Princeton University **Spring 2021**
Department of Economics **ECO 518**

Exercise 8

Version 1 (April 1, 2021)

Mikkel Plagborg-Møller

Due on Canvas by 2.59 pm on Monday, April 12

The problem set should be submitted as a single PDF file on the course website. You may turn in one problem set per group of at most 3 students. You may discuss the exercises with any of your classmates. Please attach your code to the problem set. Use any software you like, although you learn the most by avoiding canned statistics/econometrics packages. If you find any errors or require clarification, please let me know right away.

Question 1

The data set `mroz.csv`, which is available on the course website, contains 753 observations on married women with the following variables: labor market participation (`part`, where 1 means participation and 0 means non-participation), number of children less than 6 years old (`kidslt6`), age in years (`age`), education in years (`educ`), and non-wife income (that is, other income of the household aside from the salary of the woman) in 1000s of dollars (`nwifeinc`).

Let Y_i be the value of `part` and X_i be the value of the vector $(1, \texttt{kidslt6}, \texttt{age}, \texttt{educ}, \texttt{nwifeinc})'$ for individual i. Assume that the data is i.i.d. across individuals, and

$$\Pr(Y_i = 1 \mid X_i = x) = \Phi(\beta_0' x),$$

where $\beta_0 \in \mathbb{R}^k$ is an unknown parameter vector and $\Phi(\cdot)$ is the standard normal CDF. This is called the probit model.

i) Compute the MLE of β_0.

ii) Compute standard errors for your estimates in (i). Compute the standard errors in two ways: using the Hessian estimator (assuming correct specification), and using the misspecification-robust QMLE formula.

iii) Plot a 95% confidence set for the coefficients on `kidslt6` and `educ`. Use the variance estimate that is robust to misspecification.

Question 2

Suppose that you have an i.i.d. sample $\{Y_i, X_i\}_{i=1}^N$, where Y_i is the number of patents in a certain year for firm i (a non-negative integer), and X_i is a vector of covariates for firm i.

i) Assume that the distribution of Y_i given $X_i = x$ is Poisson with mean $e^{x'\beta_0}$. Derive the log likelihood $\sum_{i=1}^N \log f(Y_i \mid X_i, \beta)$.

ii) Now drop the assumption that the data follows a Poisson distribution. Let $\hat{\beta}$ be the QMLE of β_0, i.e., the maximizer of the Poisson log likelihood from (i). Argue that as long as $E[Y_i \mid X_i = x] = e^{x'\beta_0}$, then $\hat{\beta}$ is consistent for β_0 even if Y_i does not have a Poisson distribution conditional on X_i. (You do not need to give all proof details, just the basic logical steps.)

Question 3

We will estimate a demand curve for fish using data collected at the Fulton Fish Market in New York City (Angrist, Graddy & Imbens, *REStud* 2000). The data set `fish.csv`, which is available on the course website, has log of average daily prices (`logp`) and log of total quantities (`logq`) of whiting sold for each of 111 days during winter 1991–1992.

We aim to estimate the following demand curve by instrumental variables:

$$\texttt{logq}_t = \beta_0 + \beta_1 \texttt{logp}_t + \varepsilon_t. \tag{1}$$

To estimate the demand curve for whiting, we will use weather conditions at sea to construct instruments for price. The data set contains two instruments: `stormy` and `mixed`. `stormy` is an indicator of stormy weather conditions at sea and `mixed` is an indicator of mixed weather conditions at sea. We make the (heroic) assumption that the data is i.i.d. across days.

i) Estimate the parameters of equation (1) by 2SLS. Calculate standard errors (without assuming homoskedasticity).

ii) Use the 2SLS estimates from the previous question to estimate the efficient GMM weight matrix and the efficient GMM estimator. Calculate standard errors.

iii) Report the 95% Anderson-Rubin confidence set for β_1. (Use a fine grid that spans at least the interval $[-4, 1]$.) Compare with the usual 95% confidence interval that is based on the results in (ii).

iv) Compute the p-value of the over-identification test.

Question 4 (entire question is OPTIONAL)

This question provides the theoretical arguments underlying the efficient GMM weight matrix (see the "Extremum estimators I" slide deck). For square matrices A and B, use the notation "$A \geq B$" to denote that the difference $A - B$ is a positive semidefinite matrix.

i) Suppose we have two different estimators $\hat{\beta}$ and $\tilde{\beta}$ of a $k \times 1$ parameter vector β_0. Show that if $\text{Var}(\hat{\beta}) \geq \text{Var}(\tilde{\beta})$ (in the positive semidefinite sense), then $\text{Var}(\lambda'\hat{\beta}) \geq \text{Var}(\lambda'\tilde{\beta})$ (in the usual sense) for any constant $k \times 1$ vector λ.

ii) Consider the normal regression model written in vector form:

$$Y = X\beta + \varepsilon, \quad (\varepsilon \mid X) \sim N(0_{N\times 1}, \Omega).$$

Here Y is $N \times 1$, X is $N \times k$ (and has full column rank), β is $k \times 1$, ε is $N \times 1$, and Ω is $N \times N$ and symmetric positive definite. Let W be any symmetric positive definite $N \times N$ matrix and define the estimator $\hat{\beta} = (X'WX)^{-1}X'WY$ of β.

Use the Gauss-Markov theorem to argue that $\text{Var}(\hat{\beta} \mid X) \geq (X'\Omega^{-1}X)^{-1}$.

iii) Use the result from (ii) to show that

$$(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} \geq (G'\Omega^{-1}G)^{-1}$$

for any $r \times k$ matrix G (of full column rank) and any $r \times r$ symmetric positive definite matrices W and Ω.