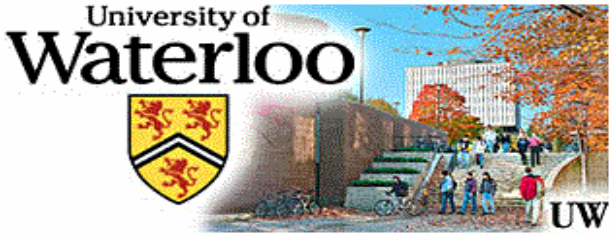
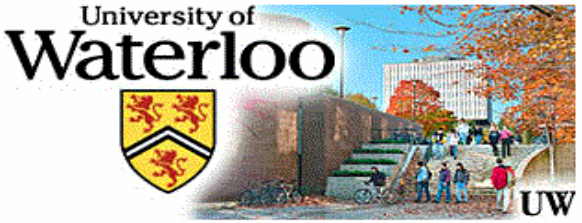
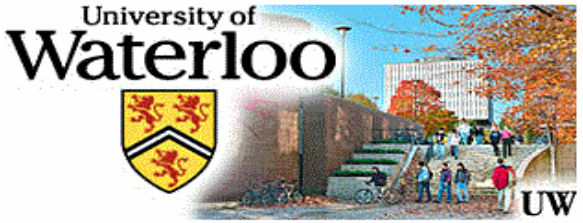




REGRESSION WITH MISSING COVARIATES: IMPORTANCE SAMPLING AND IMPUTATION

Don McLeish & Cyntha Struthers

benefited from discussions with Jerry Lawless, Yang Zhao, Chris Wild, Rod Little,...)



Outline

- Introduction: Regression with missing covariates
- Discuss possible solutions, all weighted averages of the score function with different weights: relate to importance sampling.
- Provide simulations and attempt conclusions.

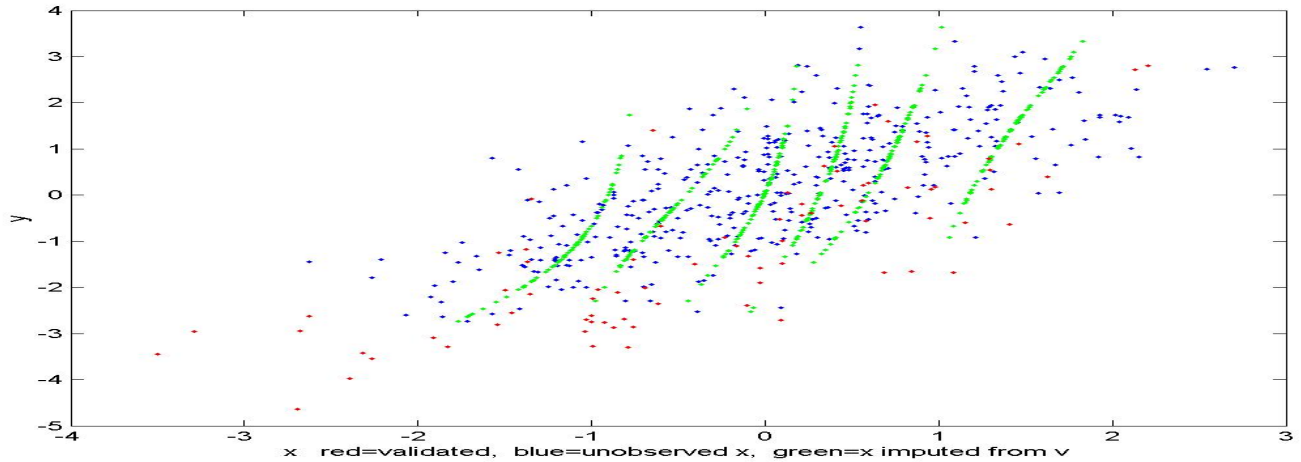
The Regression Problem

Y is response variable. Y has p.d.f. $f(y \mid x, v, \beta)$.

x, v are covariates. β is vector of unknown parameters.

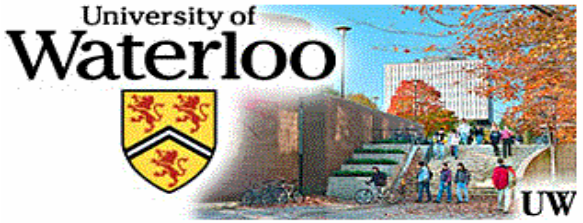
v is observed for all subjects

x is only observed for a subset (the validated sample)



Example: Low Birthweight data

- Risk factors involved in low-birthweight babies (e.g. Lawless, Kalbfleisch and Wild ,1999, Thompson et al, 2001)

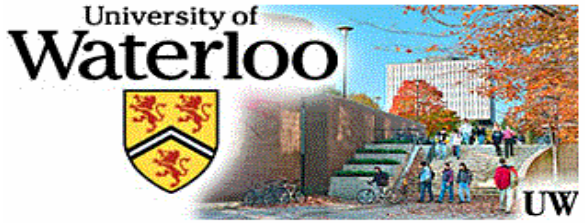


Y = Birthweight of baby

v = gestational age, sex, smoking habits and other routine hospital - collected covariates

x = covariates of interest

In simulations v is a surrogate for x.



Missingness

x is "Missing at random" (MAR) (Little & Rubin, 2002) :

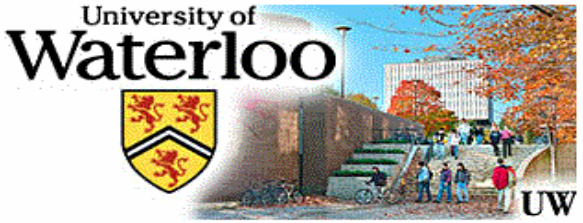
Probability of missing depends only on observed data.

$\Delta = 1$ or 0 as x is observed or not. We assume

$P(\Delta = 1 \mid \mathrm{y}, \mathrm{x}, \mathrm{v}) = \pi(\mathrm{y}, \mathrm{v})$ does not depend on x.

X may be missing by design (e.g. x=expensive covariate, v=surrogate) or by accident.

The function $\pi(\mathrm{y}, \mathrm{v})$ is known.



The ML estimating function

For complete data, might use ML estimating function of the form

$$\sum_{i=1}^{N} S(y_i \mid x_i, v_i, \beta)$$ with S the score function

$$S(y \mid x, v, \beta) = \frac{\delta}{\delta\beta} \ln f(y \mid x, v, \beta)$$

(or any unbiased estimating function - robust against misspecified $f(x,v)$)

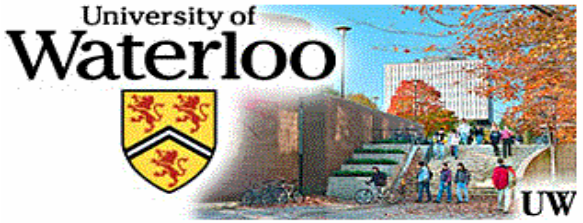
For incomplete data, project (condition) on observed data :

$$\sum_{i=1}^{N} \{\Delta_i S(y_i \mid x_i, v_i, \beta) + (1-\Delta_i) E[S(Y \mid X, V, \beta) \mid y_i, v_i]\}$$

This is MLEF for partially observed information.

Robins et. al. ('94,'95)

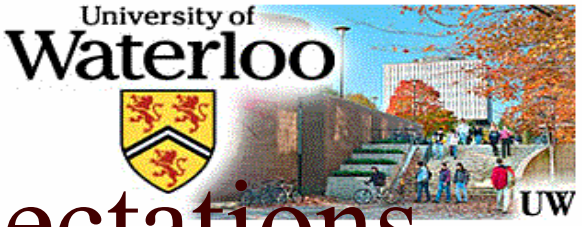
Conditional Distribution $f(x|v)$ unknown



The term $E[S(Y \mid X, V, \beta) \mid y, v]$

in estimating function is unknown since $f(x|v)$ the conditional distribution of X given V is unknown. *$f(x|v)$ is a nuisance parameter*.

Pepe and Fleming (1991), Carroll and Wand (1991) use the empirical distribution of $X|V$ for validated X only.



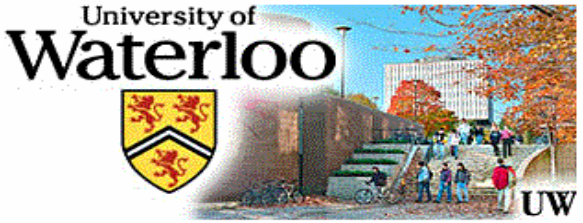
Estimating Conditional Expectations

- Suppose we want to estimate $E[g(X)|Y=y,V=v]$.
  - Average $g(X)$ over all validated observations with $Y=y,V=v$

    Problem: There may be none!
  - **Better:** Average $g(X)w(X,y,v)$ over ALL validated X

where w satisfies $E[g(X)|Y=y,V=v]=E[g(X)w(X,y,v)]$

Or impute values of X using some importance distribution.



Importance Imputation

Notice that for arbitrary joint p.d.f. $h(X, V)$ (which may depend on y, v and arbitrary density $K_v(V)$ which may depend on v, *(X,V)* generated from *h*

$$
\begin{aligned}
E[g(X)\frac{f(X|y,v,\beta)}{h(X,V)}K_v(V)] &= \int\int g(X)\frac{f(X|y,v,\beta)}{h(X,V)}K_v(V)h(X,V)dXdV \\
&= \int\int g(X)f(X|y,v,\beta)K_v(V)dXdV \\
&= \int g(X)f(X|y,v,\beta)dX \\
&= E_\beta[g(X)|y,v].
\end{aligned}
$$

Approximating Conditional Expected value

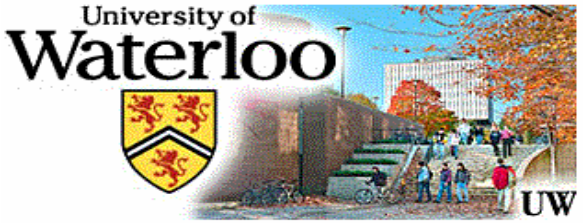
Therefore we can approximate this conditional expectation with a weighted average of the self-normalized form

$$E_\beta[g(X)|y,v] \simeq \frac{\sum w_j g(X_j)}{\sum w_j}$$

where

$$\begin{aligned} w_j &= w(x_j, v_j, y, v, \beta) \propto \frac{f(x_j|y,v;\beta)}{h(x_j,v_j)} K_v(v_j) \\ &\propto \frac{f(y|x_j,v;\beta) f(x_j|v)}{h(x_j,v_j)} K_v(v_j) \end{aligned}$$

Examples: *h* and corresponding *w*

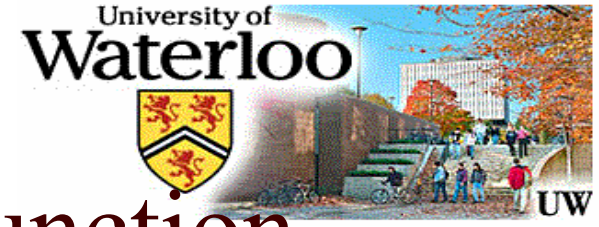


Method	$\mathbf{h}(\mathbf{x}_j, \mathbf{v}_j \mid \mathbf{y}, \mathbf{v}) \propto$	$\mathbf{w}(\mathbf{x}_j, \mathbf{v}_j, \mathbf{y}, \mathbf{v}, \boldsymbol{\beta}) \propto$
Pepe and Fleming (91)*	$\Delta_j f(x_j, v)$	$\Delta_j f(y \mid x_j, v; \beta)$
Reilly&Pepe(95): mean score*	$\Delta_j f(x_j \mid y, v; \beta)$	Δ_j
Chatterjee, Chen, Breslow(03)*	$f(x_j \mid v)\eta(x_j, v_j, \beta)$	$\frac{\Delta_j}{\eta(x_j, v, \beta)} f(y \mid x_j, v; \beta)$
No-name	$f_x(x_j) E_\beta[\Delta_j \mid x_j]$	$\frac{\Delta_j}{E[\Delta_j \mid x_j]} f(y \mid x_j, v; \beta) f(v \mid x_j) K_v(v_j)$
Quasi-profile 1*	$f(x_j \mid v)\eta(x_j, v_j, \beta)$	$\frac{\Delta_j}{\widehat{\eta}_P(x_j, v_j, \beta)} f(y \mid x_j, v; \beta)$
Quasi-profile 2 (all x)	constant	$f(y \mid x_j, v; \beta)\widehat{f}_{NP2}(x_j \mid v)$
New Profile	constant	$f(y \mid x_j, v; \beta)\widehat{f}_{NP3}(x_j \mid v)$
Regression	$f_x(x_j)\eta(x_j, v_j, \beta)$	$\frac{\Delta_j}{\eta(x_j, v_j, \beta)} f(y \mid x_j, v; \beta)\widehat{f}_N(v \mid x_j) K_v(v_j)$
Copula 1	$f_x(x_j)\eta(x_j, v_j, \beta)$	$\frac{\Delta_j}{\eta(x_j, v, \beta)} f(y \mid x_j, v; \beta)\widehat{f}_C(v \mid x_j) K_v(v_j)$

where $\widehat{f}_{NP1}, \widehat{f}_{NP2}, \widehat{f}_{NP3}$ are nonparametric MLE of the conditional distribution, differing only in the assumed supporting values of x, $\widehat{f}_N$ approximates using bivariate normality, $\widehat{f}_C$ using Copula.

the symbol $\propto$ indicating up to a factor involving y, v. The trailing factor $K_v(v_j)$ can be used to localize an approximation. Discrete cases in which $K_v(v_j)$ is assumed $I(v_j = v)$ are labelled with * and it is left out of the weight function. Modified Chatterjee, copula, profile, etc. refer to a Gaussian kernel,

$$K_v(v_j) \propto \exp\{-\mathbf{c}(\mathbf{v}_j - v)^2\}.$$



The Resulting Estimating Function

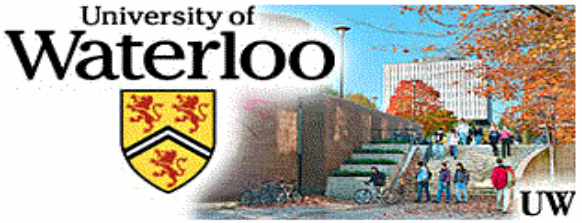
$$\sum_{i=1}^{N} \{\Delta_i S(y_i \mid x_i, v_i, \beta) + (1-\Delta_i)\hat{E}[S(Y \mid X, V, \beta) \mid y_i, v_i]\}$$

where $\hat{E}$ is an estimator of the conditional expectation of the form

$$\hat{E}[S(Y \mid X, V, \beta) \mid y, v] = \sum_j w(x_j, v_j, y, v, \beta) S(y \mid x_j, v, \beta).$$

Weights, $w(x_j, v_j, y, v, \beta)$, are normalized to have sum 1.

Iterative estimation of β



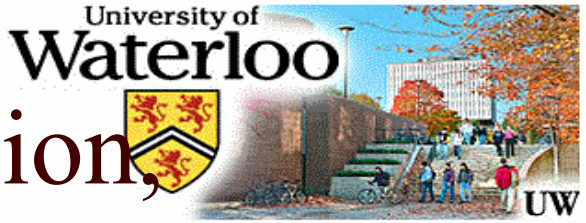
Weights $w(x_j, y, v, \beta)$ depend on β. Use iterative scheme :

β_{n-1} = estimate of β from $(n\text{-}1)'$ *st* iteration.

1) Get weights $w(x_j, y, v, \beta_{n-1})$

2) Solve estimating equation for β_n.

Simulations: Linear Regression, V is a surrogate for X.



Y_i is $N(\beta_0 + \beta_1 X_i, \sigma^2), i = 1,..., N$ indep.

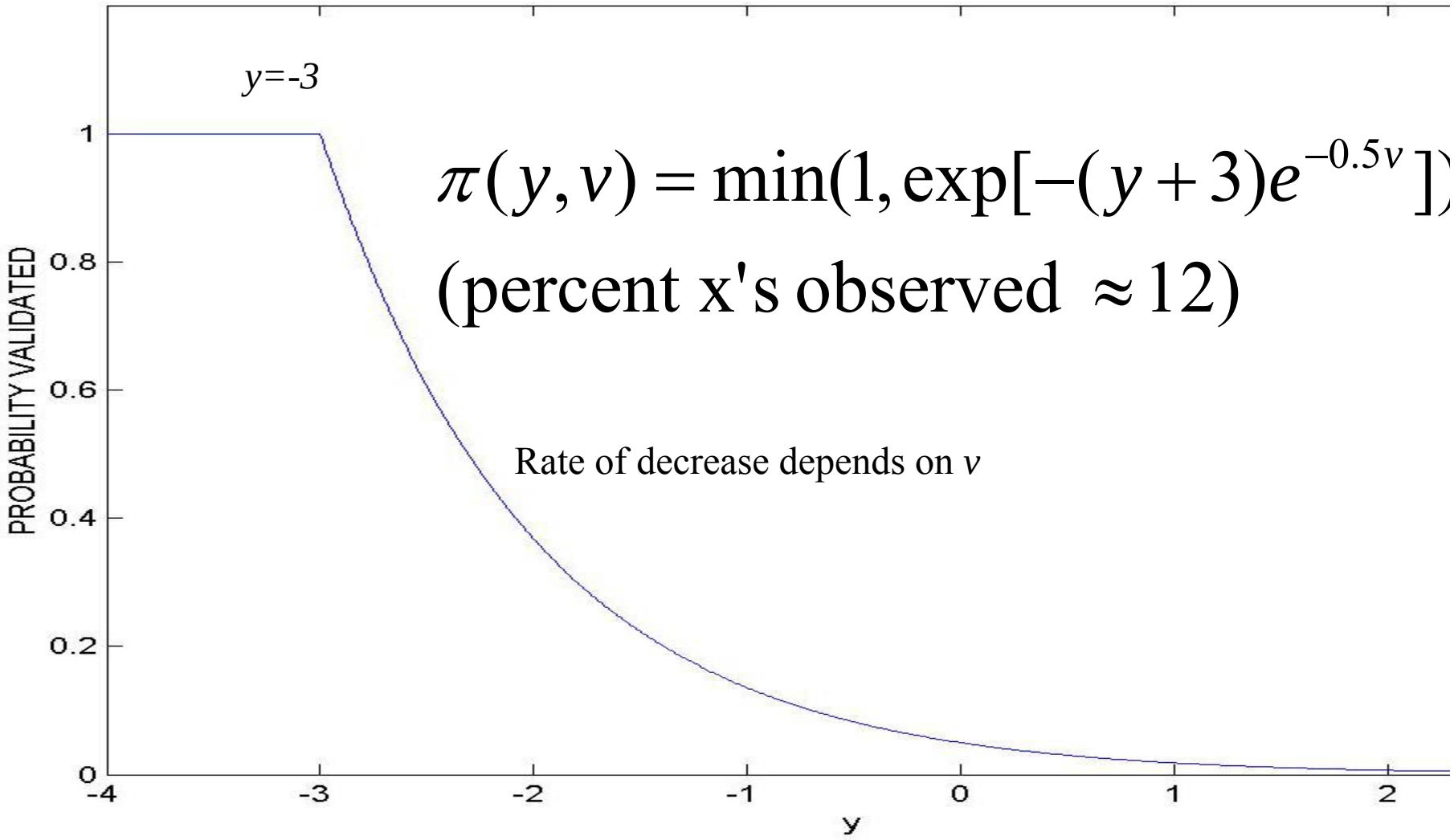
X and V are $N(0,1), Cor(X,V) = \rho$

$N = 1000$ (about 120 validated) $\beta_0 = 0, \beta_1 = 1, \sigma^2 = 1$

$\rho = 0.9,\ 0.75,\ 0.5,\ 0.25$

V discretized into 6 or 20 values or continuous.

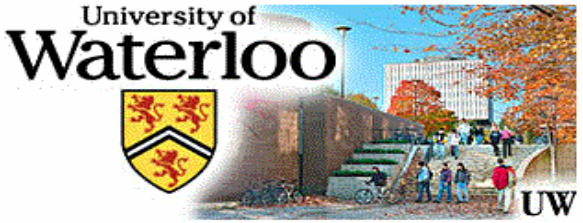
Probability *x* is fully observed



$$\pi(y,v) = \min(1, \exp[-(y+3)e^{-0.5v}])$$

(percent x's observed ≈ 12)

Rate of decrease depends on *v*



Estimators of β_0, β_1.

$$S(y \mid x, v, \beta) = -\frac{(y - \beta_0 - \beta_1 x)}{\sigma^2}\begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} N & \sum \tilde{x}_i \\ \sum \tilde{x}_i & \sum \tilde{x}_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum \tilde{x}_i y_i \end{pmatrix}$$

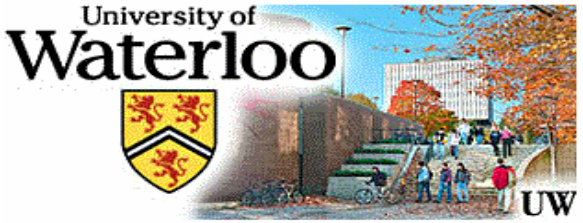
where

$$\tilde{x}_i = \Delta_i g x_i + (1 - \Delta_i)\hat{E}[X \mid y_i, v_i] \text{ and}$$

$$\tilde{x}_i^2 = \Delta_i x_i^2 + (1 - \Delta_i)\hat{E}[X^2 \mid y_i, v_i]$$

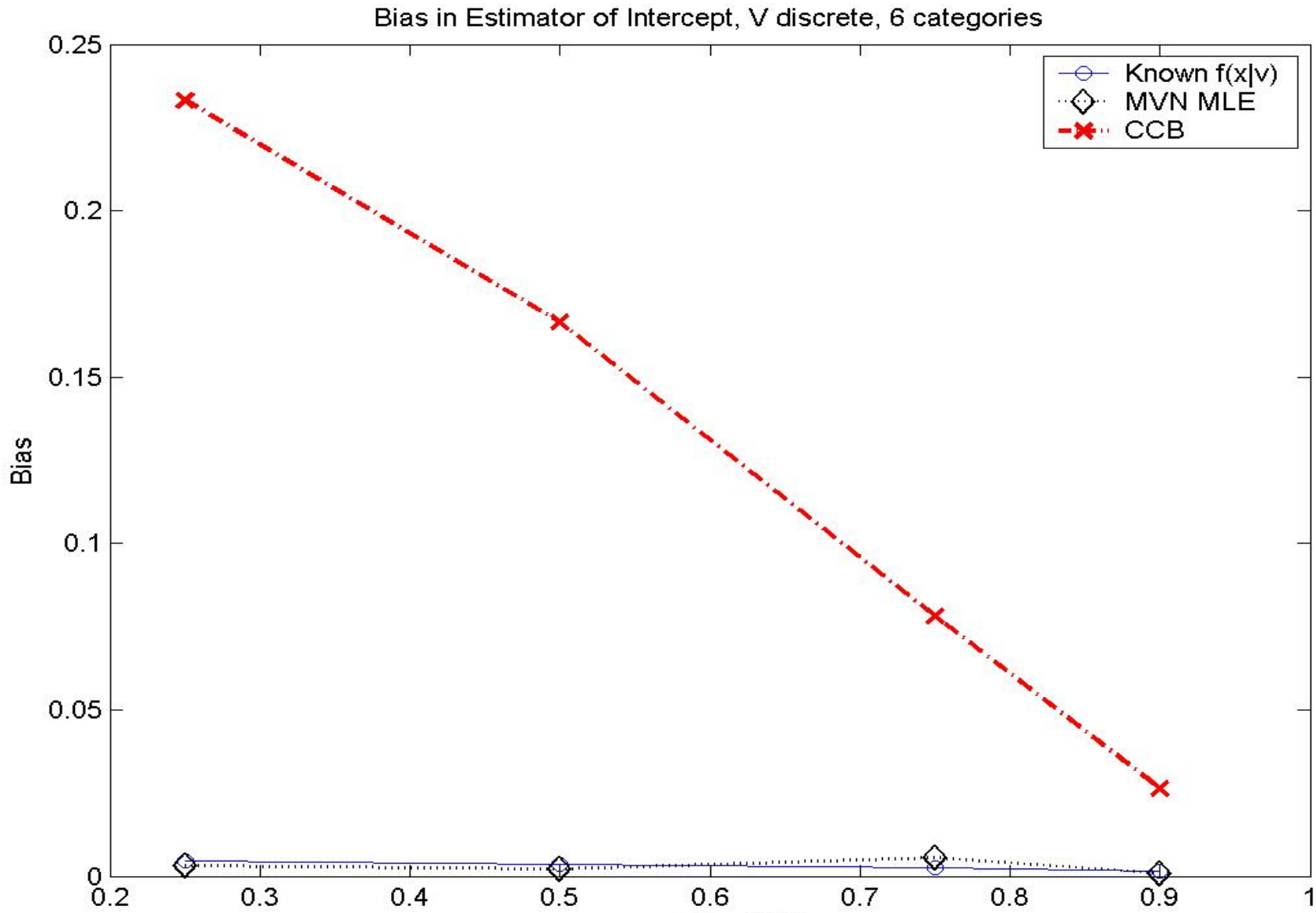
Looks like the usual least squares solution with $\tilde{x}_i$ replacing x_i.

Methods compared to known $f(x|v)$ and MVN MLE case

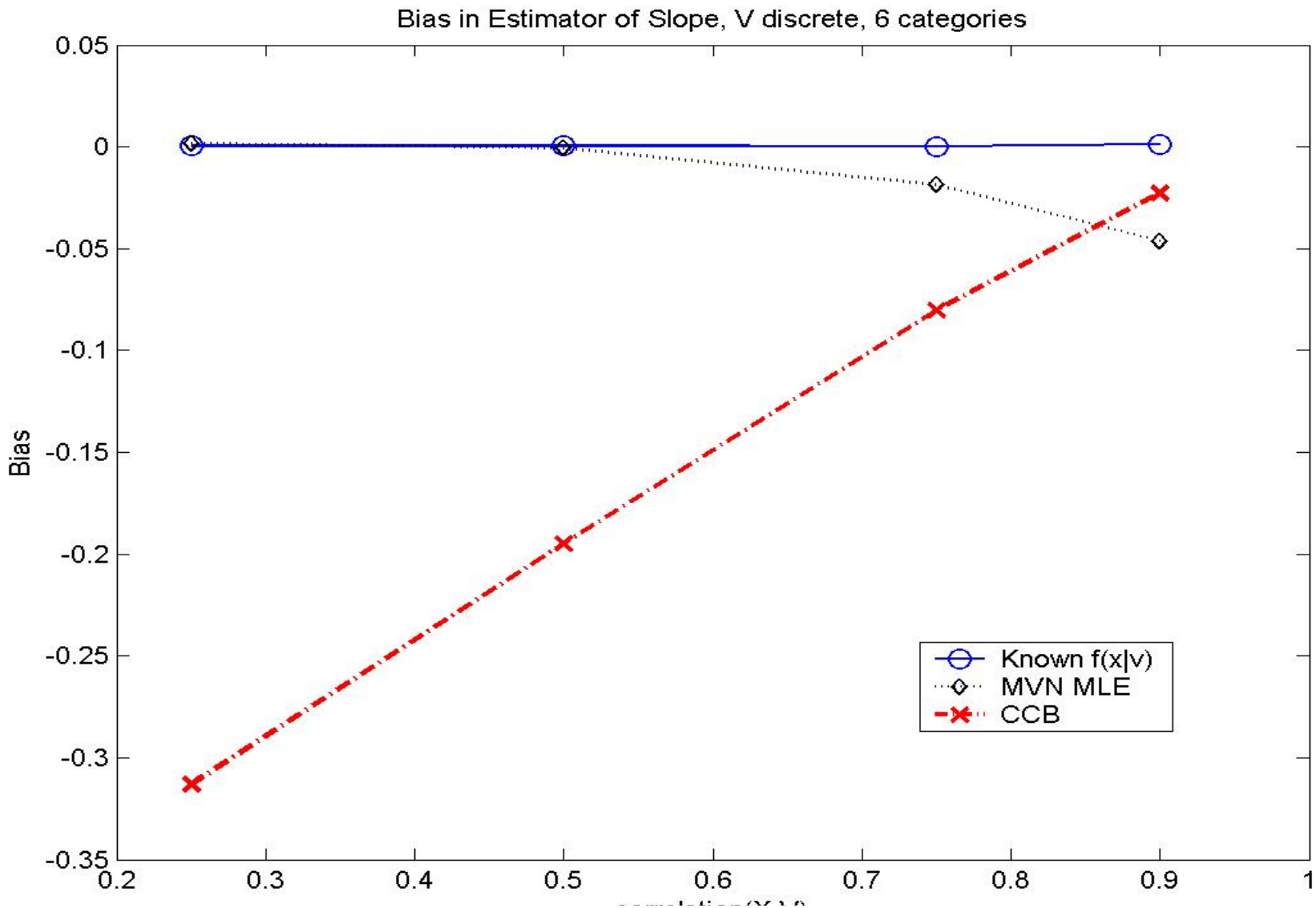


- i.e. maximum likelihood estimation of the regression parameters $\beta_0, \beta_1, \sigma^2$ assuming the conditional distribution of $X|V$ is known
- or MLE assuming trivariate normality, all parameters unknown.

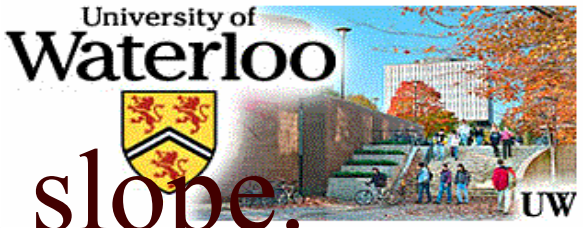
Bias in intercept estimators



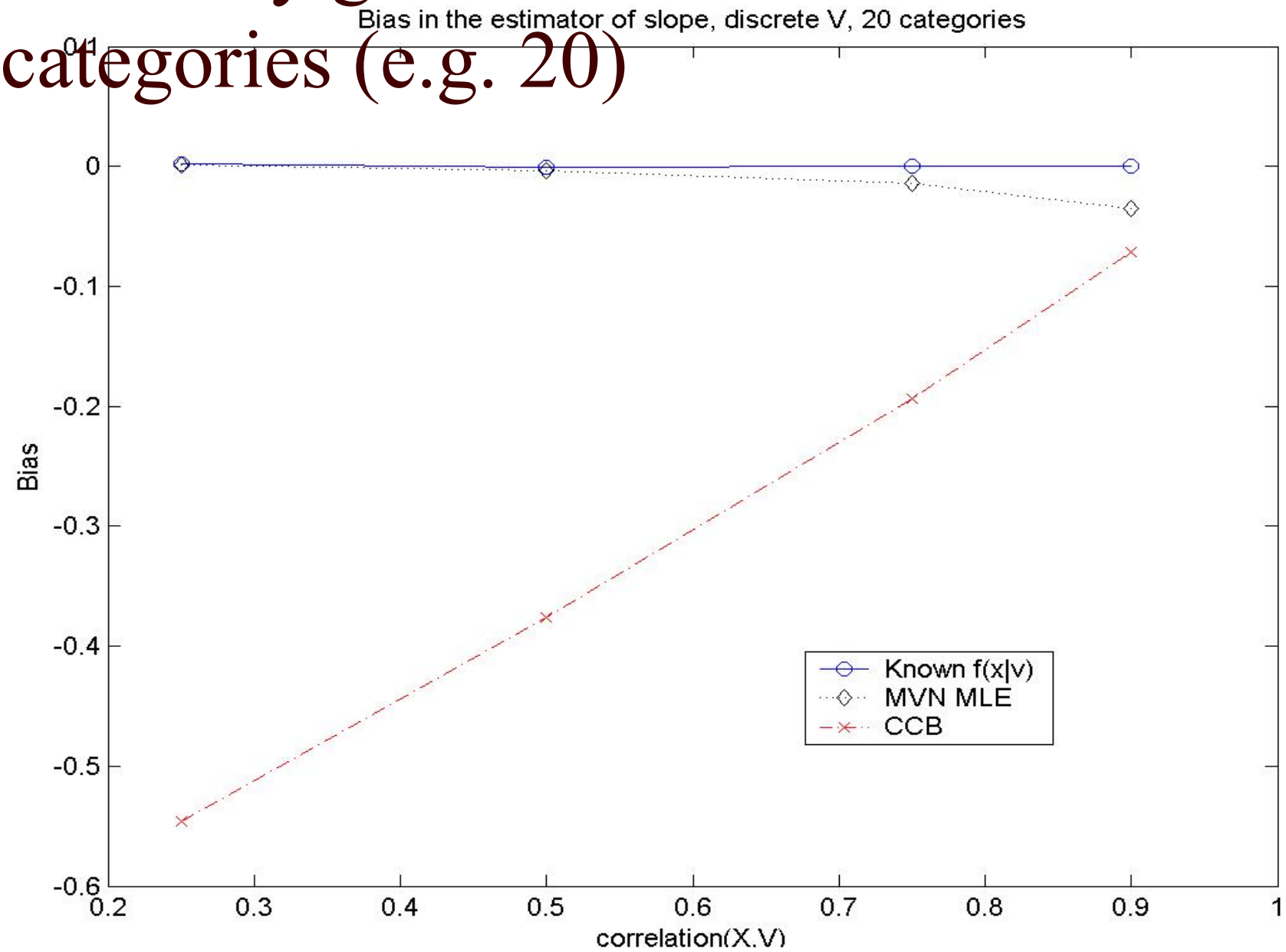
Bias in slope estimators, discrete v (6 categories)



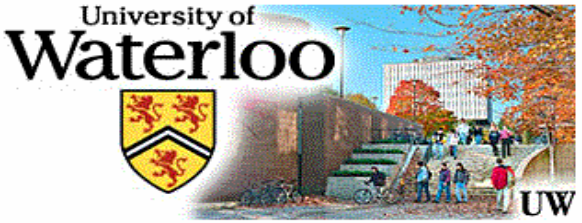
More Estimators…bias in slope.

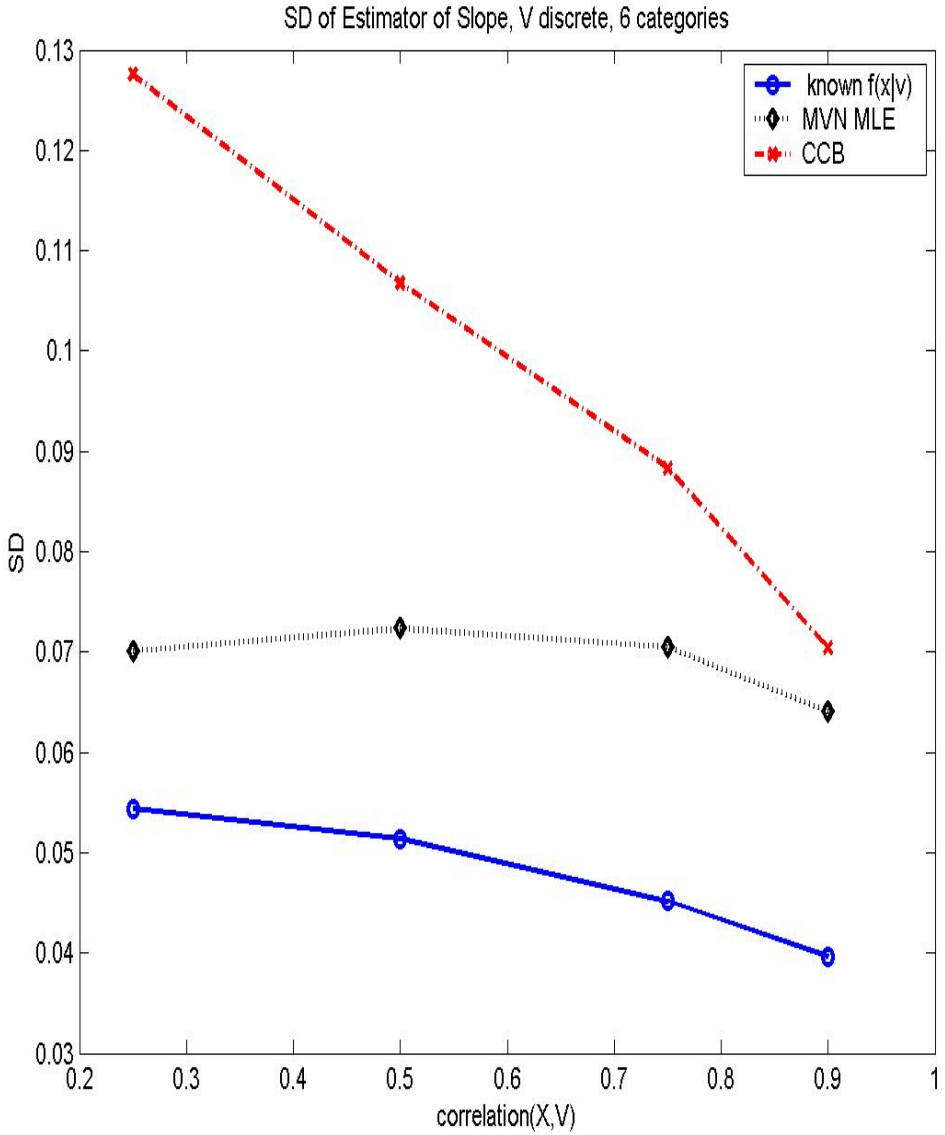


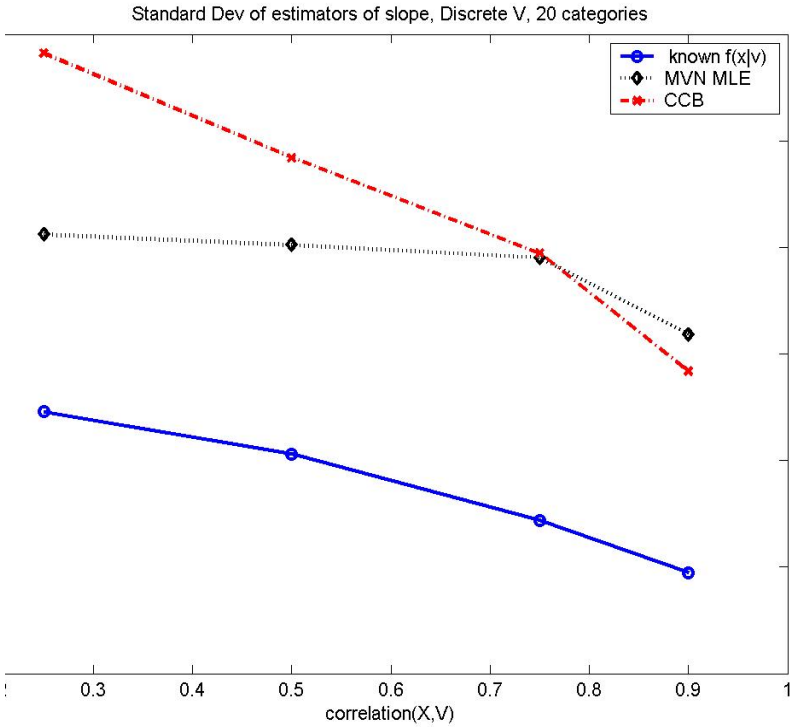
Bias only gets worse as V has more categories (e.g. 20)

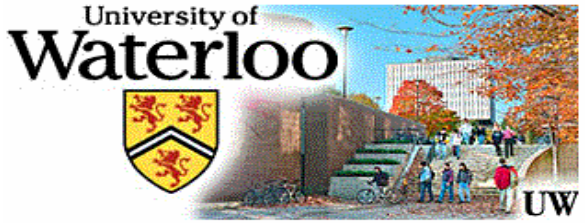


Standard Dev of Estimators



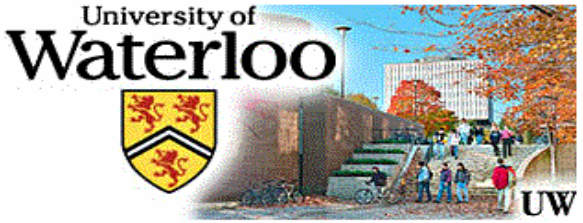




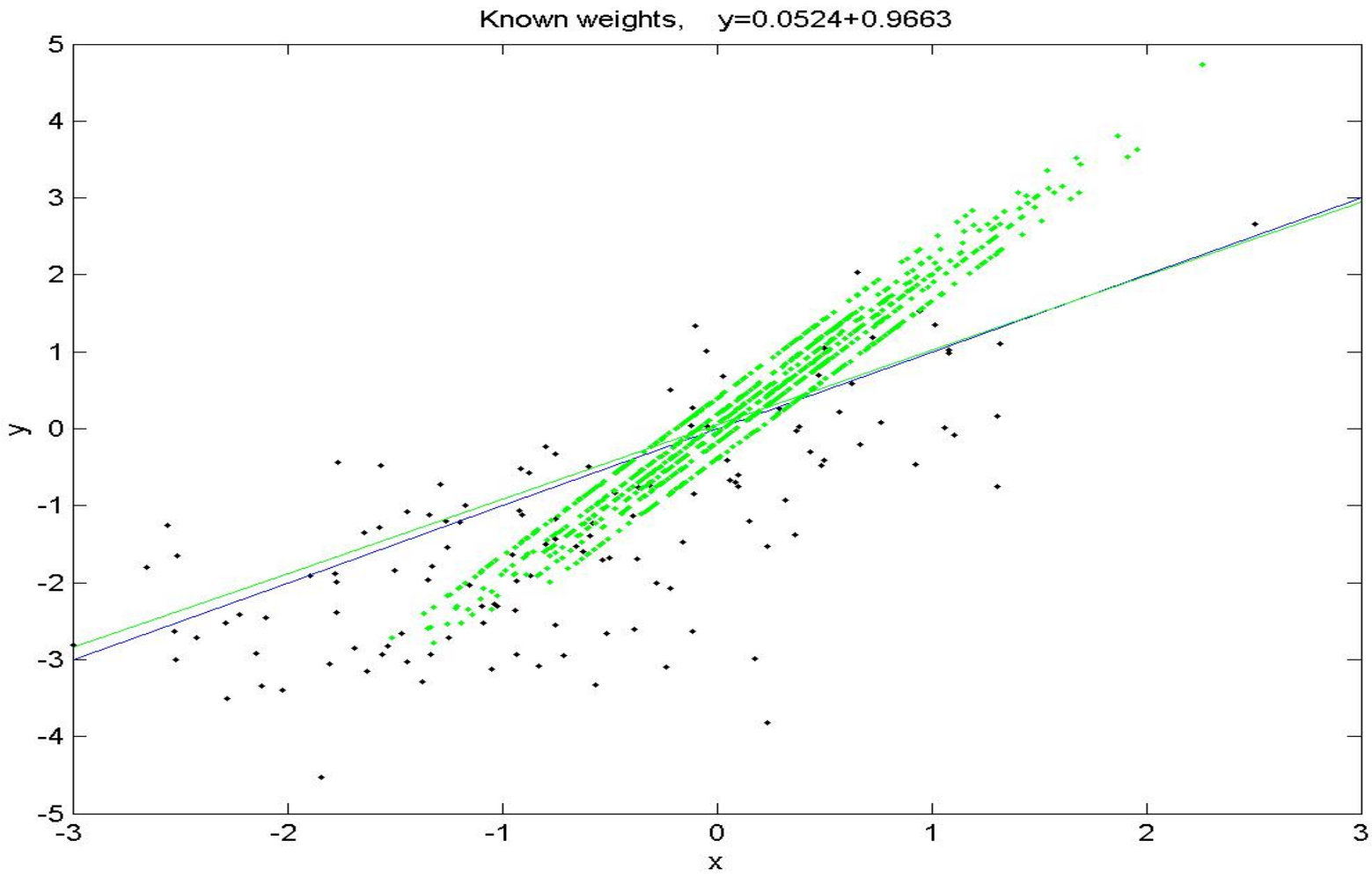


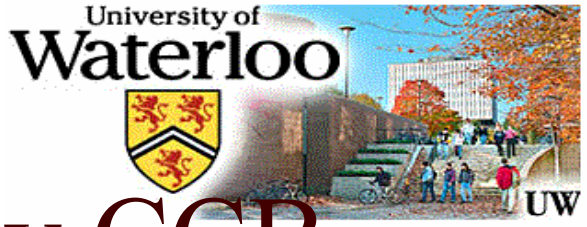
The simulations show…

- CCB has substantial bias especially for low cor(X,V). Why?
- CCB and Profile 1 puts weight only on *x*-values with corresponding *v* equal.
- Other methods requiring a model for *f(x|v)* suffer less from this problem e.g. allow weights on all validated *x* (profile 2&3)
- Problem with support for *f(x|v)* for each *v*

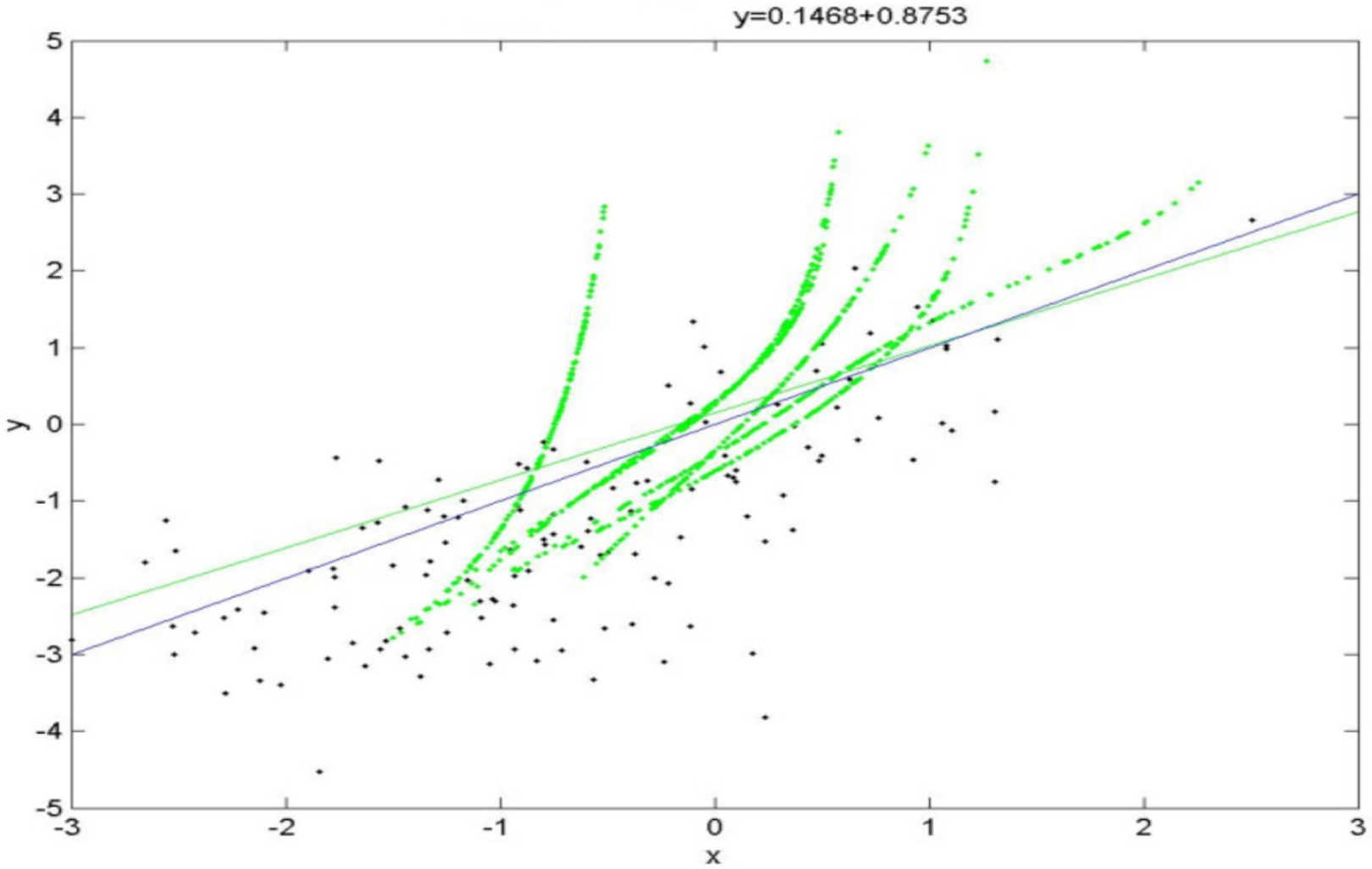


Example: 6 categories, missing values imputed using correct distribution $\rho = 0.25$

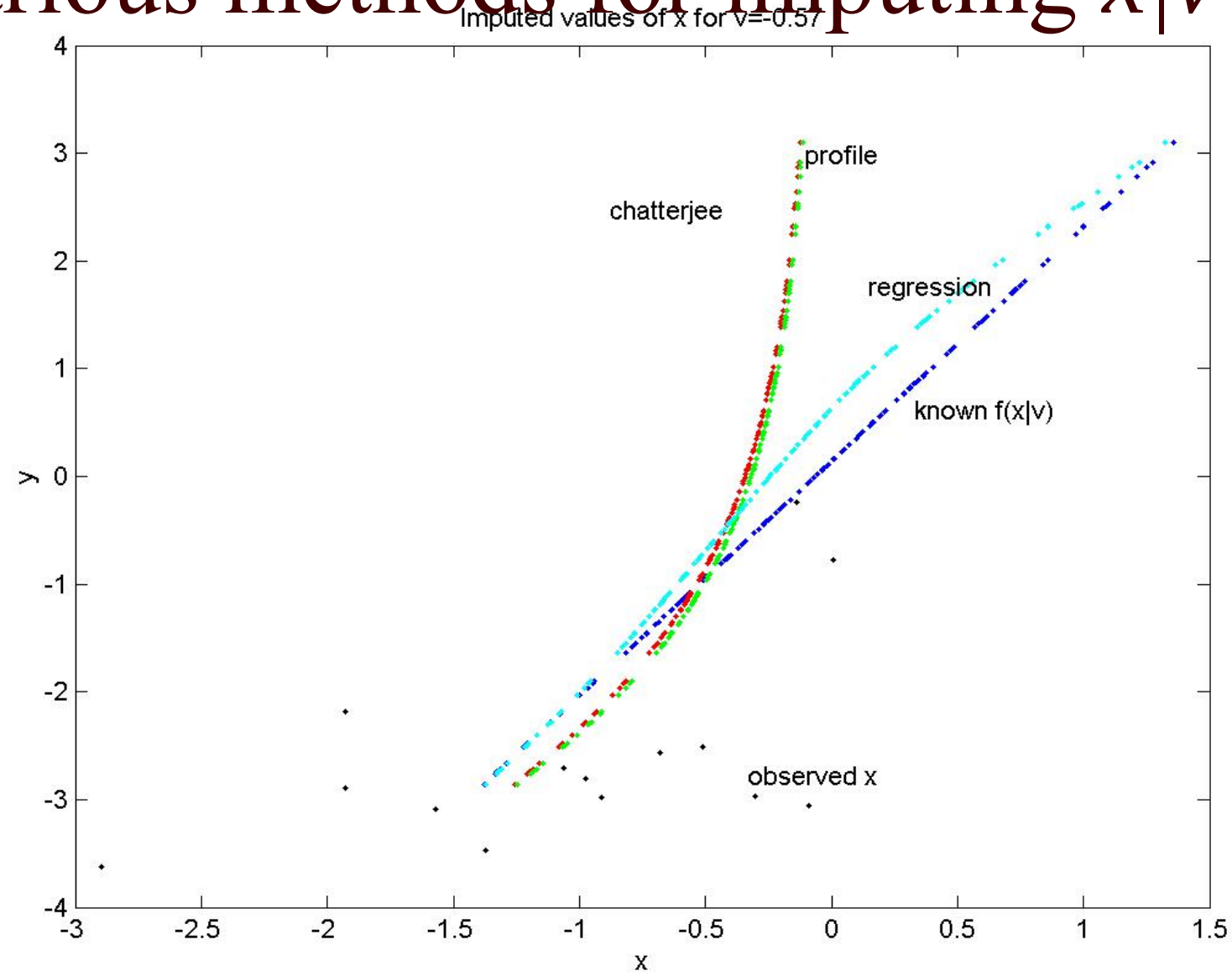




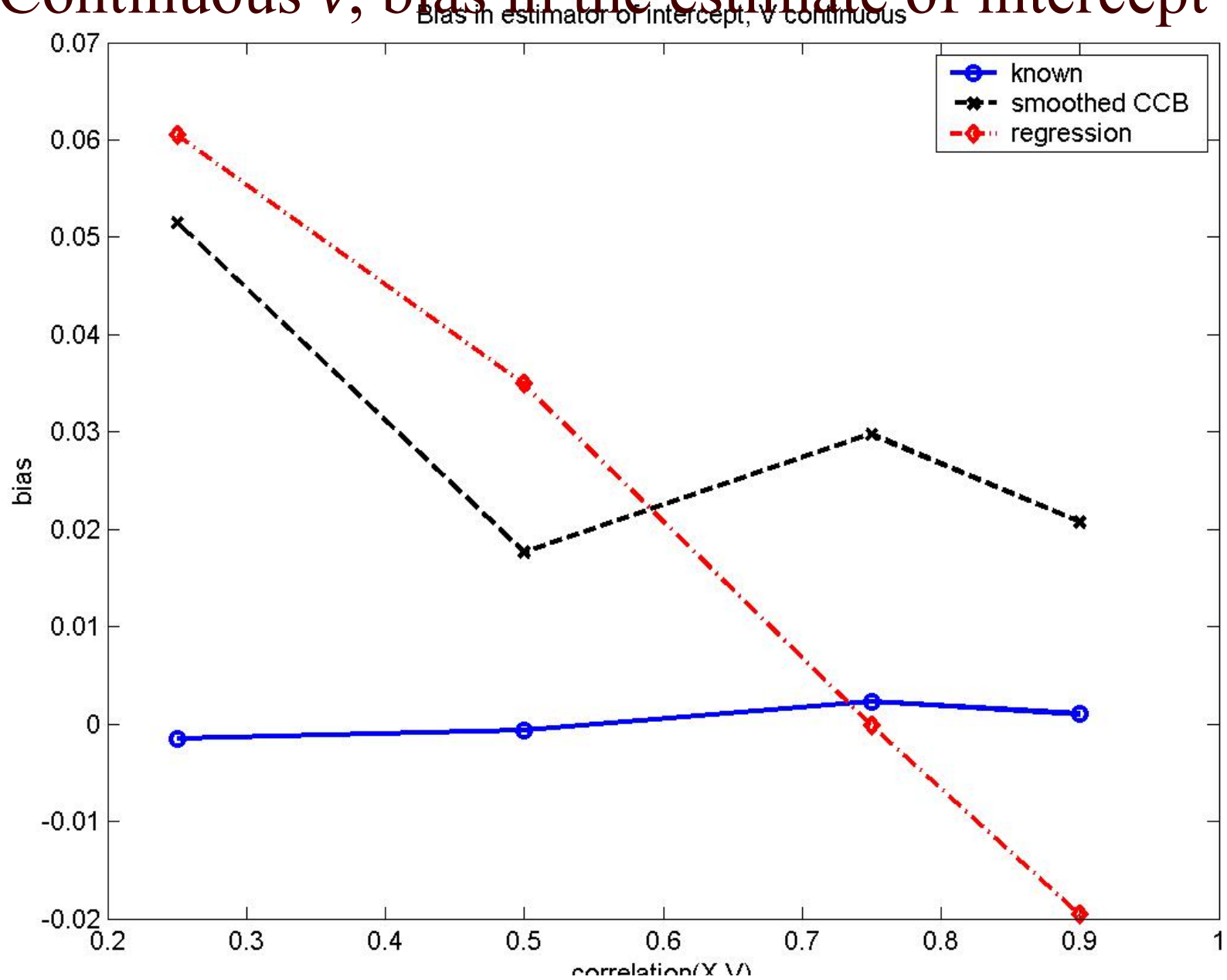
Example: Values imputed by CCB



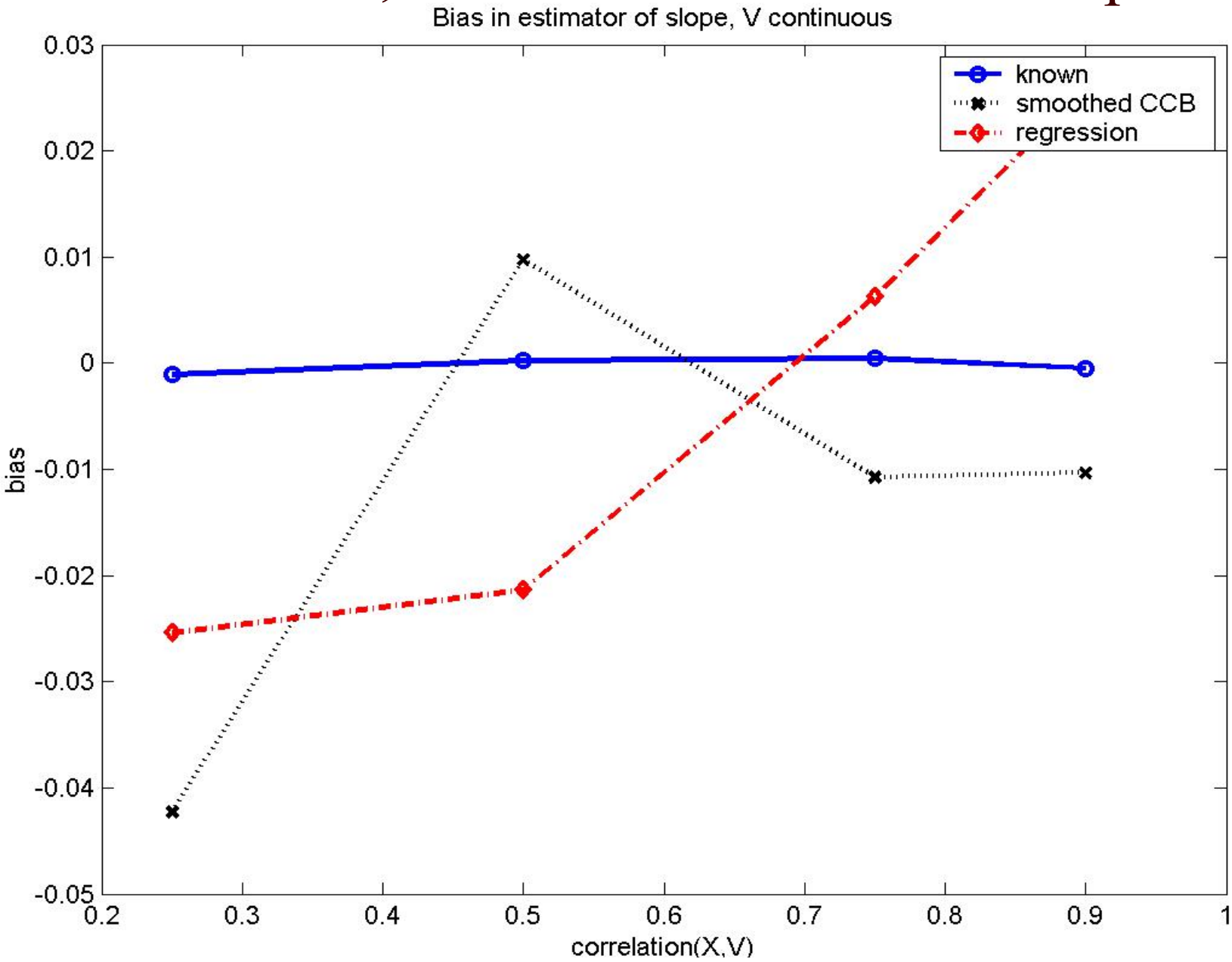
Various methods for imputing $x|v$



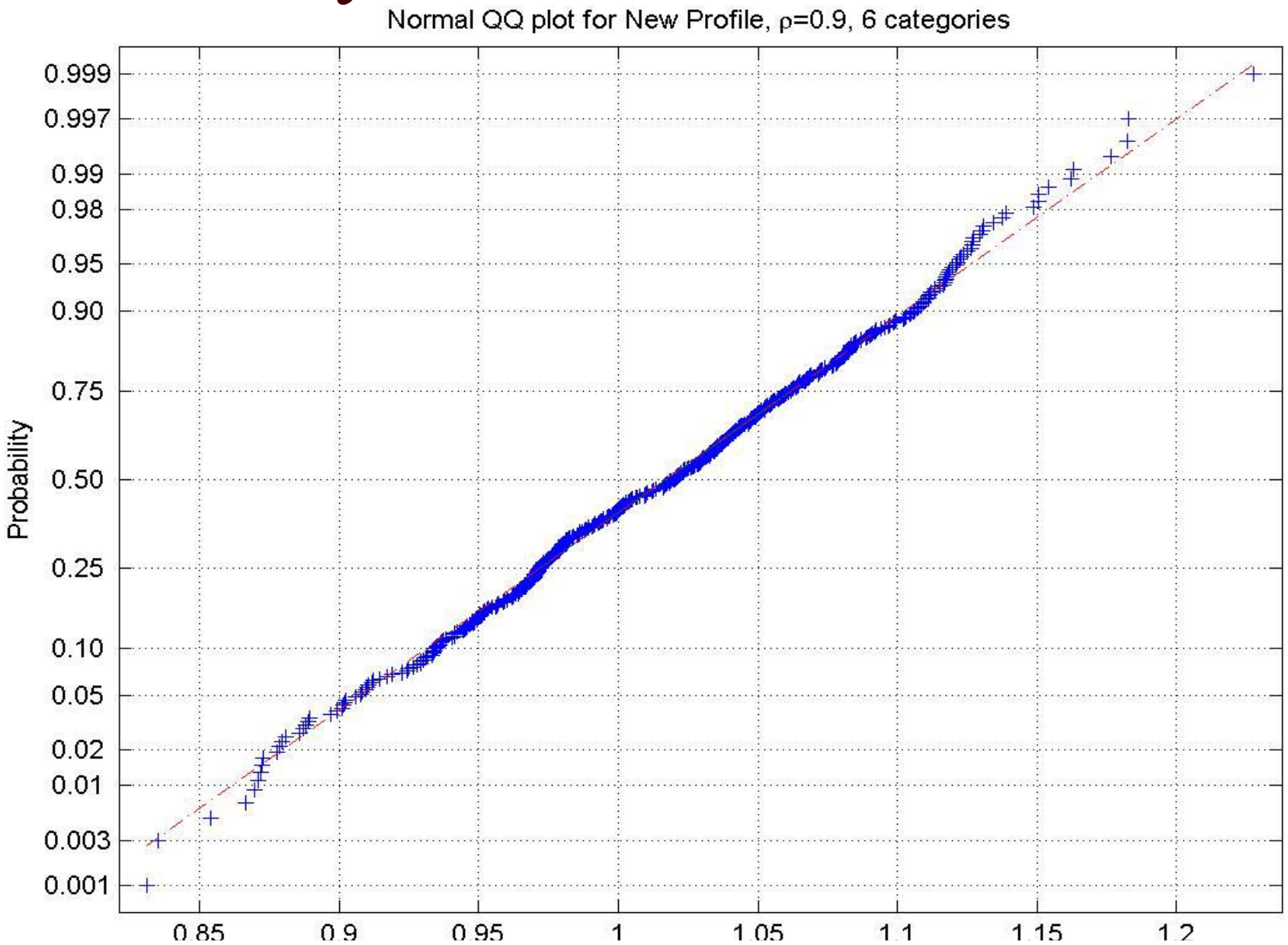
Continuous *v*, bias in the estimate of intercept



Continuous *v*, Bias in the estimator of slope



Normality. No Problem.



CONCLUSIONS: The price of ignorance

- When there is high correlation between *x* and *v*, *most* methods work reasonably well.
- Significant bias occurs in most estimators, especially for low *cor(X,V)*, especially CCB and profile if we restrict its support.
- For low *cor(X,V)*, profiles and regression weights permit reasonable efficiency, low bias. Profile robust against model failure.
- Pooling categories of *V* or smoothing over *V* reduces bias and variance.
- *High price paid (bias/variance) in not using knowledge of the conditional distribution of x\v. Is this because of additional nuisance parameters? Adjust profile likelihood?*
- *Bayesian: Gibbs sampling?*