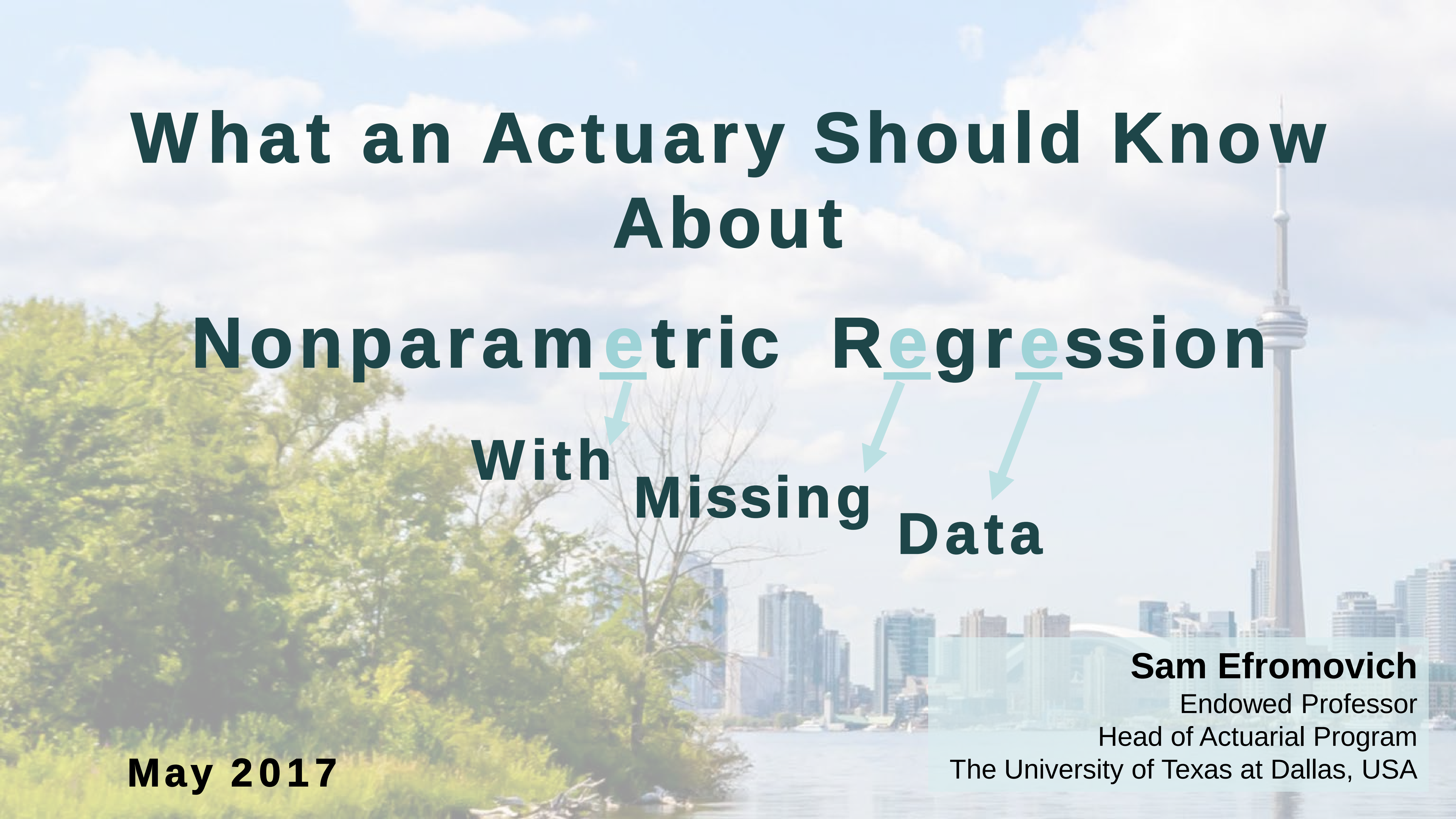


What an Actuary Should Know About

Nonparametric Regression

With Missing Data

Sam Efromovich
Endowed Professor
Head of Actuarial Program
The University of Texas at Dallas, USA

May 2017

The “Whats” of Missing Data

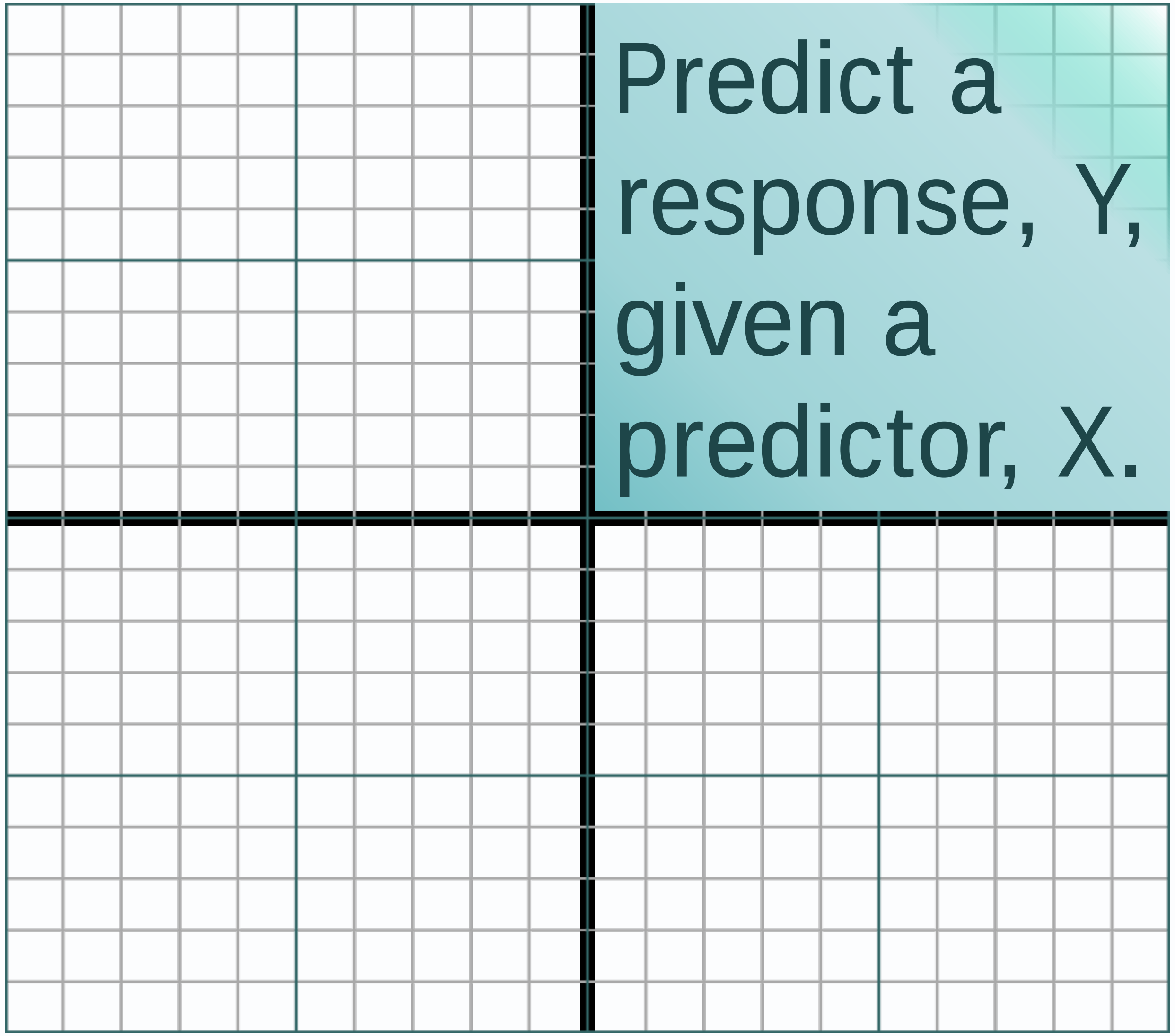
What?
So what?
Now what?

What is nonparametric regression?

What if I’m missing some data?

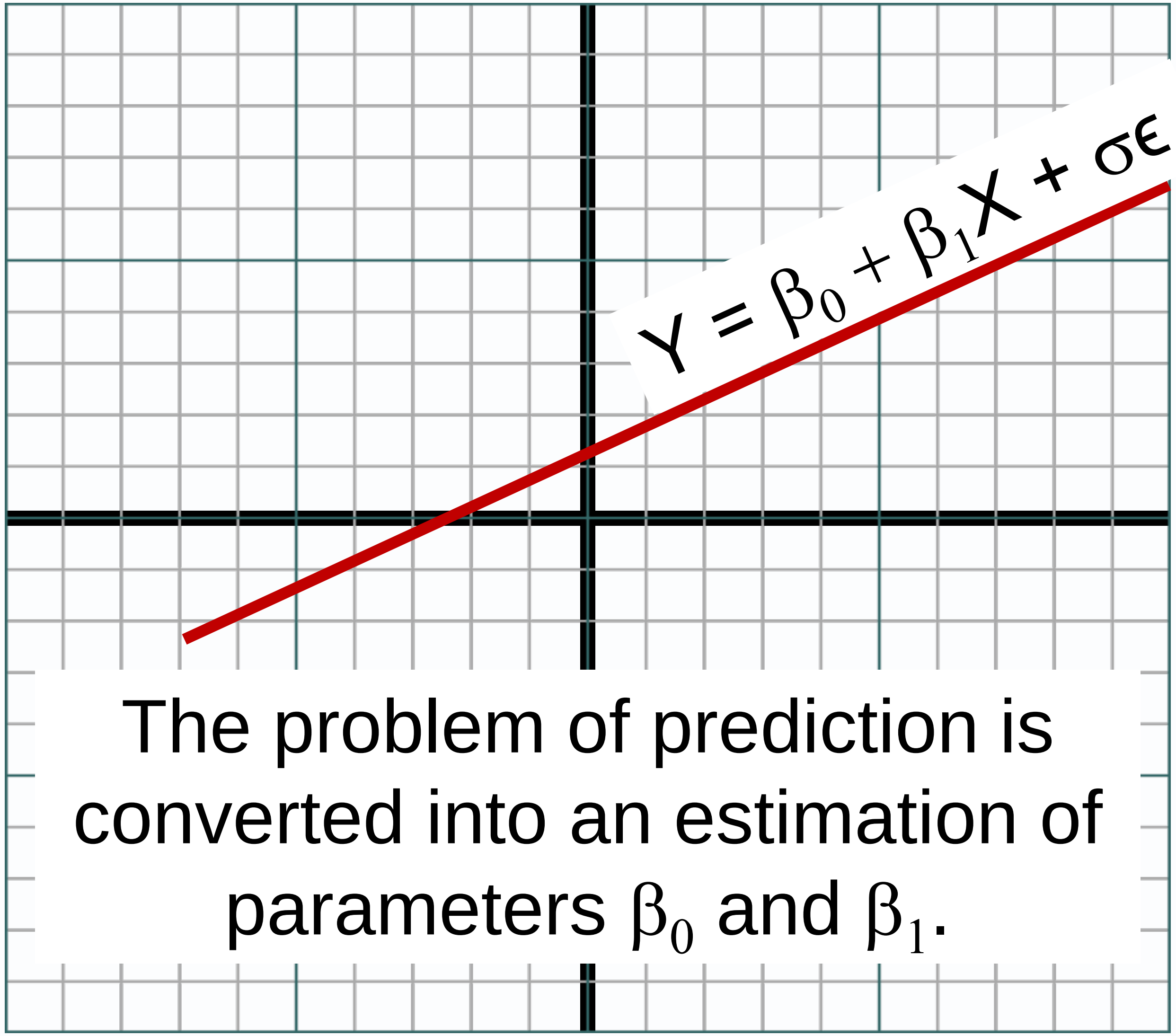
How do I deal with missing data?

Regression



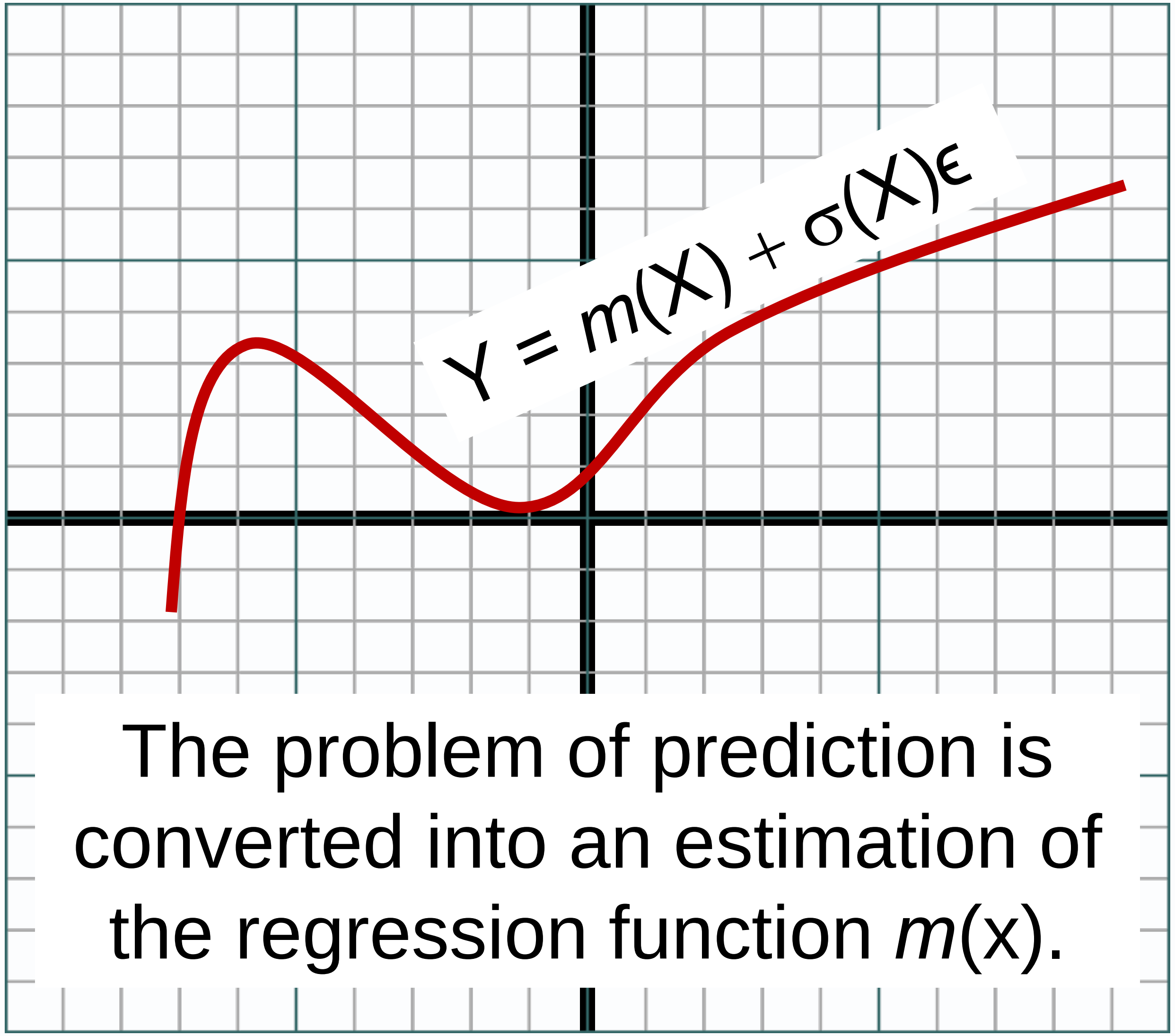
Predict a response, Y, given a predictor, X.

Parametric (Linear) Regression



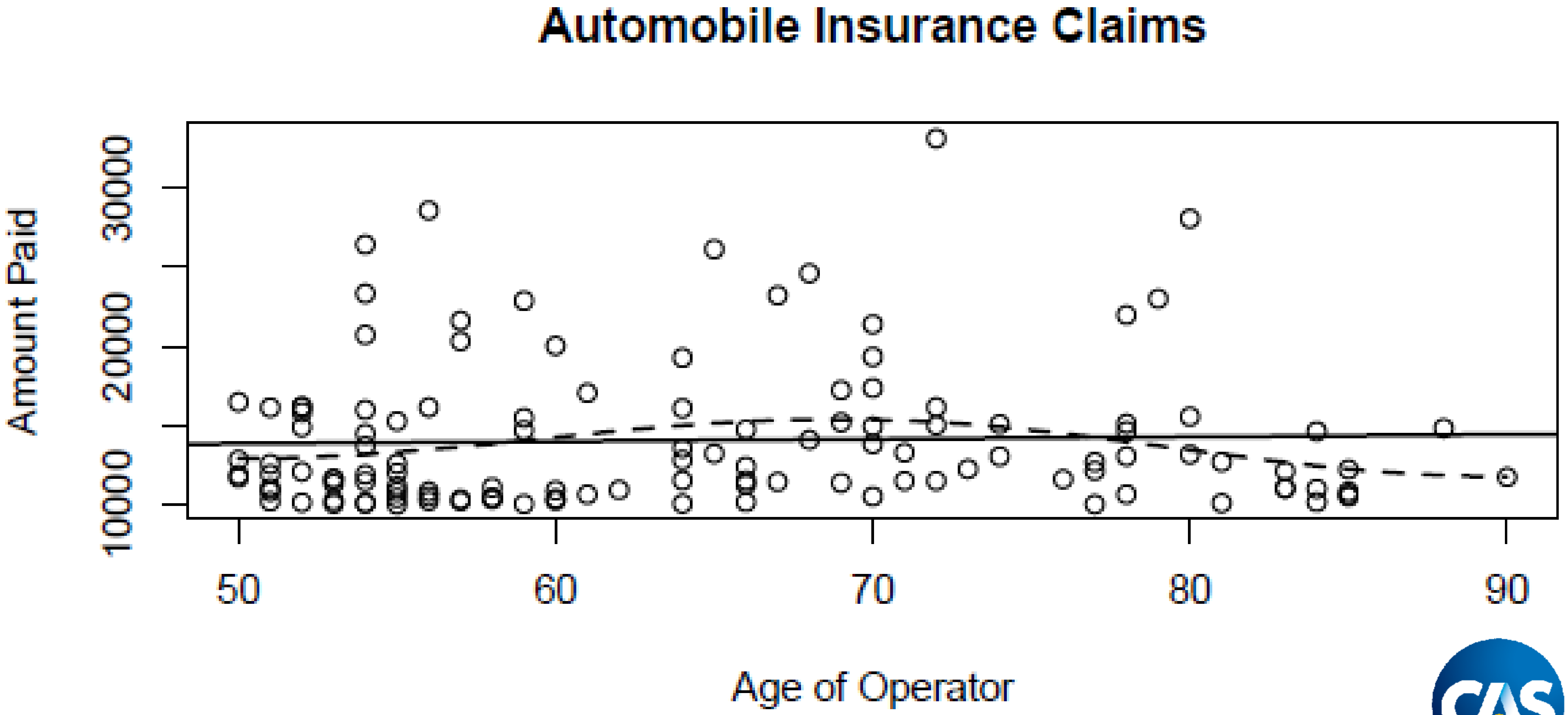
The problem of prediction is converted into an estimation of parameters β_0 and β_1.

Nonparametric Regression

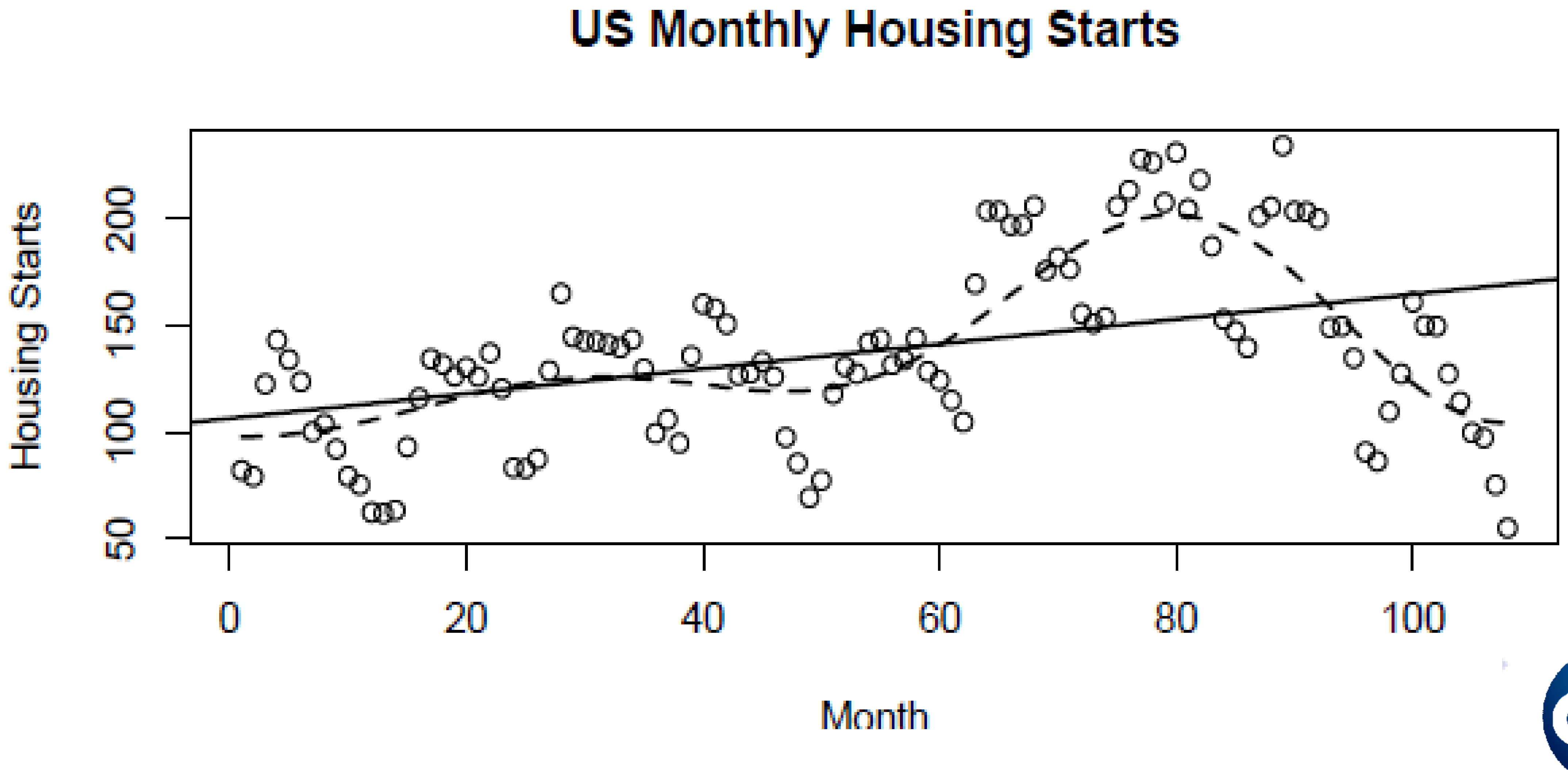


The problem of prediction is converted into an estimation of the regression function *m*(x).

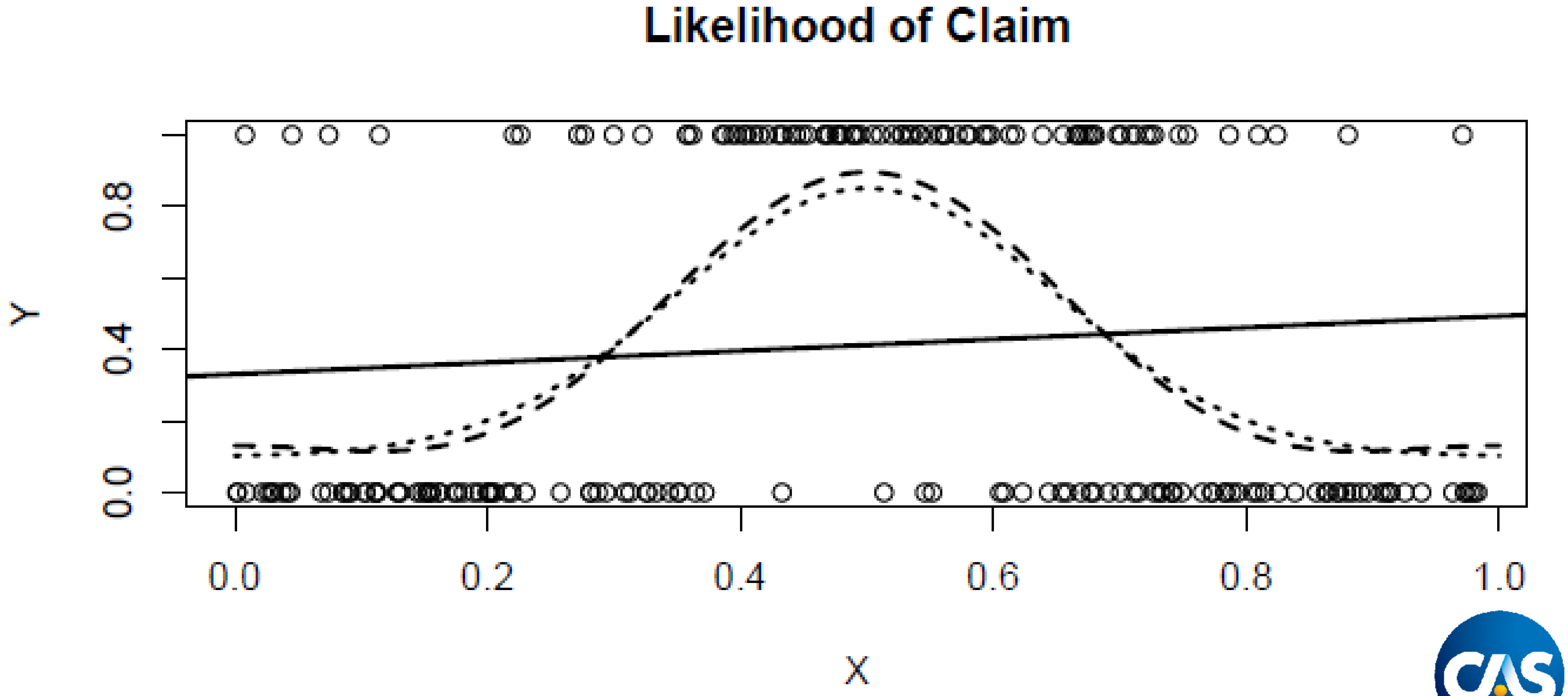
Linear vs. Nonparametric Regression



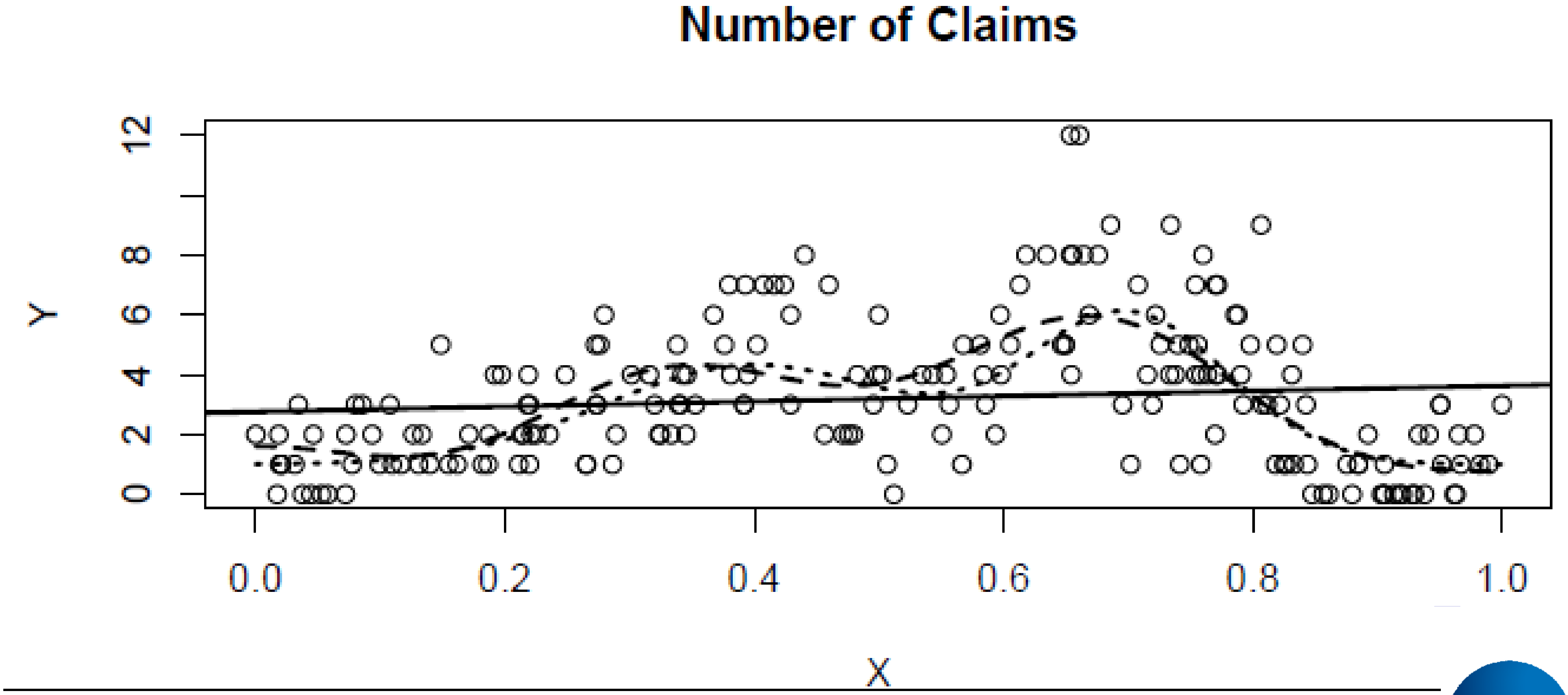
Linear vs. Nonparametric Regression



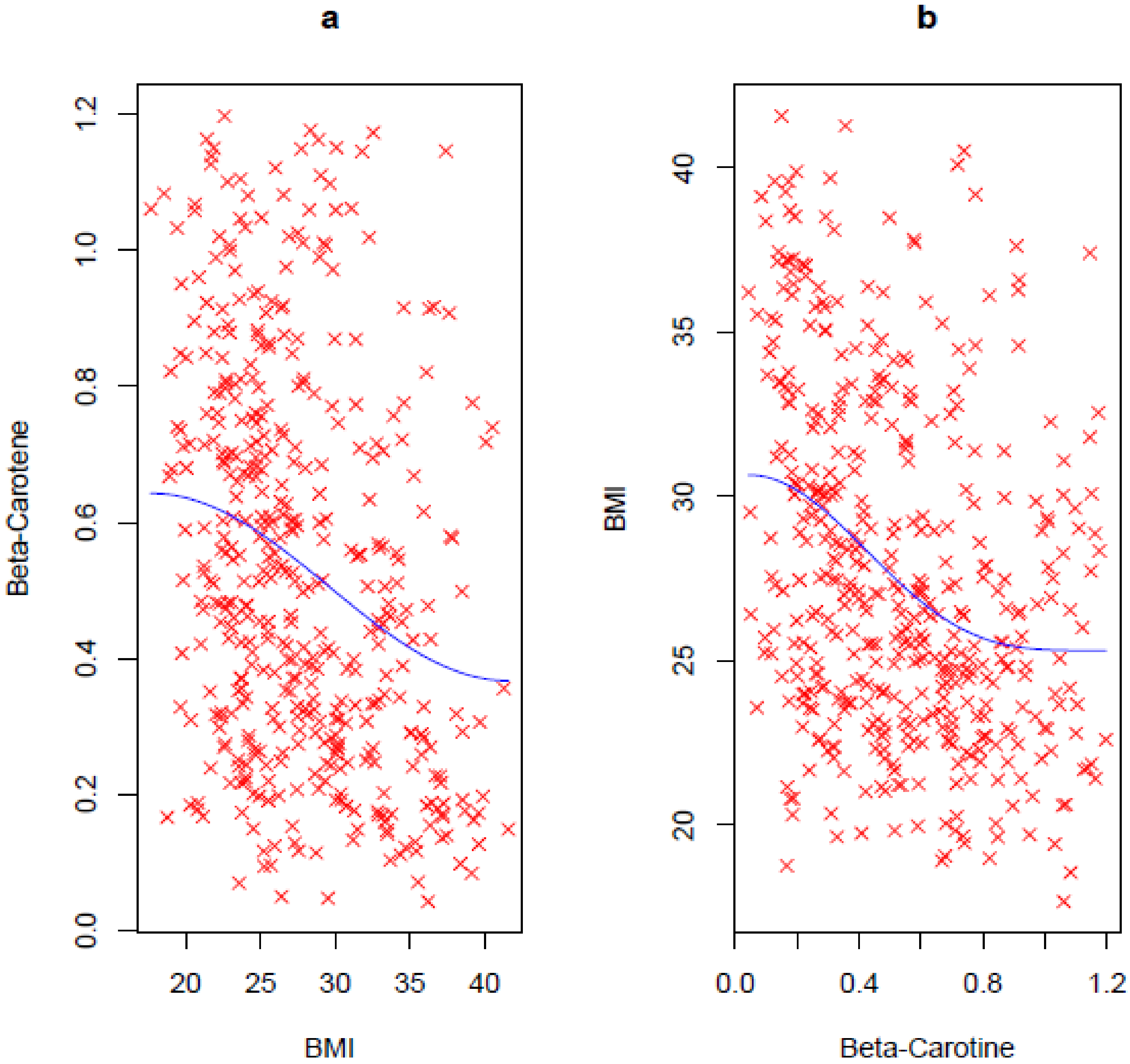
Simulated Bernoulli and Poisson Regression



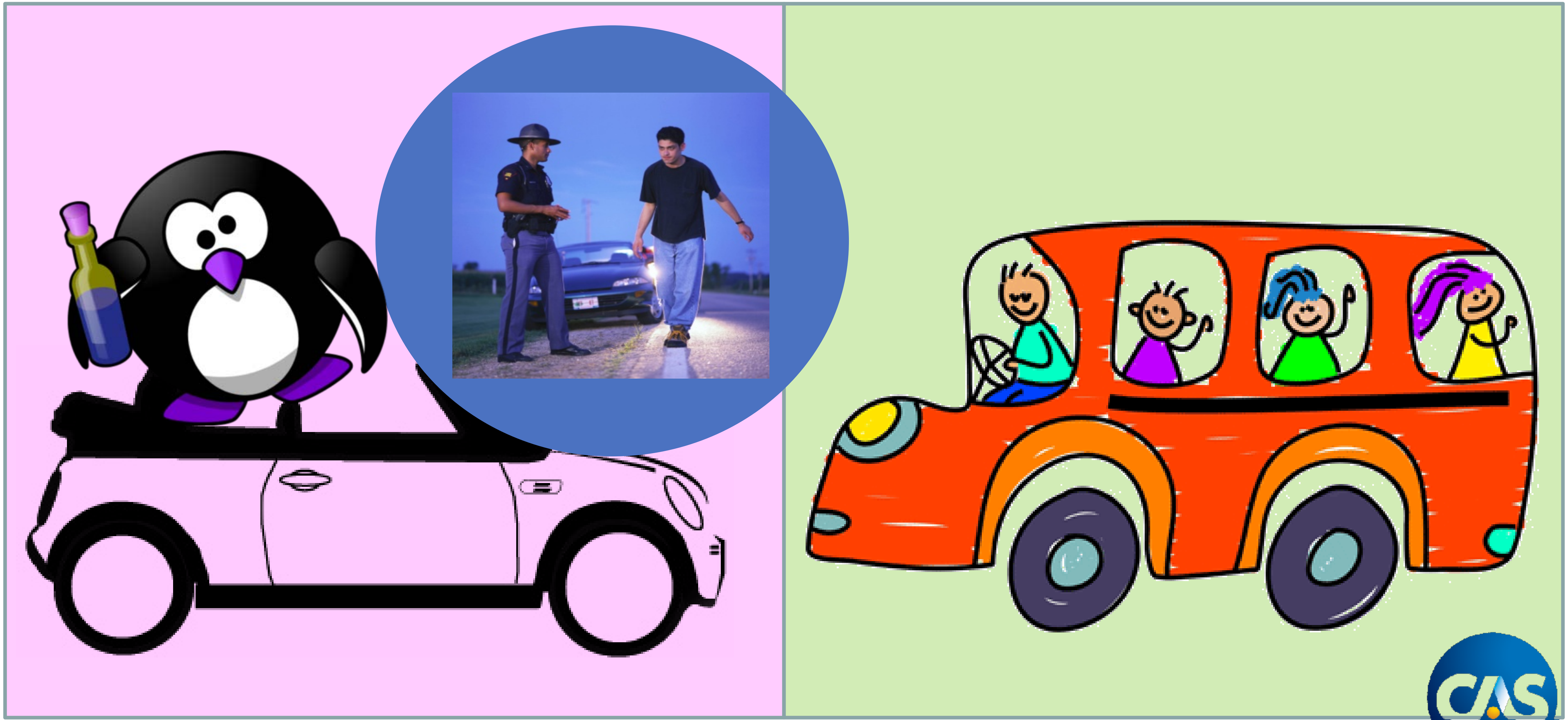
Simulated Bernoulli and Poisson Regression



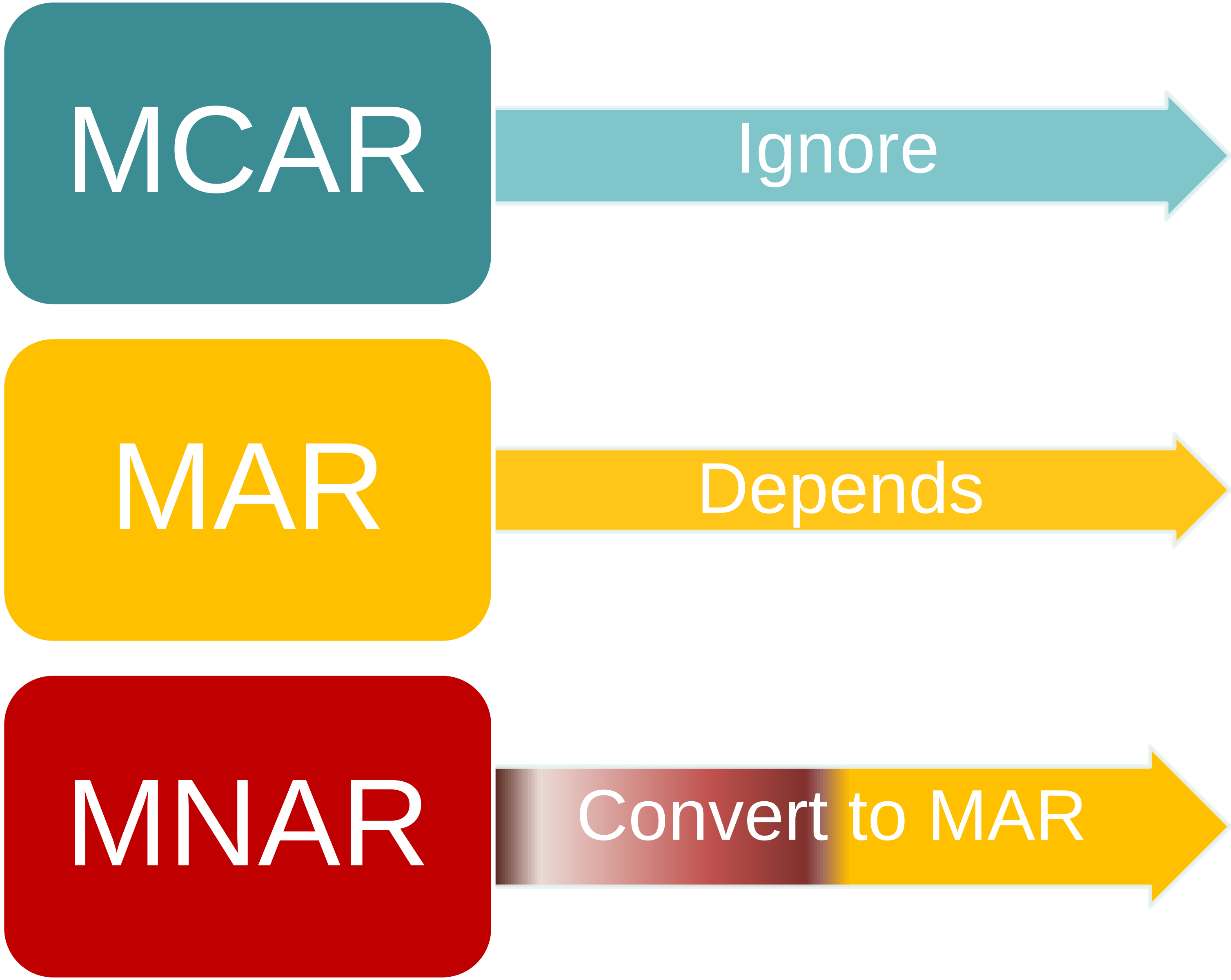
Nonparametric Regression: Body Mass Index vs. Beta Carotene



Example of Missing that Creates Biased Data



Main Types of Missing Data



MCAR → Ignore

MAR → Depends

MNAR → Convert to MAR

Regression with Missing Responses - MNAR

$$\Upsilon = m(X) + \sigma(X)\varepsilon$$

The underlying regression model

Available sample is from: $(A\Upsilon, A, X)$

A is the availability variable (Bernoulli)

Availability likelihood: □$(A=1|X,\Upsilon) = h(\Upsilon)$

Regression with Missing Responses – MNAR

$\Upsilon = \mathrm{m}(\mathrm{X}) + \sigma(\mathrm{X})\varepsilon$ ← The underlying regression model

Available sample is from: $(\mathrm{A}\Upsilon, \mathrm{A}, \mathrm{X})$

A is the availability variable (Bernoulli)

Availability likelihood: $\square(\mathrm{A}=1|\mathrm{X},\Upsilon) = h(\Upsilon)$

The joint density is (set $\psi(y,x) := h(y)f^{\Upsilon|\mathrm{X}}(y|x)$)

Regression with Missing Responses - MNAR

- The underlying regression model is

$$Y = m(X) + \sigma(X)\varepsilon.$$

- Available sample is from (AY, A, X) where: (i) The availability variable A is Bernoulli; (ii) The availability likelihood is

$$\mathbb{P}(A = 1|X, Y) = h(Y).$$

- The joint density is (set $\psi(y,x) := h(y)f^{Y|X}(y|x)$)

$$f^{X,AY,A}(x, ay, a) = [\psi(y,x)f^{X}(x)]^{a}[(1 - \int_{-\infty}^{\infty} \psi(y,x)dy)f^{X}(x)]^{1-a}$$

- We can estimate only the product $\psi(y,x) = h(y)f^{Y|X}(y|x)$, and this implies the MNAR (destructive missing) unless $h(y)$ is known.

Regression with Missing Responses - MAR

- Assume that the availability likelihood is

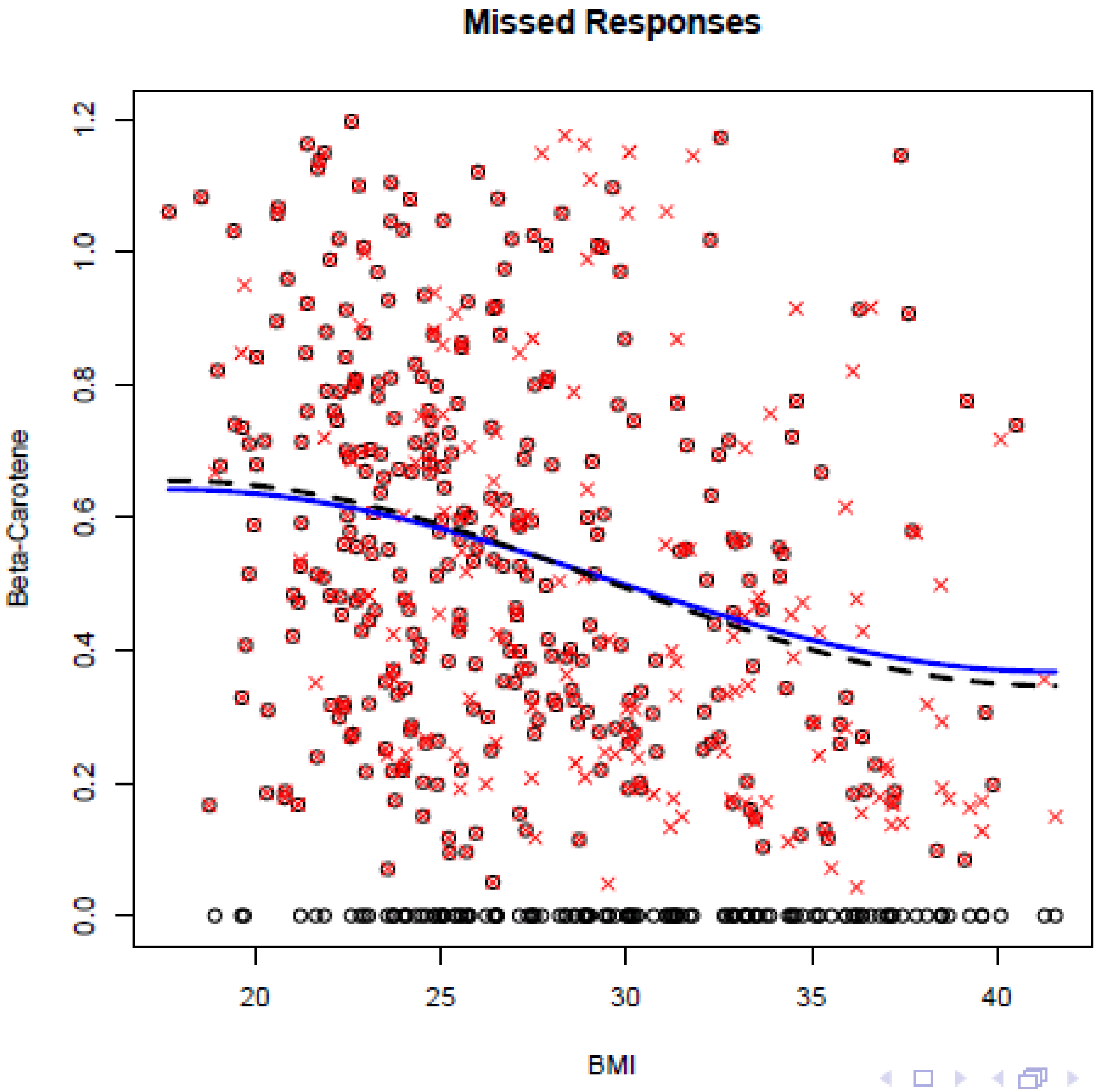
$$\mathbb{P}(A = 1|X, Y) = h(X).$$

- The joint (mixed) density of the triplet is

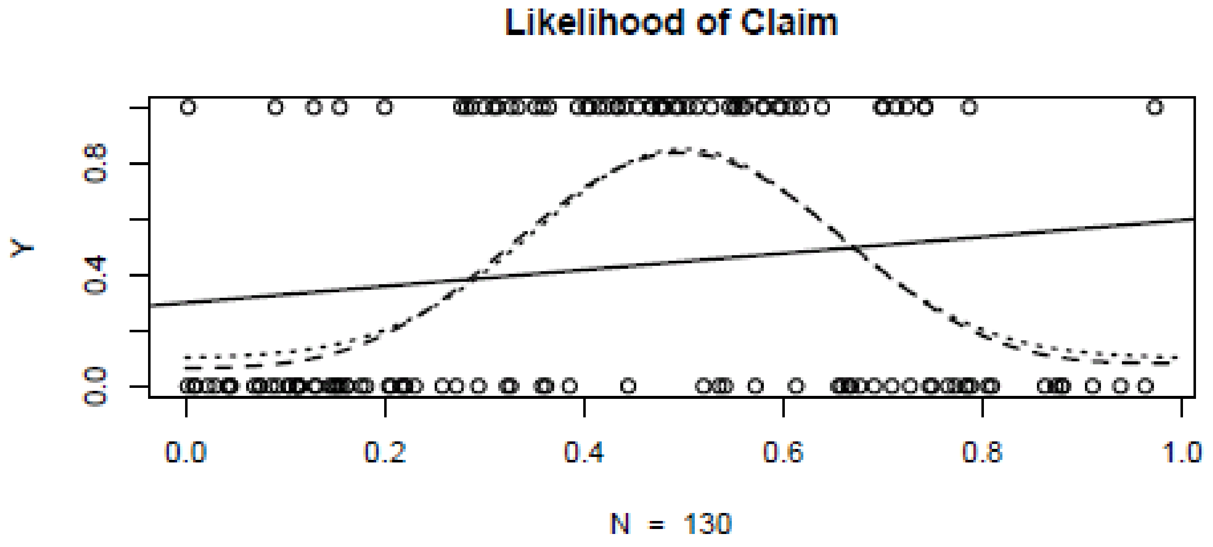
$$f^{X,AY,A}(x, ay, a) = [f^{Y|X}(y|x)h(x)f^X(x)]^a[(1 - h(x))f^X(x)]^{1-a}.$$

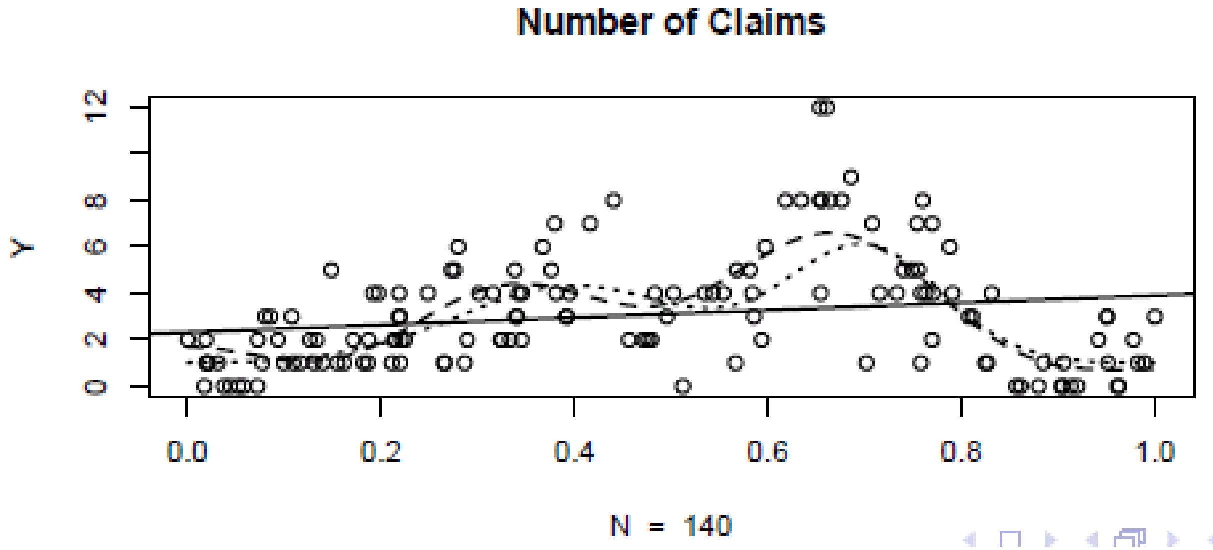
- In a subsample of complete cases the "new" design density is $g^X(x) = h(x)f^X(x)/q$, where $q := \int_0^1 h(x)f^X(x)dx = \mathbb{P}(A = 1)$. This is what allows us to use only complete cases.
- Binomial number $N := \sum_{l=1}^{n} A_l$ of complete cases; sequential estimation looks attractive.
- Traditional Methods: Imputation, Maximum Likelihood, EM, etc.; Vast Literature; Controversy.
- MAR typically does not affect rate of convergence, and the rate is the only issue that the mainstream literature is concerned about.

Regression with Missed Responses, $n = 441$, $N = 312$



Bernoulli and Poisson Regressions with Missed Responses, $n = 200$





Regression with MAR Predictors

- A sample is observed from (Y, AX, A) and the aim is to estimate $m(x) = \mathbb{E}\{Y|X = x\}$.

- It is assumed that the availability likelihood is (MAR)

$$\mathbb{P}(A = 1|X, Y) = \mathbb{P}(A = 1|Y) = h(Y).$$

- The joint density is

$$f^{AX,Y,A}(ax, y, a) = [f^{Y|X}(y|x)h(y)f^{X}(x)]^{a}[(1-h(y))f^{Y}(y)]^{1-a}, a \in \{0, 1\}.$$

- We could use only complete cases if $h(y)$ and $f^{X}(x)$ were known.

Regression Estimation for MAR Predictors

For the case of a complete case when $A = 1$,

$$f^{AX,Y,A}(x, y, 1) = f^{Y|X}(y|x)h(y)f^{X}(x).$$

Steps in regression estimation:

1. Estimate the density of response $f^{Y}(y)$ for $y = Y_l$ where $A_l = 1$.

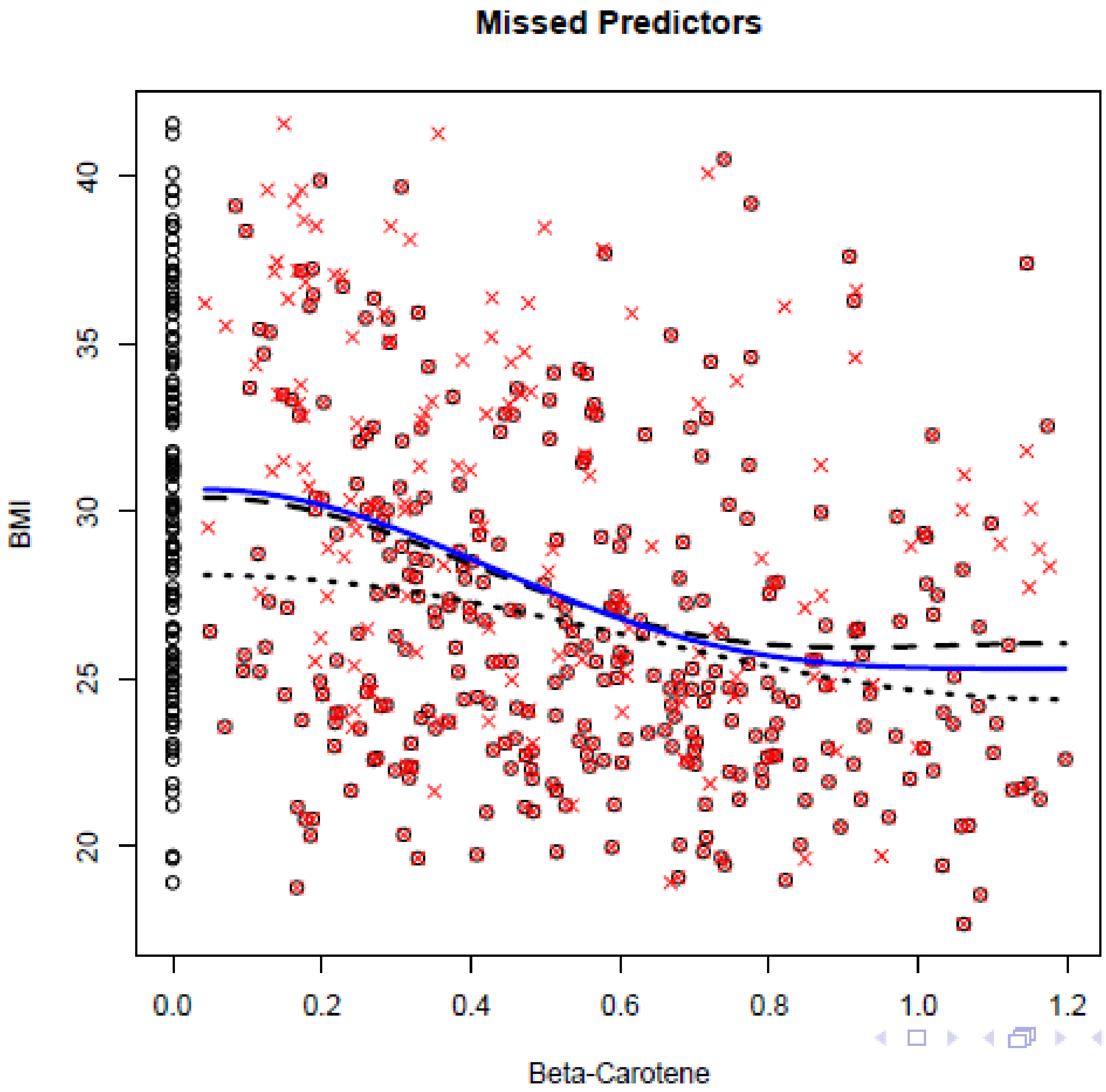
   Note: This is the only place where we need all n observations! (May use a smaller extra sample from Y.)

2. Estimate the availability likelihood $h(y)$ for $y = Y_l$ where $A_l = 1$.

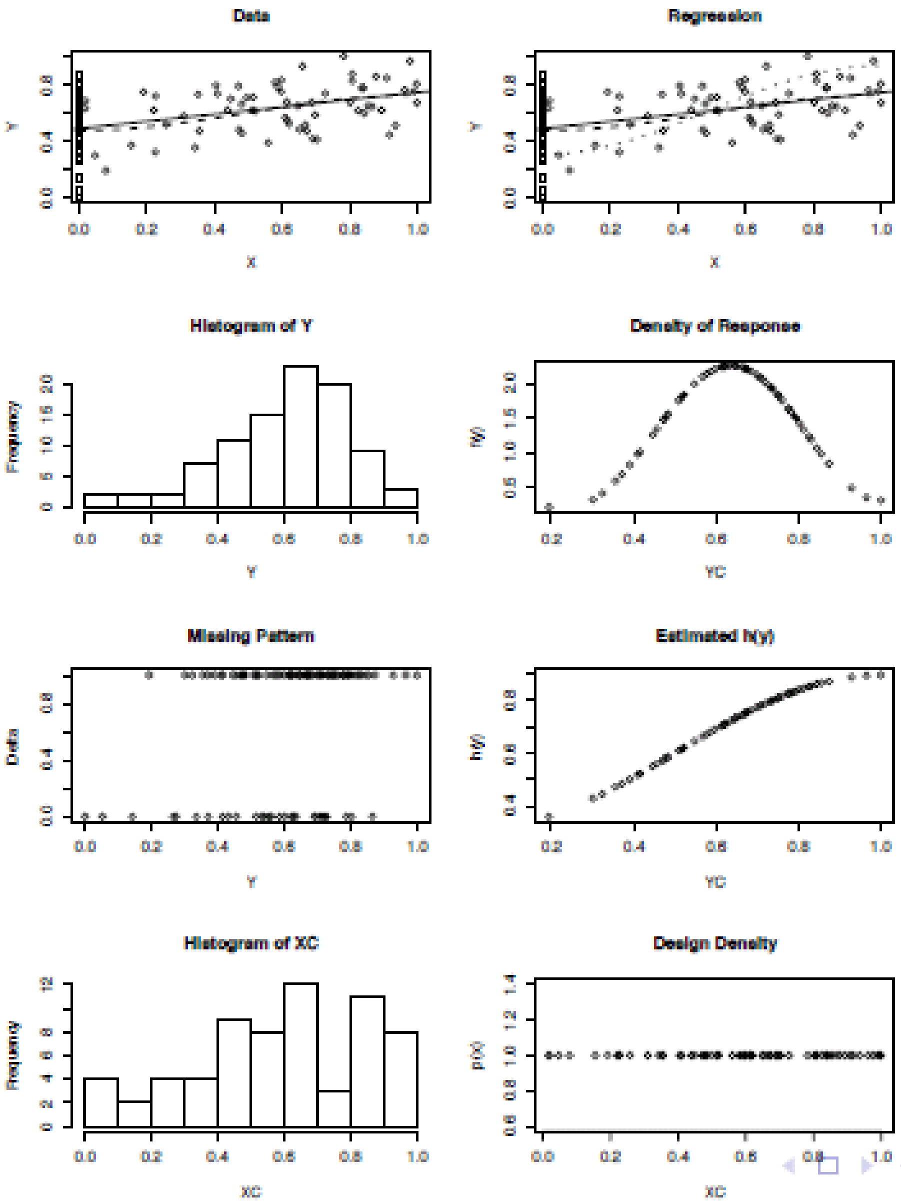
3. Estimate the design density $f^{X}(x)$ for $x = X_l$ where $A_l = 1$.

4. Estimate the regression function based on complete cases.

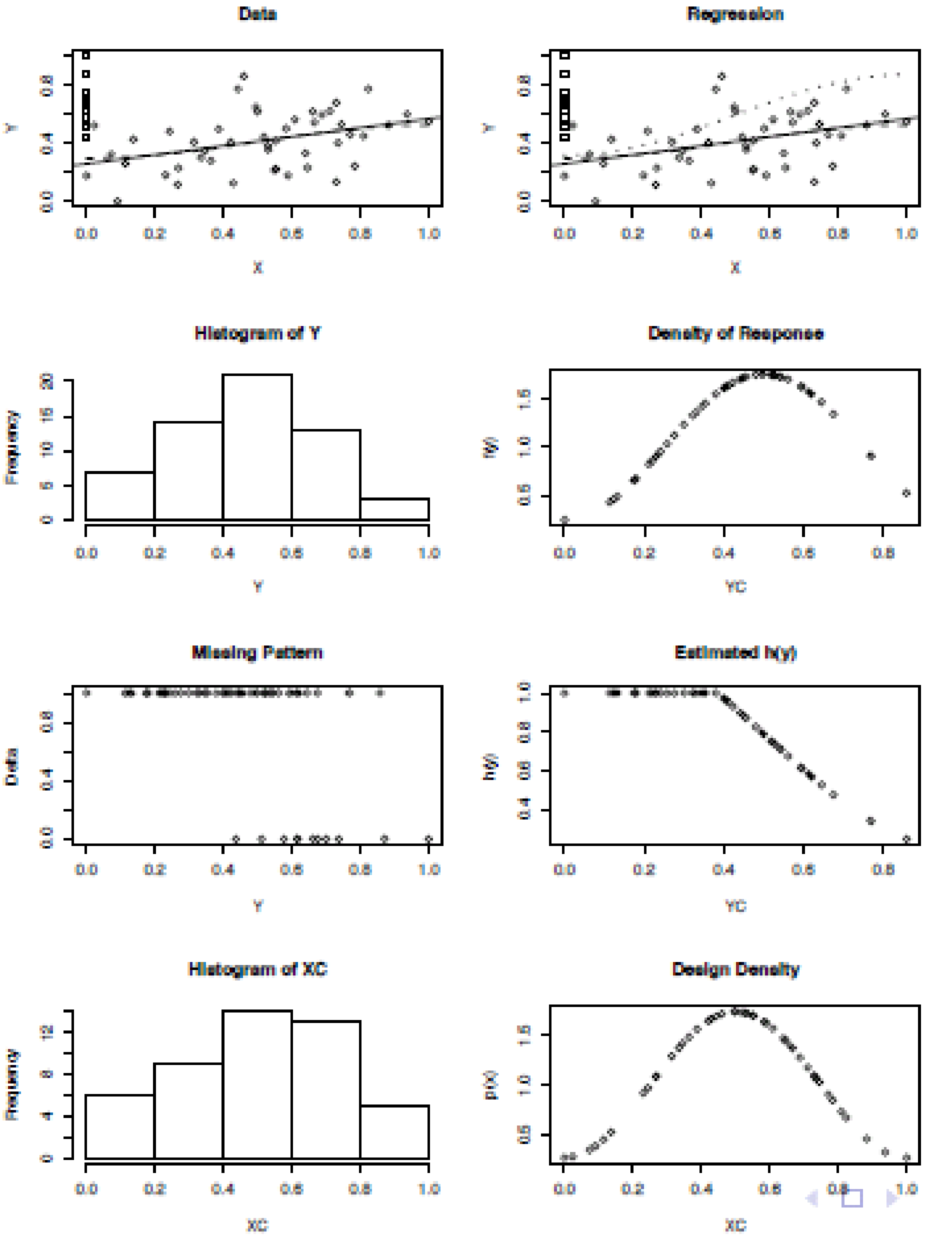
Regression with Missed Predictors, $n = 441, N = 312$



GPA (Y) versus Credit Score (X), Class A: $n = 94$, $N = 65$.



GPA (Y) versus Credit Score (X), Class B: $n = 58$, $N = 47$.



Sam Efromovich

Endowed Professor, Head of Actuarial Program
The University of Texas at Dallas, USA

efrom@utdallas.edu

(Research is supported by the NSF, NSA and CAS-SOA)

May 2017