Product selection in the presence of selectivity bias

I-Ding Wu and David J Hand

Department of Mathematics, Imperial College London

August/2007

Outline

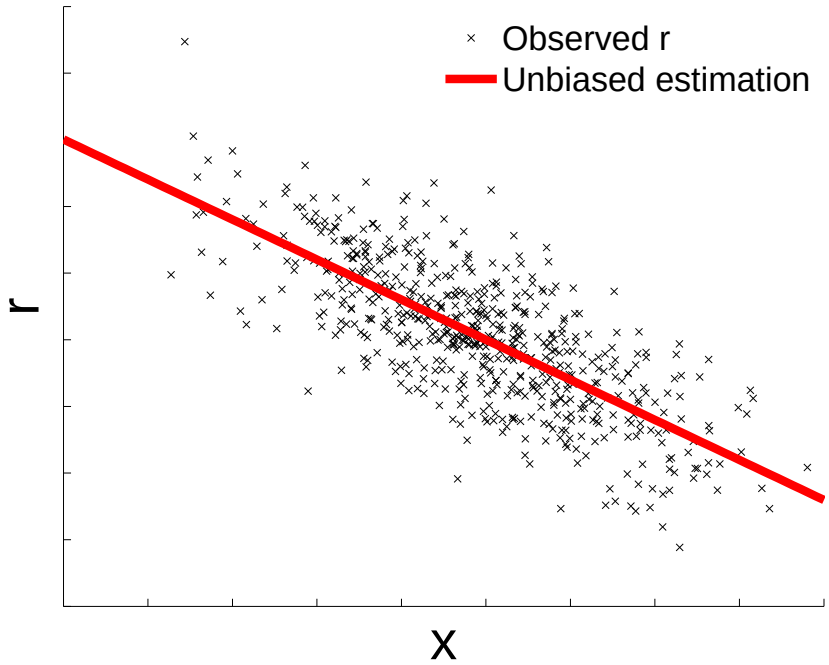
When does the bias occur?

Estimate the correlation between r and x from a historical sample.

E.g. Estimate the correlation between the profit (r) generated from a customer and the account balance (x).

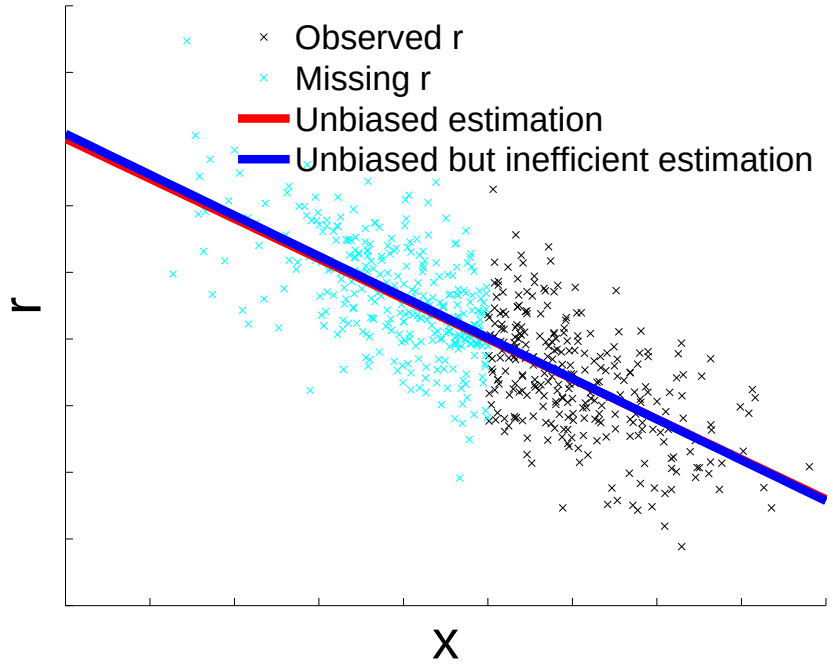
When r is completely observed

E.g. The previous decision maker granted credit to all customers without selection.



When r_i is observed only if $\gamma(x_i) = x_i\beta_{\gamma,x} > \tau$

E.g. The previous decision maker only granted credit to customers whose account balance (x) exceeded a certain amount.

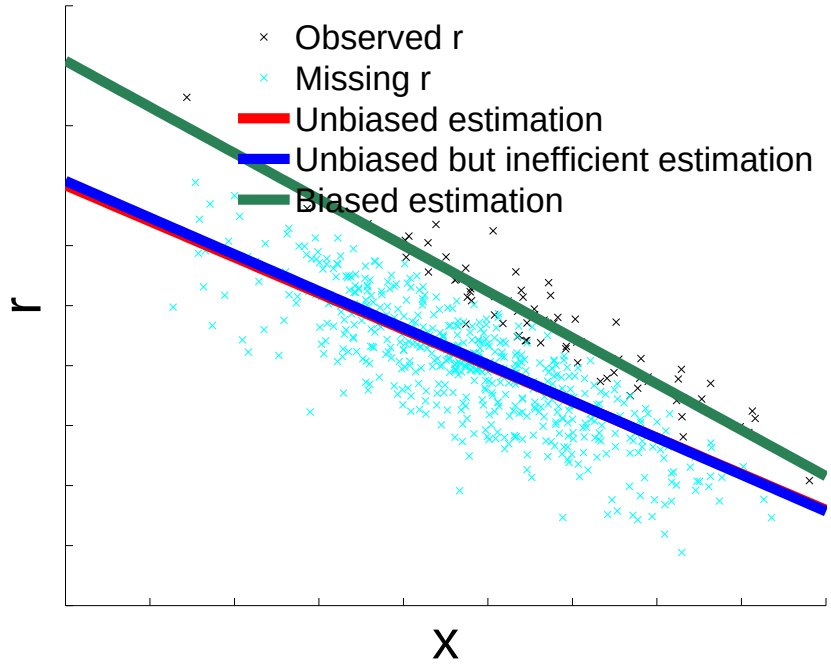


When r_i is observed only if $g(x_i, z_i) = -t + x_i\beta_{g,x} + z_i > 0$

E.g. Previous decision maker only granted credit to customers whose account balance (x) plus annual income (z) exceeded a certain amount.

Constraints:

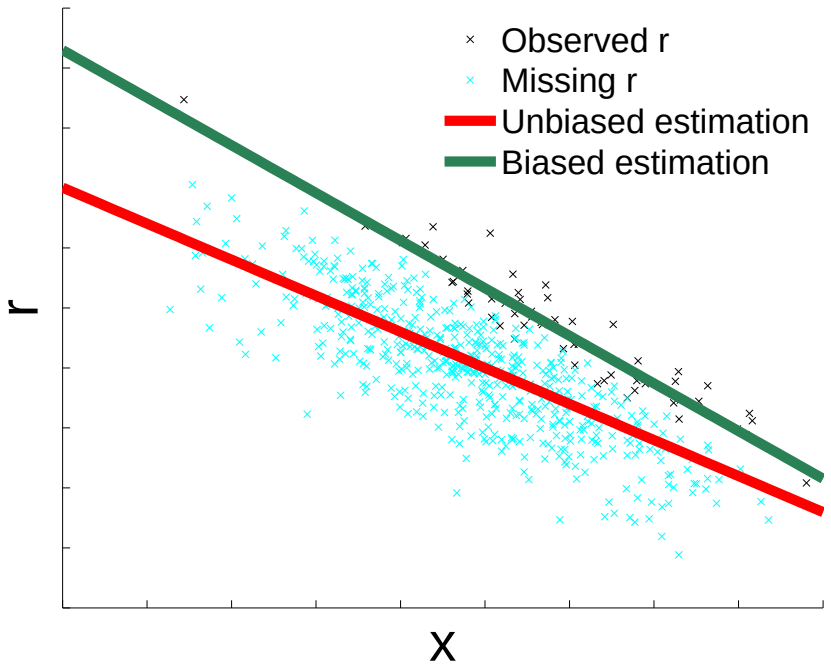
- z is correlated to r
- the information regarding z, $\beta_{g,x}$, and t is lost



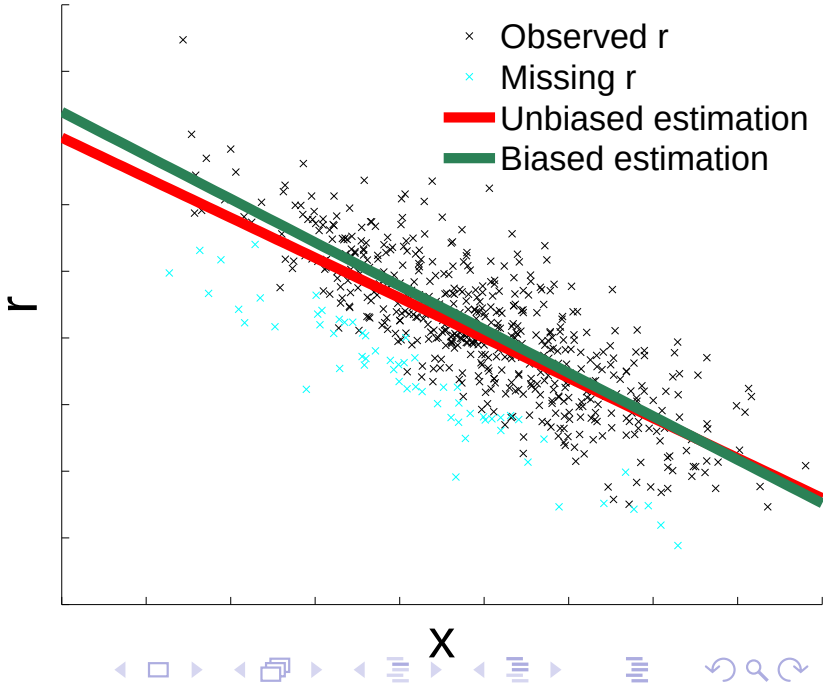
When does one adjust such bias?

Crook and Banasik: "when the rejection rate is not so large, the scope for improving a model parametrised only on those accepted appears to be very small".

If most r are missing:



If most r can be observed:



Wish to compare two responses

Consider "two" response variables: r_1 **and** r_0

E.g. r_1 can be the profit generated from a gold card, and r_0 be the profit generated from a classic card.

Wish to compare $E(r_1|X)$ **with** $E(r_0|X)$

E.g. Should we issue a gold card or a classic card to a customer with a certain account balance?

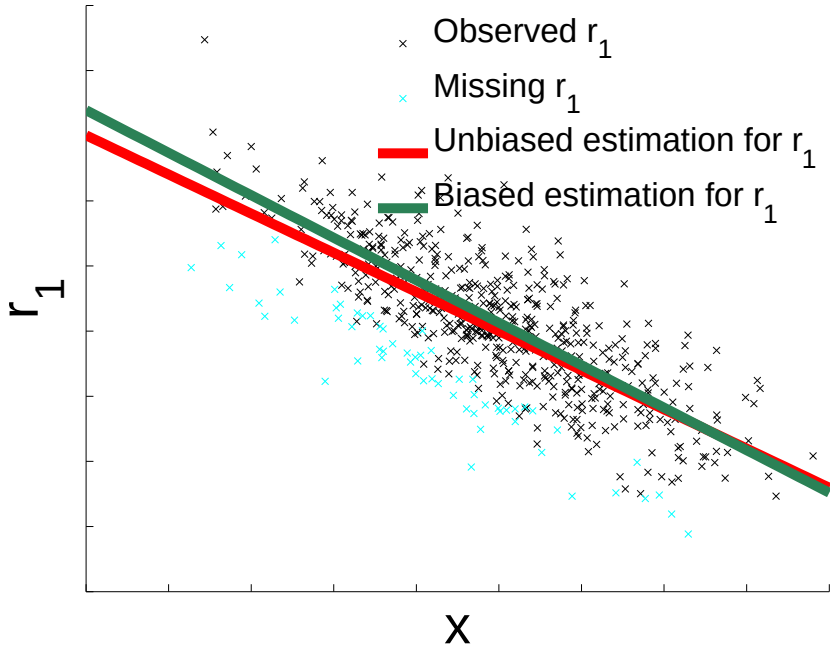
Constraints:

- if $g_i = -t + x_i\beta_{g,x} + z_i > 0$: r_{1i} is observed, r_{0i} is missing, set $a_i = 1$;
- if $g_i = -t + x_i\beta_{g,x} + z_i \leq 0$: r_{0i} is observed, r_{1i} is missing, set $a_i = 0$;
- $r_{1i} = \beta_{r_1} + x_i\beta_{r_1,x} + z_i\beta_{r_1,z} + \epsilon_{1i}$;
- $r_{0i} = \beta_{r_0} + x_i\beta_{r_0,x} + z_i\beta_{r_0,z} + \epsilon_{0i}$.

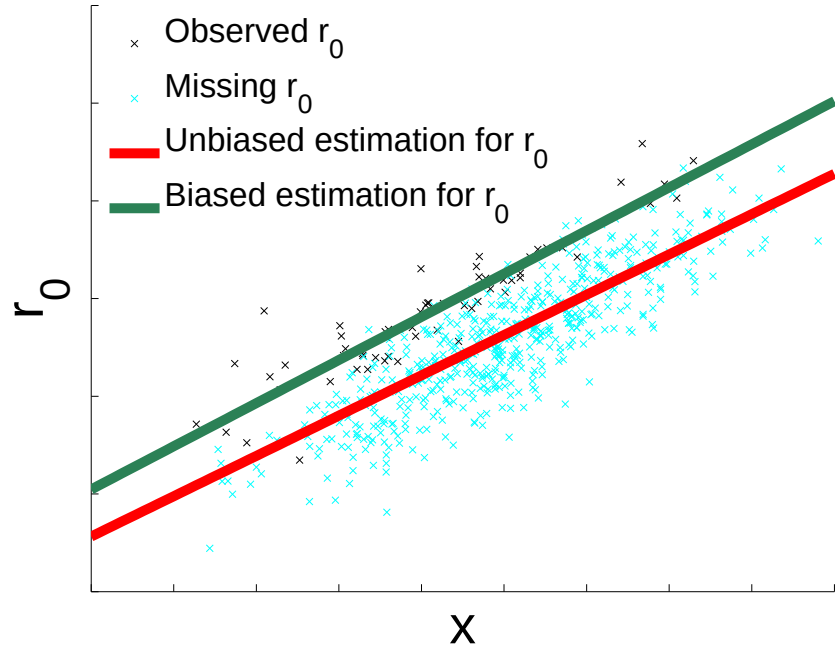
E.g. The previous decision maker only issue gold cards but not classic cards to customers whose account balance plus annual income exceeded a certain amount.

Wish to compare two responses

Good estimation for r_1:



Bad estimation for r_0:



Heckman's two-step method

The bias can be ignored if the values of z can be estimated.

- Step1: Estimate the missing z.
- Step2: Apply ordinary least squares regression.

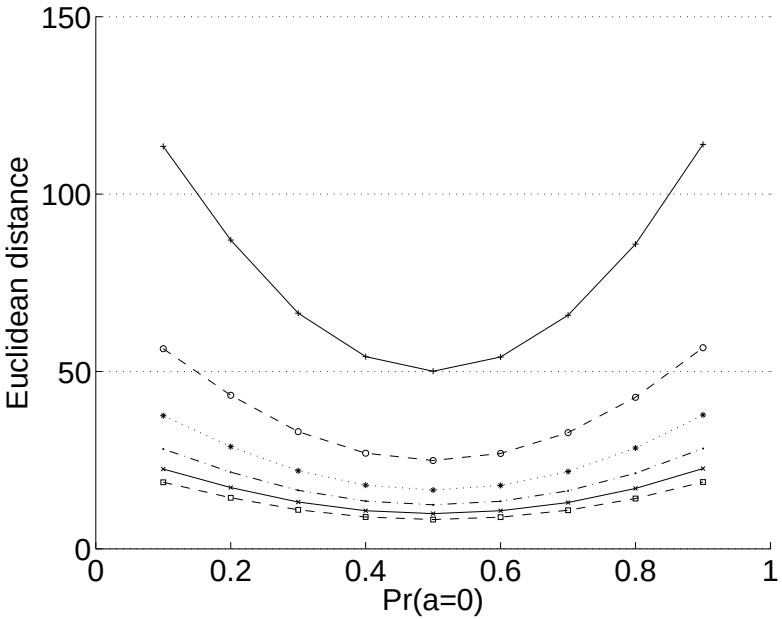
Heckman's two-step method

- **Step1:** Estimate the missing z.
  - Assume $z \sim normal(0, 1))$, and is independent of x.
  - Apply probit analysis to estimate parameters in $g = -t + x\beta_{g,x} + z$.
  - Obtain estimator $\hat{z}$ for z:
    - if $a_i = 1$, $\hat{z}_i = E(z|z > \hat{t} - x_i\hat{\beta}_{g,x})$
    - if $a_i = 0$, $\hat{z}_i = E(z|z \leq \hat{t} - x_i\hat{\beta}_{g,x})$
- **Step2:** Apply ordinary least squares regression.
  - Among $a = 1$, regress r_1 on x and $\hat{z}$ to estimate function of r_1.
  - Among $a = 0$, regress r_0 on x and $\hat{z}$ to estimate function of r_0.

Performance of Heckman's two-step method

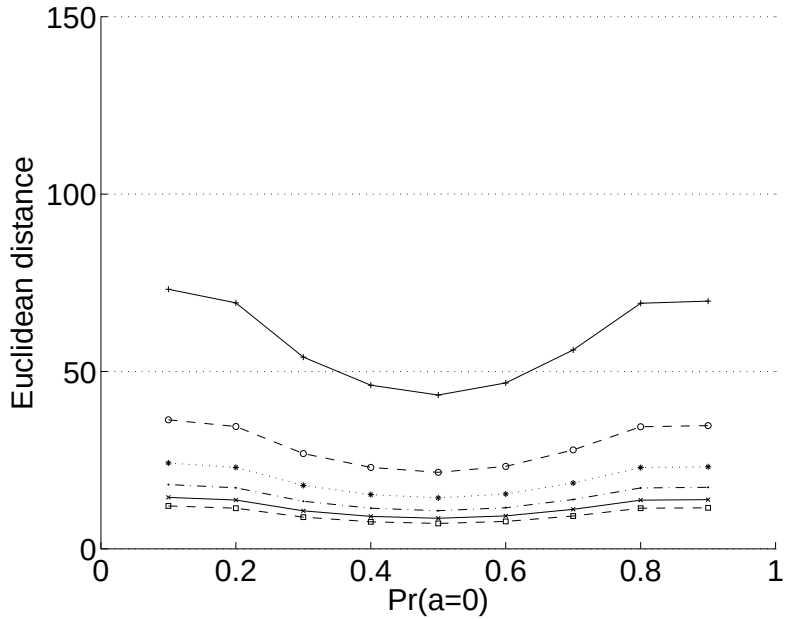
- $z \sim normal$

Insufficiency of unadjusted estimates:



(the larger the distance the more bias)

Improvement from Heckman's method:

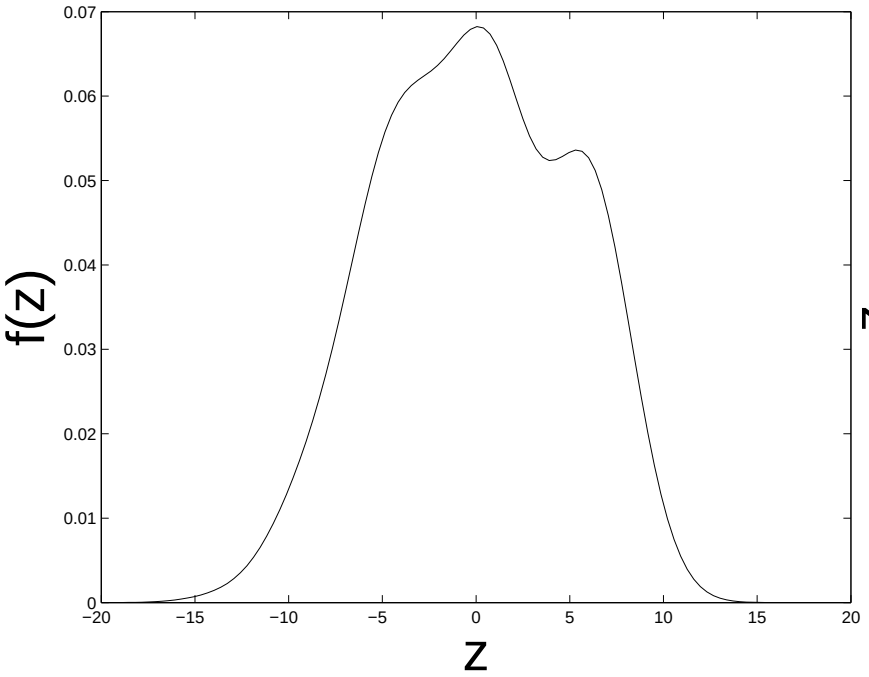


(the larger the distance the better improvement)

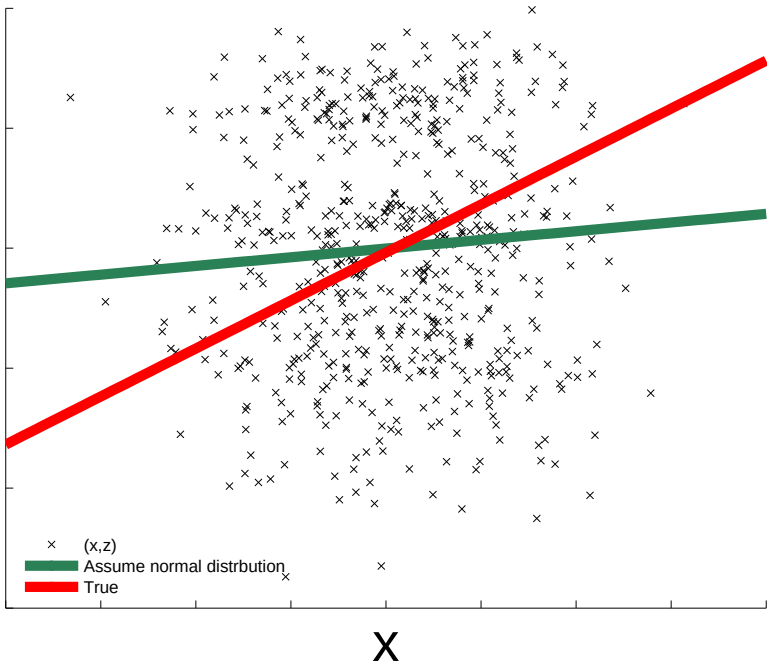
Disadvantage of Heckman's two-step method

- Assume z follows a normal distribution.

Density plot of z:



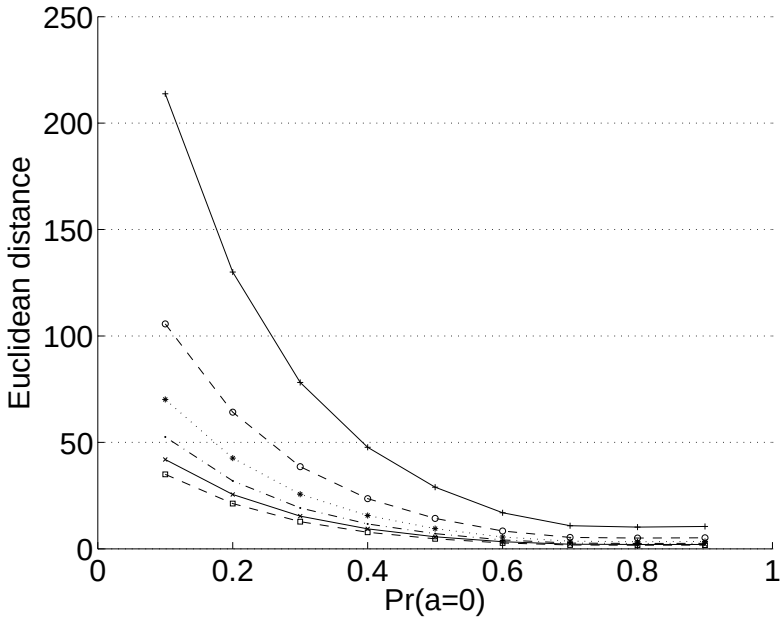
$\beta_{g,x}x + z - t = 0$:



Performance of Heckman's two-step method

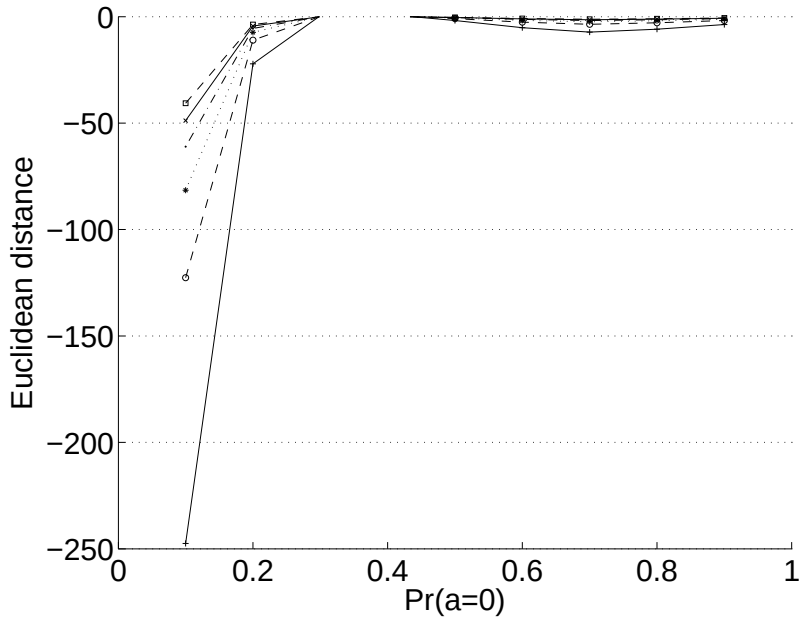
- $z \sim non - normal$

Insufficiency of unadjusted estimates:



(the larger the distance the more bias)

Improvement from Heckman's method:



(the larger the distance the better improvement)

Make use of empirical distribution

Argument: Making use of an empirical distribution, e.g. $f(\zeta)$, is better than assuming normal distribution.
Aim: Use $f(\zeta)$ instead of $normal(0, 1)$ to estimate $\beta_{g,x}$, t, and z.

$$\begin{array}{lllll} g & = & -t+ & x\ \beta_{g,x} & + z \\ ? & = & ? & ?\ x & + ? \end{array}$$

Suggestion: Consider z as a regression residual (shift the available empirical distribution to have zero mean), and iterate the following algorithm until converge.

$g = -t + x\beta_{g,x} + z$

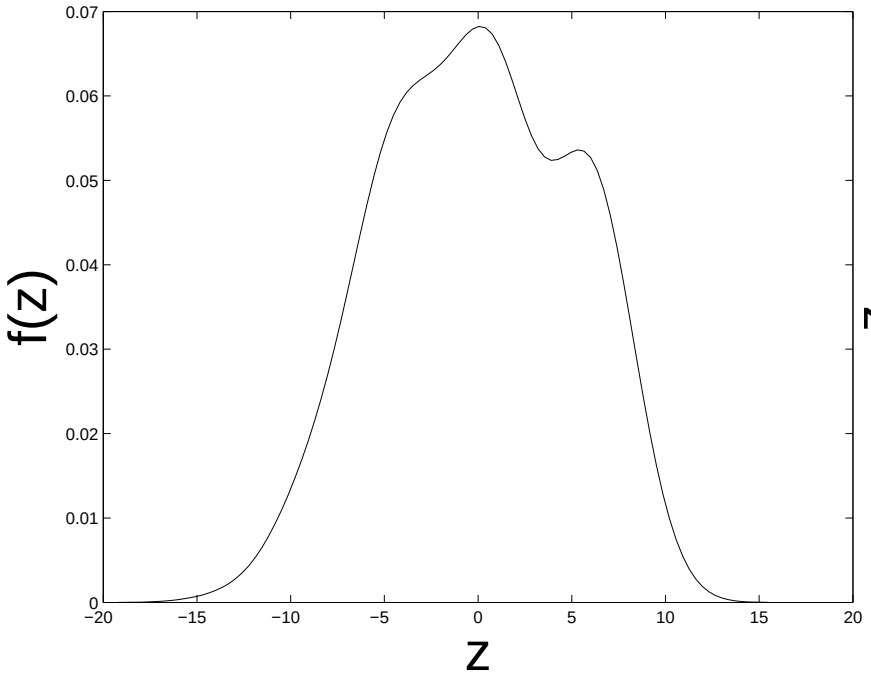
- $$\hat{g} = ? + x? + ?$$

  Regress g on x only to obtain $\hat{t}$ and $\hat{\beta}_{g,x}$.
- if
  $$-\hat{t} + x\hat{\beta}_{g,x} + ? > 0$$
  $$\hat{z}_i = E(\zeta | \zeta > \hat{t} - x_i\hat{\beta}_{g,x}) = \int_{\hat{t}-x_i\hat{\beta}_{g,x}}^{\infty} \zeta f(\zeta) d\zeta$$
- if ‘
  $$-\hat{t} + x\hat{\beta_{g,x}} + ? \leq 0.$$
  $$\hat{z}_i = E(\zeta | \zeta \leq \hat{t} - x_i\hat{\beta}_{g,x}) = \int_{-\infty}^{\hat{t}-x_i\hat{\beta}_{g,x}} \zeta f(\zeta) d\zeta$$
- $$? = -\hat{t} + x\hat{\beta}_{g,x} + \hat{z}$$
  $$\hat{g}_i = -\hat{t} + x_i\hat{\beta}_{g,x} + \hat{z}_i$$

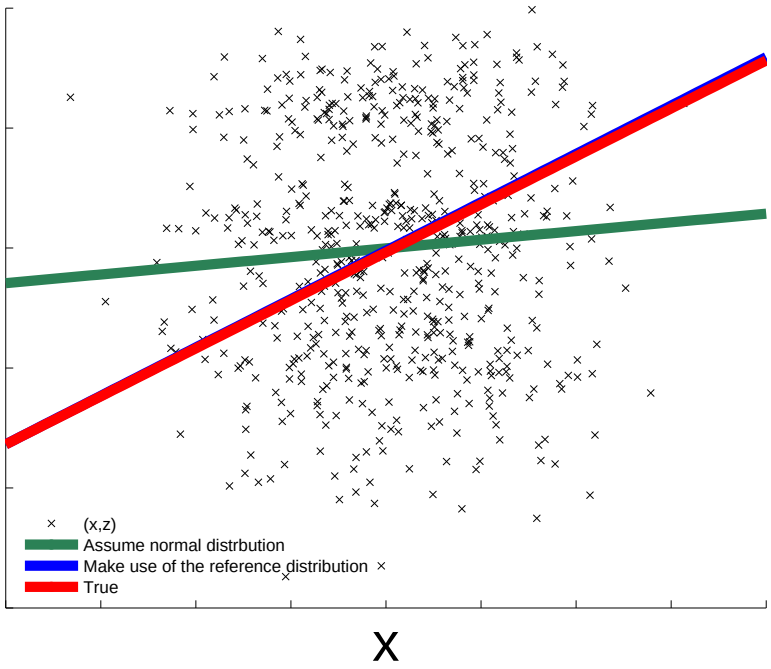
Make use of empirical distribution

- Using empirical distribution instead of assuming normal distribution.

Density plot of z:



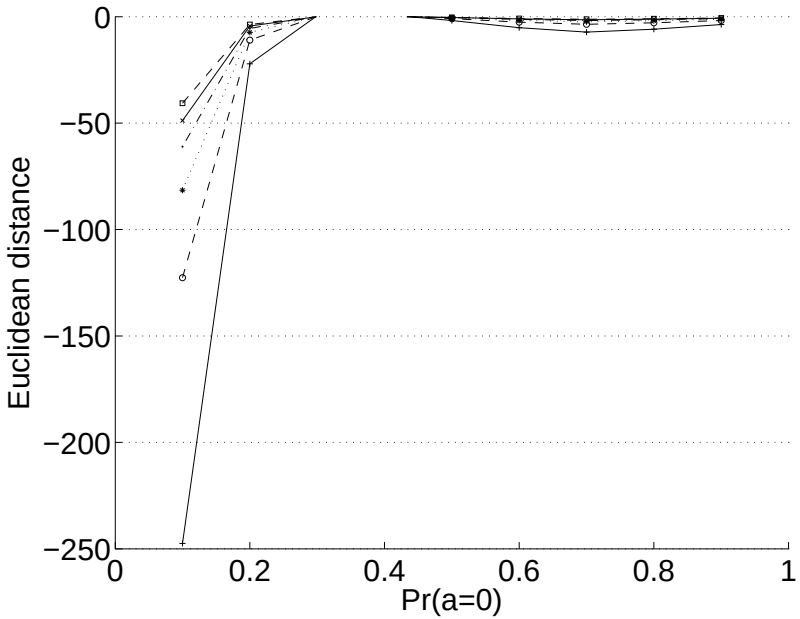
$\beta_{g,x}x' + z - t = 0$:



Make use of empirical distribution

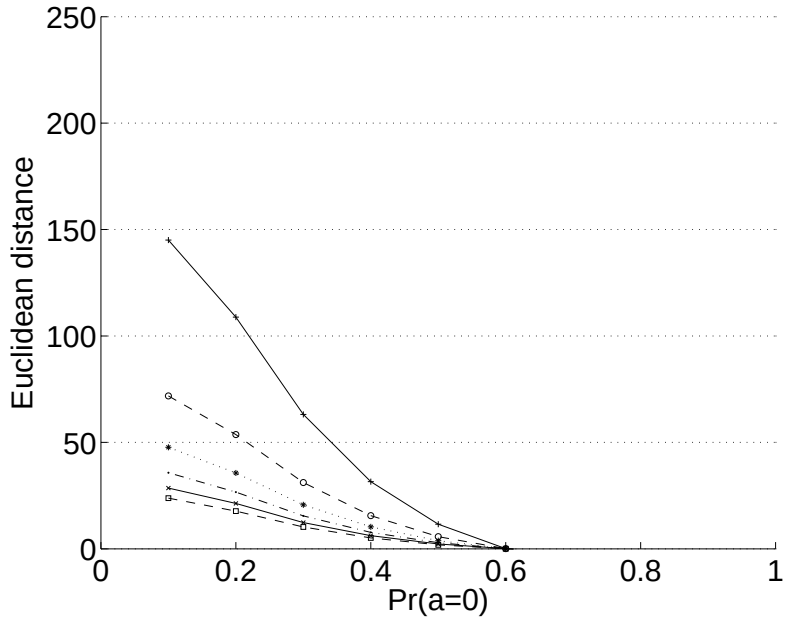
- $z \sim non - normal$

Improvement from Heckman's method:



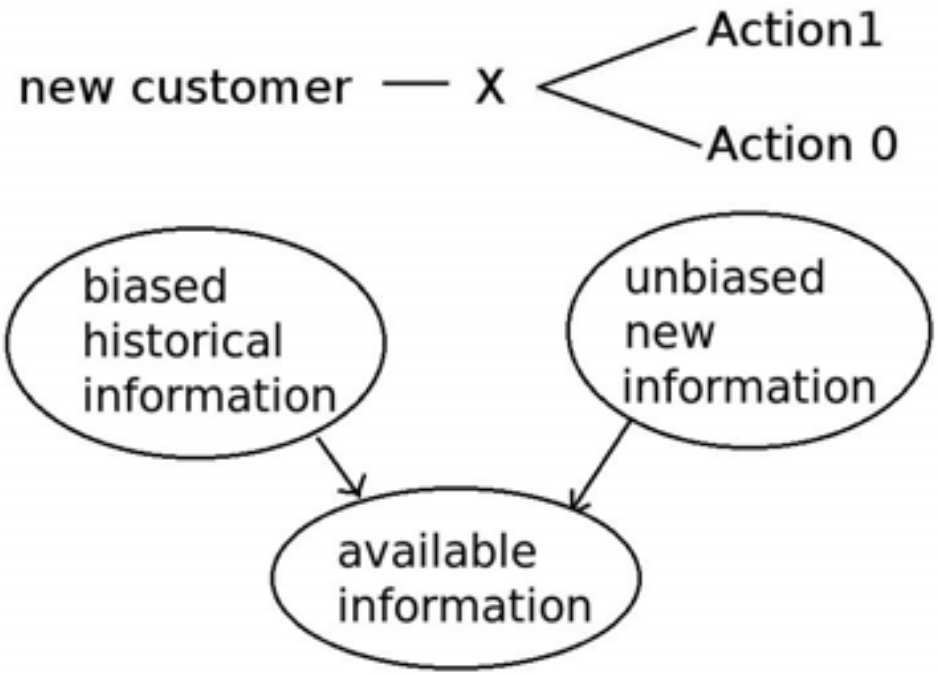
(the larger the distance the better improvement)

Improvement from using empirical distribution:



(the larger the distance the better improvement)

Make use of updated sample



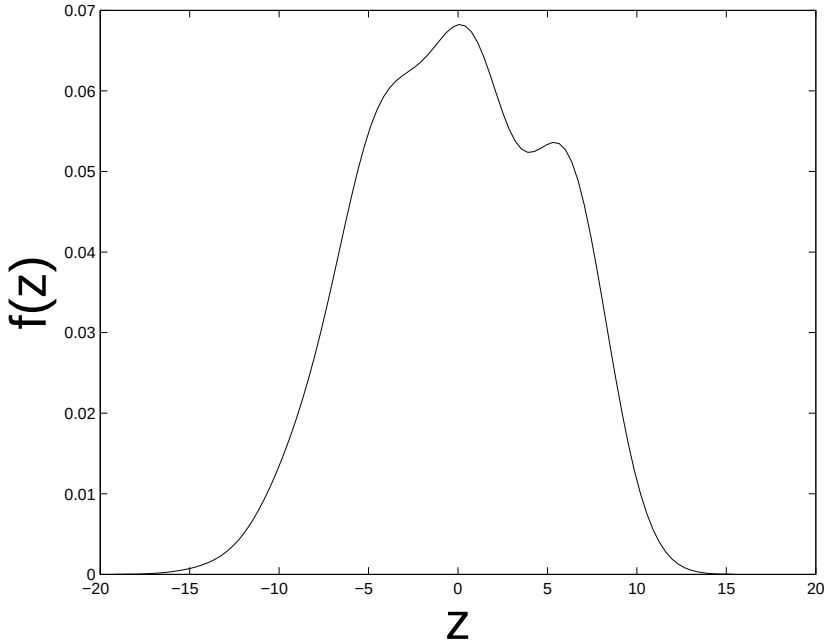
- Obtain approximation to the empirical distribution
- Adaptive method

Obtain approximation to the empirical distribution

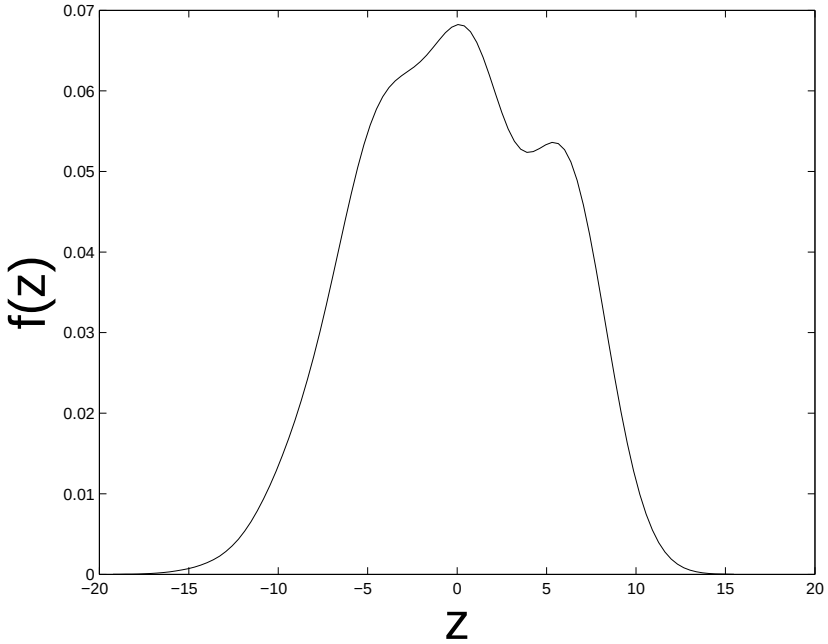
$$r_{ai} = x_i \beta_{r_a,x} + z_i \beta_{r_a,z} + \epsilon_{ai}$$

regress r_a on $x \longrightarrow$ residual: $e_i = z_i \beta_{r_a,z} + \epsilon_{ai}$
$\longrightarrow$ take $f(e)$ as $f(z)$

Density plot of z:



Density plot of residuals



Obtain approximation to the empirical distribution

Advantage: When few new customers enter, and most of them were assigned the same action, e.g. Action 1,

- Use new sample only:
  Estimation for function of r_1 might be satisfying but not that for function of r_0.
- Use both new sample and historical samples:
  Obtain approximation to the empirical distribution, estimations for both functions of r_0 and r_1 are likely to be satisfying.

Adaptive method

Regression model: $E(r_a|x) = x\beta_{r_a,x}$

Unbiased estimators: $\hat{\beta}_{r_a,x} = [X_a' X_a]^{-1} [X_a' R_a]$

where X_a and R_a include the value of x and r_a from customers who were assigned Action a.

Hist: X_a^H, R_a^H. New: X_a^N, R_a^N.

(X_a^N and R_a^N are changed whenever the number of customers in the new sample increase.)

a new customer enters

$\downarrow$

fixed X_a^H, R_a^H + updated X_a^N, R_a^N

$\downarrow$

update $\hat{\beta}_{r_a,x}$

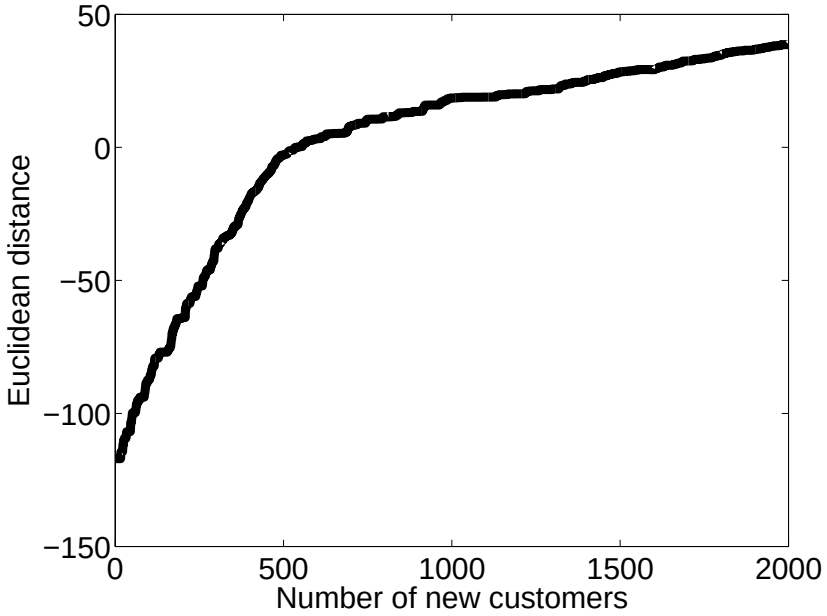
Adaptive method

$$\begin{aligned}\hat{\beta}_{r_a,x} &= \left[X_a^{H\prime}X_a^H + X_a^{N\prime}X_a^N\right]^{-1}\left[X_a^{H\prime}R_a^H + X_a^{N\prime}R_a^N\right] \\ &= \left[X_a^{H\prime}X_a^H + X_a^{N\prime}X_a^N\right]^{-1}\left[X_a^{H\prime}X_a^H\hat{\beta}^H_{r_a,x} + X_a^{N\prime}R_a^N\right] \\ &= \left[V^{-1} + X_a^{N\prime}X_a^N\right]^{-1}\left[V^{-1}\hat{\beta}^H_{r_a,x} + X_a^{N\prime}R_a^N\right]\end{aligned}$$

where $V = \sigma^{-2}Var(\beta^H_{r_a,x})$, $\sigma^2 = Var(R_a^H - X\hat{\beta}^H_{r_a,x})$.

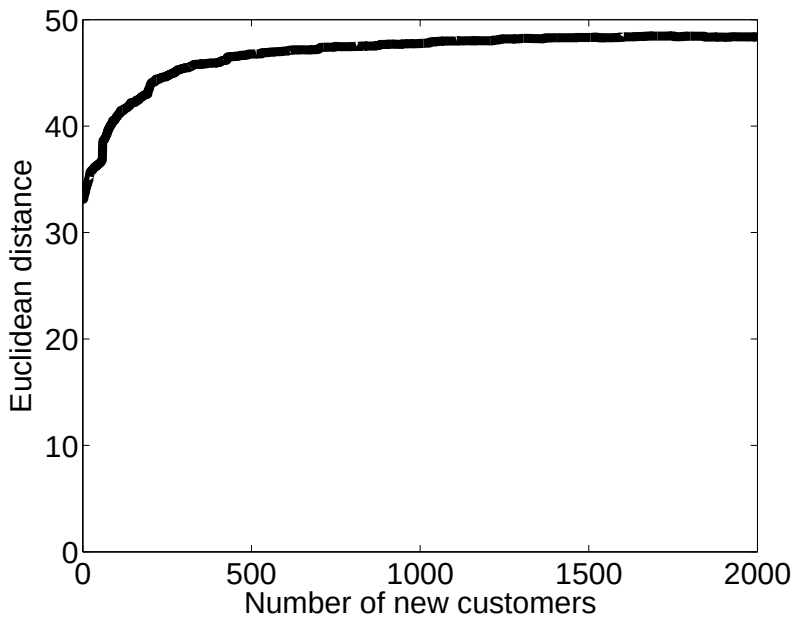
Performance of the adaptive model

Started from Heckman's two-step method:



(the larger the distance the better improvement)

Started from the method that uses empirical distribution:



(the larger the distance the better improvement)

Conclusion

- When considering more than one response variable, an adjustment is worthwhile.
- Existing methods are not sufficient because of the distribution assumption.
- It is indeed helpful to make use of empirical distribution.
- Making use of new information is worthwhile.

Conclusion

Thank you.