Urvalsmetoder och Estimation 3

Sampling and Estimation 3

2008-11-06

3. Regression type estimators

3.1 Estimation using auxiliary variables

- A population U is given
- An auxiliary variable: X_i; i € U is known
- Study variable Y_i unknown
- Take a sample S, observe Y_i; i € S
- Using Y from the sample and X from the population try to find a good estimator

Some methods to use auxiliary information

- We will now concentrate at the estimation phase
- Sample will be assumed taken and we will for simplicity assume SRS
  (Everything works for other sampling schemes too, but more complicated).
- Design-based approach
- Estimator of the population total

Some estimation techniques using auxiliary variables

- Difference estimators
- Ratio estimators
- Regression estimators
- Generalised regression estimators
- Prediction estimators

3.2 Difference estimators

- Suppose that we can make a prior guess of the unknown Y_i-value for all units using the auxiliary variables. Here we call the guess X_i.
- For example last years value or last years value plus inflation.
- Look at the difference $E_i = Y_i – X_i$
- Estimate the total difference T_e by t_e as in SRS
- Estimate the total T_y by $t_{yD} = T_x + t_e$, where T_x is known
- Estimate variance accordingly $Var^*(t_{yD}) = Var^*(t_e) = ((N-n)/N)\ \Sigma_S(E_i - \Sigma_S E_i/n)^2/(n-1)$

Are difference estimators good?

- The variance is $Var(t_{yD}) = Var(T_x + t_e) = Var(t_e) = Var(t_y - t_x) = Var(t_y) + Var(t_x) - 2Cov(t_y,t_x)$
- We gain if: $2Cov(t_y,t_x) > Var(t_x)$
- If the guess is good $Var(t_y) \sim Var(t_x)$. Then we gain if $\rho(t_y,t_x) = \rho(Y,X) > 1/2$

(The reverse martingale property allows us to omit the correction for finite population)

- Otherwise if $\rho(Y,X) > \sigma_x/2\sigma_y$

Are difference estimators good?

- Not used so often as a basic approach
- Often used in secondary analyses and longitudinal approaches
- One problem is that it often needs small changes i.e. X has the same level as Y.
- Another problem is that there often are large variances in X
- Difference estimators are good as a building stones for other estimators

3.3 Ratio estimators

- In difference estimators we looked at the difference $T_Y = T_X + T_{Y-X}$ and estimated it by $t_{yD} = T_X + t_{y-x} = T_X + t_y - t_x$
- Here we look at the ratio instead $t_{yR} = T_X * t_y / t_x$
- Sensible only for positive variables

Are ratio estimators good?

- Works best if both the mean and the variance of Y given X increase linearly with X
- May be slightly biased.
  (Problem since t_y/t_x is a convex function in t_x)
- To compute the approximate bias and variance we need the following theorem
  (Gauss approximation)

If X and Y are random variables with variances and f is differentiable, then

$$Var(f(X,Y)) \approx$$

$$\left(\frac{\partial f(m_x,m_y)}{\partial x}\right)^2 Var(X) + \left(\frac{\partial f(m_x,m_y)}{\partial y}\right)^2 Var(Y) +$$

$$\left(\frac{\partial f(m_x,m_y)}{\partial x}\right)\left(\frac{\partial f(m_x,m_y)}{\partial y}\right) Cov(X,Y)$$

and

$$E(f(X,Y)) \approx f(m_x,m_y) + \left(\frac{\partial^2 f(m_x,m_y)}{\partial x^2}\right) Var(X) +$$

$$\left(\frac{\partial^2 f(m_x,m_y)}{\partial y^2}\right) Var(Y) + \left(\frac{\partial^2 f(m_x,m_y)}{\partial x \partial y}\right) Cov(X,Y)$$

We use the second expression on
$t_{yR} = T_X * t_y / t_x$ and get Bias(t_{yR}) =

$$\approx T_x\left(\frac{Var(t_x)2E(t_y)}{E^3(t_x)} + \frac{Cov(t_y,t_x)}{E^2(t_x)}\right)$$

$$\approx \frac{Nm_x}{n}\left(\frac{\sigma_x^2 m_y}{m_x^3} - 2\frac{\sigma_{xy}}{m_x^2}\right)\frac{N-n}{N}$$

The bias is thus of order N/n and may be large in particular if the variance coefficient of x ($Var^{1/2}(X)/E(X)$) is large.

- We use the second expression on $t_{yR} = T_X * t_y / t_x$ and get Var(t_{yR}) =

$$\approx T_x^2\left(\frac{Var(t_y)}{E^2(t_x)} + \frac{Var(t_x)E^2(t_y)}{E^4(t_x)} + \frac{Cov(t_y,t_x)E(t_y)}{E^3(t_x)}\right)$$

$$\approx \frac{N^2 m_y^2}{n}\left(\frac{\sigma_y^2}{m_y^2} + \frac{\sigma_x^2}{m_x^2} - 2\frac{\sigma_{xy}}{m_x m_y}\right)\left(\frac{N-n}{N}\right)$$

Eqvivalently and easier to remember:
RelVar(t_{yR}) ~ RelVar(t_y) + RelVar(t_x) - 2RelCov(t_y t_x)

Where RelVar stands for the relative variance or coefficient of variation and RelCov for relative covariance
(The Var for difference estimator is replaced by RelVar)

Variance estimator

- One may use the above espression for the variance and replace all unknown parameters by their estimates
  m_x by t_x/n; m_y by t_y/n; σ_y^2 by $\Sigma_s(y-t_y/n)^2/(n-1)$; σ_x^2 by $\Sigma_s(y-t_x/n)^2/(n-1)$; and σ_{xy}^2 by $\Sigma_s(y-t_y/n)(x-t_x/n)/(n-1)$:
- We will show another expression later
  $Var^*(t_{Y,R}) = N(N-n)/n\ \Sigma_S (E^*_i - \Sigma_S E^*_i/n)^2/(n-1)$
  with $E_i^* = Y_i - (t_y / t_x) X_i$).

Are ratio estimators good?

- Much more often used (than difference estimator).
- One gains if correlation is larger than ½ (if the relative variances (= variation coefficients) are the same)
- Multiplicative relations are more often encountered in practice than additive (All size dependent variables)
- Variance often increases with size, often linearly.

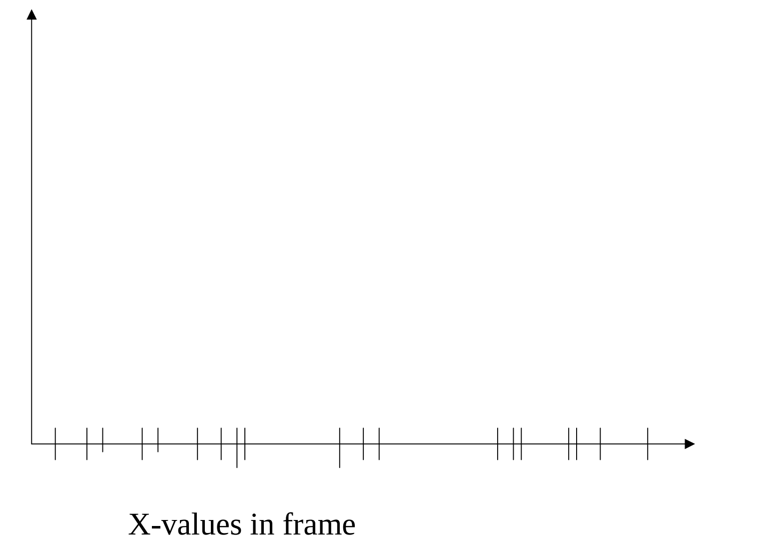
3.4 Regression estimator

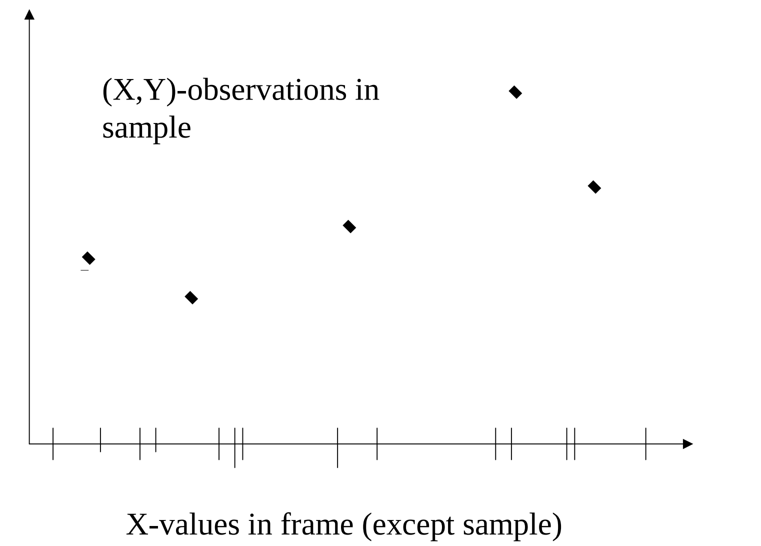
- Same set up: A population U and a known auxiliary variable: X_i; i € U
- Take a sample S, observe a study variable Y_i; i € S
- Idea: Using the sample, find a relation between X and Y. Try to use this relation and that X is known in the estimation phase

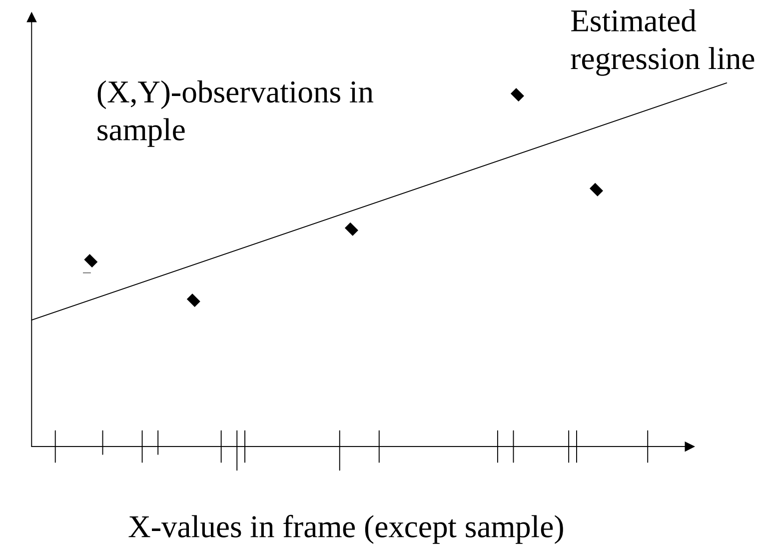
Regression estimator

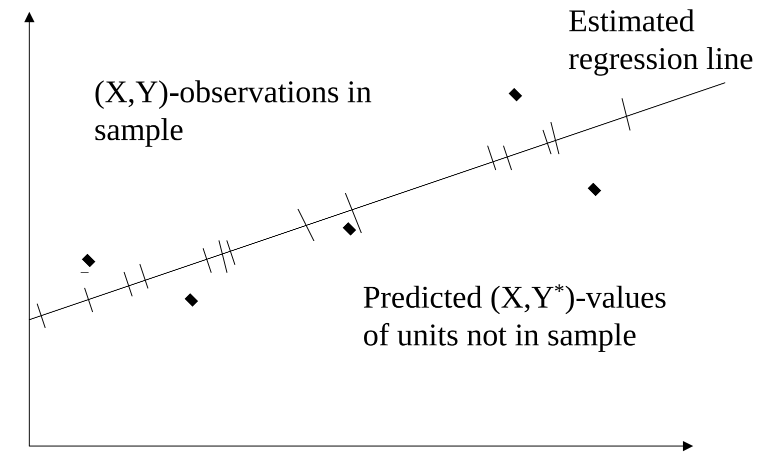
- Simple linear relation: Estimate a and b in the relation $Y_i = a + bX_i + \varepsilon_i$ by a^* and b^*
- Predict $Y_i^* = a^* + b^* X_i$; i € U-S
- Estimate population total by
  $T_Y^* = \Sigma_S Y_i + \Sigma_{U\text{-}S} Y_i^*$
  $= a^* N + b^* T_X$

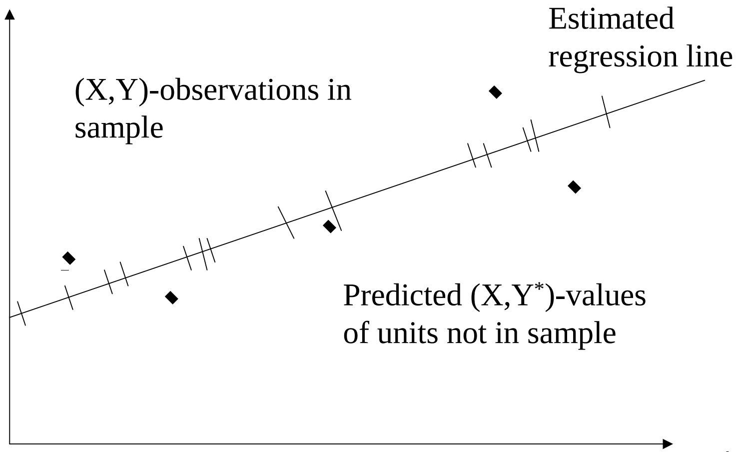
The second equality holds if a and b are estimated by Simple Linear Regression


X-values in frame


(X,Y)-observations in sample
X-values in frame (except sample)


Estimated regression line
(X,Y)-observations in sample
X-values in frame (except sample)


Estimated regression line
(X,Y)-observations in sample
Predicted (X,Y*)-values of units not in sample



Predict the total by summing all Y and Y^*-values in population

Regression estimator Further comments

- Is it sensible with Simple Linear Regression? Parameters can be estimated in other ways, and the method works then too.
- One method is the Asymptotically Optimal Regression Estimator. This approach minimises the asymptotic error variance (Montanari, (1987), ISR, 55). Based on the estimated covariance matrix of (t_y,t_x).
  $t_{y,Mont} = t_{y,} + (Cov^*(t_y,t_x)/Var^*(t_x)\ (T_x - t_x)$
- But usually only marginally better and may be much worse for small sample sizes (same for SRS)

Variance and variance estimation

- Suppose first we know a and b
- Write $E_i = Y_i - (a + b X_i)$
- Then $T_Y^* = a N + b T_X + (N/n) \Sigma_S E_i$
- In that case the variance of the estimator is just the variance of $t_e = (N/n) \Sigma_S E_i$, the standard estimate from SRS with E instead of Y. (cf difference estimators)
- $Var^*(t_e) = N(N-n)/n\ \Sigma_U (E_i - \Sigma_U E_i/N)^2/(N-1)$
- And the variance estimator is
  $Var^*(t_e) = N(N-n)/n\ \Sigma_S (E_i - \Sigma_S E_i/n)^2/(n-1)$

Variance and variance estimator (cont.)

- The variance with unknown a and b is approximately the same
  $Var(t_{Y,reg}) \sim N(N-n)/n\ \Sigma_U (E_i - \Sigma_U E_i /N)^2/(N-1)$
- In variance estimation one usually just replaces a and b by their estimates (a^* and b^*).
- The variance estimator (with known a and b) was
  $Var^*(t_e) = N(N-n)/n\ \Sigma_S (E_i - \Sigma_S E_i/n)^2/(n-1)$
- The variance can be estimated by replacing $E_i = Y_i - (a + b X_i)$ by $E_i^* = Y_i - (a^* + b^* X_i)$ i.e.
- $Var^*(t_{Y,reg}) = N(N-n)/n\ \Sigma_S (E^*_i - \Sigma_S E^*_i/n)^2/(n-1)$

Are regression estimates good?

- Improves the asymptotic variance as soon as the correlation is different from zero. (compared to mean estimate, difference or ratio estimator. The decrease depends on the the ”explained variance”)
- May be worse for small sample sizes
- The procedure works even if the ”true” relation is not linear. The estimate may be slightly biased
  (Since the estimate b* and the mean t_x may be dependent or the regression estimator can’t be computed, if all sampled x-values have the same value. Extremely unlikely but enough to destroy exact unbiasedness)
- Care must be taken when the sample is taken with varying probabilities. (Not today’s topic). (The best line may be different in sample and the remaining population)

3.5 General regression estimates (GREG) 1

- Previously: only one (continuous) X-variable
- What is said can easily be generalised to several auxiliary variables using multiple linear regression
- X-values can be discrete, categorical (dummy-variables) or derived (e.g.: x^2 or $x_1 * x_2$)
- Everything holds with $E_i = Y_i - (a + \Sigma_j b_j^* X_{ij})$
  in the approximate variance estimation expression

Weighted general regression estimates

- Different weighting in the regression (optimally weights should be proportional to $Var(y_i|x_i)$. ”Model-assisted approach”)

Ordinary: b* =

$$\Sigma(y_i\text{-ybar})(x_i\text{-xbar}) / \Sigma(x_i\text{-xbar})^2$$

Weighted: b* =

$$\Sigma((y_i\text{-ybar})(x_i\text{-xbar})/Var(y_i|x_i)) / \Sigma((x_i\text{-xbar})^2\, Var(y_i|x_i))$$

if intercept is unknown

- Other weights are sometimes used
- Difference estimators can be thought of as regression estimators with known slope (equal to one)
- Ratio estimators can be thought of as regression estimators, where the intercept is known (0) and the weights are proportional to x_i
  (Holds for Poisson distribution and often for economic variables from e.g. firms)

Are general regression estimators good?

- Asymptotically never worse than simple estimators or regression estimators with a subset of the auxiliary variables
- With many auxiliary variables the random error increases due to estimation problems of the regression coefficients. (In particular if the auxiliary variables are irrelevant or only a weak relation to Y).
- The variance estimator underestimates the true variance
- Avoid regression estimator when there are ”outliers” in the population with a set of X-values which differs much from the others or influential outliers in the sample.

An alternative expression for the variance estimator of the ratio estimator

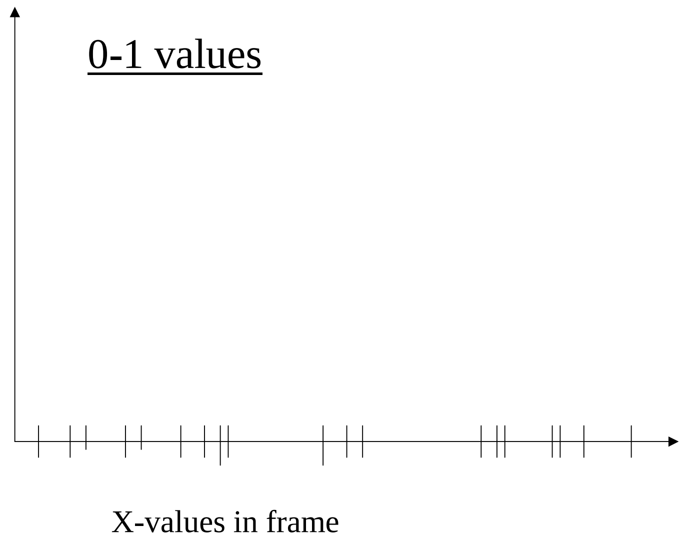
- As we said the ratio estimator can be viewed as a regression estimator with known intercept a=0.
- The variance estimator of the regr. estimator was $Var^*(t_{Y,reg}) = N(N-n)/n\ \Sigma_S (E^*_i - \Sigma_S E^*_i/n)^2/(n-1)$ with $E_i^* = Y_i - (a^* + b^* X_i)$
- Thus a variance estimator for the ratio estimator is $Var^*(t_{Y,reg}) = N(N-n)/n\ \Sigma_S (E^*_i - \Sigma_S E^*_i/n)^2/(n-1)$ with $E_i^* = Y_i - (0 + b^* X_i)$ where $b^* = t_y / t_x = \Sigma_S Y_i / \Sigma_S X_i$
(This is the altternative expression we gave above)

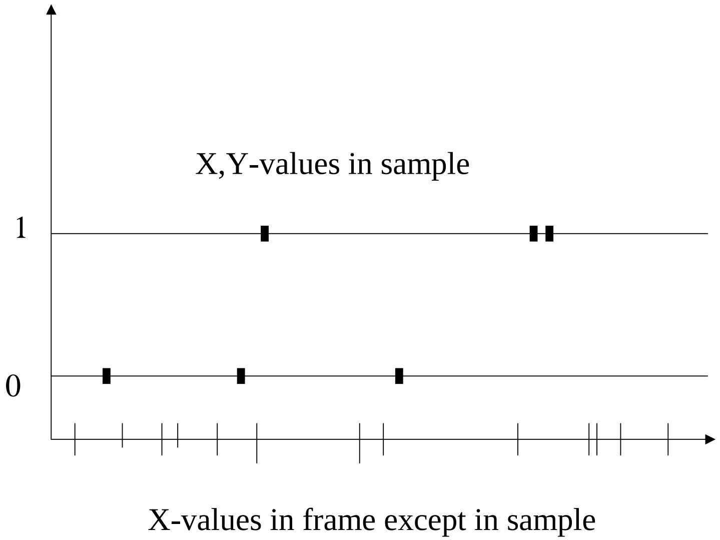
3.6 Prediction estimates

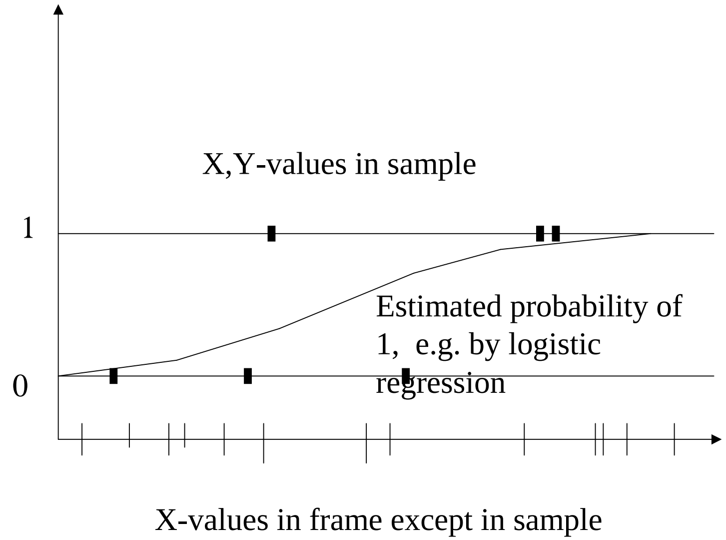
- In regression estimates we wrote $T_Y^* = \Sigma_S Y_i + \Sigma_{U-S} Y_i^*$ where Y_i^* was the ordinary best linear predictor (BLUE)
- One may try to use the same formula with other predictors.
- Sensible if the predictor is good. But a bias that is unimportant for predicting one single unit may become disastrous when summed over the whole population.

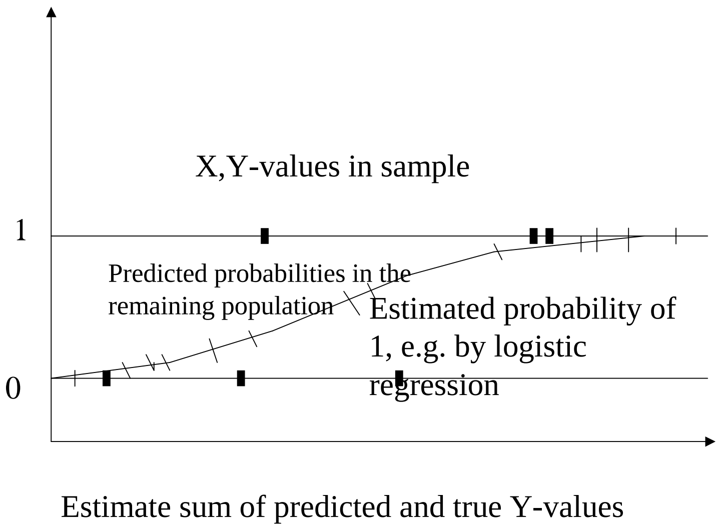
Prediction estimates

- Solution: Use the idea of a difference estimator. ”Suppose that we have some prior guess of the value for all units, X_i”
- Replace the guess X_i by Y_i^*, in the whole population
- Look at the difference $E_i = Y_i - Y_i^*$
- Estimate the total difference T_e by t_e
- Estimate the total T_y by $t_{yPred} = T_{Y*} + t_e$
- Estimate variance accordingly $Var^*(t_{yPred}) = Var^*(t_e)$ (asymptotically valid under mild restrictions)
- Example: 0-1-variables, proportions









- The prediction is not necessarily unbiased from a design-based perspective (in particular if the logistic model does not hold)
- They ought thus to be corrected, to be almost unbiased for any population and large samples $t_{Y,pred} = \Sigma_U Y_i^{*} + (N/n)\, \Sigma_S (Y_i - Y_i^{*})$
- Note that Y_i^{*} is computed also for units in the sample.
- If one believed in the logistic model and used modelbased inference the second term would not be needed.

- Prediction estimators can also be seen as approximate difference estimators
- Treat the prediction Y_i^* as the auxiliary variable (as we did for the regression estimator) and write
  $t_{Y,pred} = \Sigma_U Y_i^* + (N/n) \Sigma_S E_i$
  where $E_i^* = Y_i - Y_i^*$
- The variance of this can be estimated by
  $Var^*(t_{Y,pred}) = N(N-n)/n \, \Sigma_S (E^*_i - \Sigma_S E^*_i/n)^2/(n-1)$
  (exactly as for the regression and ratio estimators)

Are prediction estimators good?

Similar comments as for regression estimators:

- Asymptotically never worse than ordinary mean
- But if the sample is small and the estimated function uncertain one may loose efficiency.
- The better the model fit the better the estimator
- ...