

ECON 623
Answers to Problem Set 2
John Rust, Fall 2007

I. I spoke of *nonparametric* estimation methods in class.

1. If we have N IID observations, $\{X_1, \dots, X_N\}$ from a an unknown distribution $F(x)$ and we are interested in estimating its mean, $\mu = \int xF(dx)$, is the *sample mean*

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i \quad (1)$$

a non-parametric estimator of μ ?

Answer Yes, the sample mean is a non-parametric estimator since we do not have to make any assumptions about the functional form of the distribution $F(x)$ to implement this estimator, and as long as the distribution has finite mean and variance, the sample mean will be a consistent estimator of the true mean, $\mu = \int xF(dx)$.

2. Use the law of large numbers and the central limit theorem to determine the *asymptotic properties* of $\hat{\mu}$. Under what conditions is $\hat{\mu}$ a *consistent estimator* of μ ? What is the asymptotic distribution of $\hat{\mu}$?

Answer Under the *IID* assumption, we can apply the Law of Large Numbers and the Central Limit Theorem to determine the asymptotic properties of $\hat{\mu}$. The LLN implies that $\hat{\mu}$ converges to μ almost surely (i.e. with probability 1 as $N \rightarrow \infty$), and the CLT implies that

$$\sqrt{N}[\hat{\mu} - \mu] \implies N(0, \sigma^2), \quad (2)$$

where $\sigma^2 = \text{var}(X_i) = \int (x - \mu)^2 F(dx)$. These results will hold as long as μ and σ^2 are finite. (Check the assumptions required to prove the LLN and CLT).

3. Suppose that the true distribution $F(x)$ is a Normal distribution with mean μ and variance σ^2 . What is the *exact finite sample distribution* of $\hat{\mu}$?

Answer Since sums of normal random variables are normal, we have

$$\hat{\mu} \sim N(0, \sigma^2/N) \quad (3)$$

where the notation $\hat{\mu} \sim N(0, \sigma^2/N)$ is a shortcut for “distributed as as Normal random variable with mean zero and variance σ^2/N .”

4. Suppose $F(x)$ is a *Cauchy distribution*. Can you determine the exact finite sample distribution of $\hat{\mu}$ in this case? If so (or even if not) can you determine whether the sample mean $\hat{\mu}$ is a consistent estimator of μ in this case?

Answer We can use *characteristic functions* to show that for each N $\hat{\mu}$ has the same distribution as \tilde{X}_1 , i.e. it is the same Cauchy distribution as each of the observations. To see this, note that the density of a Cauchy distribution with location parameter μ and scale parameter σ is

$$f(x|\mu, \sigma) = \frac{1}{\pi\sigma \left(1 + \left[\frac{(x-\mu)}{\sigma}\right]^2\right)}, \quad (4)$$

and the characteristic function for a Cauchy distribution is

$$\begin{aligned} \Psi_{\tilde{X}_1}(t) &= E \{ \exp\{it\tilde{X}\} \} \\ &= \int_{-\infty}^{\infty} \left[\frac{\exp\{itx\}}{\pi\sigma \left(1 + \left[\frac{(x-\mu)}{\sigma}\right]^2\right)} \right] dx \\ &= \exp\{i\mu t - \sigma|t|\}. \end{aligned} \quad (5)$$

Now using the fact that the characteristic function of a *sum* of *IID* random variables is the *product* of the individual characteristic functions (i.e.

$$E \left\{ \exp\left\{it \sum_{j=1}^N \tilde{X}_j\right\} \right\} = \prod_{j=1}^N E \{ \exp\{it\tilde{X}_j\} \} = [\Psi_{\tilde{X}}(t)]^N. \quad (6)$$

In the case of the Cauchy we have that the Characteristic function of the sum is

$$[\Psi_{\tilde{X}}(t)]^N = \exp\{iN\mu t - N\sigma|t|\}. \quad (7)$$

Dividing the sum by N just amounts to recaling, and you can easily work out that if \tilde{X} is a Cauchy with parameters (μ, σ) , then for any scalar λ we have $\lambda\tilde{X}$ is a Cauchy with parameters $(\lambda\mu, \lambda\sigma)$ since its characteristic function is

$$\Psi_{\lambda\tilde{X}}(t) = E \{ \exp\{it\lambda\tilde{X}\} \} = \Psi_{\tilde{X}}(\lambda t) = \exp\{i\lambda\mu t - \lambda\sigma|t|\}. \quad (8)$$

So substituting $\lambda = 1/N$, we find that the characteristic function of $\hat{\mu}$ is

$$\Psi_{\hat{\mu}}(t) = \Psi_{\tilde{X}}(t), \quad (9)$$

which implies that for each N $\hat{\mu}$ is distributed as a Cauchy with parameters (μ, σ) . So the finite sample distribution of the sample mean is a Cauchy with parameters (μ, σ) . Clearly this cannot be a consistent estimator for μ since for every N , and even in the limit as $N \rightarrow \infty$, we have $\hat{\mu}$ is always a Cauchy distribution. A consistent estimator must converge to μ with probability 1, but in this estimator does not converge to anything with probability 1, it is always a Cauchy random variable, and so it only converges in distribution, but not almost surely. Can you determine what distribution $\hat{\mu}$ converges in distribution to?

5. Suppose that instead of trying to estimate μ we are interested in estimating the distribution $F(x)$ itself. How could we estimate this distribution non-parametrically?

Answer We can estimate $F(x)$ using the *empirical distribution function* $\hat{F}(x)$ given by

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^N I\{X_i \leq x\}. \quad (10)$$

6. Use the law of large numbers and the central limit theorem to determine the *asymptotic properties* of the estimator $\hat{F}(x)$ given by

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^N I\{X_i \leq x\} \quad (11)$$

Under what conditions is $\hat{F}(x)$ a consistent estimator of $F(x)$ and what is its asymptotic distribution?

Answer The conditions to establish consistency and asymptotic normality of the empirical distribution are significantly weaker than the conditions necessary to establish consistency and asymptotic normality of the sample mean. The reason is that the random variables entering the sum defining $\hat{F}(x)$, $I\{\tilde{X}_i \leq x\}$, are *Bernoulli random variables* and thus bounded, (in particular, they take only two possible values, 0 and 1), *even if the random variables \tilde{X}_i have no finite moments, such as the Cauchy distribution*. Let $B_i(x) = I\{\tilde{X}_i \leq x\}$. Clearly we have

$$\begin{aligned} E\{B_i(x)\} &= F(x) \\ \text{var}(B_i(x)) &= F(x)[1 - F(x)]. \end{aligned} \quad (12)$$

The latter result follows because for a Bernoulli random variable which takes on the value $B_i(x) = 1$ with probability

$$p = F(x) = \Pr\{I\{\tilde{X}_i \leq x\} = 1\} = \Pr\{\tilde{X}_i \leq x\}. \quad (13)$$

It follows that the mean of the Bernoulli is just $p = F(x)$ and the variance of the Bernoulli is $p(1 - p) = F(x)[1 - F(x)]$, since we have

$$\text{var}(B_i(x)) = E\{(B_i(x) - p)^2\} = E\{[B_i(x)]^2\} - [E\{B_i(x)\}]^2 = p(1 - p) = F(x)[1 - F(x)]. \quad (14)$$

Thus, $\hat{F}(x)$ is an unbiased estimator of $F(x)$ and it is asymptotically normal, with asymptotic variance $\sigma^2(x) = F(x)[1 - F(x)]$. In summary, the LLN implies that $\hat{F}(x)$ converges to $F(x)$ with probability 1, and that

$$\sqrt{N}[\hat{F}(x) - F(x)] \longrightarrow N(0, F(x)[1 - F(x)]) \quad (15)$$

under very weak conditions, which are just that $\{\tilde{X}_i\}$ is an *IID* sample from the distribution F , but F is not required to have *any* finite moments for this to hold.

7. Suppose $F(x)$ is known to be a normal distribution, but with an unknown mean μ and variance σ^2 . What is a reasonable *parametric estimator* of the quantity $F(x)$ at some point x ?

Answer Since a normal distribution is entirely determined by the two parameters (μ, σ) , we can estimate just these two parameters by their “sample analogs” the sample mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, and these estimators turn out to be the same as the maximum likelihood estimators of

these parameters as we will see below. Then the natural “parametric estimator” of $F(x)$ would be

$$F(x, \hat{\theta}) = \Phi\left(\frac{(x - \hat{\mu})}{\hat{\sigma}}\right), \quad (16)$$

where Φ is the standard normal CDF, where $\hat{\theta}' = (\hat{\mu}, \hat{\sigma})$.

8. Compare the *asymptotic variance* of the estimator of $F(x)$ in part 6 with the estimator you determined in part 7. Which of these is a more *efficient estimator*?

Answer We would expect the parametric estimator $F(x, \hat{\theta})$ given above to be more efficient than the non-parametric estimator (the empirical CDF, $\hat{F}(x)$) since the parametric estimator uses the additional *prior information* that the data are draws from a normal distribution. Thus, we would expect that the asymptotic variance of $F(x, \hat{\theta})$ to be lower than the asymptotic variance of $\hat{F}(x)$, which is just

$$\Phi\left(\frac{(x - \mu)}{\sigma}\right) \left[1 - \Phi\left(\frac{(x - \mu)}{\sigma}\right)\right] \quad (17)$$

in this case (since the true distribution is $N(\mu, \sigma)$ whose CDF is $\Phi((x - \mu)/\sigma)$). To compute the asymptotic distribution of $F(x, \hat{\theta})$ we need to apply the *delta theorem*. In its general form, it states the following:

Theorem Suppose that $\sqrt{N}[\hat{\theta} - \theta] \rightarrow N(0, \Sigma)$ and that $H(\theta)$ is a continuously differentiable function of θ . Then we have

$$\sqrt{N}[H(\hat{\theta}) - H(\theta)] \rightarrow N(0, \nabla H(\theta)\Sigma\nabla H(\theta)'), \quad (18)$$

where $\nabla H(\theta)$ is the gradient of $H(\theta)$ with respect to θ .

Proof: Expand $H(\hat{\theta})$ in a Taylor series about θ to get

$$H(\hat{\theta}) = H(\theta) + \nabla H(\tilde{\theta})(\hat{\theta} - \theta), \quad (19)$$

where $\tilde{\theta}$ is a point on the line segment joining $\hat{\theta}$ and θ . Since $\sqrt{N}[\hat{\theta} - \theta]$ converges in distribution, it is easy to see that $\hat{\theta}$ itself must converge with probability 1 to θ . So the *continuous mapping theorem* implies that $\nabla H(\tilde{\theta})$ converges in probability to $\nabla H(\theta)$ and further it also implies that when we rewrite the Taylor series expansion for $H(\hat{\theta})$ as

$$\sqrt{N}[H(\hat{\theta}) - H(\theta)] = \nabla H(\tilde{\theta})\sqrt{N}[\hat{\theta} - \theta] \quad (20)$$

the continuous mapping theorem and the continuity of $\nabla H(\theta)$ in θ imply that the right hand side of the above equation converges in distribution to

$$\nabla H(\tilde{\theta})\sqrt{N}[\hat{\theta} - \theta] \rightarrow N(0, \Omega) \quad (21)$$

where Ω is given by

$$\Omega = \nabla H(\theta)\Sigma\nabla H(\theta)'. \quad (22)$$

It follows that $\sqrt{N}[H(\hat{\theta}) - H(\theta)]$ has this same asymptotic distribution, and thus the asymptotic covariance matrix of $H(\hat{\theta})$ is Ω . Now consider the application at hand: we have

$$H(\theta) = \Phi(x, \theta) = \Phi\left(\frac{(x-\mu)}{\sigma}\right), \quad (23)$$

where $\theta' = (\mu, \sigma)$. So in this case $\nabla H(\theta)$ is given by

$$\nabla H(\theta) = \begin{bmatrix} \frac{\partial}{\partial \mu} \Phi\left(\frac{(x-\mu)}{\sigma}\right) \\ \frac{\partial}{\partial \sigma} \Phi\left(\frac{(x-\mu)}{\sigma}\right) \end{bmatrix} = \begin{bmatrix} \frac{-1}{\sigma} \phi\left(\frac{(x-\mu)}{\sigma}\right) \\ \frac{-(x-\mu)}{\sigma^2} \phi\left(\frac{(x-\mu)}{\sigma}\right) \end{bmatrix}. \quad (24)$$

Now the only remaining issue is to determine the asymptotic covariance matrix Σ of $\hat{\theta} = (\hat{\mu}, \hat{\sigma})'$. This is not too hard to do, either directly, by applying the CLT to the formulas for $\hat{\mu}$ and $\hat{\sigma}$, or by using the fact that these are the *maximum likelihood estimators* of the true parameters (μ, σ) . We know that asymptotically, if $\hat{\theta}$ is a maximum likelihood estimator that

$$\sqrt{N}[\hat{\theta} - \theta] \longrightarrow N(0, [I(\theta)]^{-1}), \quad (25)$$

where $calI(\theta)$ is the *information matrix* given by

$$I(\theta) = E \left\{ \left[\frac{\partial}{\partial \theta} \log[f(\tilde{x}|\theta)] \right] \left[\frac{\partial}{\partial \theta} \log[f(\tilde{x}|\theta)] \right]' \right\}. \quad (26)$$

In this case, $\log[f(x|\theta)]$ is given by

$$\log[f(x|\theta)] = \frac{-1}{2} \log(2\pi) - \log(\sigma) - \left[\frac{(x-\mu)}{\sigma} \right]^2. \quad (27)$$

Calculating the derivatives of this with respect to μ and σ and taking expectations, we get

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \quad (28)$$

so it follows that

$$\Sigma = [I(\theta)]^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}. \quad (29)$$

Now, using this formula for Σ and the formula above for $\nabla H(\theta)$, we can derive the “sandwich formula” for Ω in this case

$$\Omega = \nabla H(\theta) \Sigma \nabla H(\theta)' = \phi\left(\frac{(x-\mu)}{\sigma}\right)^2 \left[1 + \frac{1}{2} \left(\frac{(x-\mu)}{\sigma}\right)^2 \right]. \quad (30)$$

I leave it to you to check specific values of x and (μ, σ) to see whether Ω is lower than the variance of the empirical CDF estimator, $\hat{F}(x)$ given by

$$\Phi\left(\frac{(x-\mu)}{\sigma}\right) \left[1 - \Phi\left(\frac{(x-\mu)}{\sigma}\right) \right]. \quad (31)$$

9. Suppose $F(x)$ is a Cauchy distribution. Does knowing this fact affect your ability to consistently estimate $F(x)$ using either the estimator $\hat{F}(x)$ in part 6, or some other “parametric estimator” where you make use of your prior information that $F(x)$ is a Cauchy distribution?

Answer The answer is no. As we discussed in the answer to part 6, the empirical CDF $\hat{F}(x)$ still has the same distribution in the Cauchy case, even though the mean and variance of a Cauchy distribution do not exist. Furthermore, since the parameters of the Cauchy are $\theta = (\mu, \sigma)'$ the same parametric estimator works in this case, i.e. $F(x, \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ . One can use the delta theorem to calculate the asymptotic distribution of the parametric $F(x, \hat{\theta})$ just as in the normal case, and the same reasoning leads to the conclusion that the parametric estimator has a lower asymptotic covariance matrix than the non-parametric estimator.

10. Suppose instead of just trying to estimate $F(x)$ at a single point x , we want to estimate it at a *vector of points* (x_1, x_2, \dots, x_m) (note: the lower case x_i are *fixed points* and should be distinguished from the *observations* which are upper case letters, e.g. X_j). Consider the $m \times 1$ vector estimator $(\hat{F}(x_1), \dots, \hat{F}(x_m))$. Using a multivariate version of the central limit theorem, describe the asymptotic distribution of this vector estimator of F at these m points (x_1, \dots, x_m) .

Answer We apply a multivariate version of the central limit theorem, stacking the empirical CDF estimator at the m points as an $m \times 1$ vector we denote as $\hat{F}(x)$, where x is not interpreted as the $(m \times 1)$ vector of points $x = (x_1, \dots, x_m) \in \mathbb{R}^m$. $\hat{F}(x)$ is now just a normalized sum of *vectors of indicator functions* and it is easy to see that

$$\sqrt{N}[\hat{F}(x) - F(x)] \longrightarrow N(0, \Sigma) \quad (32)$$

where Σ is the $(m \times m)$ covariance matrix of the vector of indicators functions

$$\Sigma = \text{var}(I\{\tilde{x} \leq x\}) = \text{var} \begin{pmatrix} I\{\tilde{x} \leq x_1\} \\ I\{\tilde{x} \leq x_2\} \\ \dots \\ I\{\tilde{x} \leq x_{m-1}\} \\ I\{\tilde{x} \leq x_m\} \end{pmatrix}. \quad (33)$$

It is easy to see that the i^{th} diagonal entry of Σ is just $F(x_i)[1 - F(x_i)]$ since this is just the “marginal” asymptotic variance of $\hat{F}(x_i)$, the empirical CDF evaluated at the single point x_i which we calculated above. So the only thing to worry about is the covariance between two indicators,

$$\Sigma_{ij} = \text{cov}(I\{\tilde{x} \leq x_i\}, I\{\tilde{x} \leq x_j\}). \quad (34)$$

But recall that the covariance of two random variables is defined as

$$\text{cov}(\tilde{X}, \tilde{Y}) = E\{\tilde{X}\tilde{Y}\} - E\{\tilde{X}\}E\{\tilde{Y}\}. \quad (35)$$

We know that for the two Bernoulli random variables $B(x_i) = I\{\tilde{x} \leq x_i\}$ and $B(x_j) = I\{\tilde{x} \leq x_j\}$ we have $E\{B(x_i)\} = F(x_i)$ and $E\{B(x_j)\} = F(x_j)$. So we only need to compute

$$E\{B(x_i)B(x_j)\} = E\{I\{\tilde{x} \leq x_i\}I\{\tilde{x} \leq x_j\}\} = E\{\tilde{x} \leq \min(x_i, x_j)\} \quad (36)$$

since the product of the two Bernoulli's, $B(x_i)B(x_j)$ is just another Bernoulli that takes the value 1 if *both* the events happen, i.e. if $\tilde{x} \leq x_i$ *and* $\tilde{x} \leq x_j$. But both of these will happen if and only if $\tilde{x} \leq \min(x_i, x_j)$. Thus we conclude that

$$E\{B(x_i)B(x_j)\} = E\{I\{\tilde{x} \leq x_i\}I\{\tilde{x} \leq x_j\}\} = E\{\tilde{x} \leq \min(x_i, x_j)\} = F(\min(x_i, x_j)). \quad (37)$$

so

$$\Sigma_{ij} = F(\min(x_i, x_j)) - F(x_i)F(x_j). \quad (38)$$

11. **Extra credit, harder question.** Can you write a formula for the *exact finite-sample distribution* of $\hat{F}(x)$ where x is a single (scalar) point in the support of $F(x)$?

Answer Note that $N\hat{F}(x)$ is a sum of independent Bernoulli random variables. Sums of N independent Bernoulli random variables have a *binomial distribution* (you can show this easily using characteristic functions, for example). Thus $\hat{F}(x)$ will take one of the $N + 1$ values $(0, \frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}, 1)$ with probability given by

$$\text{Prob} \left\{ \hat{F}(x) = \frac{j}{N} \right\} = \binom{N}{j} F(x)^j [1 - F(x)]^{N-j}. \quad (39)$$

II. Consider *parametric estimation* by the method of *maximum likelihood*. Consider the general case where $\{X_1, \dots, X_N\}$ are N IID observations from a *density* $f(x|\theta)$ where θ is a $K \times 1$ vector of unknown parameters to be estimated.

1. Define in the general case, the *maximum likelihood estimator* of θ .
2. Suppose that θ is the 2×1 vector $\theta = (\mu, \sigma)$ and $f(x|\theta)$ is given by

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -(x - \mu)^2 / 2\sigma^2 \right\}, \quad -\infty < x < \infty. \quad (40)$$

Can you provide explicit formulas for the maximum likelihood estimators of μ and σ in this case?

3. Using the law of large numbers and central limit theorem, can you describe the *asymptotic properties* of the maximum likelihood estimator $\hat{\theta}$ in the general case? What “regularity conditions” do you need to impose on $f(x|\theta)$ in order to establish these properties?
4. Now consider the specific case in part 2, where $f(x|\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ . Can you be more specific about the asymptotic distribution in this case?
5. What if $f(x|\theta)$ is a *Cauchy distribution* given by

$$f(x|\theta) = \frac{1}{\pi [1 + (x - \theta)^2]}. \quad (41)$$

Can you write down what the maximum likelihood estimator is for θ and determine its asymptotic properties in this case? In particular, is the maximum likelihood estimator consistent and asymptotically normal?

II. Consider *Bayesian estimation* of the parameter θ where $\{X_1, \dots, X_N\}$ are N IID observations from a *density* $f(x|\theta)$ where θ is a $K \times 1$ vector of unknown parameters to be estimated. Here we assume that you have a *prior distribution* over θ represented by a density $p(\theta)$.

1. Write a formula for the *posterior density* of θ given the observations $\{X_1, \dots, X_N\}$.

2. Suppose $f(x|\theta)$ is a normal distribution with an unknown mean μ but a *known* variance $\sigma^2 = 1$. Suppose your prior distribution for μ is Normally distributed with mean of 1 and a variance of 1. What is the posterior distribution for θ in this case?
3. Use what you know so far about asymptotic theory to show what happens to the posterior distribution as $N \rightarrow \infty$. Does the posterior distribution converge in some sense to the “true” parameter θ^* (the true parameter is the parameter that “nature” uses for the “data generating mechanism”, that is the data are assumed to be draws from the density $f(x|\theta^*)$). Do you need to make any assumptions on your prior $p(\theta)$ when you determine these asymptotic properties of the Bayesian estimator? In particular, what happens if θ^* is not in the support of $p(\theta)$: can the posterior distribution converge to the true θ^* in this case?

III. Prove that the *conditional expectation* $E\{\tilde{y}|\tilde{X}\}$ is the *best predictor* of a random variable \tilde{y} using *any measurable function of \tilde{X}* $f(\tilde{X})$. **Hint: you need to show that for any measurable function $f(x)$**

$$E\{[\tilde{y} - f(\tilde{X})]^2\} \geq E\{[\tilde{y} - E\{\tilde{y}|\tilde{X}\}]^2\}. \quad (42)$$

Another hint: use the Law of Iterated Expectations to show that the inequality above holds, and additionally using the “orthogonality property” of the residual $\tilde{\varepsilon}$ given by

$$\tilde{\varepsilon} = \tilde{y} - E\{\tilde{y}|\tilde{X}\}. \quad (43)$$

IV. Derive the Mill’s ratio “selectivity bias” correction for the two equation model of female labor supply

$$\begin{aligned} w &= X\beta + \varepsilon \\ u &= W\gamma + \eta \end{aligned} \quad (44)$$

where (ε, η) are a bivariate normal distribution with mean 0, and both are independent of (X, W) but (ε, η) could be mutually correlated with correlation coefficient $\rho \in (0, 1)$. The first equation is the *wage equation*, and the second equation is the *net utility from participation in the labor force*. That is, a woman works only if the net utility from participating is positive, $u \geq 0$. Assume, naturally, that we only observe wages for women who *participate* in the labor market but not women who choose not to participate. If we believe the wage equation is a true regression (i.e. ε is independent of X and so if we could observe the *hypothetical wage* (or wage offer) that would have been offered to women who chose not to participate had they decided to search and get a job offer, then regression *on this full population* would uncover (i.e. consistently estimate) the true parameters β of the wage equation. However show that if we try to estimate the wage equation only for the subsample of women who participate using ordinary least squares, we will generally not get consistent estimates of β . Derive a formula for the conditional expectation of w for the subpopulation of participants that we actually observe, i.e. derive a formula for $E\{w|X, u \geq 0\}$ and show it involves an extra term, the *Mill’s ratio*. Are there any conditions where OLS will be a consistent estimator of β ? Otherwise, what type of estimator would you suggest we use if we are interested in estimating β but only have wage data on women who participate in the job market?