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

efrom@utdallas.edu

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

May, 2017

Learning Objectives

1. Nonparametric Regression
2. Missing May Cause Data to Be Biased
3. Biased Data
4. Types of Missing Data (Typically Require Different Actions)
5. Missing May Be Destructive
6. How to Deal with Missing
7. Examples

Regression

- Given a sample from (X, Y), the problem of regression is to predict a response Y given a predictor X.
- Parametric (Linear) Regression:

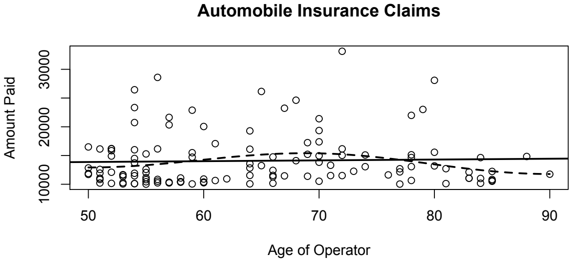
  $$Y := \beta_0 + \beta_1 X + \sigma\epsilon \text{ where } \mathbb{E}\{\epsilon|X\} = 0.$$

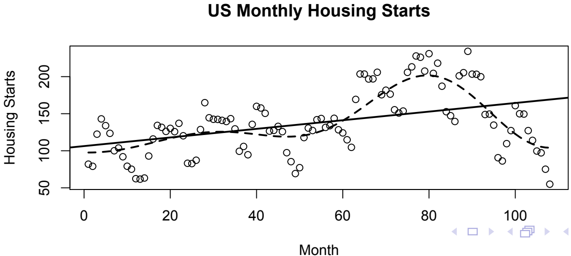
  The problem of prediction is converted into estimation of parameters β_0 and β_1.
- Nonparametric Regression:

  $$Y := m(X) + \sigma(X)\epsilon \text{ where } m(x) := \mathbb{E}\{Y|X = x\}.$$

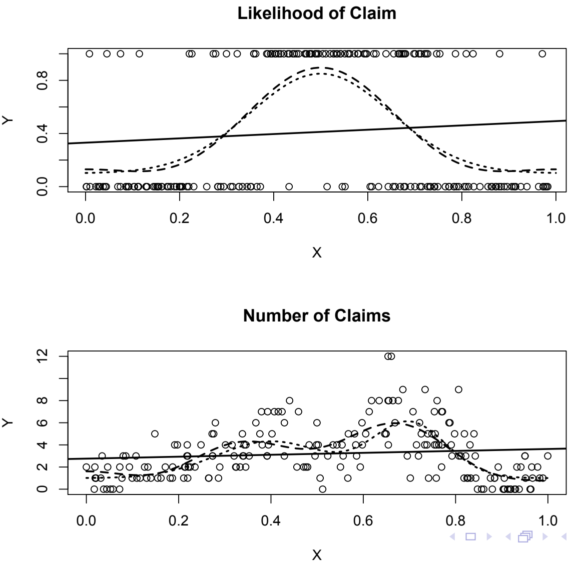
  The problem of prediction is converted into estimation of the regression function $m(x)$.

Linear and Nonparametric Regression, $n_{claims} = 124$, $n_{months} = 108$.

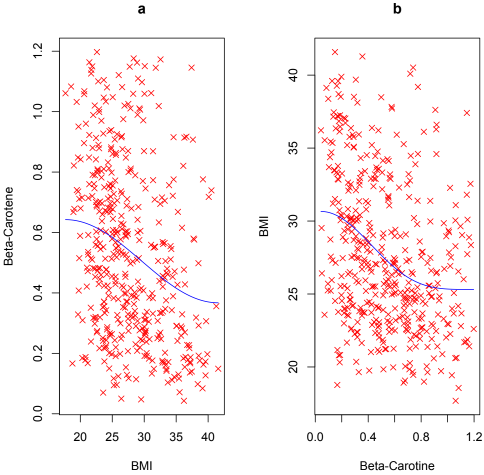




Simulated Bernoulli and Poisson Regression, $n = 200$



Nonparametric Regression: Body Mass Index versus Beta-Carotene



Example of Missing That Creates Biased Data

- Suppose that a researcher would like to know the distribution of the ratio of alcohol in the blood of liquor-intoxicated drivers traveling along a particular highway. The data are available from routine police reports on arrested drivers charged with driving under the influence of alcohol. Because a drunk driver has a larger chance of attracting the attention of the police, it is clear that the data are length-biased toward higher ratios of alcohol in the blood. Thus, the researcher should make an appropriate adjustment in estimation of an underlying density of the ratio of alcohol in the blood of all intoxicated drivers.
- The available data are created by a missing mechanism which is unknown.

Three Main Types of Missing Data

1. Missing completely at random (MCAR)
   Missing a value occurs by a chance that does not depend on the missing variable. No destruction of information occurs. Typically “ignore missing” (complete case) approach is optimal.
2. Missing at random (MAR)
   Missing a value occurs by a chance that depends only on always observed variables. No “complete” destruction of information containing in data occurs. Method of optimal estimation depends on a model.
3. Missing not at random (MNAR)
   Missing a value occurs by a chance depending on data. MNAR implies destruction of information contained in underlying data. Additional data/information, which converts MNAR into MAR, is required for consistent estimation.

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

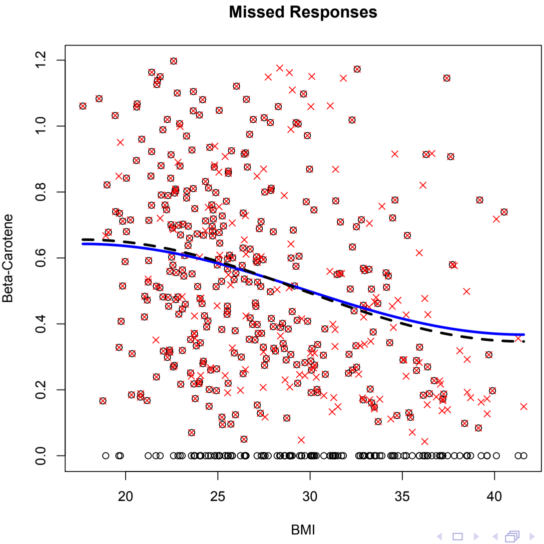
$$\mathbb{P}(A=1|X,Y)=h(X).$$

- The joint (mixed) density of the triplet is

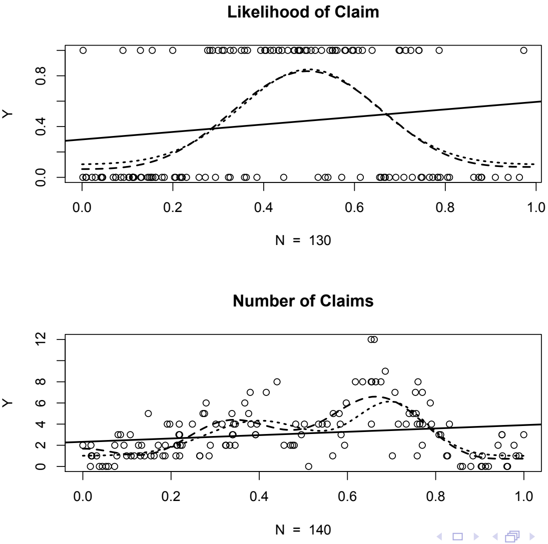
$$f^{X,AY,A}(x,ay,a)=[f^{Y|X}(y|x)h(x)f^X(x)]^a[(1-h(x))f^X(x)]^{1-a}.$$

- 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

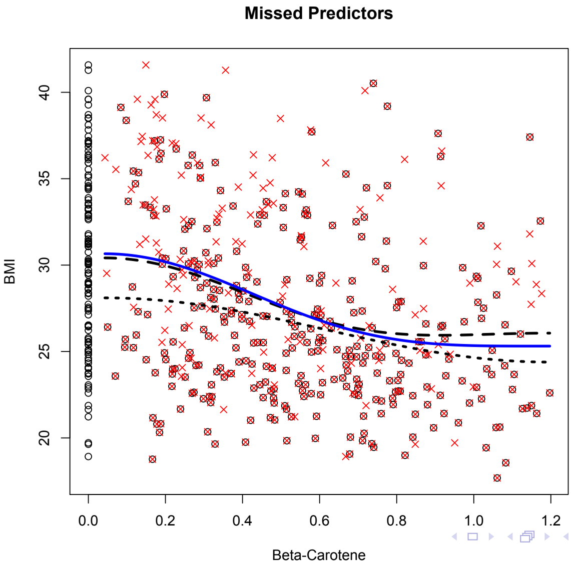
For the case of a complete case when $A = 1$,

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

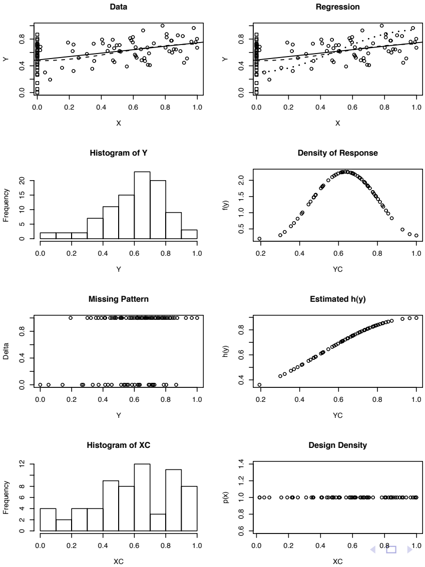
Steps in regression estimation:

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.

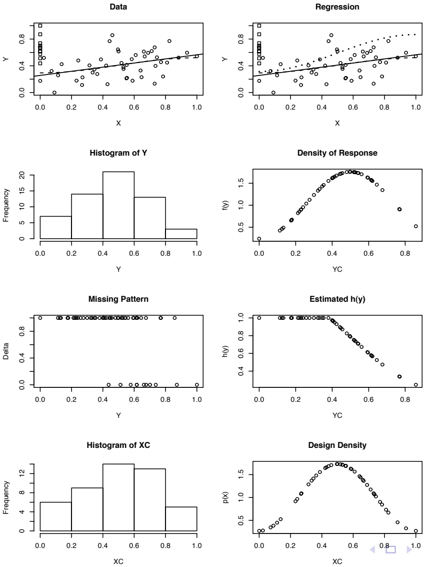
Regression with Missed Predictors, $n = 441$, $N = 312$



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



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


Data
Y
X
Regression
Y
X
Histogram of Y
Frequency
Y
Density of Response
f(y)
YC
Missing Pattern
Delta
Y
Estimated h(y)
h(y)
YC
Histogram of XC
Frequency
XC
Design Density
p(x)
XC