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

Exercise 6

Version 1 (March 19, 2021)

Mikkel Plagborg-Møller

Due on Canvas by 2.59 pm on Monday, March 29

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

Consider the data set `Guns.xlsx` (or equivalently `Guns.dta`) on Canvas. A data documentation file is also available on Canvas. Assume that the data is i.i.d. across the 51 U.S. states (plus D.C.), but not necessarily independent across time. Consider a linear fixed effects regression of the logarithm of `vio` on `shall`, controlling for `incarc_rate`, `density`, `avginc`, `pop`, `pb1064`, `pw1064`, and `pm1029`.

i) Compute an appropriate bootstrap standard error for the coefficient on `shall`. [Note: You are welcome to use Stata's built-in bootstrap functionality, as long as you take care to understand how it works.]

ii) The Stata log file `guns_518.log` on Canvas contains output from five different regressions at the bottom. Briefly compare the bootstrap standard error from part (i) with the various analytical standard errors computed by Stata. Theoretically speaking, which of the analytical standard errors are comparable to the bootstrap standard error in terms of the assumptions needed for their validity?

iii) Compute two 95% bootstrap confidence intervals for the coefficient on `shall`: one that uses Efron's method and one that uses the percentile-t method.

Question 2

We are interested in predicting an individual's retirement age Y using that individual's total earnings Z^* between ages 30–34. We observe data on retirement age and earnings for a random sample of N individuals, but unfortunately the earnings data is *top coded* (censored at the top), so that total earnings above \$1,000,000 are recorded as "$\geq$ \$1,000,000". However, we also observe a vector X of individual characteristics, such as education level. To run the regression of Y on Z^*, we therefore first *impute* actual earnings for the censored observations.

To do the imputation, we assume a *censored normal regression model*:

$$Z^* = \gamma' X + \varepsilon, \quad (\varepsilon \mid X) \sim N(0, \sigma^2).$$

Due to top coding, we do not observe Z^* itself, but rather we observe $Z = \min\{Z^*, 10^6\}$. We can estimate the coefficient vector γ and the error variance σ^2 using Maximum Likelihood with the conditional likelihood function

$$\hat{\ell}_N(\gamma, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(Z_i - \gamma' X_i)^2/(2\sigma^2)} \right)^{\mathbb{1}(Z_i < 10^6)} \left(1 - \Phi\left(\frac{10^6 - \gamma' X_i}{\sqrt{\sigma^2}} \right) \right)^{\mathbb{1}(Z_i = 10^6)},$$

where $\Phi(\cdot)$ is the standard normal CDF. (You do not have to derive this likelihood, but if you are interested, you can read chapter 8.2 in the Hayashi (2000) textbook.) Once we have computed the MLE $(\hat{\gamma}, \hat{\sigma}^2)$, we can impute the missing earnings data as $\hat{Z}_i^* = \hat{\gamma}' X_i$ for individuals with $Z_i = 10^6$.

In summary, we perform the following estimation steps on the data $\{Y_i, Z_i, X_i\}_{i=1}^N$:

1. Compute the MLE $(\hat{\gamma}, \hat{\sigma}^2)$ in the censored normal regression of Z_i on X_i.
2. Define imputed earnings $\hat{Z}_i^* = Z_i \mathbb{1}(Z_i < 10^6) + (\hat{\gamma}' X_i) \mathbb{1}(Z_i = 10^6)$, $i = 1, \dots, N$.
3. Run an OLS regression of retirement age Y_i on a constant and $\hat{Z}_i^*$. Denote the coefficient estimates on the constant and $\hat{Z}_i^*$ by $\hat{\alpha}$ and $\hat{\beta}$, respectively.

We wish to compute standard errors for the regression in step (3), taking into account the estimation error in the imputation steps (1)–(2).

i) Give a step-by-step guide to bootstrapping valid standard errors for $\hat{\beta}$. You do not need to write any code or formulas, but the guide should be clear enough that it could be used directly by a research assistant who has taken ECO 517+518 and who is well-versed in R/Matlab and numerical optimization.

ii) Describe how to compute valid standard errors for $\hat{\beta}$ using asymptotic theory. You do not need to provide detailed formulas, but all the steps in the derivation should be described so that the afore-mentioned research assistant (who is also good at taking derivatives) could carry them out. [Hint: Stack the first-order conditions for maximization of the log likelihood and for the OLS estimator. The goal of this exercise is to review the Taylor expansion asymptotic calculations you did for basic maximum likelihood in ECO 517. If you can't quite crack it, don't worry – we will discuss two-step estimators in more detail later.]

Question 3

Consider the possibly heteroskedastic linear regression model

$$Y = \alpha_0 + \beta_0 X + u, \quad u = \sigma_0(X)\varepsilon,$$

where X is a scalar regressor, $\sigma_0(\cdot)$ is a function, and ε is independent of X.

Consider testing the null hypothesis $H_0\colon \beta_0 = 0$ against the alternative $H_1\colon \beta_0 \neq 0$. We will consider tests based on the OLS estimator $\hat{\beta}$ of β_0 from a regression of Y on X and a constant. Let t_N denote the absolute value of the t-statistic that uses Eicker-Huber-White (EHW) robust standard errors. We will consider three tests that use this statistic, with three different critical values: (a) asymptotic normal critical value, (b) nonparametric bootstrap critical value, and (c) residual bootstrap critical value. We will in addition consider two tests that reject when the statistic $\tilde{t}_N = |\sqrt{N}\hat{\beta}|$ exceeds a critical value, namely: (d) nonparametric bootstrap critical value for $\tilde{t}_N$, and (e) residual bootstrap critical value for $\tilde{t}_N$. Notice that the bootstrap calculations for cases (d) and (e) do not explicitly compute EHW standard errors.

i) Conduct a Monte Carlo simulation study of the rejection frequency of the five tests (a)–(e). When simulating the data, assume $\alpha_0 = 1$, $\beta_0 = 0$, $\sigma_0(X) = 1$ (a constant), $X \sim N(0, 4)$, $\varepsilon \sim t(5)$ (the t-distribution with 5 degrees of freedom), sample size $N = 50$, and test significance level 5%. Use at least 1,000 Monte Carlo replications, with at least 500 bootstrap draws per replication. Report the results in a nicely formatted table.

ii) Repeat the exercise from part (i), but now set $\sigma_0(X) = 0.5X$.

iii) Comment on the results in parts (i) and (ii). Be brief and concrete.