Econometrics HW6

Michael B. Nattinger*

April 22, 2021

1 Question 27.1

$$\begin{aligned} m(X) = E[Y|X] &= E[\max\{X'\beta + e, 0\}|X] \\ &= E[(X'\beta + e)1\{X'\beta + e > 0\}|X] \\ &= X'\beta\left(1 - \Phi\left(\frac{-X'\beta}{\sigma}\right)\right) + \sigma\phi\left(\frac{-X'\beta}{\sigma}\right) \\ &= X'\beta\Phi\left(\frac{X'\beta}{\sigma}\right) + \sigma\phi\left(\frac{X'\beta}{\sigma}\right). \end{aligned}$$

$$\begin{aligned} m^{\#}(X) = E[Y^{\#}|X] &= E[X'\beta + e|X'\beta + e > 0] \\ &= X'\beta + \sigma\lambda\left(\frac{X'\beta}{\sigma}\right). \end{aligned}$$

2 Question 27.2

OLS is not consistent for β. β will be biased towards zero as the best-fit line through Y^* will be flatter than it would have been through Y.

3 Question 27.3

3.1 Part A

Yes, $\hat{\beta}$ will be consistent for β, which we can show. Let n be the number of observations which satisfy the condition that $X_1 > 0$, and let X, Y be the corresponding data which satisfies the condition.

$$\begin{aligned} \hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i e_i\right) \\ &\to_p \beta + E[X_i X_i'|X_{1i} > 0]^{-1} E[X_i e_i|X_{1i} > 0] \\ &= \beta. \end{aligned}$$

*I worked on this assignment with my study group: Alex von Hafften, Andrew Smith, and Ryan Mather. I have also discussed problem(s) with Emily Case, Sarah Bass, Katherine Kwok, and Danny Edgel.

3.2 Part B

No, this is not consistent. The mathematical argument is the same but in the probability limit the conditioning on the error is different and induces bias:

$$\begin{aligned}\hat{\beta} &\to_p \beta + E[X_i X_i' | Y_i > 0]^{-1} E[X_i e_i | Y_i > 0] \\ &= \beta + E[X_i X_i' | X_i\beta + e_i > 0]^{-1} E[X_i e_i | X_i\beta + e_i > 0] \\ &\neq \beta.\end{aligned}$$

This does not equal β in the probability limit because $E[X_i e_i | X_i\beta + e_i > 0] \neq 0$ in general.

4 Question 27.4

The NLLS identification would be the sum of squared errors of the conditional probability distribution given a parameterization and the Y data:

$$(\hat{\beta}, \hat{\sigma}) = \arg\min_{\beta,\sigma} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\Phi\left(\frac{X_i'\beta}{\sigma}\right) - \sigma\phi\left(\frac{X_i'\beta}{\sigma}\right)\right)^2$$

5 Question 27.8

$$\begin{aligned}E[Y|X, Z, S = 1] &= X'\beta + E[e|X, Z, S = 1] \\ &= X'\beta + E[e|u > -Z'\gamma] \\ &= X'\beta + \rho_{2,1} E[u|u > -Z'\gamma] \\ &= X'\beta + \rho_{2,1}\lambda(Z'\gamma).\end{aligned}$$

6 Question 27.9

Regression results (point estimates) for the various parts of the question follow.

	A	C	D	E
cons	49.725	50.53	49.512	46.007
inc	-42.761	-42.816	-44.672	-44.439
Dinc	42.767	42.856	44.638	44.426

Note that tinkind was originally negative over 10 % of the time and therefore was not originally properly censored, so I set any negative values of tinkind to 0. After this change, our results for part B are that 22.46% of the observations are censored. This is sufficiently high to expect censoring bias to be a problem in this example.

Interpreting the differences in results, we first note that the coefficients are all relatively similar. Next, note that the magnitudes of the coefficients are almost exactly the same for inc and Dinc, but the sign is different - this is because ind and Dinc are really similar series, with Dinc just being inc shifted to the left, except for the individuals who were already at 0.

Next, note that the CLAD and Tobit regressions yield very similar results. This is because these are unbiased estimates of very similar concepts. The two OLS estimates yield different results because they are biased towards zero.

Below are Matlab codes for the regressions. Note that I had to write code to estimate the CLAD regression, which I have included, as I could not find a pre-written CLAD routine online.

```
clear; close all; clc
exload = 0;
if exload
[x,names] = xlsread('CHJ2004','Sheet1');
save data
else
load data
end
tinkind = x(:,18)./1000;
tinkind(tinkind<0,:) = 0;
income = x(:,1)./1000;
Dincome = (income-1).*(income>1);
X1 = [income Dincome];
mdl1 = fitlm(X1,tinkind);
prop = mean(tinkind==0);
ind2 = (tinkind>0);
Y2 = tinkind(ind2,:);
X2 = X1(ind2,:);
mdl2 = fitlm(X2,Y2);
[mdl3,~,cov3] = TOBIT(tinkind,X1,0,inf,1);
[mdl4,cov4] = CLAD(tinkind,X1,0,inf,1);
tab=table(mdl1.Coefficients.Estimate,mdl2.Coefficients.Estimate,mdl3,mdl4, ...
    'VariableNames',{'A' 'C' 'D' 'E'},'RowNames',{'cons' 'inc' 'Dinc'});
table2latex(tab,'table1',5)
function [B,cov] = CLAD(Y,X,lb,ub,add_constant)
if nargin<5; add_constant = 0; end
if nargin<4; ub =  inf; end
[n,k] = size(X);
if add_constant; X = [ones(n,1) X]; k=k+1; end
options = optimoptions('fminunc','Display','off','StepTolerance',1e-12, ...
    'MaxIterations',1e6,'OptimalityTolerance',1e-12);
[B,~,~,~,~,H] = fminunc(@(b)ob(Y,X,lb,ub,b),X\Y,options);
cov = sqrt(diag(inv(H)));
end

function obj = ob(Y,X,lb,ub,B)
Yh = X*B;
Yh(Yh<lb) = lb;
Yh(Yh>ub) = ub;
obj = sum(abs(Y-Yh));
end
```

7 Question 28.12

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AIC	4766.03	4381.03	4390.18	4756.74	4370.96	4379.71	4760.68	4374.95	4383.66
BIC	4814.09	4435.1	4468.28	4816.81	4437.04	4469.82	4832.77	4453.04	4485.79

Above we can see the AIC, BIC of all of the models. The models we should choose should have the lowest AIC, BIC. The model with the lowest AIC is model (5), followed somewhat closely by model (8). The model with the lowest BIC is model (2), followed closely by model (5). My preferred choice would be model (5) as it performs well under both the AIC and BIC criterion, so it can robustly be considered a good choice in a wide variety of contexts.

```
clear; close all; clc
read = 0;
if read
    cd('..\PS4')
    [x,xt] = xlsread('cps09mar.xlsx','Sheet1');
    save 'data'
    cd('..\PS6')
else
    cd('..\PS4')
    load 'data'
    cd('..\PS6')
end

hisp = x(:,3);
women = x(:,2);
hiwo = logical(hisp.*women);
mlist = 1:9;
x = x(hiwo,:);
exp = x(:,1) - x(:,4) - 6;
edu = x(:,4);
n = size(x,1);
Y = log(x(:,5)./(x(:,6).*x(:,7)));
married = double(x(:,12)<4);
region = zeros(n,3);
for i=1:3
    region(x(:,10)==i,i) = 1;
end
AIC = 0*mlist;
BIC = AIC;
for mnum = mlist
    if mnum<4 % exp,sq
        X_exp = exp.^(1:2);
    elseif mnum<7 % exp - 4
        X_exp = exp.^(1:4);
    else % exp - six
        X_exp = exp.^(1:6);
    end
    if mod(mnum,3)==1 % college
        educ = double(edu>=16);
    elseif mod(mnum,3)==2 % spline
        educ = [edu zeros(n,1)];
        educ(edu>9,2) = edu(edu>9,1) - 9;
    else % dummy
        educ = zeros(n,6);
        educ(edu == 12,1) = 1;
        educ(edu == 13,2) = 1;
        educ(edu == 14,3) = 1;
        educ(edu == 16,4) = 1;
        educ(edu == 18,5) = 1;
        educ(edu == 20,6) = 1;
    end
    X = [married region X_exp educ];
    mdl = fitlm(X,Y);
    AIC(mnum) = mdl.ModelCriterion.AIC;
    BIC(mnum) = mdl.ModelCriterion.BIC;
end
mat = [AIC;BIC];
tab = table(mat(:,1),mat(:,2),mat(:,3),mat(:,4),mat(:,5),mat(:,6),mat(:,7),mat(:,8),mat(:,9), ...
    'VariableNames',{'(1)' '(2)' '(3)' '(4)' '(5)' '(6)' '(7)' '(8)' '(9)'}, ...
    'RowNames',{'AIC' 'BIC'});
table2latex(tab,'table2',6)
```