Summer Research 2019 Week 2

Joe Sato, Katherine Liu

June 26, 2019

Review: Prediction

Prediction

- Assumes: data follow $y_i = f(x_i) + \epsilon_i$
    - $y_i = f(x_i)$: **true pattern**
    - ϵ_i : **noise** (random)
- Seeks **true pattern** from data, differentiating it from **noise**
    - then use it to predict

Ways to Estimate the True Pattern

Linear Regression

- Can fit **ANY** functions, if coefficients are **linear**
- Can produce **prediction intervals**
- **Must specify** the function in advance

Random Forests

- Works for **ANY** underlying patterns, even for **non-linear** coefficients
- **No need to specify** the function
- Needs: **large** data set (for training)
- **Maybe** can produce prediction interval **?**

Review: 95% Prediction interval

95% Prediction Interval

- W.r.t. a **single data point** yet to be observed
- Say: we estimated the true pattern from \({(x_1,f(x_1)),...,(x_n, f(x_n))}\)
- 95% PI of \(f(x_{n+1})\): [a,b] such that
    - \(f(x_{n+1})\) \(\in\) [a,b] by 95%

95% Confidence Interval

- W.r.t. the **entire set** of points of the same true pattern
- 95% CI: [a,b] such that
    - (avg. of all points in the entire set) \(\in\) [a,b] by 95%

Objective

To figure out:

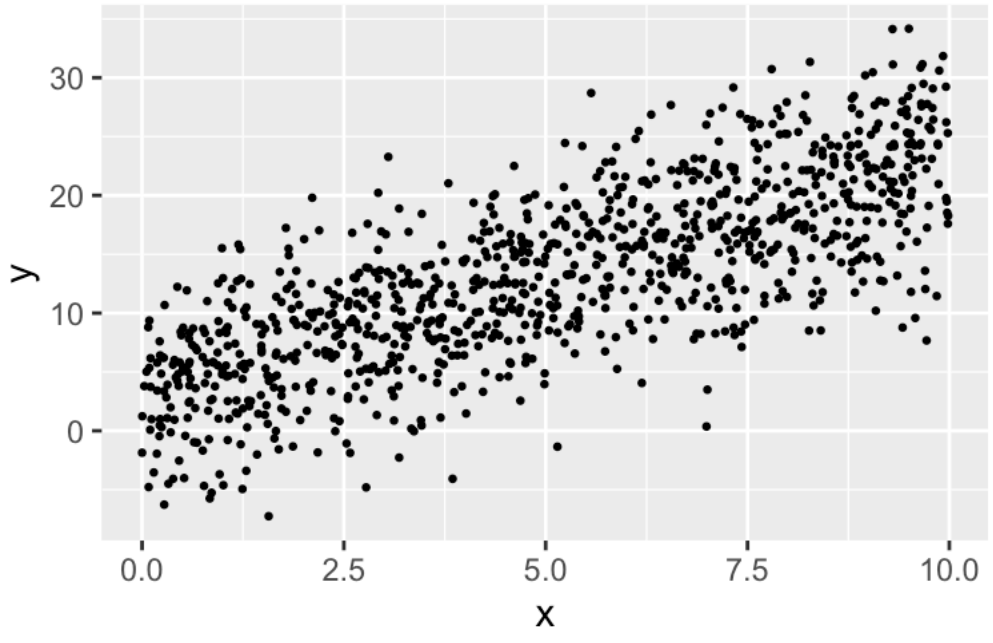
- If **Random Forests** can accurately produce **Prediction intervals**
- **For what kind of data LR & RF are appropriate ?**

Comparison: Method

1. **Generated** datasets using a specific \(y=f(x)\) \(+\) \(\epsilon\)
   - Linear & Single-variable
   - Non-Linear & Single-variable
2. Applied LR & RF and examined:
   - Accuracy of PI-s: **Coverage rate**
   - Prediction Accuracy: **MSE**
   - How informative PI-s are: **width of PI-s**

Simulations

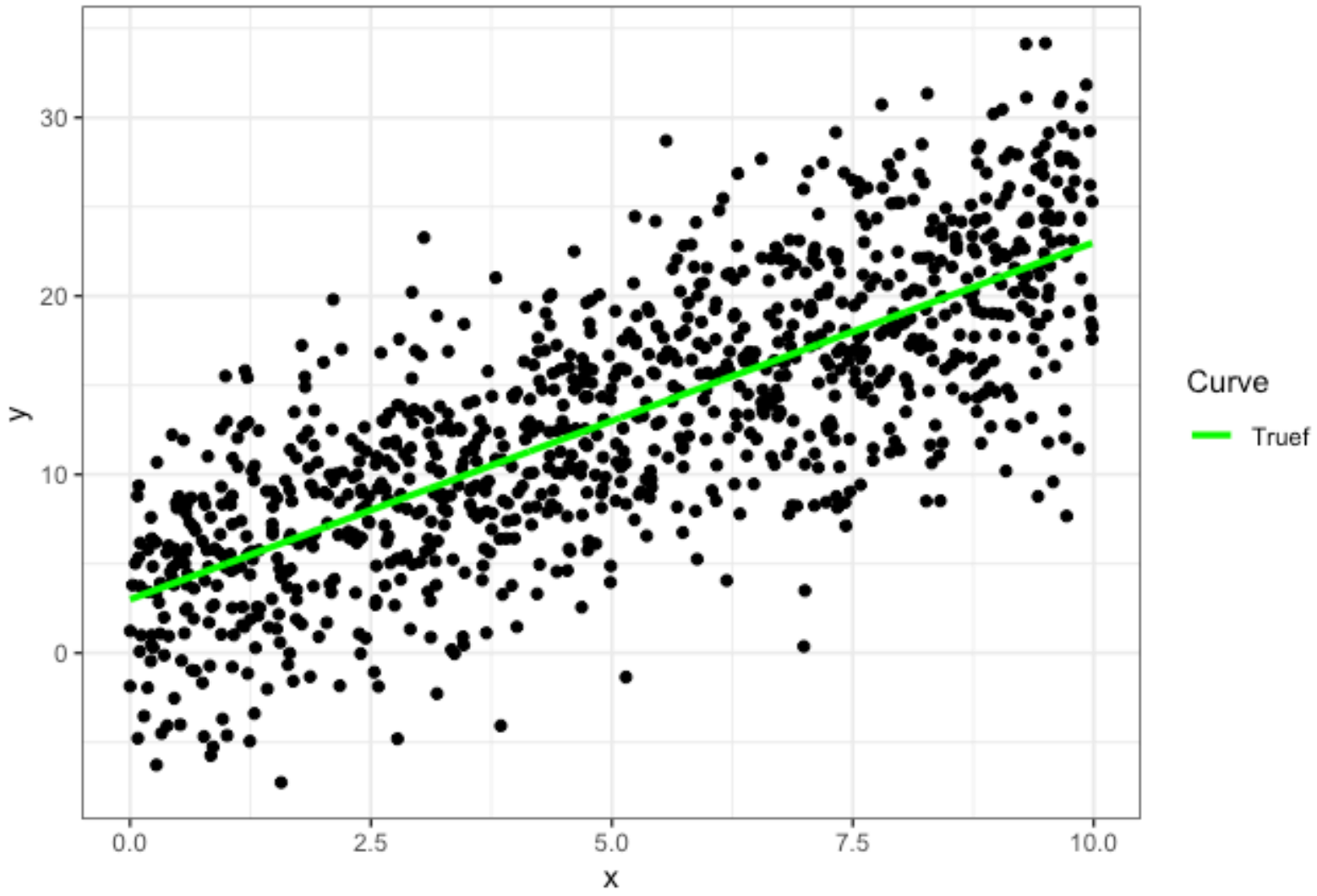
Simulation 1: $Y_i=f(X_i)+\epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 5^{2})$



What is form of f? - linear - quadratic - trigonometric - ....

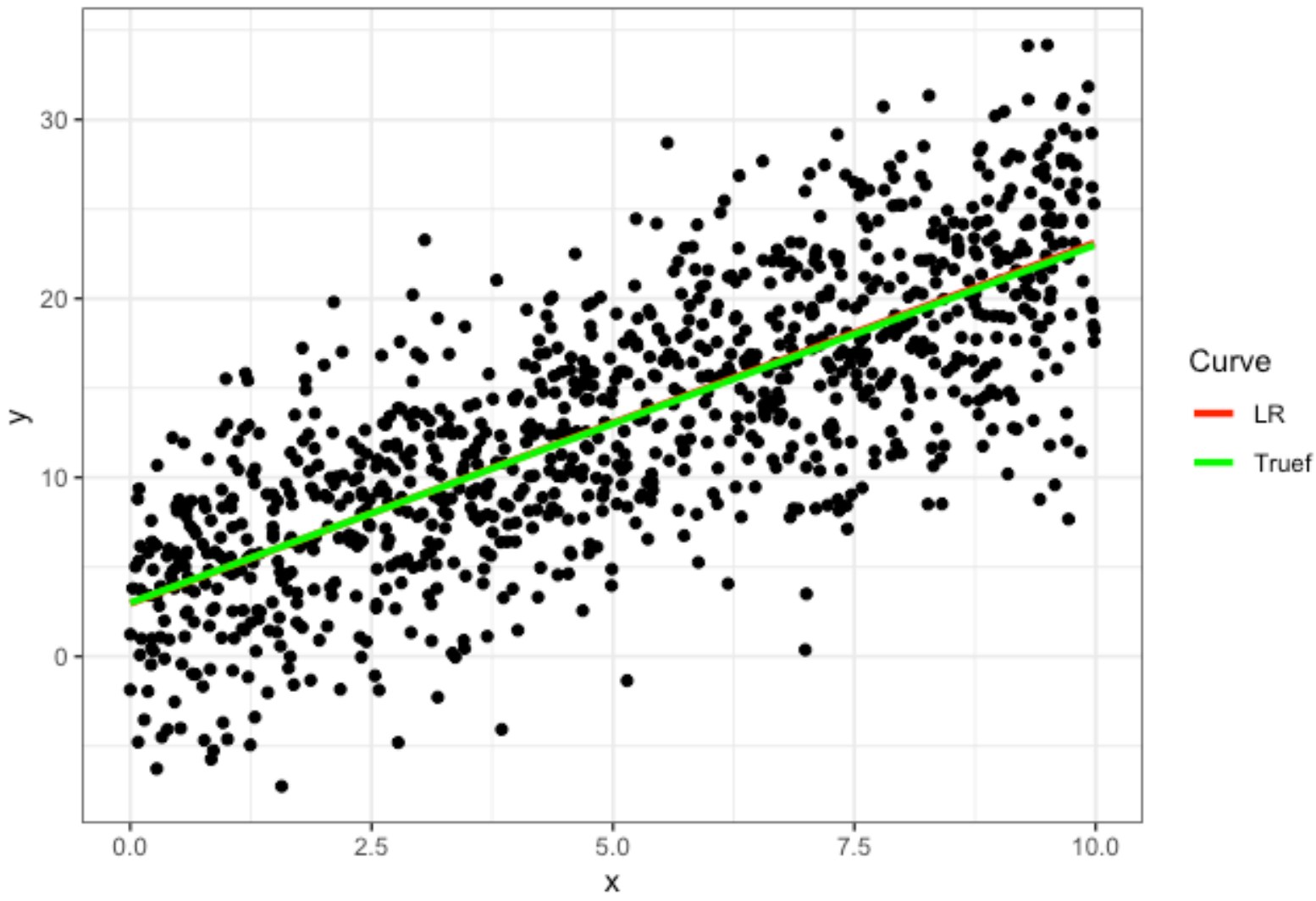
Simulation 1

Simulation 1: $Y_i = 2X_i + 3 + \epsilon_i$



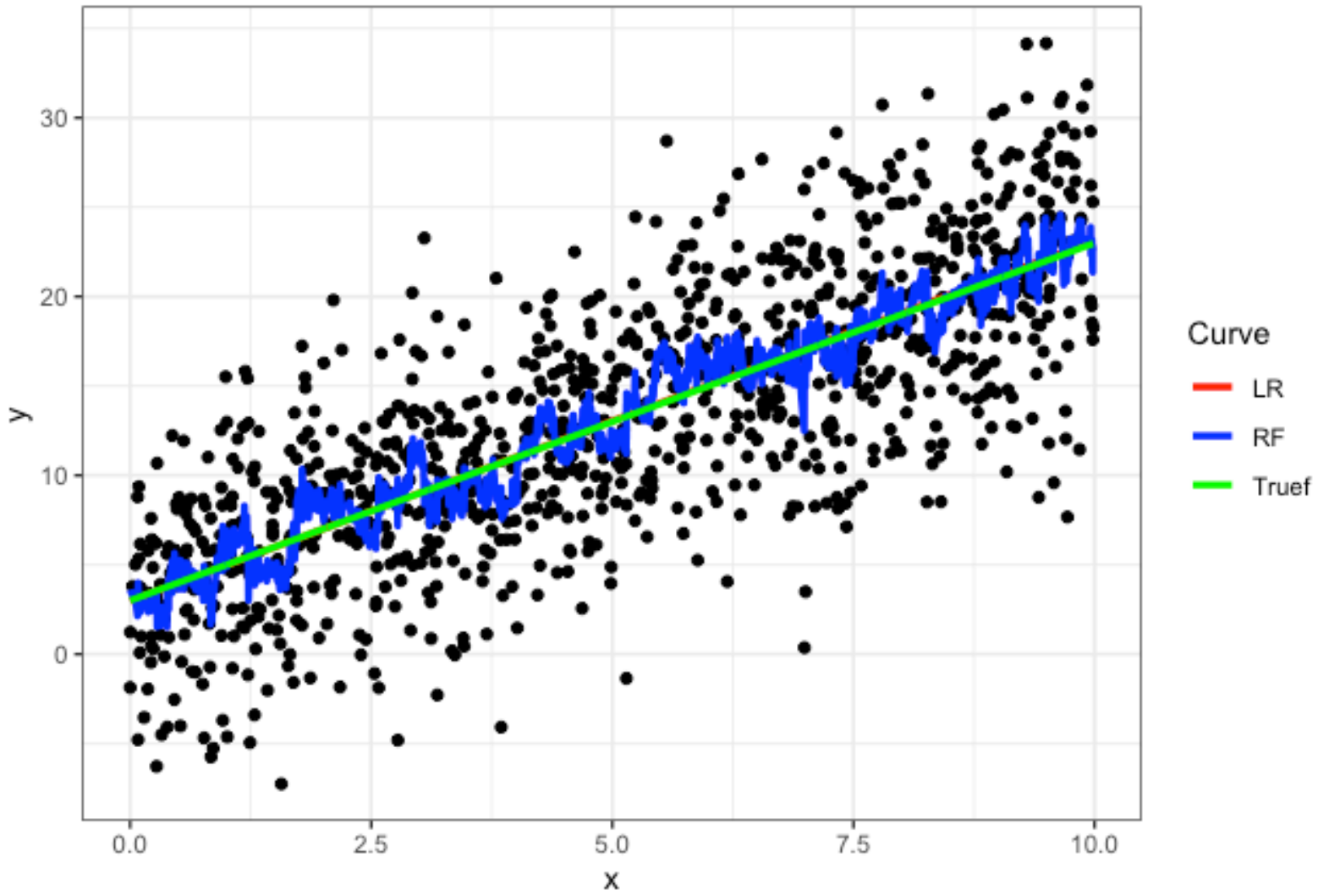
Simulation 1

True function: $Y_i = 2X_i + 3 + \epsilon_i$
Linear Regression Model: $Y_i = 2.01658 X_i + 2.96383+ \epsilon_i$



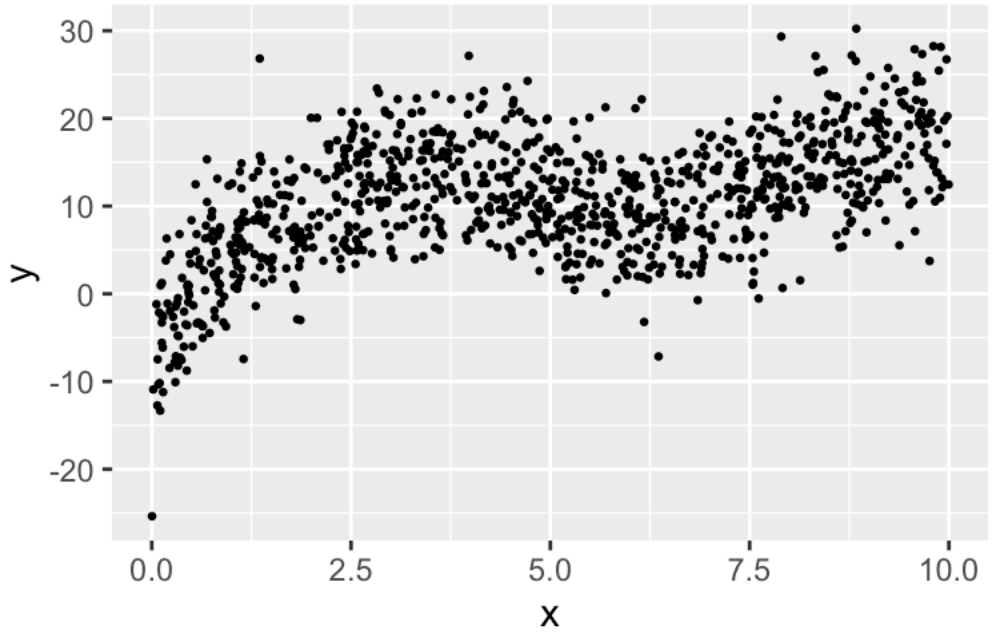
Simulation 1 Radom Forest

-nodesize = 10



Simulation 2

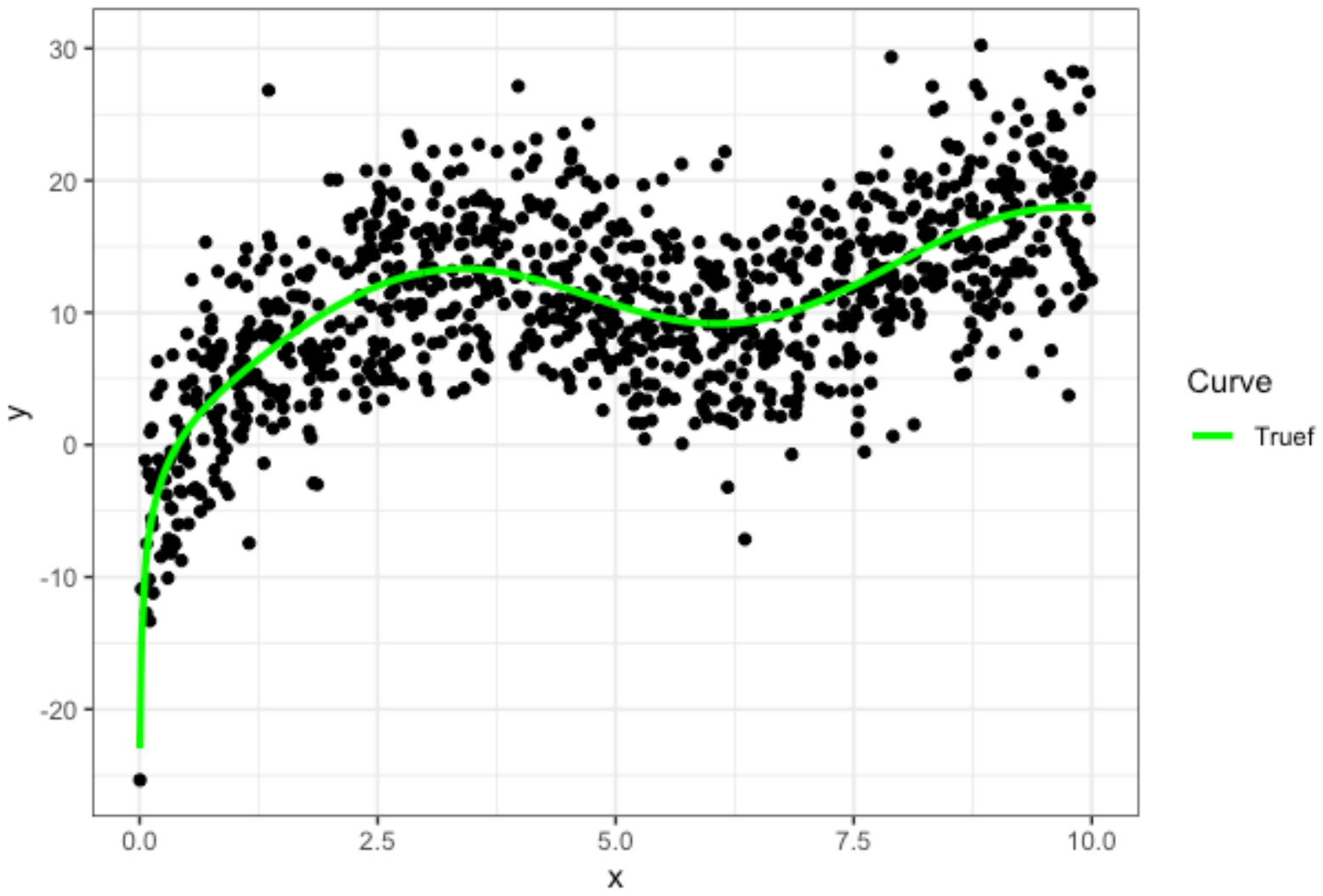
Simulation 2: $Y_i=f(X_i)+\epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 5^{2})$



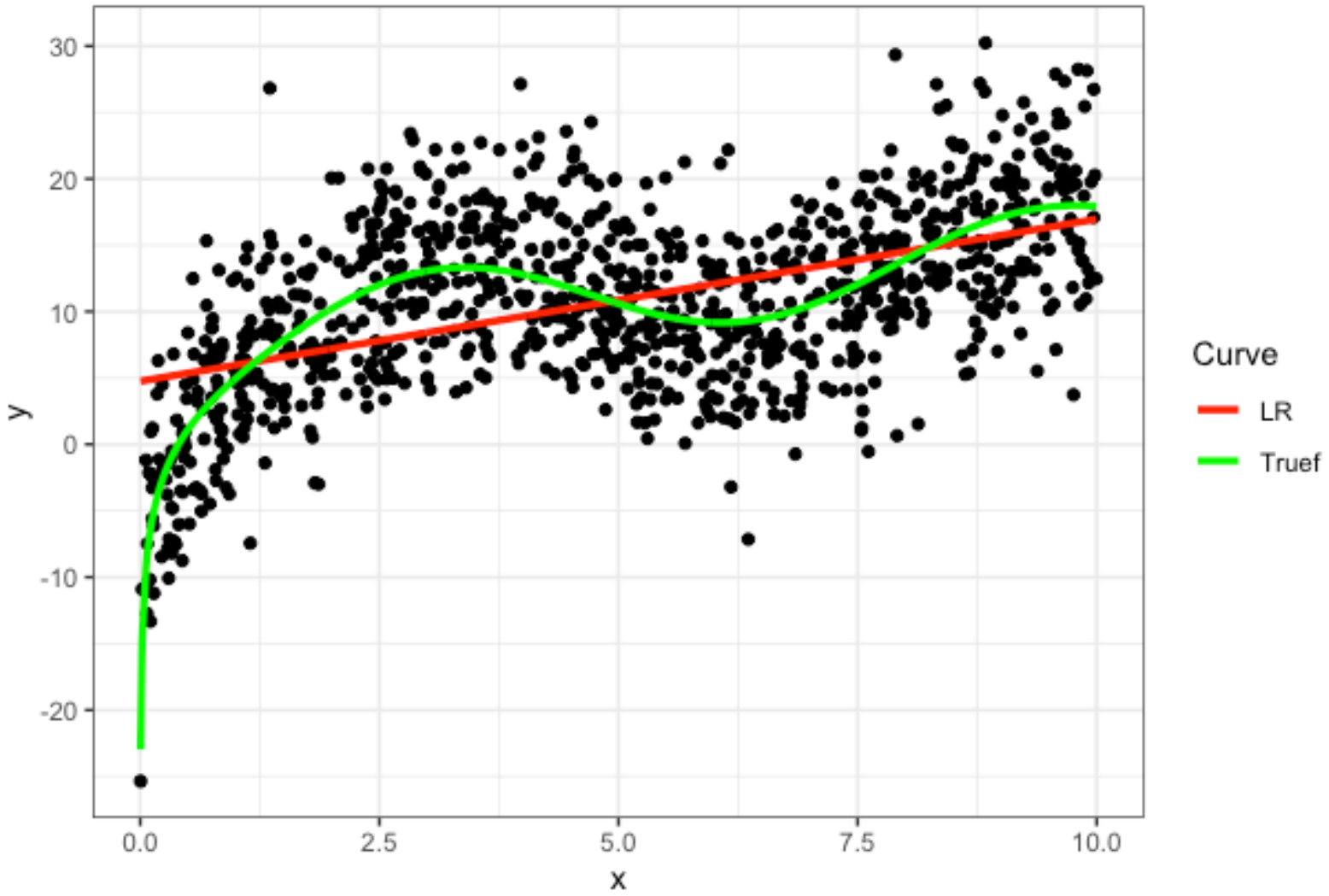
What is form of f? - linear - quadratic - trigonometric - ....

Simulation 2

True function: \(Y_i = 0.1(x-7)^2 - 3cos(x) + 5 log(|x|) + 3 + \epsilon_i\)

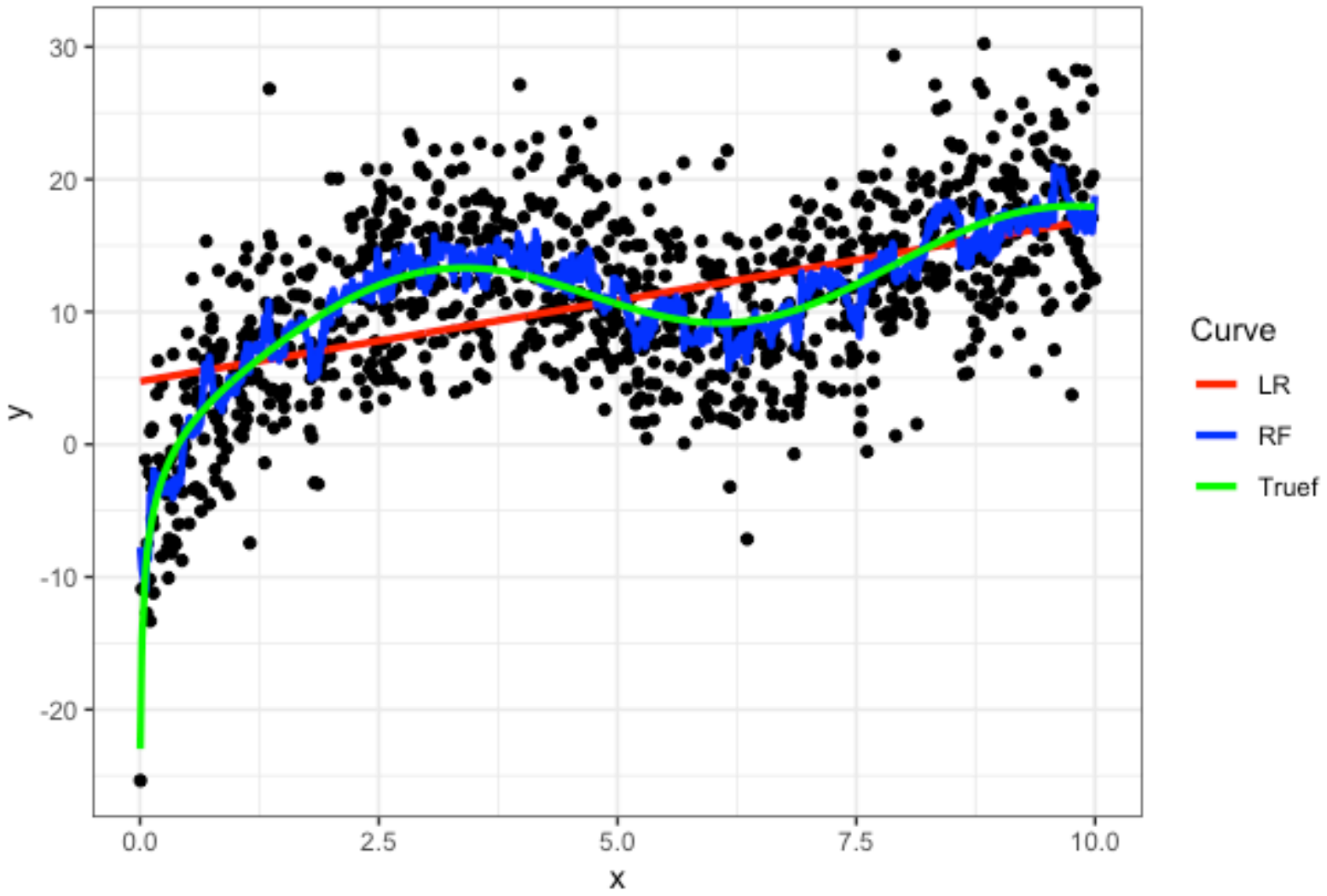


Simulation 2 Linear Regression Model



Simulation 2 Radom Forest

-nodesize = 10



Simulations 3 & 4

Simulation 3(Multivariate Linear): $Y_i =2X_{1i}+3X_{2i}+4X_{3i}-3X_{4i}+ X_{5i} +\epsilon_i$

Simulation 4(Multivariate Nonlinear): $Y_i = (X_{1i}-6)^2 + 12\cos(X_{2i}) + (X_{3i}-5)*(X_{4i}-3) + 0.02(X_{5i}-5)^5+ \epsilon_i$

Results

	MSPE	PIWidth	CoverageRate
Sim1 LR	25.45479	19.65005	0.9491
Sim1 RF	27.19786	18.84627	0.9032
Sim2 LR	26.56509	20.11342	0.9496
Sim2 RF	27.24953	18.88412	0.9027
Sim3 LR	25.41646	19.72029	0.9496
Sim3 RF	29.19473	19.42861	0.9042
Sim4 LR	58.75659	24.97114	0.9489
Sim4 RF	36.76771	21.11574	0.9040

Next Steps

- tune parameters in RF
- learn new PI method
- apply to real datasets