New York University
Applied Statistics and Econometrics II
Fall 2016

Problem Set 3

(I) Consider a model about the probability of a deterioration in the government structural balance. We will define a deterioration in the structural balance as a period with negative structural balance growth, whenever this is already negative ($g_y < 0$ and $y < 0$). Define the variable:

$$Y = \mathbb{1}_{(g_y<0)}\mathbb{1}_{(y<0)} \tag{1}$$

Then a probit model is proposed to explain the ocurrence of a a deterioration in the structural balance:

$$\mathbb{P}\left(Y = 1|\mathcal{Z}\right) = \Phi(\mathcal{Z}\beta) \tag{2}$$

where $\mathcal{Z}$ is a set of explanatory variables and β are the corresponding coefficients.

1. Suppose $z_1 \in \mathcal{Z}$ is a continuous variable. Derive an expression to calculate the partial effect of z_1 on the probability.
2. Now take $\delta_0 \in \mathcal{Z}$, where δ_0 is now a dummy variable. How would you calculate the partial effect of ths variable on the probability we are studying?
3. Using the Delta method, derive an expression to calculate standard errors on your estimate of the partial effect of z_1 on $\mathbb{P}\left(Y = 1|\mathcal{Z}\right)$.

(II) Suppose now we want to apply the previous model to real data. Download from the IMF, World Economic Outlook Database, October 2015 (from: `https://www.imf.org/external/pubs/ft/weo/2015/02/weodata/index.aspx`) data for all the countries belonging to the OECD in 2015 plus rest of Europe and Americas to estimate your model. To construct your dependent variable, use "General government structural balance" as % of GDP. Include in your matrix of explanatory variables:

1. A constant
2. Exports as as % of GDP
3. GDP growth
4. General Government net lending as % of GDP
5. Government revenue as % of GDP
6. LIBOR
7. Output gap as % of GDP
8. Gross national savings as % of GDP

Now do the following:

1. Estimate a probit model using the variables mentioned. Report your results in a table including standard deviations for each coefficient and the t-Ratios.
2. Analyze the specification of the model. Are the any variables you would remove from the model. Are there any variables that you would obviously include in your estimation?
3. Estimate the marginal effect of a decrease of the GDP on the probability of a deterioration in the structural balance.
4. Calculate standard errors for the estimated probability above.
5. Estimate the marginal effect of being a European country on the probability of a deterioration in the structural balance.

(III) Consider again the database in problem 1. Compute the total sample average across time and geographical areas for the crime rate. We want to predict if the crime rate is above or below the total average. Treat this as a categorical variable.

1. Using the same predictors as in the linear model, use LDA to forecast if the crime rate will be above or below the total average in 1978.
2. Compare your results with a QDA. Which one delivers better predictions?
3. Fit now a KNN and compare your results to the previous two methods. Justify your choice for k.

(IV) We are given a large sample of students. Our database includes the variable $noshow = 1$ if the student missed a recitation in the last two weeks. We also have a vector $\mathbf{X}_i$ including individual characteristics. In particular, we have $beer_i$ as the average beer consumption per day for individual i.

1. We want to examine the effect of beer consumption in the class lectures lost. What is the logic of the model we want to estimate?
2. One proposed model to tackle the problem is:

$$\mathbb{P}\left(noshow = 1|\mathbf{X}, beer, q_1\right) = \Phi(\mathbf{X}\beta + \gamma_1 beer + q_1) \tag{3}$$

   where q_1 is an unobservable that is possibly correlated with *beer*. What happens if we just ignore q_1 and go forward an estimate the probit on *beer* and $\mathbf{X}$?
3. Can *beer* have a conditional normal distribution in the population?
4. Explain how to test if *beer* is exogenous. Does this rely on *beer* having a conditional normal distribution?
5. Suppose that some students live in neighborhoods where bars are forced to close by midnight. Does this suggest an instrumental variable for *beer*?

(V) Use the data included in *smoke.txt* to answer the following questions:

1. Define the binary variable *smokes* if the woman smokes during pregnancy. Estimate a probit model relating smokes to *motheduc*, and $log(faminc)$. At $faminc$ evaluated at the average in the sample, what is the estimated difference in the probability of smoking for a woman with 16 years of education and one with 12 years of education?
2. Do you think $faminc$ is exogenous in the smoking equation? What about *motheduc*?
3. Test the null hypothesis that $log(faminc)$ is exogenous.