Assignment #5
(due to Friday, December 9)

Answers have to be submitted in the PDF format. The homework should contain on its front page the name(s) of the student(s) as well as the SNR number(s).

The data set used in this assignment is labelled “school” can be downloaded from the Canvas (in the DTA, CSV, XLS formats). It comes from Pindyck and Rubinfeld (1998) “Econometric Models and Economic Forecasts” and contains the following variables: a dummy whether children attend private school (*private*), number of years the family has been at the present residence (*years*), log of property tax (*logptax*), log of income (*loginc*), and whether one voted for an increase in property taxes (*vote*). There are two dependent variables of interest – private and vote – which can be modeled individually by two univariate probit models or jointly by one bivariate probit model, for instance.

To solve this exercise, you can use software of your choice, but – in the case you use Stata – the help is provided in this text. For example, you will have to find the following methods in Stata help, download them, and install them:

kdens enhanced kernel density estimator;

sml semiparametric binary-choice regression.

If you do not use the indicated methods in Stata, please note in each question, what software, package, and function you use. For example in the case of R, there are several different implementations of nonparametric estimators. For example, the following functions (packages) can be used in place of the Stata commands mentioned in the text:

biprobit gjrm (GJRM); the bivariate probit is provided as an example in the help of this function;

hetprob hetglm (glmx); the heteroscedastic;

kdens(ity) npudens, npudensbw (np); the unconditional kernel density estimation;

lpoly npreg, npregbw (np); the nonparametric regression method;

probit glm (stats); family binomial(link = "probit") corresponds to the probit model;

sml npindex, npindexbw (np); estimation of single-index models.

Question 1

Using the school data described above, estimate the bivariate probit model explaining the choices *private* and *vote* by variables *years*, *logptax*, and *loginc*. By bivariate probit, we understand the model based on two latent linear equations

$$\begin{aligned} y_{i1}^* &= x_{i1}^\top \beta_1 + \varepsilon_{i1}, \\ y_{2i}^* &= x_{i2}^\top \beta_2 + \varepsilon_{i2}, \end{aligned}$$

observed responses $y_{i1} = I(y_{i1}^* > 0)$, $y_{2i} = I(y_{2i}^* > 0)$, and the error distribution

$$(\varepsilon_{i1}, \varepsilon_{i2}) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

1. Derive and write down the likelihood function for the bivariate probit model.

2. Estimate the bivariate probit model using the likelihood function in point 1 (command `biprobit`). Report the results and note whether there are significant relationships found in both equations or not?

Question 2

Since we are able to predict only the voting behavior, let us now estimate the binary-choice model $y_i = I(x_i^\top \beta + \varepsilon_i > 0)$ explaining the choices *vote* by variables *years*, *logptax*, and *loginc*. Keep in mind that this is a rather small sample of only 95 observations.

1. Let us first look at the income distribution. Since it is often assumed that income follows approximately log-normal distribution, *loginc* should be approximately normally distributed. Plot a kernel density estimate, the corresponding normal distribution function, and 95% confidence intervals. What do you conclude about the normality of *loginc*? Hint: use `kdens` instead of standard `kdensity` to obtain confidence intervals.

2. As a reference point, estimate the probit model (command `probit`) and report its output. Which variables are significant and which not?

3. To verify the normality assumption of the probit, store the linear index $x_i^\top \beta$ (command `predict <varname>, xb`) and the probabilities predicted by probit (command `predict <varname>, pr`). Further, estimate nonparametrically by the Nadaraya-Watson or local polynomial regression the

conditional probabilities $P(y_i = 1|x_i^\top \beta)$ and plot them along with the corresponding confidence intervals and the probit predictions (see command `lpoly`). What does the result imply about the probit assumptions?

4. Estimate the probit again, while accounting for heteroscedasticity (command `hetprob`). Assume that there is heteroscedasticity driven by and only by variable *years*. Report the results. Do you find significant heteroscedasticity? Are the slope coefficients or their significance substantially different compared to probit?

5. Repeat the analysis in point 3 for the heteroscedastic model (including the plot). Do you reject the heteroscedastic probit specification or not? Is the probit link function, that is, the plot of predicted probabilities from the heteroscedastic model, smooth as in point 3 or not? Explain why.

6. Estimate the binary-choice model by the method of Klein and Spady (1993) (command `sml`). Can you directly compare the estimated coefficients with those of probit or heteroscedastic probit? What needs to be done to facilitate the comparison? Are the estimates similar or rather different across the three estimated models?

7. Plot $P(y_i = 1|x_i^\top \beta)$ using the parameters estimated by the Klein and Spady method. Is it similar to the heteroscedastic probit output in point 5? Assuming that the parameters estimated in point 4 and 6 are close to each other, does it mean that we can use the heteroscedastic probit to draw conclusions about $P(y_i = 1|x_i^\top \beta)$ and marginal effects?

8. What causes the strange shape of the link function $F(x_i^\top \beta) = P(y_i = 1|x_i^\top \beta)$? Try to find it out in the data set.

9. Repeat the estimation in points 2-7 for data with variable *years* smaller than 25 years. Do the estimates become closer to each other? Which method would you prefer now? Report outputs supporting your claims.

10. For the β estimates of your preference and variable *years* smaller than 25 years, estimate $P(y_i = 1|x_i^\top \beta)$ by the Nadaraya-Watson, local linear, and local quadratic regression. Plot the results and discuss similarities and differences across the graphs.