Homework #2

Due in class 9/23/16
Answers Must be Typed
Work in Groups of 2-4

1) Consider the following Table from Fehr and Goette (2007) reporting descriptive statistics from an individual-level randomized experiment:

TABLE 1—DESCRIPTIVE STATISTICS

		Participating messengers		Difference groups A and B	Nonparticipating messengers, Veloblitz	Messengers, Flash
		Group A	Group B			
Four-week period prior to experiment	Mean revenues	3,500.67 (2,703.25)	3,269.94 (2,330.41)	241.67 [563.19]	1461.70 (1,231.95)	1637.49 (1,838.61)
	Mean shifts	12.14 (8.06)	10.95 (7.58)	1.20 [1.75]	5.19 (4.45)	6.76 (6.11)
	N	21	19		21	59
Treatment period 1	Mean revenues	4,131.33 (2,669.21)	3,005.75 (2,054.20)	1,125.59 [519.72]	844.21 (1,189.53)	1,408.23 (1,664.39)
	Mean shifts	14.00 (7.25)	9.85 (6.76)	4.15 [1.53]	3.14 (4.63)	6.32 (6.21)
	N	22	20		21	65
Treatment period 2	Mean revenues	2,734.03 (2,571.58)	3,675.57 (2,109.19)	−941.53 [513.2]	851.23 (1,150.31)	921.58 (1,076.47)
	Mean shifts	8.73 (7.61)	12.55 (7.49)	−3.82 [1.65]	3.29 (4.15)	4.46 (4.74)
	N	22	20		24	72

Notes: Standard deviations in parentheses, standard error of differences in brackets. Group A received the high commission rate in experimental period 1, group B in experimental period 2.

Focus on treatment period 1 where the wages of bicycle messengers in Group A were raised by 25% relative to group B. It is useful to convert the estimated impact on revenues to an elasticity for comparison with the earlier literature on labor supply.

a) Use the Delta method to construct a standard error estimate for the quantity:

$$\widehat{\eta} \equiv \frac{\overline{Y}^A - \overline{Y}^B}{\overline{Y}^B}$$

where $\overline{Y}^A = 4131.33$ is the mean revenue of Group A and $\overline{Y}^B = 3005.75$ is the mean revenue of Group B. Assume the data are *i.i.d.*.

b) Let $\eta \equiv plim\ \hat{\eta}$. Then we may approximate the revenue/wage elasticity as $\eta/0.25$. Use your answer in a) to construct a confidence interval for $\eta/0.25$.

2) Let Y_i be a dichotomous 0/1 random variable and X_i a $K \times 1$ vector of predictor variables. Consider the Logit log-likelihood:

$$\begin{aligned} l_i &= Y_i \ln \Lambda\left(X_i'\beta\right) + (1 - Y_i) \ln\left[1 - \Lambda\left(X_i'\beta\right)\right] \\ \Lambda\left(X_i'\beta\right) &= \frac{\exp\left(X_i'\beta\right)}{1 + \exp\left(X_i'\beta\right)} \end{aligned}$$

Denote the maximizer of $E[l_i]$ as β_{ML} and the maximizer of $\frac{1}{N}\sum_i l_i$ as $\widehat{\beta}_{ML}$.

a) Derive the score of the logit log-likelihood

b) Using iterated expectations provide a discussion of the moment conditions identifying β_{ML}.

c) Define

$$\beta_{NLLS} = \arg\min_{\beta} E\left[\left(Y - \Lambda\left(X_i'\beta\right)\right)^2\right]$$

Derive the population moment conditions identifying β_{NLLS}

d) Under what conditions will β_{NLLS} and β_{ML} coincide?

e) Derive an expression for the asymptotic variance of $\widehat{\beta}_{ML}$ under the assumption of proper model specification

f) Derive an expression for the asymptotic variance without the assumption of proper specification but with the assumption of independence across observations.

g) Suppose the data are independent across but not necessarily within clusters. Propose a cluster robust estimator of the asymptotic variance of $\widehat{\beta}_{ML}$.

3) Use Matlab to generate 500 observations from the following Probit DGP:

$$\begin{aligned} Y_i &= I\left[X_i + 0.1X_i^2 + \varepsilon_i > 0\right] \\ (X_i, \varepsilon_i) &\sim N\left(0, I_2\right) \end{aligned}$$

Programming Tips: Start by setting the simulation seed to 1234 so that your results can be replicated. Do this by typing "rng(1234)". Then generate X_i by typing "X=randn(500,1)". Then generate ε_i by typing "e=randn(500,1)". Y_i is then generated by typing "Y=(X + .1*X.^2 + e)>0".

a) Write a program Q.m taking the observed data matrix (Y_i, X_i) and a parameter vector as inputs and returning the negative log-likelihood of the data.

b) Use the "fminunc" command to estimate the following misspecified Probit model by ML using Q.m

$$P\left(Y_i = 1 | X_i\right) = \Phi\left(\alpha + bX_i\right)$$

Report your ML point estimates of $\left(\widehat{\alpha}, \widehat{b}\right)$ along with estimates of their asymptotic standard errors.

c) Conduct a score test of the hypothesis that $b_2 = 0$ in the expanded model

$$P\left(Y_i = 1 | X_i\right) = \Phi\left(\alpha + bX_i + b_2 X_i^2\right)$$

Hint: It is possible to do this via an auxiliary regression involving the restricted model's "generalized residuals". Report the value of your test statistic and the associated p-value.

d) Use the "fminunc" command to estimate the unrestricted model in c). Conduct a Wald test that $b_2 = 0$. How does the p-value compare to the implied p-value from your LM test in c)?

4) Use Stata to generate 100 clusters and 100 observations within each cluster from the following DGP:

$$\begin{aligned} Y_{ic} &= \nu_c + D_{ic}\eta_c + \varepsilon_{ic} \\ D_{ic} &= I\left[D_{ic}^* > .5\right] \\ D_{ic}^* &\sim U\left[0,1\right] \\ \eta_c &\sim N\left(1, \sigma_\eta^2\right), \nu_c \sim N\left(0,1\right) \\ \varepsilon_{ic} &\sim N\left(0,1\right) \end{aligned}$$

This DGP can be thought of as representing the effect of some randomized treated D_{ic} assigned at the individual level with probability $\frac{1}{2}$ which has heterogeneous effects η_c across clusters (sites) but an average effect of 1. There is also site heterogeneity in untreated outcomes as captured by the ν_c.

a) Simulate data from this DGP setting $\sigma_\eta^2 = 1$. Regress Y_{ic} on D_{ic} in Stata. What is the standard error of the estimated coefficient on D_{ic}? What is the "robust" standard error? What is the "clustered" standard error?

Programming Tips: Start by setting the simulation seed to 1234 so that your results can be replicated. When simulating the DGP generate the cluster level random variables η_c and v_c first. Then create a unique cluster identifier by typing "gen clusterid=_n". Then use the "expand" command to create 100 micro-level observations within each cluster. Then generate the unit level random variables $(D_{ic}^*, \varepsilon_{ic})$ and form Y_{ic}.

b) Now set seed back to 1234 and repeat step a) setting $\sigma_\eta^2 = 0$. What is the "robust" standard error? What is the "clustered" standard error?

c) Comment on the differences between your answers in a) and b).

d) Now use the "simul" command to perform a Monte Carlo exercise generating data from the above DGP 1000 times when $\sigma_\eta^2 = 1$. Each time perform a t-test of the (true) null that the population regression coefficient on D_c equals one based upon: a) the robust standard error from the microlevel regression, b) the clustered standard error. Use a significance level of 5% and in each simulation record whether the test in question rejects. What fraction of the time does each test reject the null that the regression coefficient equals one?

Programming Tip: To accomplish this exercise you will need to take your previous code and use it to define an "rclass" program. Do this by typing "program define dgp, rclass" at the beginning of a do-file and then pasting your code below. When you run tests you will need to "return" the result of the test so that the simul

command can read it. If you use the "test" command to perform your test then the relevant part of your code would read as follows:

test D=1

return scalar reject=r(p)<.05

This fragment tests whether the coefficient on the regressor named D equals one and then posts a scalar in r(reject) which equals one if the p-value of the test was smaller than .05.

e) Now conduct a Monte Carlo where the $\{\eta_c\}_{c=1}^{100}$ are fixed in repeated draws as would be the case if we considered re-running the experiment on the same 100 sites in our study. What fraction of the time does each test reject the true null that the regression coefficient equals $\frac{1}{100}\sum_{c=1}^{100}\eta_c$?

f) Compare and contrast your answers from parts d) and e). What does this say about the appropriate level of clustering in randomized experiments?

5) Download the CPS extract and do-file from the course webpage.

a) Use the "xi:" command to regress lnwage on dummies for the categories of the variables agecat, educcat, and sex. Use main effects only – i.e. no interactions. Note: another way to accomplish the same task is to use the "anova" command with the ", regress" suffix.

b) Use the "collapse" command to construct a dataset of means and number of observations of the variable lnwage within cells defined by combinations of the variables agecat, educcat, and sex. To keep things simple don't use the person weights.

c) Run a WLS regression of cell wages on dummies for age, education, and sex (main effects only) using the number of observations in each cell as a weight. How do the coefficients compare to what you got in a)?

d) If wages are *iid* within cell, what is the variance of each cell mean?

e) Use Stata to construct an estimate of the variance of each cell mean. Do not assume the data are homoscedastic across cells.

f) Run a WLS regression of cell wages on dummies for age, education, and sex (main effects only) using the inverse of your estimate of the cell variance as a weight.

g) Compute a chi-squared test from the weighted sum of squared residuals of this regression. (Hint: be careful about the difference between iweights and aweights in Stata. Aweights will force the weights to sum to 1 which affects computation of the residual sum of squares.) What do you conclude about the fit of the main effects model?

h) Run a WLS regression of cell wages on main effects for age, education, and sex plus all two-way interactions (e.g. age×education, education×sex, and age×sex).

i) Compute the chi-squared test. Can you reject this model at the 5% level?

j) Compute the predicted cell means from the model you estimated in h). Graph the predicted and actual cell means against age by subgroup. Look for signs of misspecification.

k) Can you find a simple regression specification that passes the chi-squared test?

6) Consider the following random coefficient binary choice model:

$$Y_i = 1\left[b_i X_i + \varepsilon_i > 0\right]$$

where $\varepsilon_i \sim Logistic(0,1)$ and $b_i \sim N\left(\mu, \sigma^2\right)$.

a) Write an expression for the simulated log likelihood and its derivatives with respect to μ and σ.

b) Load the Matlab dataset logit.mat. The first column of the matrix "data" gives observations on Y_i and second on X_i. Set the seed to 1234 with the rng() command and generate a 2000x1000 matrix V of random normal variates with the command "V=randn(2000,1000)". The columns of this matrix index simulation draws while the rows index observations.

c) Write a program SML.m that takes as arguments the data vectors, the matrix V of random variates, and the parameters $(\mu, \ln\sigma)$ and returns the negative of the simulated log likelihood and its analytic gradient.

Programming tip: speed up your program by vectorizing your code and avoiding for-loops wherever possible. Two matrices can be multiplied elementwise with the ".*" command (e.g., A.*B). A matrix can be multiplied elementwise by a conformable vector with the "bsxfun" function (e.g., bsxfun(@times,A,v)).

d) Use the fminunc command to maximize the simulated likelihood and to compute the Hessian of the SML criterion. Make sure that fminunc uses your analytic gradient expression by setting options=optimset('GradObj','on'). Report your SML parameter estimates of $(\mu, \ln\sigma)$. Verify that your estimates are not sensitive to starting values.

e) Report standard errors for $(\mu, \ln\sigma)$ computed under the assumption of proper specification.

f) Report robust standard errors based upon the sandwich matrix.