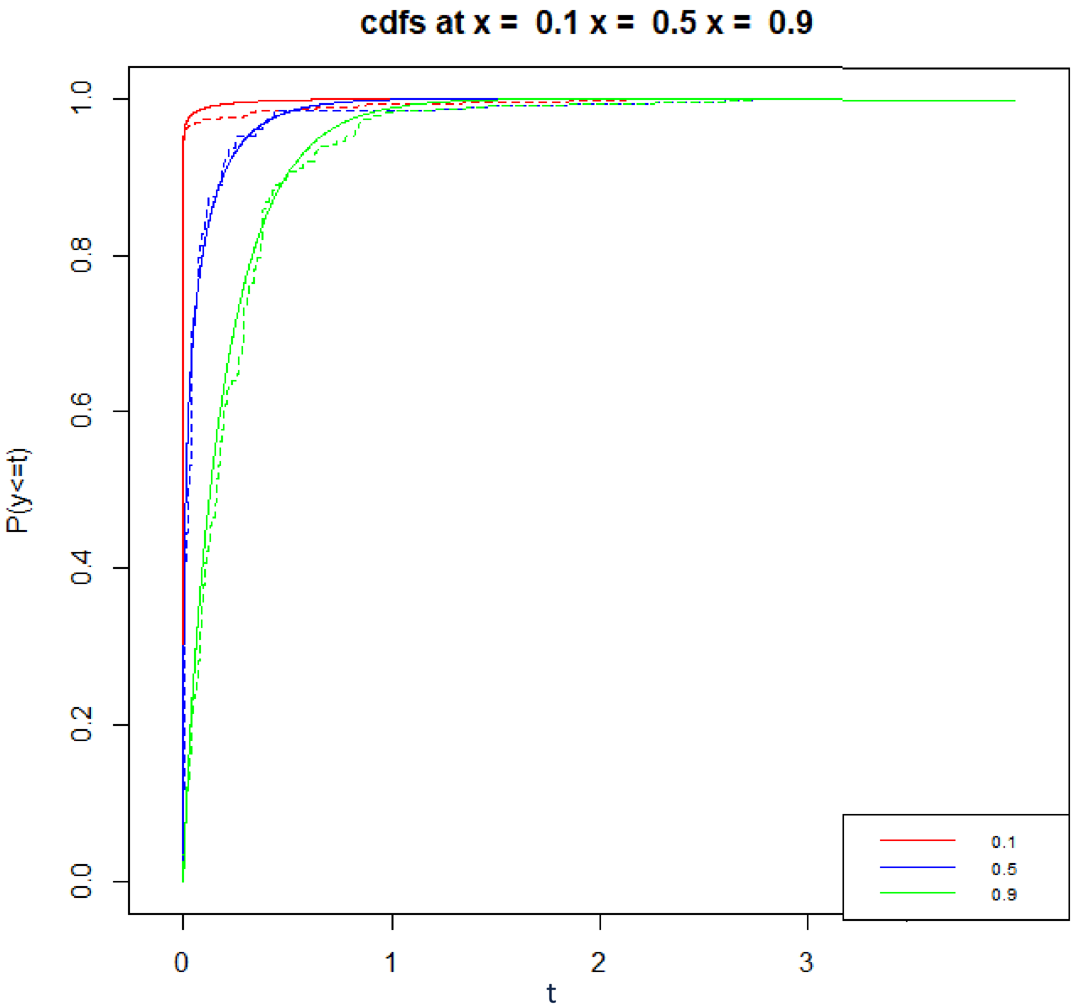
- For each t, create a binary data vector $1(Y_1 \leq t, \ldots, Y_n <= t)$
- Perform Cross-Validated Dyadic CART on this binary vector to get an estimate $\hat{F}(t, x)$ of $F(t, x)$ as a function of x
- At this stage, $\hat{F}(t, x)$ viewed as a function of t for a given x is not necessarily monotone
- We do a postprocessing sorting step to arrive at a final $\hat{F}(t, x)$

CDF Estimation - Discrete



- X ~ **Uniform**(0,1)
- Y|X = x ~ **Binomial**(10, probability = x^2)
- Sample size: 2048
- MSE: 0.0028

CDF Estimation - Continuous



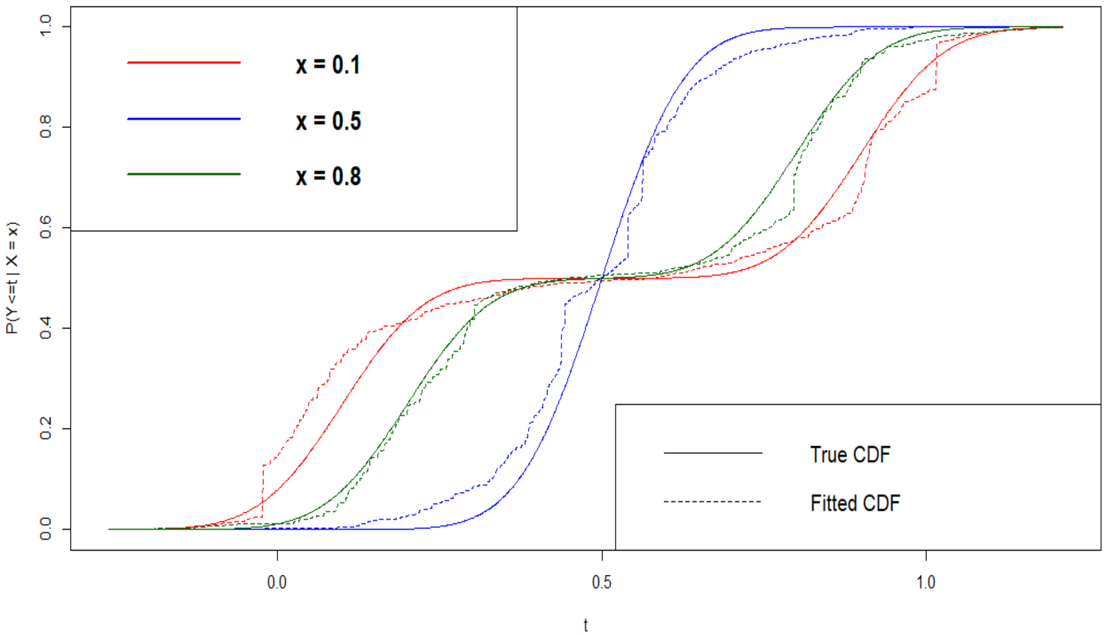
- X ~ **Uniform**(0,1)
- Y|X = x ~ **Gamma**(x^2, f2(x))
- F2(x) produced either 1, 2, or 4
- x = 0.1, 0.5, 0.9
- Sample size: 2048
- Average MSE: 0.004

Distributional Regression

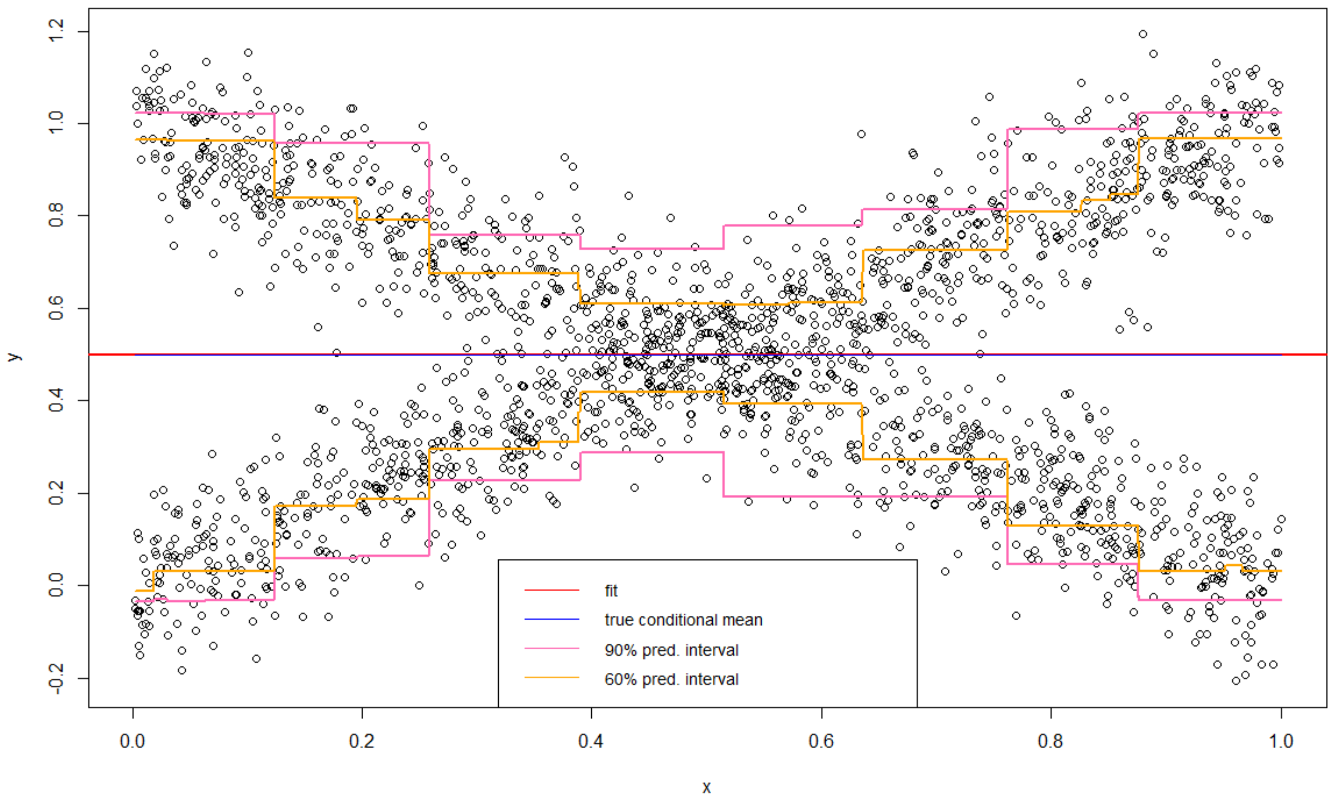
With distributional regression, we can estimate the conditional CDF at every x, which allows us to build prediction intervals around our conditional mean estimate

First we can plot the conditional CDFs of a few x's, as a function of t

From here we can take the region between our desired prediction bounds, for example .05 and .95 for a 90% prediction interval, and plot those for every x to obtain a better picture of the data

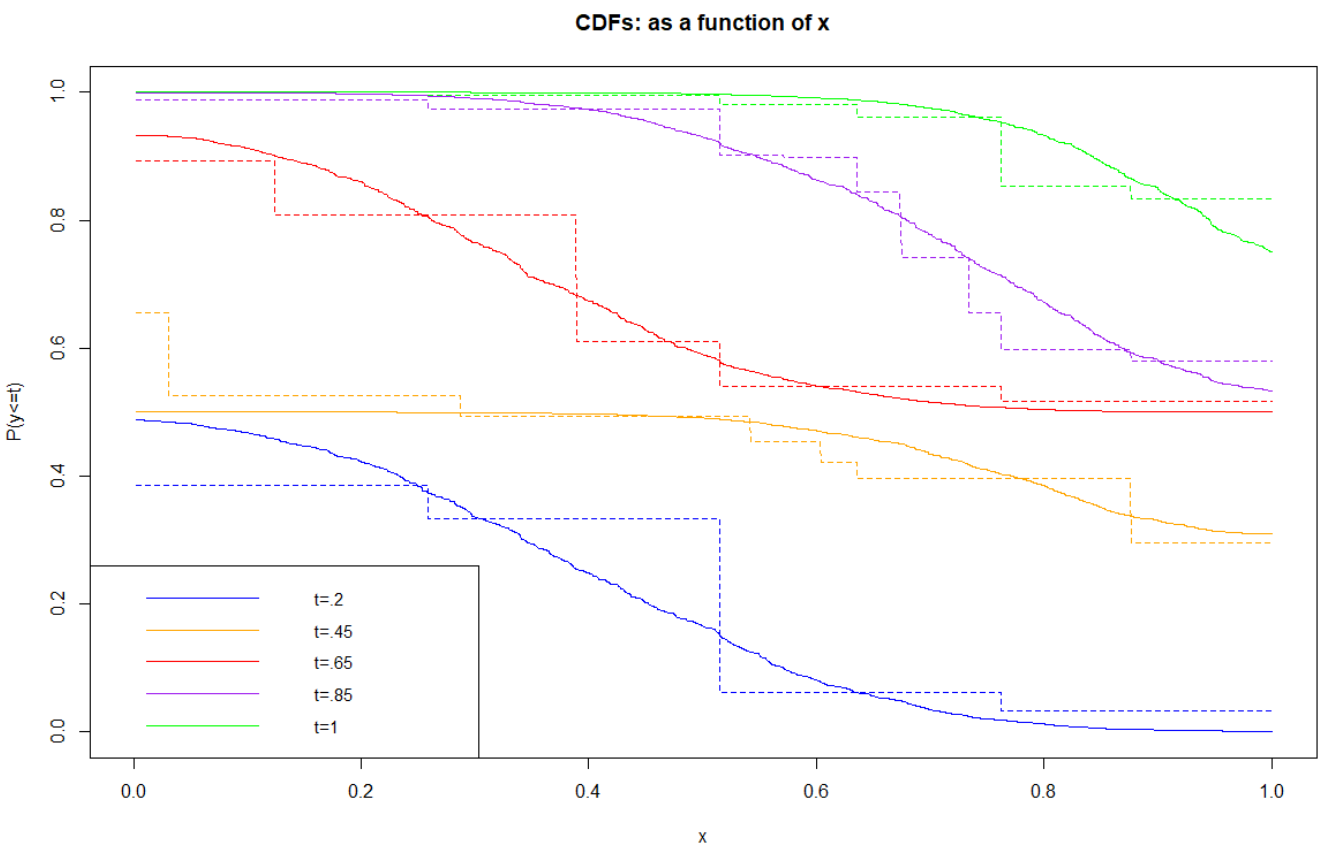


Distributional Regression



With distributional regression, theoretically we can obtain much more information from data

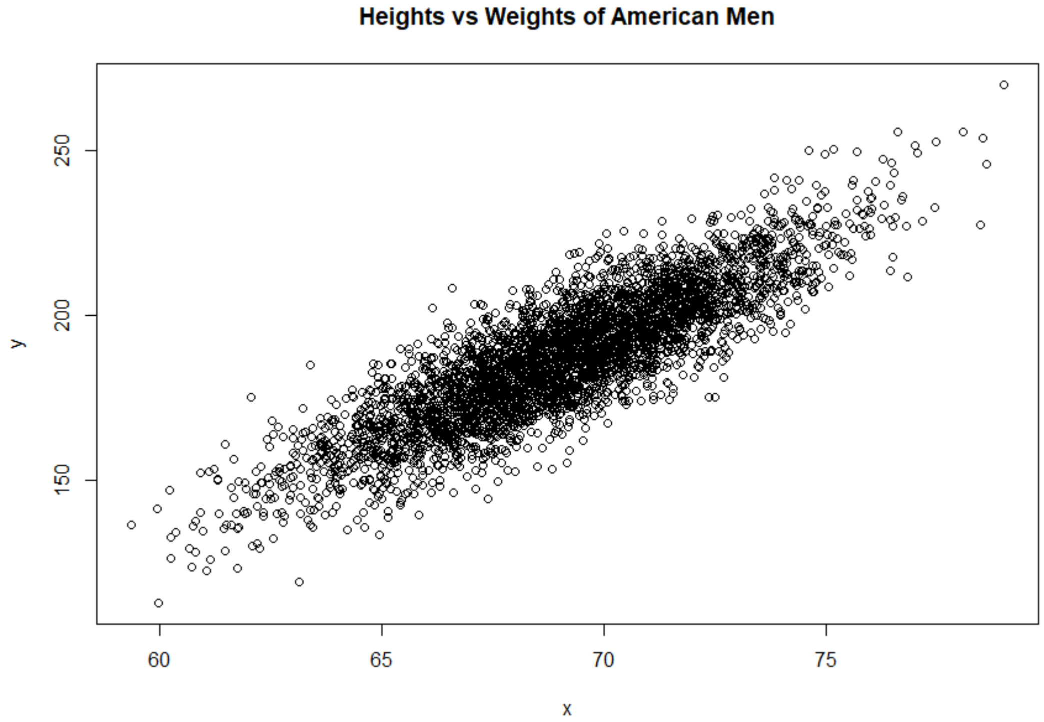
CDF Estimation – as a function of x



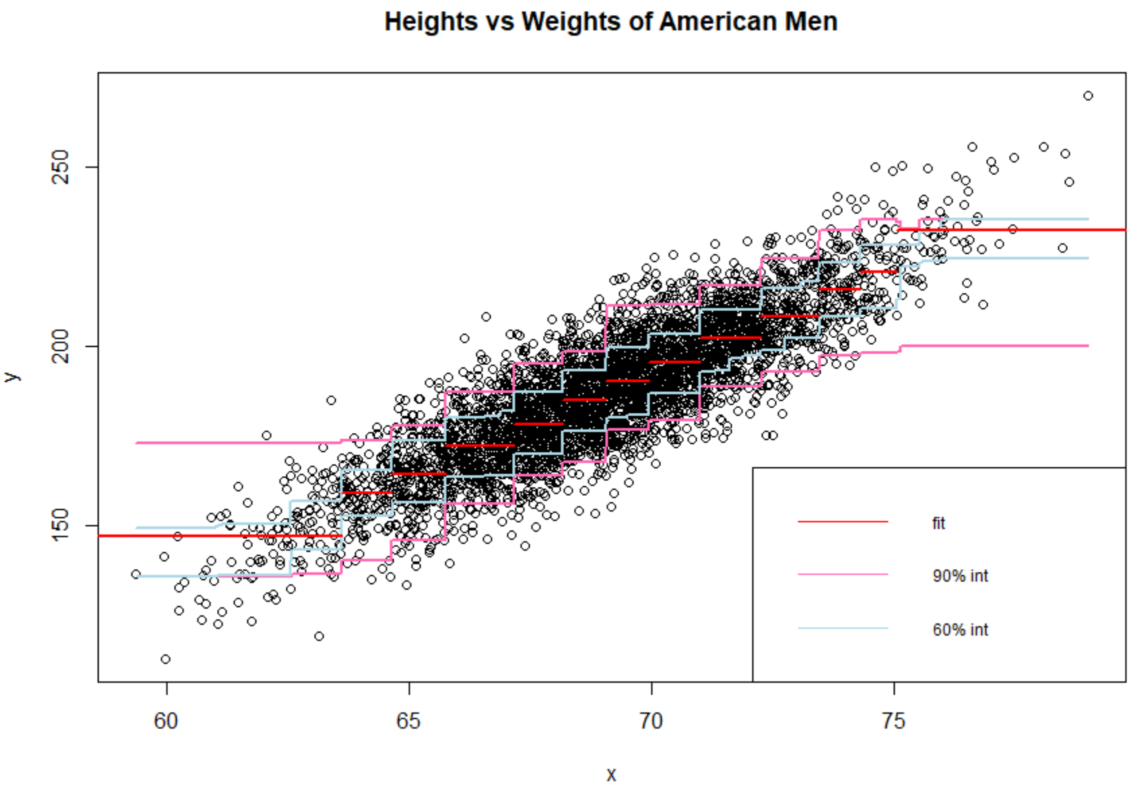
- X ~ **Uniform**(0,1)
- Y|X = x ~ **Normal**(x, 0.1) or **Normal**(x, 1-x)
- Sample size: 2048
- Average MSE: 0.02

Male Height vs. Weight

- Dataset: 4096 observations of the heights and weights of American men (x = height, in inches, y = weight, in pounds)
- Question: Can we predict an American man's weight from his height?

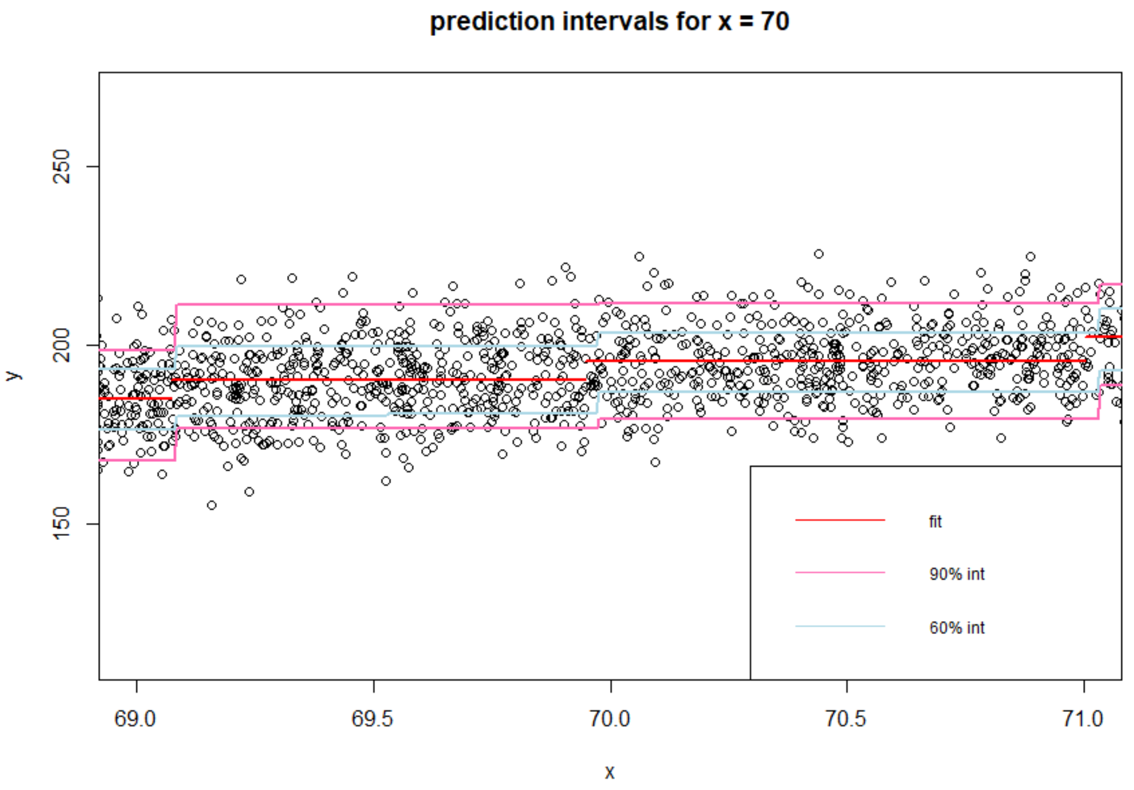


Making Predictions



We use the explained process before, and can obtain a graph of the data with prediction bounds to predict future men's weights given their height

Making Predictions



As an example, if a man told us he wanted to predict his weight, and he was 5'10" (70 inches), our model would conclude that a man's true weight at this height will fall within a range of (180, 212) 90% of the time. 60% of the time, his weight will fall in (186, 203).

Thank You!

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN