# Problem Set III
Econometrics III

Nurfatima Jandarova

February 20, 2017

## 1

### 1.1

**1.1.1** Rewrite the probability $P(Y = 1|X)$ as

$$P(Y = 1|X) = P(Y^* \geq 0|X) = P(\varepsilon \geq -X\theta|X) \overset{\varepsilon \perp X}{=} P(\varepsilon \geq -X\theta) = P(\varepsilon \leq X\theta) = \Phi(X\theta)$$

where $\Phi(\cdot)$ is a cdf from a standard normal distribution.
Since we are given that $P(Y = 1|X)$ for every $X$, we can use the above to write

$$\begin{aligned} X\theta &= \Phi^{-1}(P(Y = 1|X)) \\ X'X\theta &= X'\Phi^{-1}(P(Y = 1|X)) \\ \mathbb{E}(X'X)\theta &= \mathbb{E}\left[X'\Phi^{-1}(P(Y = 1|X))\right] \Rightarrow \\ \theta &= \mathbb{E}(X'X)^{-1}\mathbb{E}\left[X'\Phi^{-1}(P(Y = 1|X))\right] \qquad \text{if } \mathbb{E}(X'X) \text{ is invertible} \end{aligned}$$

**1.1.2** Now, the probability $P(Y = 1|X)$ is expressed as

$$P(Y = 1|X) = P\left(\frac{\varepsilon}{\sigma_\varepsilon} \leq \frac{X\theta}{\sigma_\varepsilon}\right) = \Phi\left(\frac{X\theta}{\sigma_\varepsilon}\right)$$

If we follow the same steps as in 1.1.1:

$$\begin{aligned} \frac{X\theta}{\sigma_\varepsilon} &= \Phi^{-1}\left(P(Y = 1|X)\right) \\ \frac{\theta}{\sigma_\varepsilon}\mathbb{E}(X'X) &= \mathbb{E}\left[X'\Phi^{-1}\left(P(Y = 1|X)\right)\right] \\ \underbrace{\frac{\theta}{\sigma_\varepsilon}}_{\text{unobserved}} &= \underbrace{\mathbb{E}(X'X)^{-1}\mathbb{E}\left[X'\Phi^{-1}\left(P(Y = 1|X)\right)\right]}_{\text{observed}} \end{aligned}$$

Since $\sigma_\varepsilon^2$ is unknown, we can only identify $\frac{\theta}{\sigma_\varepsilon}$, and not $\theta$ individually.

**1.2** Assuming that the sample is iid and using the notation given in the problem $\pi_i = P(y_i = 1|X_i)$, write the likelihood function:

$$\begin{aligned} L &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \Rightarrow \\ \mathcal{L} &= \sum_{i=1}^{n} \left[y_i \ln(\pi_i) + (1 - y_i)\ln(1 - \pi_i)\right] \end{aligned}$$

Notice that if one assumes $\pi_i = \frac{\exp(\alpha+\beta x_i+\gamma x_i^2)}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}$, then

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \ln\left(\exp(\alpha + \beta x_i + \gamma x_i^2)\right) = \alpha + \beta x_i + \gamma x_i^2$$

which is exactly what we are given in the problem set. Substitute it in the log-likelihood to obtain

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\exp(\alpha+\beta x_i+\gamma x_i^2)}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}\right)+(1-y_i)\ln\left(\frac{1}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}\right)\right]= \\ &= \sum_{i=1}^{n}\left[y_i(\alpha+\beta x_i+\gamma x_i^2)-y_i\ln\left(1+\exp(\alpha+\beta x_i+\gamma x_i^2)\right)-(1-y_i)\ln\left(1+\exp(\alpha+\beta x_i+\gamma x_i^2)\right)\right]= \\ &= \sum_{i=1}^{n}\left[y_i(\alpha+\beta x_i+\gamma x_i^2)-\ln\left(1+\exp(\alpha+\beta x_i+\gamma x_i^2)\right)\right]\end{aligned}$$

$$\text{FOC: } \frac{\partial\mathcal{L}}{\partial\alpha}: \sum_{i=1}^{n}\left[y_i-\frac{\exp(\alpha+\beta x_i+\gamma x_i^2)}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}\right]=0$$

$$\frac{\partial\mathcal{L}}{\partial\beta}: \sum_{i=1}^{n}\left[y_i x_i-\frac{x_i\exp(\alpha+\beta x_i+\gamma x_i^2)}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}\right]=0$$

$$\frac{\partial\mathcal{L}}{\partial\gamma}: \sum_{i=1}^{n}\left[y_i x_i^2-\frac{x_i^2\exp(\alpha+\beta x_i+\gamma x_i^2)}{1+\exp(\alpha+\beta x_i+\gamma x_i^2)}\right]=0$$

$$H=-\sum_{i=1}^{n}\begin{bmatrix}\frac{\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i^2\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2}\\ \frac{x_i\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i^2\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i^3\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2}\\ \frac{x_i^2\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i^3\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2} & \frac{x_i^4\exp(\alpha+\beta x_i+\gamma x_i^2)}{(1+\exp(\alpha+\beta x_i+\gamma x_i^2))^2}\end{bmatrix}$$

Hence, the estimate of the variance is given by $-H^{-1}$ since $AVar(\sqrt{n}(\hat{\theta}-\theta))=-(\mathbb{E}H)^{-1}$.
For the second part of the exercise, we also need to obtain the estimate of the variance of $\beta\gamma := g(\theta)$. Notice that

$$g'(\theta)=\begin{bmatrix}0\\ \gamma\\ \beta\end{bmatrix}\neq 0$$

Then, according to Delta Method, $AVar(\sqrt{n}(g(\hat{\theta})-g(\theta)))=-(g'(\theta))'(\mathbb{E}H)^{-1}g'(\theta)$. Hence, we can use

$$-\begin{bmatrix}0 & \hat{\gamma} & \hat{\beta}\end{bmatrix}H^{-1}\begin{bmatrix}0\\ \hat{\gamma}\\ \hat{\beta}\end{bmatrix}$$

as the estimate of the variance of $\beta\gamma$.
The test statistics computed are presented in a table below

| | Wald test | LM test | LR test |
|---|---|---|---|
| $H0:\beta=2$ | 60.1704 | 46.9379 | 37.8047 |
| $H0:\beta\gamma=\frac{1}{20}$ | 13.7725 | 8.8800e+05[1] | 460.5020[2] |

Table 1: Computed test statistics

Since all of these test statistics are distributed according to $\chi_1^2$, all of them should be compared to the same threshold $c=3.8415$ at 95% confidence level. Without further delay, one can conclude that both null hypotheses could be rejected at 95% confidence level.

# 2

**2.1** The regression results are given in Table 2. Recall from lecture notes that the estimated coefficient could no longer be interpreted as marginal effects. However, in case of binary dependent

[2]Could be because the variance is so small that at small deviations the slope of the likelihood function becomes extremely steep.

[2]Perhaps, this is due to the fact that *loglikelihood* $=-678$ is wrong. With the constrained value of $\hat{\theta}$ I get a likelihood of -2.8283e+03, which is puzzling.

variable, the sign of the estimated coefficients tells us the direction of the marginal effect. For example, probability of being arrested is negatively related to the share of reported infractions leading to an arrest. This result sounds somewhat counter intuitive. If the variable *pcnv* reports

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Probit | Probit Marginal | Logit | Logit Marginal | | LPM |
| VARIABLES | arr86 | effects | arr86 | effects | LPM | robust |
| pcnv | -0.553*** | -0.178*** | -0.926*** | -0.175*** | -0.154*** | -0.154*** |
| | (0.0721) | (0.0230) | (0.124) | (0.0231) | (0.0209) | (0.0190) |
| avgsen | 0.0127 | 0.00410 | 0.0196 | 0.00369 | 0.00350 | 0.00350 |
| | (0.0212) | (0.00684) | (0.0352) | (0.00664) | (0.00634) | (0.00589) |
| tottime | -0.00765 | -0.00246 | -0.0115 | -0.00216 | -0.00206 | -0.00206 |
| | (0.0169) | (0.00544) | (0.0282) | (0.00531) | (0.00489) | (0.00423) |
| ptime86 | -0.0812*** | -0.0262*** | -0.134*** | -0.0252*** | -0.0216*** | -0.0216*** |
| | (0.0180) | (0.00576) | (0.0309) | (0.00579) | (0.00447) | (0.00275) |
| inc86 | -0.00463*** | -0.00149*** | -0.00825*** | -0.00156*** | -0.00122*** | -0.00122*** |
| | (0.000478) | (0.000151) | (0.000877) | (0.000159) | (0.000127) | (0.000114) |
| black | 0.467*** | 0.150*** | 0.774*** | 0.146*** | 0.162*** | 0.162*** |
| | (0.0720) | (0.0232) | (0.118) | (0.0223) | (0.0235) | (0.0255) |
| hispan | 0.291*** | 0.0938*** | 0.499*** | 0.0942*** | 0.0893*** | 0.0893*** |
| | (0.0654) | (0.0210) | (0.110) | (0.0206) | (0.0206) | (0.0211) |
| born60 | 0.0112 | 0.00361 | 0.0120 | 0.00226 | 0.00287 | 0.00287 |
| | (0.0557) | (0.0179) | (0.0941) | (0.0177) | (0.0172) | (0.0172) |
| Constant | -0.314*** | | -0.491*** | | 0.361*** | 0.361*** |
| | (0.0513) | | (0.0859) | | (0.0161) | (0.0167) |
| Observations | 2,725 | 2,725 | 2,725 | 2,725 | 2,725 | 2,725 |
| R-squared | | | | | 0.082 | 0.082 |

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Estimation results

percent of "successful" infractions up to year 1986, then it could be that a person has already been serving a sentence in 1986 for past crime (high *pcnv*), so the probability of being arrested in 1986 is lower. However, this effect should have been captured by including *ptime*86, if I understand the variable correctly. So, the estimation results are still puzzling and I have no other ideas how one could explain this.

**2.2** With low % of reported infractions leading to an arrest ($pcnv = 0.25$) the probability of being arrested is 0.4, while with high % of reported infractions leading to an arrest ($pcnv = 0.25$) the probability of being arrested falls to 0.3 (Figure 1). Again, as noted above this could be due to a person already serving a sentence.

**2.3** The percent correctly predicted is 72.6972%. Percent correctly predicted when $arr86 = 0$ is 96.5990%, and when $arr86 = 1$ is 10.3311%. So, the model is more successful in predicting probability of not being arrested, than otherwise.

**2.4** One can see by comparing first and third columns in Table 2 that the coefficients from probit and logit estimations are different in magnitude. However, the magnitudes of coefficients are uncomparable, because they assume different specification of the conditional probability. But one can see that the sign and significance of the coefficients is the same. Marginal effects from two estimations are also very close to each other. This could be due to the fact that we estimate marginal effects at the mean, where the two cdfs are more likely to have a similar slope.

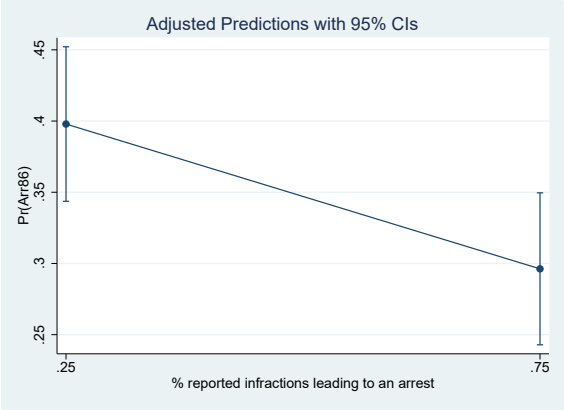


Figure 1: Estimated probability of arrest at $pcnv = \{0.25, 0.75\}$

**2.5** The marginal effects of $inc86$ are shown in Table 3 for $inc86 = \{50, 100, 150, 200\}$, respectively. As one can observe from the table, the marginal effect of legal income o the probability of being arrested is different at different points in the distribution of income, i.e., is not constant. This is one of the reasons to use probit/logit models instead of LPM, which explicitly assumes marginal effects are constant at all points in the distribution.

| VARIABLES | (1) Probit | (2) Logit |
|---|---|---|
| 1bn._at | -0.00152*** | -0.00159*** |
| | (0.000156) | (0.000165) |
| 2._at | -0.00127*** | -0.00126*** |
| | (0.000106) | (9.94e-05) |
| 3._at | -0.00102*** | -0.000954*** |
| | (5.72e-05) | (4.66e-05) |
| 4._at | -0.000768*** | -0.000692*** |
| | (2.77e-05) | (2.66e-05) |
| Observations | 2,725 | 2,725 |

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Marginal effects of $inc86$ for different values of $inc86$

**2.6** The estimation results of LPM both with usual (i.e., homoskedasticity assumed) and heteroskedasticity robust standard errors are reported in Table 2. When using LPM, one should get robust standard errors, because the error term of linear regression model is necessarily heteroskedastic. For illustration purposes, suppose we have a binary variable $y_i$, which we try to fit with the following model: $y_i = x_i\beta + u_i$, where $x_i$ is $1 \times k$ vector of regressors and $\mathbb{E}(u_i|x_i) = 0$. Then,

$$\begin{aligned} Var(u_i|x_i) &= \mathbb{E}(u_i^2|x_i) = \mathbb{E}((y_i - x_i\beta)^2|x_i) = \mathbb{E}(y_i^2|x_i) - 2\beta' x_i' \mathbb{E}(y_i|x_i) + x_i\beta\beta' x_i' \\ &= P(y_i = 1|x_i)(1 - 2\beta' x_i') + x_i\beta\beta' x_i' = x_i\beta(1 - 2\beta' x_i') + x_i\beta\beta' x_i' \\ &= x_i\beta(1 - \beta' x_i') \end{aligned}$$

i.e., conditional variance of an error term is a function of $x_i \Rightarrow$ heteroskedastic.

**2.7** Since we have computed marginal effects in table 2 at mean values of variables, and cdf at the mean tends to be close to a linear function, coefficients from LPM are relatively close the computed marginal effects. What I find puzzling is that the coefficient of $inc86$ in LPM is closer to the marginal effects from probit/logit at $inc86 = 100$, despite the fact that sample mean of

*inc*86 is 54.9671. Frankly, I do not know how to explain this because I would have expected again that the LPM coefficient would be closer to the marginal effect at the mean.
To sum up all of the above, we have

i) non-constant marginal effects as witnessed from probit and logit models;

ii) fitted probabilities from LPM lower than 0 or above 1 (Figure 2);

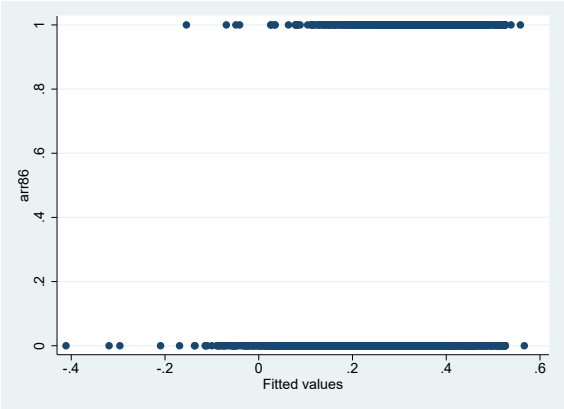


Figure 2: Fitted vs. actual values of $arr86$

iii) I have also computed % correctly classified by LPM model using same cut-off value. Basically, employing the definitions by Stata, the predicting power of the models could be tabulated as

| | Probit | Logit | LPM |
|---|---|---|---|
| Sensitivity | 10.33% | 11.92% | 7.68% |
| Specificity | 96.60% | 94.87% | 97.41% |
| Correctly classified | 72.70% | 71.89% | 72.55% |

Using this information, I am more inclined to conclude that probit/logit are more appropriate to use here. In fact, the two (probit and logit) produce quite similar results with logit performing marginally better at predicting when $arr86 = 1$.

## 3

In general, Altonji, Elder, and Taber (2005) are trying to assess validity of different instruments, such as proximity to Catholic schools, religious affiliation, etc., that had been used by researchers to identify the effect of Catholic schools on education quality. In particular, they observe that in some cases 2SLS estimates were much more noisy and had values hard to give economic interpretation, while probit estimations were more precise and sensible. Then, authors note that "the linearity and normality assumptions of the model are sufficient, and an exclusion restriction is not necessary." (Altonji, Elder, and Taber 2005).
Therefore, they estimate probit model with two additional terms inside: predicted probability holding $X_i$ constant at its mean ($\Phi(\bar{X}_i\hat{\beta} + Z_i\hat{\lambda})$) and predicted probability holding the vector of instruments $Z_i$ constant at its mean ($\Phi(X_i\hat{\beta} + \bar{Z}_i\hat{\lambda})$). Therefore, if the identification comes from exclusion restriction, i.e., variation in the instrumental variables, then the coefficient in front of $\Phi(\bar{X}_i\hat{\beta} + Z_i\hat{\lambda})$ should be the same as the coefficient obtained by running the usual specification of probit. On the other hand, if the nonlinearity assumption embedded in probit is the main identification tool, then the coefficient in front of $\Phi(\bar{X}_i\hat{\beta} + Z_i\hat{\lambda})$ would differ from the usual specification of probit. By doing such comparisons, authors conclude that religious identification seems to be crucial for identification of the effect in a sample of urban white students, but not for urban minorities. In addition, other instruments do not seem powerful for either of the subsamples. This drives them to conclude that most of the identifications in probit models comes from non-linearity assumption.