Microeconometrics - PS2

Pietro Bianchini, Marco Valentino Caravella and Elia Gioele Larcinese

17th December 2024

1 Part 1: Probit Model

1) Estimate Model 2 using ML. Are the regressors jointly significant? How can you interpret the estimated coefficients?

We first sort the dataset by year, pooling together the cross-sections corresponding to the three waves. This pooled cross sections dataset can then be analysed using the same techniques as with a pure cross section, but including dummies to account for aggregate changes over time. We thus create the time dummies corresponding to years 2011, 2013 and 2015.
To choose which variables to include in the vector $x_{i,t}$, we start with a baseline model that includes only the must-have regressors. We then iteratively expand the number of controls, grouping candidate controls in categories that we believe may have a significant impact on life satisfaction: household composition, household wealth and individual health. Each time a new control is added, we evaluate whether it significantly improves goodness of fit using the pseudo R-squared and check whether it leads to changes in the sign or significance of any of the regressors already included. Throughout the analysis, the estimated coefficients' magnitude is not directly interpretable as a partial effect. However, we interpret the coefficients' sign - as it gives the direction of the associated partial effect - and significance level - since it is the same as that of the partial effect. We also select controls being careful to avoid multicollinarity and transform some of them to have a clearer and more meaningful interpretation. Besides these criteria, our choice of regressors is guided by the intention to portray the most complete socio-demographic and health-related picture of the individual.
The starting baseline model only includes as explanatory variables *income*, the must-have regressors - *age*, *gender*, *yesu*, *mstat*, *hstatus* and *gali* - and the dummies for years 2013 and 2015. We thus choose 2011 as the reference year.
Life satisfaction is likely to be affected by perceived stability, and two important factors affecting this dimension are employment status and housing situation. We are mainly interested in whether homeowners are more or less satisfied that those who do not own a house, whether employed or self-employed people are more or less satisfied that those that are jobless, and whether retired people are more or less satisfied than those who are still working. Therefore, we transform the categorical variable *otrf* in a dummy *owner* that equals 1 if the individual owns his house and 0 otherwise, and we transform the categorical variable *cjs* in two dummies: one called *employed* that equals 1 if the person is employed and 0 otherwise; and one called *retired* if the person is retired and 0 otherwise. We then add these three dummies to the baseline model, noting that the pseudo R-squared increases, as well as the percentage of cases accurately predicted. We also observe that including them leads the coefficient on *gender* in the new model to change sign from negative to positive and to become significant. The resulting model, which we call *model 2*, is thus the following:

$$Pr[life\_sat\_i, t = 1 | income_{i,t}, \mathbf{x_{i,t}}, \lambda_t] = \Phi(\alpha + \gamma income_{i,t} + \mathbf{x_{i,t}}\beta + \lambda_t \iota) \tag{1}$$

where $\mathbf{x_{i,t}}$ is a vector of the following variables:

$$\underset{1\times 9}{\mathbf{x_{i,t}}} = \begin{pmatrix} age & gender & yedu & mstat & hstatus & gali & owner & employed & retired \end{pmatrix} \tag{2}$$

and $\lambda_\mathbf{t}$ includes the time dummies for 2013 and 2015.
We then expand *model 2* iteratively, adding controls that are grouped by category. First, we turn to those controls under the label of household composition: *htype*, *nchild*, *ngrchild* and *nursinghome*. First, in order to construct a model that is as parsimonious as possible we discard *htype* since including it does not imply substantial modifications to the pseudo R-squared nor to the sign or significance of coefficients. Moreover,

mstat already captures the relationship status of the individual and at least partially whether the individual lives alone or with a partner. We also discard *nursinghome* because including it does not increase the pseudo R-squared. Then, even though they lead to only small increases in the pseudo R-squared, we decide to include in the model *nchild* and *ngrchild* since we believe life satisfaction is likely to be affected in either direction by whether the individual successfully reproduces.
We then turn to covariates in the category of household wealth: *thexp*, *fahc*, *fohc*, *hprf* and *hnetw*. First, we believe that the net worth of the household could be a good measure of financial stability, which is conducive to life satisfaction. We thus include it in our model, observing an increase in the pseudo R-squared. Second, as higher income or higher household net worth do not necessarily translate into higher spending, we are interested in whether spending has an impact on life satisfaction net of the effect of other variables. To this regard, *thexp* is likely to be a more complete measure than *fahc*, *fohc* or *hprf*. We thus decide to include it, discarding the other three. Including these controls for household wealth increases the pseudo R-squared and reduces the significance of the coefficients on income and the number of children.
Finally, we turn to the covariates in the health category: *phinact*, *bmi*, *esmoked*, *doctor* and *hospital*. We believe that being phisycally active is key for a healty and productive life. Therefore, we start by including *phinact* in the model, noting that this leads to an increase in the pseudo R-squared. This is the only health-related covariate we end up including, as the others do not lead to substantial modifications in the pseudo R-squared nor in the sign or significance of the coefficients. Therefore, in the final model that we estimate, $\mathbf{x_{i,t}}$ is a vector of the following variables:

$$\underset{14\times 1}{\mathbf{x}_{\mathbf{i,t}}^{\top}} = \begin{pmatrix} age \\ gender \\ yedu \\ mstat \\ hstatus \\ gali \\ owner \\ employed \\ retired \\ nchild \\ ngrchild \\ thexp \\ hnetw \\ phinact \end{pmatrix} \tag{3}$$

The regressors in the final model are jointly significant. Indeed, as indicated by the result of the Wald test, we reject the null hypothesis that their coefficients are all equal to zero. The estimated coefficients and their robust standard errors are reported in table 1.

Table 1: Probit Estimation of Life Satisfaction

	(1) life_sat
life_sat	
Income level	0.0682*
	(0.0294)
Age	0.0119***
	(0.00175)
Gender	0.0649*
	(0.0267)
Years of education	0.00734*
	(0.00320)
Marital status	-0.0773***
	(0.00731)
Self-perceived health status	0.276***
	(0.0145)
Limitation with activities	-0.161***
	(0.0295)
owner	0.185***
	(0.0378)
employed	0.216***
	(0.0395)
retired	0.186***
	(0.0333)
N. of children	0.0248*
	(0.0117)
N. of grandchildren	-0.00660
	(0.00591)
Total household expenditure	0.0000141***
	(0.00000315)
Household net worth	0.000000302***
	(6.03e-08)
Physical inactivity	-0.180***
	(0.0311)
d2013	-0.119***
	(0.0310)
d2015	-0.0220
	(0.0306)
Constant	-1.744***
	(0.149)
Observations	12244

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2) Compute manually the partial effect at the average of the continuous variable income and interpret your findings.

Consider the following response probability equation:

$$P(Y = 1|x) = G(x\beta) = P(x) \tag{4}$$

This equation is called *index model* because the probability of observing $Y = 1$ depends only on the *index* $x\beta$. The partial effect is:

$$\frac{\partial P(x)}{\partial x_j} = \frac{\partial G(x\beta)}{\partial (x\beta)} \times \frac{\partial (x\beta)}{\partial x_j} = g(x\beta) \cdot \beta_j \neq \beta_j \tag{5}$$

where $G(.)$ and $g(.)$ are respectively the cdf and the pdf of a random variable. In 5 we notice that the coefficient β_j alone is not the partial effect of variable x_j. This is why we will follow some steps before computing the partial effect. Notice that we want to evaluate the partial effect at the average:

$$\frac{\partial \hat{P}(\bar{x})}{\partial x_j} = g(\bar{x}\hat{\beta}) \cdot \hat{\beta}_j \tag{6}$$

Since we use a probit model, we assume that $G(.) = \Phi(.)$, which is the cdf of a standard normal random variable. Hence, $g(.)$ is the pdf of a standard normal.
To compute the PEA, we compute the index using the average of all explanatory variables. For categorical variables, the argument could be made that the mode is a better measure of central tendency. Moreover, in the case of discrete quantitative variables - such as *nchild* and *ngrchild*, the mode would allow us to use meaningful and actually observed values. However, given that the intention is computing the partial effect of income for an artificially created "average" individual, we discard the option of using the mode, losing interpretability in favor of generality. For instance, if we used the mode of gender - which is 2 - instead of its average, we would compute the partial effect of income for the average woman. Once we have found the averages of all explanatory variables, we compute the index in the following way:

$$\begin{aligned} z = {} & \hat{\beta}_0 + \hat{\gamma} \cdot 0.3110063 + \hat{\beta}_{age} \cdot 66.90983 + \hat{\beta}_{gender} \cdot 1.55178 + \hat{\beta}_{yedu} \cdot 8.722231 + \hat{\beta}_{mstat} \cdot 1.916204 \\ & + \hat{\beta}_{hstatus} \cdot 2.751878 + \hat{\beta}_{gali} \cdot 0.4097517 + \hat{\beta}_{owner} \cdot 0.8371447 + \hat{\beta}_{employed} \cdot 0.218066 + \hat{\beta}_{retired} \cdot 0.513231 \\ & + \hat{\beta}_{nchild} \cdot 1.955162 + \hat{\beta}_{ngrchild} \cdot 2.032424 + \hat{\beta}_{thexp} \cdot 8403.044 + \hat{\beta}_{hnetw} \cdot 249674.9 \\ & + \hat{\beta}_{phinact} \cdot 0.2114505 + \hat{\iota}_{2013} \cdot 0.3484156 + \hat{\iota}_{2015} \cdot 0.3879451 = 0.22022214 \end{aligned} \tag{7}$$

We now plug z into the pdf function in 6 and multiply it by $\hat{\gamma}$:

$$\frac{\partial P(x)}{\partial income} = \phi(z) \cdot \hat{\gamma} = \phi(.22022214) \cdot .0681992 = .02655573 \tag{8}$$

Thus, ceteris paribus, a unit increase in income leads to an increase of 2.66 percentage points in the probability that the average individual is really satisfied with her life - which is statistically significant at the 5% level against a two-sides hypothesis, according to the std error of $\hat{\gamma}$ in the final model. As expected, the partial effect of income is positive and statistically significant, providing suggestive evidence that income plays an important role in determining a satisfactory life.

3) Compute manually the partial effect at the average of gali and interpret your findings.

As in the previous point, to find the partial effect of *gali* at the average, we compute the index using the averages of all the regressors. However, the procedure to compute the partial effect of a binary variable is slightly different. Namely, the partial effect is given by:

$$G(\hat{\beta}_0 + \cdots + \hat{\beta}_6 + \cdots + \hat{\lambda}_2 \cdot d2015) - G(\hat{\beta}_0 + \cdots + \hat{\beta}_5 \cdot hstatus + \hat{\beta}_7 \cdot owner + \cdots + \hat{\lambda}_2 \cdot d2015) \tag{9}$$

where $\hat{\beta}_6$ is the estimated coefficient on *gali*; in the first cdf *gali* takes value 1 and in the second it takes value 0; and all other explanatory variables enter the index with their average. The computed partial effect is $-.0628773$, indicating that ceteris paribus an average individual that has limitations with activities is 6.29 percentage points less likely to be really satisfied with her life with respect to an average individual that does not have limitations with activities. This partial effect is statistically significant at 1% against a two-sided hypothesis according to the standard error of the estimated coefficient in the final model, indicating that being free of any form of limitation in performing daily activities is key for a satisfactory life.

4) Use the command margins to compute both the average partial effects (APE) and the partial effects at the average (PEA) of all explanatory variables. Explain the difference between the two.

We now introduce a "new" partial effect: the Average Partial Effect (APE). The APE of x_j is given by:

$$\frac{1}{N}\sum_{1=1}^{N} g(x\hat{\beta})\hat{\beta}_j \tag{10}$$

We can see that the APE is different from the Partial Effect at the Average (PEA). This is because the PEA of x_j is the marginal effect of x_j on the predicted probability evaluated at the average values of all the covariates. While, for a given variable x_j, the APE is the average computed across all the observations in the dataset of the marginal effects of that variable on the predicted probability. Hence, we expect them to be different.
We use *margins* to find these two values. APE results are shown in table 2, while those for PEA are in table 3.

Table 2: Average Partial Effects

	(1) life_sat
Income level	0.0240*
	(2.32)
Age	0.00418***
	(6.81)
Gender	0.0229*
	(2.43)
Years of education	0.00259*
	(2.30)
Marital status	-0.0273***
	(-10.72)
Self-perceived health status	0.0974***
	(19.90)
Limitation with activities	-0.0568***
	(-5.49)
owner	0.0653***
	(4.90)
employed	0.0763***
	(5.49)
retired	0.0657***
	(5.62)
N. of children	0.00874*
	(2.11)
N. of grandchildren	-0.00233
	(-1.12)
Total household expenditure	0.00000499***
	(4.49)
Household net worth	0.000000107***
	(5.04)
Physical inactivity	-0.0634***
	(-5.81)
d2013	-0.0418***
	(-3.83)
d2015	-0.00776
	(-0.72)
Observations	12244

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Partial Effects at the Average

	(1) life_sat
Income level	0.0266*
	(2.32)
Age	0.00461***
	(6.78)
Gender	0.0253*
	(2.43)
Years of education	0.00286*
	(2.30)
Marital status	-0.0301***
	(-10.57)
Self-perceived health status	0.108***
	(19.05)
Limitation with activities	-0.0627***
	(-5.47)
owner	0.0721***
	(4.89)
employed	0.0842***
	(5.47)
retired	0.0726***
	(5.59)
N. of children	0.00966*
	(2.11)
N. of grandchildren	-0.00257
	(-1.12)
Total household expenditure	0.00000551***
	(4.48)
Household net worth	0.000000118***
	(5.02)
Physical inactivity	-0.0700***
	(-5.79)
d2013	-0.0462***
	(-3.83)
d2015	-0.00857
	(-0.72)
Observations	12244

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5) How good are the predicted levels of life satisfaction? Show the sample type I (false positive) and type II (false negative) error.

By using table 4 we can compute the ability of our model to predict accurately the level of life satisfaction.

	Predicted Life_Sat		
Observed Life_Sat	0	1	Total
0	2423	2720	5143
	19.79	22.21	42.00
1	1381	5720	7101
	11.28	46.72	58.00
Total	3804	8440	12244
	31.07	68.93	100.00

Table 4

Type I error (false positive) is the rejection of a true null hypothesis: we predict that the individual is satisfied ($\widehat{life\_sat} = 1$) while she is not. Whereas, the Type II error (false negative) is the failure to reject a false null hypothesis: the model predicts that the individual is not satisfied ($\widehat{life\_sat} = 0$), but she actually is. The computed Type I and II errors are:

- Type I error: $= \frac{\#(\widehat{life\_sat}=1|life\_sat=0)}{\#(life\_sat=0)} = \frac{2720}{5143} = .5288742$

- Type II error: $\frac{\#(\widehat{life\_sat}=0|life\_sat=1)}{\#(life\_sat=1)} = \frac{1381}{7101} = .19447965$

These results show that our model predicts better the probability of observing an individual which is not fully satisfied.
Moreover, by summing up the correctly predicted dependent variable and dividing it by the total number of observations, we get a measure of the goodness of the model:

$$\frac{\#(\widehat{life\_sat} = 1|life\_sat = 1) + \#(\widehat{life\_sat} = 0|life\_sat = 0)}{N} = \frac{2423 + 5720}{12290} = .6625712 \qquad (11)$$

That is, the 66.26% of the predictions are correct.

6) Consider now the following model:

$$Pr[life\_sat_{i,t} = 1|income_{i,t}] = \Phi(\alpha + \gamma income_{i,t}) \qquad (12)$$

Estimate it for males and females separately. Is the average partial effect of income on life satisfaction the same across gender? Does it vary across levels of income? Represent in a scatter plot the predicted probabilities as a function of income for males and females separately. Comment on your findings.

The average partial effect for males is equal to .0701868. It means that, ceteris paribus, a unit increase in income increases on average the probability of being fully satisfied by 7 percentage points for males. This APE is significant at the 1% level against a two-sided hypothesis. Instead, the APE for females is .0827362, which indicates that, ceteris paribus, a unit increase in income increases on average the probability of being fully satisfied by 8.27 percentage points for females. This APE is significant at 1%. We conclude that the APE of income on life satisfaction is not the same across gender, and is higher for women.
We now verify whether a unit increase in income has a different APE on life satisfaction at different starting levels of income. To do so, we use income quartiles for males and females as starting levels - in practice, since the last quartile is much higher than the third one, we use the 99th percentile in its place. For both genders, the APE of income on the probability of an individual being fully satisfied is roughly constant over the first three quartiles (it only decreases very slightly) and then drops significantly at the 99th percentile. However, throughout all levels, it remains positive and statistically significant at the 1% level. This may indicate that below a certain threshold income has roughly constant marginal returns to satisfaction, while after that threshold income has decreasing marginal returns, although it keeps contributing significantly to

the probability of being fully satisfied even at higher starting levels. This may provide suggestive evidence that, as income increases, other factors possibly play a more important role in determining a satisfactory life. Another thing to notice is that the APE of income is higher for females at any level. These results are confirmed by the scatter plot in figure 1, which represents the predicted probability of an individual being fully satisfied as a function of income for males and females separately. If we fit a curve through the points in the scatter plot, we would observe that the slope at any point is steeper for females, suggesting that the partial effect of income is higher than for males at any income level. This may suggest that, at any given level, women care more than men about an increase in income, and explains the convergence between males and females in the predicted probability of being fully satisfied as income increases.

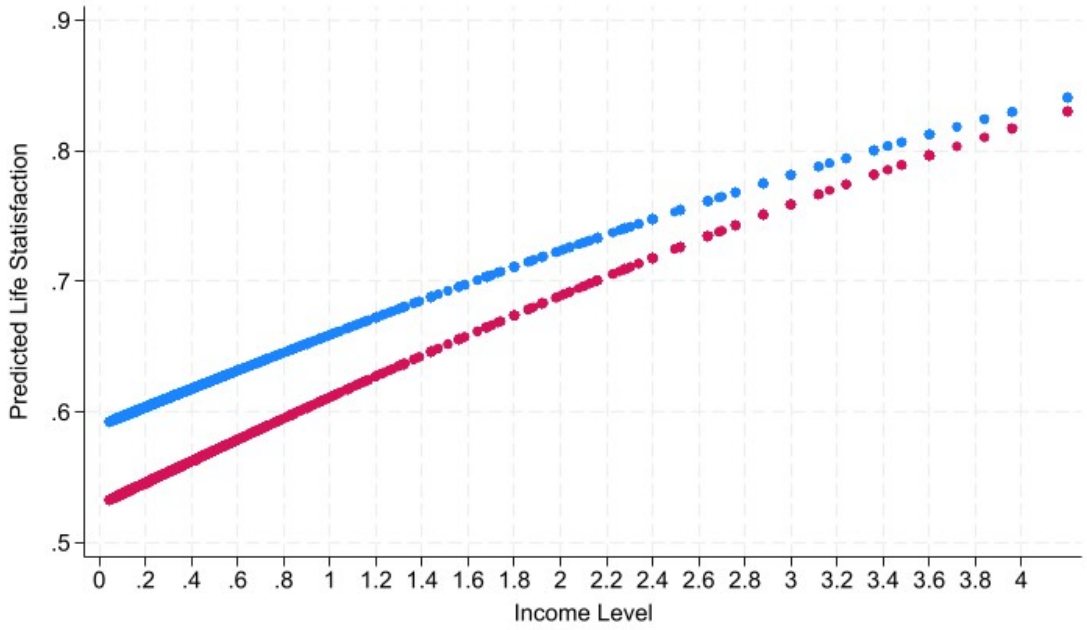

Figure 1: Predicted Life Satisfaction for males (blue) and females (red)

bonus. How would you repeat the previous point including covariates? Would it be so easy to represent it graphically?

We now repeat the previous point including the rest of the covariates. The APE results are:

- for males: .0275018. Ceteris paribus, a unit increase in income increases on average the probability of a male being fully satisfied by 2.75 percentage points. This APE is significant at the 10% level against a two-sided hypothesis.

- for females: .0191103. Ceteris paribus, a unit increase in income increases on average the probability of a female being fully satisfied by 1.91 pp. This APE is not statistically significant.

Therefore, once we account for other factors that may affect the probability of being fully satisfied, the APE of income is much lower for both genders than in the previous point, and marginally significant only for males.
Consider now the APE across income quartiles. By looking at the results we notice that, like in the previous point, the APE of income is roughly constant over the first three quartiles (decreasing only slightly) and then drops at the 99th percentile, although the drop is smaller than in the previous point. However, we also notice that the APEs are much lower than before, and not significant for females. Moreover, contrary to the findings of the previous point, the APEs for females are smaller than that those for males at any level of income. In this case, the relationship between income and the predicted probability of the individual being fully satisfied is not easy to represent graphically. The reason is that, once we add other covariates besides income, we account for the fact that individuals differ along many dimensions that may impact the probability of being fully satisfied. Therefore, we would need a scatter plot with many dimensions, which is hard to construct. While this high dimensionality cannot be represented meaningfully in a two dimensional scatterplot, we can still try to extract some relevant information. Indeed, as shown in figure 2, we observe that for the same level of income observations display significant variability in terms of the probability of being fully satisfied. This may indicate that other factors besides income play a key role in determining a

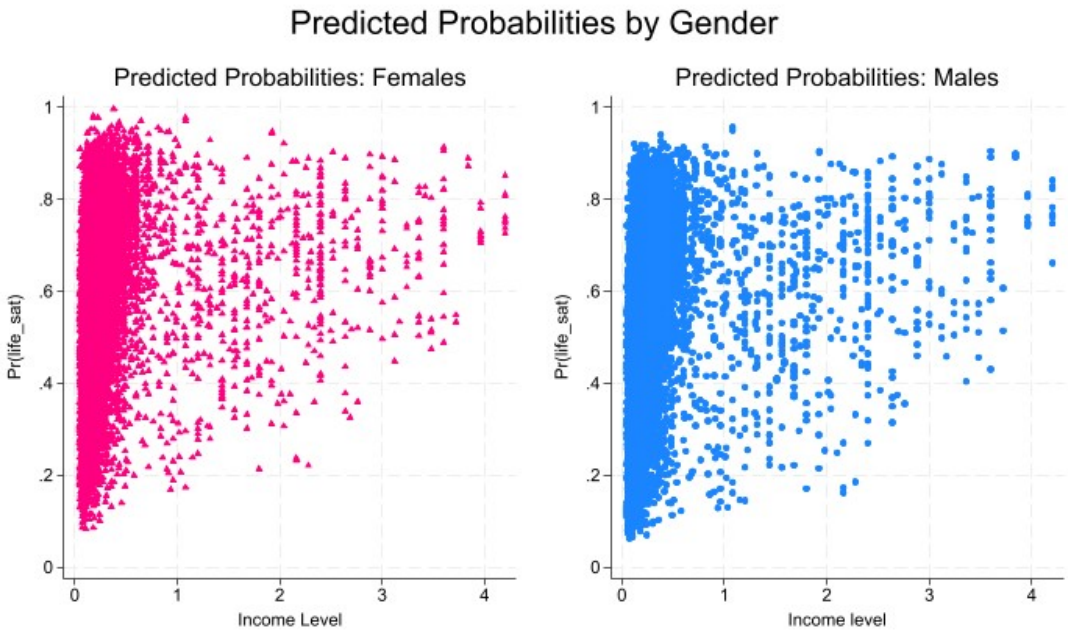

Figure 2

satisfactory life, and this holds for any level of income. Moreover, someone with a very low income, may well be more likely to be satisfied with his life than someone with a very high level. The fact that the estimated APE of income tends to be lower for women than for men at any level of income may also be taken as suggestive evidence that, once we account for other factors that determine a satisfactory life, men care about income more than women.