ECON 626: Empirical Microeconomics

Problem Set 1

Department of Economics
University of Maryland
Fall 2016

Problem Set 1 is due at 5pm on Thursday, September 15.

1. **Review: sum of normals.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$ denote a random variable X that is normally distributed with mean μ and variance σ^2. The sum of two independent normally distributed random variables is also a normally distributed random variable, with mean equal to the sum of the means of the originals and variance equal to the sum of the variances of the originals. Let independent random variables $X_i \sim \mathcal{N}(\mu, \sigma^2) \forall i \in \{1, 2, 3\}$. Fill in the blanks:

   (a) $A = X_1 + X_2$.
   $A \sim \mathcal{N}(\_, \_)$.

   (b) $B = X_1 + X_2 + X_3$.
   $B \sim \mathcal{N}(\_, \_)$.

   (c) Are A and B independent?

   (d) $C = 2X_1$.
   $C \sim \mathcal{N}(\_, \_)$.

   (e) $D = A + B$.
   $D \sim \mathcal{N}(\_, \_)$.

2. **Review: normal distribution.** Let $F_X(x)$ denote the cumulative distribution function (CDF) of random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$. Let random variable $Z \sim \mathcal{N}(0, 1)$ (mean 0, variance 1); let $Y \sim \mathcal{N}(6, 4)$ (mean 6, variance 4).

   For each of the following, write an expression for the value as a function $\Phi()$, and, using whatever software is handy, compute its value to at least a few decimal places:

   (a) $\Pr[Z \leq 0]$

   (b) $\Pr[Z \leq 1]$

   (c) $\Pr[Z \leq 2]$

   (d) $\Pr[|Z| > 2.57583]$

   (e) $\Pr[Y \leq 2]$

   (f) $\Pr[Y \leq 6]$

   (g) $\Pr[|Y - 6| > 4]$

   (h) $\Pr[|Y - 6| > 5.15166]$

3. **Review: continuous uniform distribution.** Denote that continuous random variable X is uniformly distributed between a and b by writing $X \sim \mathcal{U}(a, b)$. Its density is:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

It is straightforward to show that $E[X] = \frac{b+a}{2}$ and that $Var(X) = \frac{(b-a)^2}{12}$. Let $W \sim \mathcal{U}(0, 1)$; let $T \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ *so, approximately,* $\mathcal{U}(-1.732, 1.732)$; for each of the following, using whatever software is handy (where needed), compute its value to at least a few decimal places, or in the form of a simple fraction if that is convenient:

(a) $E[W]$
(b) $Var(W)$
(c) $\Pr[W \leq 0.5]$
(d) $\Pr[W \leq 2]$
(e) $E[T]$
(f) $Var(T)$
(g) $\Pr[T \leq 0]$
(h) $\Pr[T \leq 1]$
(i) $\Pr[T \leq 2]$
(j) $\Pr[|T| > 2.57583]$

4. **Simulation of random variables.** For this problem, you will use Stata. You should write a ".do" file that produces your answers; this file, which should be well-commented, should be part of your submitted work, but the key commands you use should be included in the LaTeX-formatted PDF that represents your full solutions to the problem set. (In your .do file, use comments to indicate where you are answering each part of the question.)

(a) First, using any reasonable combination of Stata functions (likely including `uniform()`, `invnorm()`, or `rnormal()`), generate a dataset of 10,000 observations with variables as follows. Be sure to `set seed` so that your code produces identical results when run twice.
   i. Generate four variables, Z_1 through Z_4, that are independently drawn from the distribution $\mathcal{N}(0, 1)$.
   ii. Generate four variables, T_1 through T_4, that are independently drawn from the distribution $\mathcal{U}(-\sqrt{3}, \sqrt{3})$.

(b) What are the sample mean and variance of Z_1 ?

(c) What are the sample mean and variance of T_1 ?

(d) For what fraction of observations is...
   i. ...$Z_1 \leq 0$?
   ii. ...$Z_1 \leq 1$?

iii. ...$Z_1 \leq 2$?
iv. ...$|Z_1| > 2.57583$?

(e) For what fraction of observations is...
i. ...$T_1 \leq 0$?
ii. ...$T_1 \leq 1$?
iii. ...$T_1 \leq 2$?
iv. ...$|T_1| > 2.57583$?

(f) Now, generate two additional variables, $Zmean = (Z_1 + Z_2 + Z_3 + Z_4)/4$ and $Tmean = (T_1 + T_2 + T_3 + T_4)/4$.

(g) What are the sample mean and variance of $Zmean$?

(h) What are the sample mean and variance of $Tmean$?

(i) For what fraction of observations is...
i. ...$Zmean \leq 0$?
ii. ...$Zmean \leq 0.5$?
iii. ...$Zmean \leq 1$?
iv. ...$|Zmean| > 1.28791$?

(j) For what fraction of observations is...
i. ...$Tmean \leq 0$?
ii. ...$Tmean \leq 0.5$?
iii. ...$Tmean \leq 1$?
iv. ...$|Tmean| > 1.28791$?

5. **Regressions in Stata.** The Stata data set `MalariaData2.dta` contains data from the *AER* paper "Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial" by Jessica Cohen, Pascaline Dupas, and Simone Schaner. The study estimates the effects of price subsidies for malaria medication in Kenya. The study looked at over-treatment and under-treatment of malaria under different drug subsidies.

   The variable `took_act_first` is of the key outcomes in the study. It indicates whether, during the first malaria episode a household experiences after starting to participate in the experiment, a sick person took antimalarial drugs. Use Stata to estimate three regressions of `took_act_first` on baseline levels of education (the variable `B_head_edu`) and malaria knowledge (the variable `B_knowledge_correct`). Make a journal-ready (i.e. neat, organized, and self-contained) table of the following regressions. In Column (1), report the results of an OLS regression using the default homoskedastic errors. In Column (2), report the results of an OLS regression using the `robust` option to generate heteroskedasticity-robust standard errors. In Column (3), report the results of a probit regression using the default (homoskedastic) standard errors. What do the results suggest about the relationship between education, information, and malaria treatment?

6. **OLS regressions in MATLAB.** Write a simple MATLAB program to replicate your OLS results (Column 1). In other words, write a MATLAB program that estimates

$$\hat{\beta}_{ols} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

and calculates the associated standard errors using the formula

$$V\left[\hat{\beta}_{ols}\right] = \left(\frac{\sum_i \hat{u}_i^2}{N-K}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Confirm that your results line up with your answer to (5). (To complete this problem, you will need to export the relevant piece of the Stata data set so that it can be used by MATLAB. We recommend exporting the relevant variables as a `csv` file.)

7. **OLS regressions in MATLAB, continued.** Now, write a MATLAB program that estimates $\hat{\beta}_{ols}$ and calculates robust standard errors in MATLAB, using the formula for the sandwich estimator of the variance of $\hat{\beta}_{ols}$:

$$V\left[\hat{\beta}_{ols}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

where $\hat{\mathbf{\Omega}} = \text{Diag}\left[\hat{u}_i^2\right]$ normalized by $N/(N-K)$. Confirm that your Stata and MATLAB results line up

8. **Probit regressions in MATLAB.** The MATLAB program `PS1exampleFminunc.m` estimates $\hat{\beta}$ for a Logit model via maximum likelihood using MATLAB's `fminunc` command. It depends on the very simple `nllLogit.m` program for the likelihood function, and on the `PS1discreteOutcomeTwovar.csv` file for example data. You can confirm that the results are correct (using the logit model) in Stata using the `PS1stataequivalent.do`.

   Recall that, in a probit model, the probability that y is equal to one is given by

$$\Pr(y=1) = \Phi(-\mathbf{X}'\beta)$$

   and the likelihood function for β is given by

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n}\left(y_i \ln\left[\Phi(\mathbf{X}'\beta)\right] + (1-y_i)\ln\left[1-\Phi(\mathbf{X}'\beta)\right]\right)$$

   Modify the MATLAB likelihood function program to reflect the probit likelihood function. Use it to estimate probit coefficients on the Malaria dataset rather than the example csv data. After estimating your probit coefficients, confirm that your answers line up with your Stata results.