Metrics, final assignment

Filip Mellgren

2019-01-02

This is the last assignment for the course 5304 Econometrics. It is ungraded, which is why I could do it in R instead of Stata. Note that this is my first R markdown document.

Importing the data

We begin by reading the stata file. Turns out we need the library "readstata13" to do this.

```
library(readstata13)
df <- read.dta13("/Volumes/GoogleDrive/Min enhet/Learning/MSc/Econometrics/Metrics/IHDS_2012_Data.dta")
```

```
head(df)
```

```
##            STATEID          DISTRICT  IDPSU       IDHH     IDPERSON EW6
## 1    Tamil Nadu 33 Tiruchchirappalli 331509 3315090911 331509091102  49
## 2   Maharashtra 27            Satara 273108 2731081701 273108170102  47
## 3   Maharashtra 27          Yavatmal 271401 2714010311 271401031104  28
## 4 Uttar Pradesh 09         Chandauli  96603 0966031911 096603191104  32
## 5   Maharashtra 27             Jalna 271804 2718040211 271804021104  24
## 6   Maharashtra 27          Yavatmal 271401 2714010811 271401081102  42
##            EW8 FH5CK FP5 SPRO5 SPED5                    SPED6
## 1  3rd class 3     4   2    53  No 0              6th class 6
## 2 Secondary 10     3   2    50  No 0            Secondary 10
## 3 Bachelors 15     2   2    32  No 0 1 year post-secondary 13
## 4  9th class 9     3   2    39  No 0        High Secondary 12
## 5       none 0     2   2    27  <NA>                  none 0
## 6       none 0     5   2    48  No 0                  none 0
```

We have a few variables containing demographic information on individuals in India (2012).

```
library(tidyverse)
df <- df %>% filter(EW6 <= 60) %>% filter(EW6 >= 20)
df <- df %>% mutate(three_plus = FH5CK >= 3)
```

We need to reformat the variable EW8 before it is usable in any regression:

```
df$EW8 <- as.character(df$EW8)
df <- df %>% mutate(EW8 = substr(EW8, nchar(EW8) - 1,nchar(EW8)))
df$EW8 <- strtoi(df$EW8)
```

Linear probability model

The first task is to use the lpm to estimate marginal probabilities of having three or more children using woman's age and education level.

The linear probability model is simply the regression dependent variable on independent variables. We use the age of the women (EW6), and their years of education (EW8) to predict the probability that they have at least 3 children. We'd expect the unbiased estimates to be positive for age and negative for years of education. I use the option na.exlude so that predictions will insert NA were no prediction could be made.

```
library(pander)
lpm <- lm(three_plus ~ EW6 + EW8, data = df, na.action = na.exclude)
pander(lpm)
```

Table 1: Fitting linear model: three_plus ~ EW6 + EW8

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1081	0.01604	6.742	1.619e-11
EW6	0.0153	0.0003842	39.82	0
EW8	-0.0346	0.0007262	-47.64	0

The marginal effect associated with being one year older is 0.015 percentage points more likely to have three or more children and the marginal effect associated with having one more year of schooling is -0.035 percentage points less likely to have three or more children. The signs are what we expected them to be.

Next, let's find out the range of the predictions:

```
df <- df %>% mutate(pred_lpm= predict(lpm, type = "response", na.action = na.exclude))
pct_in_range <- round(sum(df$pred_lpm >= 0 & df$pred_lpm <= 1, na.rm=T)/length(df$pred_lpm),3)*100
```

The range of the observations is [-0.12, 1.03] and 96.8% of the observations lie inside the unit interval. It is not ideal to predict probabilities higher than 1 or smaller than 0, so let's move on to theoretically justified models.

Probit

We perform the same steps using a probit model. This comes with the advantage of not yielding any predictions outside the unit interval. However, it is somewhat more tedious to obtain marginal effects as these can't be read from the coefficients and the predicted values are similar to the lpm anyway.

```
probit <- glm(three_plus ~ EW6 + EW8, family = binomial(link = "probit"), data = df, na.action = na.exc
pander(summary(probit))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.214	0.05051	-24.04	1.158e-127
EW6	0.04643	0.001257	36.95	6.703e-299
EW8	-0.1031	0.002376	-43.37	0

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	20979 on 15159 degrees of freedom
Residual deviance:	16713 on 15157 degrees of freedom

Predict values and obtain range of predicted values:

```
df <- df %>% mutate(pred_probit= predict(probit, type = "response", na.action = na.exclude))
pct_in_range_p <- round(sum(df$pred_probit >= 0 & df$pred_probit <= 1, na.rm=T)/sum(!is.na(df$pred_probi
```

The range of the predictions is [0.03, 0.94] and 100% of the observations lie inside the unit interval.

```
library(margins)
margin_probit <- margins(probit, at = list(EW6 = mean(df$EW6), EW8 = mean(df$EW8, na.rm = TRUE)))
margin_probit_spec <- margins(probit, at = list(EW6 = 20, EW8 = 12))
pander(summary(margin_probit)[0:7])
```

factor	EW6	EW8	AME	SE	z	p
EW6	36.66	5.642	0.01844	0.0004991	36.96	5.668e-299
EW8	36.66	5.642	-0.04094	0.0009423	-43.45	0

```
pander(summary(margin_probit_spec)[0:7])
```

factor	EW6	EW8	AME	SE	z	p
EW6	20	12	0.005815	0.0001719	33.82	8.847e-251
EW8	20	12	-0.01291	0.0004711	-27.4	2.601e-165

The AME column gives the marginal effects at values specified by EW6, and EW8.

Logit

We are also told to do the same thing using the logit, essentially repeating the steps from above once more, but with the logit instead of the probit. These two yield very similar predictions, but have different properties that can be useful depending on application.

Theoretically, the choice between probit and logit should depend on a distributional assumption of an error term. If it's assumed to be normally distributed, probit should be used. If it's assumed to be logisticallyy distributed, the logit should be be used. In a binary choice framework, the coefficients for EW6 and EW8 and the variables' intensity for a given observation provide indication what choice will be made (a constant also weigh in). The only way that choice is not what was indicated, is if the unobserved error term outweigh the other evidence. The probability that the unobserved error term outweigh the combined evidence is given by the distribution we assumed the error term follows. So if the sum of the sum of the coefficients multiplied by the data for some observation is 2 and we use a probit model, the probability of observing 0 is the probability the error term is -2 std deviations away from its mean (about 5% in a normal distribution).

```
logit <- glm(three_plus ~ EW6 + EW8, family = binomial(link = "logit"), data = df, na.action = na.exclu
pander(summary(logit))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.023	0.08473	-23.88	5.128e-126
EW6	0.07746	0.002153	35.97	2.08e-283
EW8	-0.1709	0.00407	-42	0

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	20979 on 15159 degrees of freedom
Residual deviance:	16722 on 15157 degrees of freedom

The range of the predictions is given by:

```
df <- df %>% mutate(pred_logit= predict(logit, type = "response", na.action = na.exclude))
pct_in_range <- round(sum(df$pred_logit >= 0 & df$pred_logit <= 1, na.rm=T)/sum(!is.na(df$pred_logit)),3
```

The range of the predictions is [0.04, 0.93] and 100% of the observations lie inside the unit interval.

```
margin_logit <- margins(logit, at = list(EW6 = mean(df$EW6), EW8 = mean(df$EW8, na.rm = TRUE)))
margin_logit_spec <- margins(logit, at = list(EW6 = 20, EW8 = 12))
pander(summary(margin_logit)[0:7])
```

factor	EW6	EW8	AME	SE	z	p
EW6	36.66	5.642	0.01926	0.0005352	35.98	1.485e-283
EW8	36.66	5.642	-0.0425	0.001009	-42.13	0

```
pander(summary(margin_logit_spec)[0:7])
```

factor	EW6	EW8	AME	SE	z	p
EW6	20	12	0.005315	0.0001516	35.06	2.9e-269
EW8	20	12	-0.01173	0.0004102	-28.6	7.479e-180

The AME column gives the marginal effects at values specified by EW6, and EW8.

To summarize the marginal effect at the mean value of both variables:

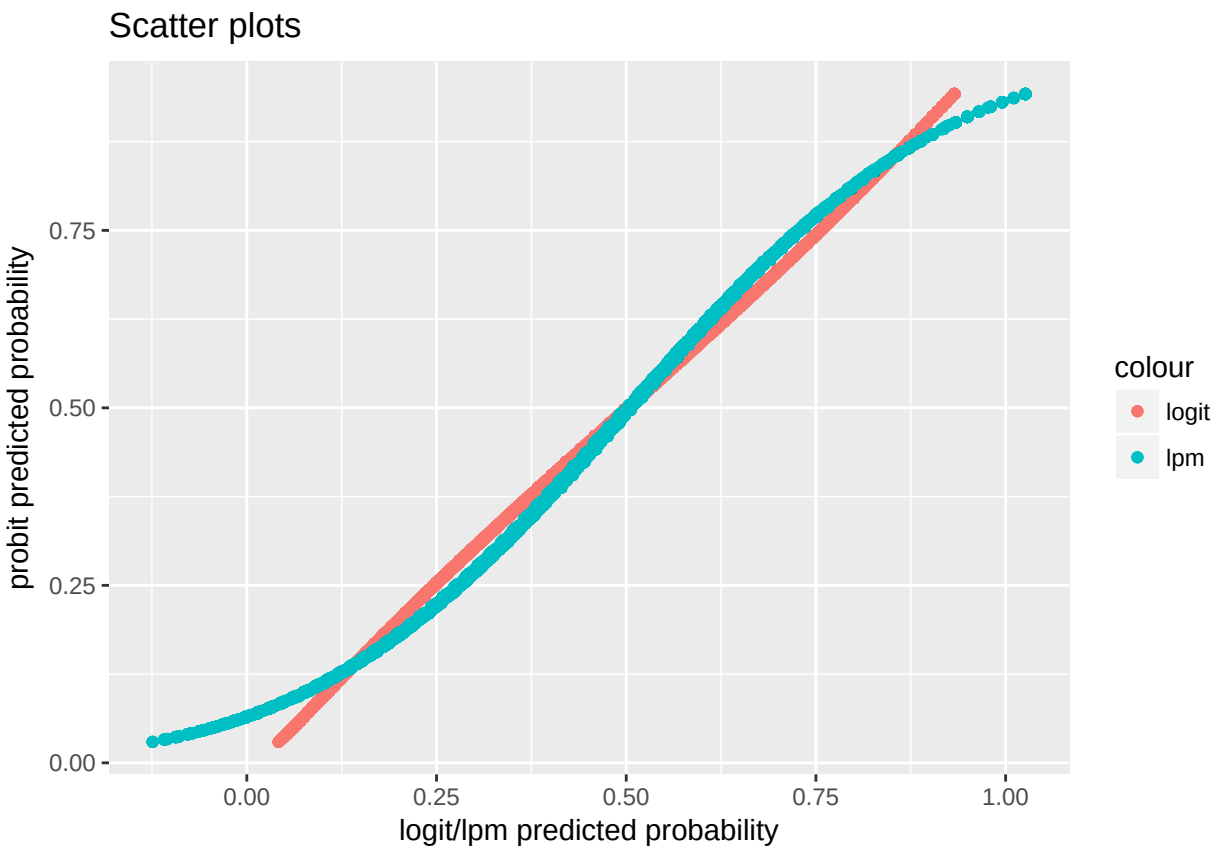
- LPM: 0.015 and -0.035
- Probit: 0.0184439, -0.040941
- Logit: 0.0192591, -0.0425004

Probit and logit look very similar indeed. And LPM is not far off either, suggesting this simple model might be preferable if we only care about marginal effects, especially around the mean.

Scatter plot

Before wrapping up, I plot both the logit and the lpm against the probit to see how the predictions differed.

```
df %>%
ggplot() +
geom_point(aes(pred_logit, pred_probit, colour = "logit")) +
geom_point(aes(pred_lpm, pred_probit, colour = "lpm")) +
xlab("logit/lpm predicted probability") +
ylab("probit predicted probability") +
ggtitle("Scatter plots")
```



The plot confirms the intuition that the results are similar around the mean. However, it is clear that the lpm yields bad predictions around the extreme values.