Problem Set 1

ECON 31703

due date: April 26, 2012

Exercise 1

Let $(X_1, \ldots, X_{99}, \varepsilon) \in \mathbb{R}^{100}$ be independent standard normal random variables. Define

$$Y = X_1 + \varepsilon.$$

Consider an i.i.d. sample from $(Y, X_1, \ldots, X_{99})$ of size N, which we denote $(Y_i, X_{i1}, \ldots, X_{i,99})$ for $i = 1, \ldots, N$. We express the sample in the following matrix form:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}, \qquad \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{99}) = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1,99} \\ X_{21} & X_{22} & \cdots & X_{2,99} \\ \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{N,99}. \end{bmatrix}.$$

(a) Consider a linear model

$$\mathbf{Y} = \mathbf{X}_1 \beta + \mathbf{e}.$$

We know $\beta = 1$ from the knowledge of the data generating process. Let $\hat{\beta}^{OLS}$ be the OLS estimator of β. This is the estimate when we know the true selection of regressors.

Question. Simulate $10,000$ samples of size $N = 102$, compute $\hat{\beta}^{OLS}$ for each sample and compute its variance. Check that the variance is very close to 0.01 which is the true value.

(b) Let $p \leq 99$ and consider the linear model

$$\mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \ldots + \mathbf{X}_p \beta_p + \mathbf{e}.$$

We know that $\beta = 1$ and $\beta_2 = \ldots = \beta_p = 0$ from the data generating process. Let $\hat{\beta}_1^{OLS}$ be the OLS estimator of β_1.

Question. For each $p = 2, 5, 10$, simulate $10,000$ samples of size $N = 102$, compute $\hat{\beta}_1^{OLS}$ for each sample and compute its variance. Report the variances and comment on the result.

(c) **Question.** Repeat **(b)** with $p = 90, 95, 99$. Report the variances and comment on the result.

Exercise 2

Download the dataset based on Sala-i Martin (1997). Use read.csv function to load the csv file. Compared to the raw data, we have dropped all the countries with at least one missing variables, and the variable AGE (average age of population) is missing from the dataset. Refer to this document (pp 26-27) for description of the key variables. Note that the variable Growth in the document is given as gamma in our dataset.

(a) Regress gamma, the growth rate, on GDPSH60, LIFEE060, and P60. Do the coefficients' signs make sense?

(b) Regress gamma on all the other covariates in the dataset. Note that the lm function will use the first 35 covariates in order to make $p = n$ and produce a perfect fit. Comment on the changes in the coefficients on GDPSH60, LIFEE060, and P60 compared to **(a)**.

(c) Consider the regression with 5 regressors (and a constant), namely GDPSH60, LIFEE060, and P60 and then two from the rest of the variables. Run regressions with all the possible choices of the two variables, and choose the one with the largest R-squared. Report the resulting choice of the two variables and check if the coefficients' signs make sense.

(d) Report how many regressions you have run in **(c)**. How many regressions you would have to run if we choose five or six additional variables from the rest, instead of two?

Exercise 3

(a) **Question.** For each $N = 100, 200, 500, 1000$, repeat the following exercise $S = 10$ times with $p = 90$ and $\rho = 0$. Report the average of the smallest eigenvalues for each N.

- Generate $X_1, \ldots, X_N \in \mathbb{R}^p$, which are i.i.d. draws of $N(0, \Sigma_p)$ where Σ_p is a $p \times p$ matrix such that

$$\Sigma_p = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

- Let

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_N' \end{pmatrix}.$$

- Compute the smallest eigenvalue of $X'X$.

(b) **Question.** Repeat **(a)** with $\rho = 0.5$ and $\rho = 0.9$. Comment on the result.

(c) **Question.** Now let p grow with N, namely $p_N = 0.9N$. Repeat **(a)** and **(b)**. Comment on the result.

(d) **Question.** Now let

$$p_N = \lfloor 19.55 \times \ln N \rfloor$$

where $\lfloor \cdot \rfloor$ is the floor function. Repeat **(a)** and **(b)**. Comment on the result.

References

Sala-i Martin, Xavier X. 1997. "I just ran two million regressions." *American Economic Review* 87 (2):178–83.