HA 4
Due on Friday, April 1, 19:00

Details:

(a) Students can work in **groups** of maximum **2** people. You need to submit one HA per group.

(b) I encourage you to discuss all the problems with your classmates, but you should write your own write-up. Cheating won't be tolerated.

(c) You should submit your work **electronically** to d.i.arkhangelsky at gmail dot com. Store of the results in the zip file with name "ha_4_"your_name".zip".

Problem 1 (Double selection)

In this problem you will implement double-selection algorithm that we discussed in lectures 13-14.

(a) First you generate the data using the following guidelines:

(1) Let $n = 1000$ and $p = 500$ and $s_1 = 10$ and $s_2 = 10$. Generate $n \times p$ matrix of $\mathcal{N}(0, 1)$ random numbers. Denote it by X_0

(2) Generate $p \times p$ matrix of $\mathcal{N}(0, 1)$ random numbers. Denote it by U_0. Let $\Sigma := U_0 U_0^T$, compute $\Sigma^{\frac{1}{2}}$ and compute $X := X_0 \Sigma^{\frac{1}{2}}$. What is the logic behind this step?

(3) Let $\mu_1 = 5$ and $\sigma_1^2 = 2$ and generate s_1-dimensional vector of $\mathcal{N}(\mu_1, \sigma_1^2)$ random numbers. Denote it by β_1.

(4) Let $\mu_2 = 3$ and $\sigma_1^2 = 1$ and generate s_2-dimensional vector of $\mathcal{N}(\mu_2, \sigma_2^2)$ random numbers. Denote it by β_2.

(5) Let $\sigma(x) := \frac{\exp(x)}{1+\exp(x)}$. Select random s_2 columns of matrix X, denote them by X_2 and compute $\pi = \sigma(\gamma + X_2\beta_2)$ – n-dimensional vector of numbers between zero and one, where γ is n-dimensional vector of constants. Adjust γ in such way that $\bar{\pi} := \frac{\sum_{i=1}^n \pi_i}{n} \approx \frac{1}{2}$.

(6) Generate n independent Bernoulli random variables with probabilities π. Denote the corresponding n-dimensional vector by D.

(7) Let $\alpha_0 = 5$, $\tilde{X} := (D, X)$ and compute

$$s := \frac{1}{n}(\alpha_0, \beta_1)^T \tilde{X}^T \tilde{X}(\alpha_0, \beta_1)$$

Let $\sigma^2 = 0.3s$.

(8) Select s_1 random columns from matrix X, denote them by X_1 and generate $Y := \alpha_0 D + X_1\beta_1 + \varepsilon$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

(9) As a result you should have the following data: n-dimensional vectors Y and D, $n \times p$ matrix X. This is the **only data** that you should use in what follows.

(10) Comment on the joint distribution of (Y, D, X). Does it satisfy (informally) the sparsity assumption that we discussed?

(b) Estimate the following simple regression model by OLS:

$$Y = \theta_0 + \theta_1 D + u_1 \tag{1}$$

report θ_1 and compare it with α_0.

(c) Estimate the full OLS model:

$$Y = \theta_0 + \theta_1 D + X\theta_2 \tag{2}$$

report θ_1 and compare it with α_0.

(d) Estimate the model (2) using the double-selection algorithm that we discussed in lectures 13-14. Use 10-fold CV to select tuning parameter for each step. Report θ_1 and compare it with α_0.

(e) Comment on the differences between the result in cases (b)-(d).

(f) (*Bonus*) Estimate standard errors of all three estimates (you may assume homoskedasticity if you want).

Problem 2 (Demand estimation)

In this problem you will implement several demand estimation procedures that we discussed during lectures 15-16.

(a) First you generate the data using the following guidelines:

(1) Let $\mu_1 = (10, 10)$ and $\mu_2 = (8, 8)$. Let $n = 50$. For $i = 1, 2$ do the following:

(i) Generate $n \times 4$ matrix of $\mathcal{N}(0, 1)$ random numbers and call it U_i. Let u_{ij} be the j-th column of the matrix U_i.

(ii) Let $X_i := 2u_{i1} + \mu_{i1}$, let $\xi_i := 3u_{i2}$, let $P_i := \mu_{i2} + 2(\frac{1}{2}u_{i2} + \frac{\sqrt{3}}{2}u_{i3})$, let $Z_i := \frac{1}{3}u_{i3} + \frac{\sqrt{8}}{3}u_{i4}$.

(2) At the end you should have a data matrix $D = (X_1, P_1, \xi_1, Z_1, X_2, P_2, \xi_2, Z_2)$ with dimensions (50×8). Argue that random vector D_i (i-th row of matrix D) has a joint normal distribution. Compute its mean and covariance matrix.

(3) Let $\beta_0 = (-5, 10)^T$ and $\beta_1 = (-3, 9)^T$. Generate n random numbers from $U[0, 1]$ and let π be the corresponding n-dimensional random vector.

(4) For each $i = 1, \ldots, n$ compute the following shares:

$$s_{ik} = \begin{pmatrix} \frac{\exp\{(\beta_k^T,1)(P_{1i},X_{1i},\xi_{1i})^T\}}{1+\exp\{(\beta_k^T,1)(P_{1i},X_{1i},\xi_{1i})^T\}+\exp\{(\beta_k^T,1)(P_{2i},X_{2i},\xi_{2i})^T\}} \\ \frac{\exp\{(\beta_k^T,1)(P_{2i},X_{2i},\xi_{2i})^T\}}{1+\exp\{(\beta_k^T,1)(P_{1i},X_{1i},\xi_{1i})^T\}+\exp\{(\beta_k^T,1)(P_{2i},X_{2i},\xi_{2i})^T\}} \end{pmatrix}^T \tag{3}$$

$$s_i = \pi_i s_{i0} + (1 - \pi_i) s_{i1} \tag{4}$$

(6) At the end you should have the following three data matrices: D, $S = (s_1, s_2)$ and π. Essentially D describes characteristics of two goods in n markets, S describes shares of the goods and π describes some demographic characteristic of each market. These are **the only data** that you can use in what follows.

(b) First, you assume that there is no randomness in coefficients (all consumers are the same). Under this assumption you estimate typical multiple logit model following these steps:

(1) Compute $s_0 = 1 - s_1 - s_2$.

(2) Compute mean utilities: $\delta_1 = \ln(s_1) - \ln(s_0)$ and $\delta_2 = \ln(s_2) - \ln(s_0)$.

(3) Estimate $\theta = (\theta_0, \theta_1, \theta_2)$ in the following model using OLS:

$$\delta_{ji} = \theta_0 + \theta_1 X_{ji} + \theta_2 P_{ji} + \xi_{ji} \tag{5}$$

report θ_2, compare it with price coefficients that you used to construct the data. Comment.

(4) Now, assume that P_{ji} is correlated with ξ_{ji} and use Z_{ji} as an instrument to estimate θ. Report θ_2 and compare it with correct price coefficients. Argue that Z_{ij} can be used as an instrument for price.

(c) *(Bonus)* Argue that there is not enough information to estimate β_0 and β_1. Generate additional instruments (valid and relevant!) that will allow you to estimate β_0 and β_1 and estimate them.