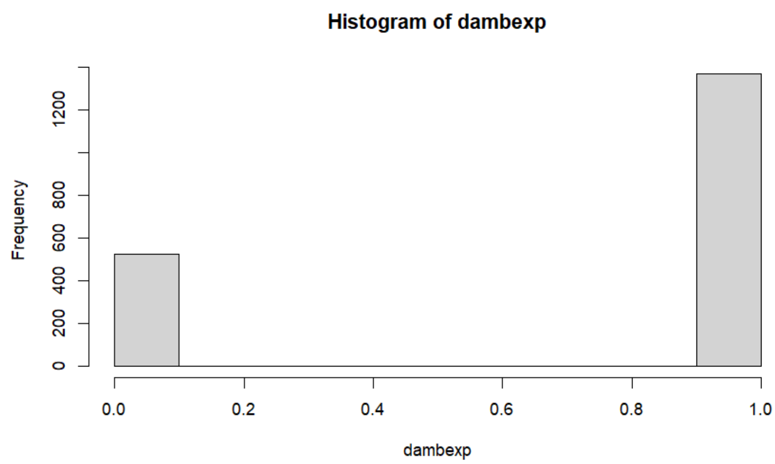
ADVANCED ECONOMETRICS

HOMEWORK 4

Fabio Marcaurelio, 599027

1 A- Create the new variable dambexp = 1 if ambexp > 0, 0 otherwise. What is the share of individuals with positive expenditure?

as we can see in the picture below the share of people with positive expenditure is 1367 (73 %):



before start the analysis we need to do some consideration. we want to study the relation between the number of individuals with a positive expenditure in ambulatory and the number of chronic diseases, age and education with a non linear model (probit and logit). i think that we can have some problem in the case of zero chronic diseases, since we expect that these individuals could have a positive (but not high) of expenditure for other medical issues, or expenditure for other people (a parents with a child) for example. these are the premises, now over to study these relationship we will pay attention to the fitting of the model given the previous doubts.

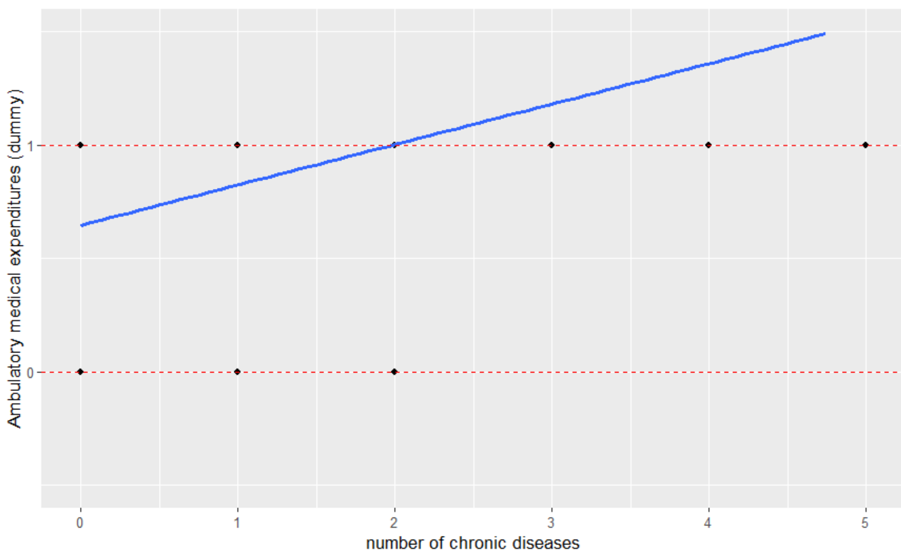
2 B- Estimate a linear probability model (LPM) for the probability of positive AM expenditure as a function of the number of chronic diseases (totchr), age (age) and education (educ). Are the single coefficients statistically different from zero? Are they jointly different from zero?

The predicted values generated by an LPM are interpreted as the predicted probability that $y = 1$, and, for a given regressor x_i, β_i is the change in that predicted probability for a unit change in x_i

Table 1:

	Dependent variable:
	dambexp
totchr	0.157***
	(0.014)
age	0.043***
	(0.009)
educ	0.031***
	(0.004)
Constant	0.068
	(0.060)
Observations	1,893
R^2	0.127
Adjusted R^2	0.125
Residual Std. Error	0.419 (df = 1889)
F Statistic	91.456*** (df = 3; 1889)
Note:	*p<0.1; **p<0.05; ***p<0.01

starting from the intercept, this has no interpretation in this framework since we can not assume that *age* is zero in this case, we can not talk about medical expenditure of an individual with zero years old, in fact it is not significant (it can became if we shift the intercept to 21 value, dataset$age-21 since 21 is the minimum age but we are not interested about this). considering the other variables in this model we can interpret the estimated coefficients as the change in probability to have a positive expenditure given a unit variation of a variable. in this model then given a change of one unit in *totchr* leads to an increase of the probability of positive expenditure by 15%. in addiction all the coefficients have a positive sine since if age increase we expect to have more medical expenditure (ageing) and with an higher level of education we expect to have an higher wage and a more possibility to access healthcare costs.



The model indicates that there is a positive relation between the number of chronic diseases and the probability of a positive expenditure. Individuals with a high value of *totchr* are more likely to have positve expenditure. The regression line (blue) is fitted outside the possible range of *ambex*. already here we see that the zero *rorchr* does not fit very well.
as we can see in the summary robust to heteroskedasticity all the coefficients are statistically different from zero, the result is more clear in the F tests.

3 C- Manually compute the fitted values from the regression in b) and report summary statistics. Comment on the results: can you properly interpret the fitted values?

we can manually compute the fitted value by:

$$fittedlpm < -coef[1] + coef[2] \cdot totchr + coef[3] \cdot age + coef[4] \cdot educ$$

and summary shows that this model can not be interpreted as a probability measure since violate the basic assumption of probabilities (constrained in 0,1) since the max value is above 1 (1.5658). then the relationship between *totchr* and *ambex* can not be linear, then we have to consider other models like probit and logit.

4 D- Estimate the model in b) using the Probit model. What is the effect of totchr on the probability of positive AM expenditure?

One drawback of the LPM is that it assumes that the probability of $Y = 1$ is linear. This issue can be solved by modeling the probability with a nonlinear functional form. Another difference to the LPM or OLS estimation is that the parameters of a logit or probit model are computed using Maximum Likelihood Estimation (MLE).

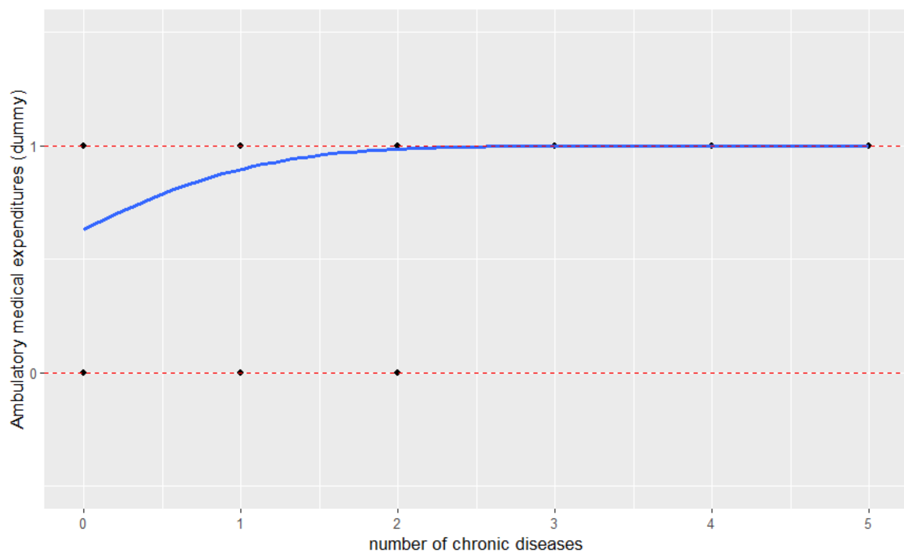
Table 2:

	Dependent variable:	
	dambexp	
	OLS	*probit*
	(1)	(2)
totchr	0.157***	0.898***
	(0.014)	(0.080)
age	0.043***	0.145***
	(0.009)	(0.030)
educ	0.031***	0.105***
	(0.004)	(0.013)
Constant	0.068	−1.600***
	(0.060)	(0.208)
Observations	1,893	1,893
R^2	0.127	
Adjusted R^2	0.125	
Log Likelihood		−961.380
Akaike Inf. Crit.		1,930.760
Residual Std. Error	0.419 (df = 1889)	
F Statistic	91.456*** (df = 3; 1889)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

The probit regression coefficients give the change in the z-score or probit index for a one unit change in the independent variable.

$$F(x\beta) = \Phi(x\beta) = \int_{-\infty}^{x\beta} \phi(\epsilon) d\epsilon$$

as we can see in the above comparison all the coefficient remain statistically different from zero but they are not probabilities. we need to compute the the fitted value as in the LPM model and we can see that now model estimate a probability measure (constrained in 0,1). starting to say that for a one unit increase in *totchr*, the z-score increases by 0.898.



as we can see in the picture above from 2 chronic diseases we start to have 100% probability to have a positive expenditure. also in this framework we denote a propensity of individuals to have a positive expenditure without chronic diseases.

5 E- Manually compute the average marginal effect at the mean and the average marginal effect of the variable totchr from the model in d).

Opposed to linear (OLS) regression, the probit regression model has no single measure which expresses the influence of an independent on the dependent variable. However, one candidate is the average marginal effect. Average marginal effects show the change in a probability when the predictor or independent variable is increased by one unit. Since a probit is a non-linear model, that effect will differ from individual to individual. What the average marginal effect does is compute it for each individual and than compute the average.

$$\frac{\partial E[y_i|x_i]}{\partial x_{ij}} = \phi(x_i\beta)\beta_j$$

as we said before The β coefficients do not have a direct interpretation but is enough to know the sine. by the summary we have an AME with 0.25 of mean and bounded in 0 and 0.35, then for one unit more of *totchr* (one chronic disease more) the probability to have the a positive expenditre increase by 25%.

6 F- Assess the goodness of fit of the model estimated in d). Include comments on the pseudo.R2 and compute the share of correctly predicted outcomes

for test the fit of the model we can use two methods, Pseudo-R^2 and the share of correctly predicted outcomes.

1. pseudo-R^2:

$$1 - \frac{inL}{inL_0}$$

   starting consider the unrestricted probit model and take the log likelihood, this will give us a -961.37 $in(L)$, now do the same for the restricted model with only constant and it give us -1118.62 $in(L_0)$ this is the value without independent variables. from this we see that log likelihood improve if we add some independent variables. then calculate the pseudo-R^2 now is easy:

$$1 - \frac{-961.37}{-1118.62} = 0.140569$$

   since this value is not very high we can say that this model does not fit very well the data, in general if the fit is very high defined we will see an $in(L)$ close to zero.

2. starting defining the true prediction as the fitted values greater than 0.5. goal is to compare when the fitted values and the real observation gives the same result, then evaluate the accuracy of the model. then we define a matrix:

```
           dambexp
dambexphat          0          1
     FALSE 0.04648706 0.03328051
     TRUE  0.23137876 0.68885367
```

   this matrix above say that when true value are zero and the prediction is zero we have 4% observation in the sample that predict correctly, the same for one, in this case we have a 70%. then model works very well for positive expenditure but has some problem of fitting in the zero case. in fact, we have 23% of observation that incorrectly give us one. then we can say the probit model predict correctly the 73% of observation correctly.

7 G- Estimate the model in b) using logit. Compare the (average) marginal effect of totchr obtained with LPM, probit and logit.

The logit model is based on a cumulative standard logistic distribution.

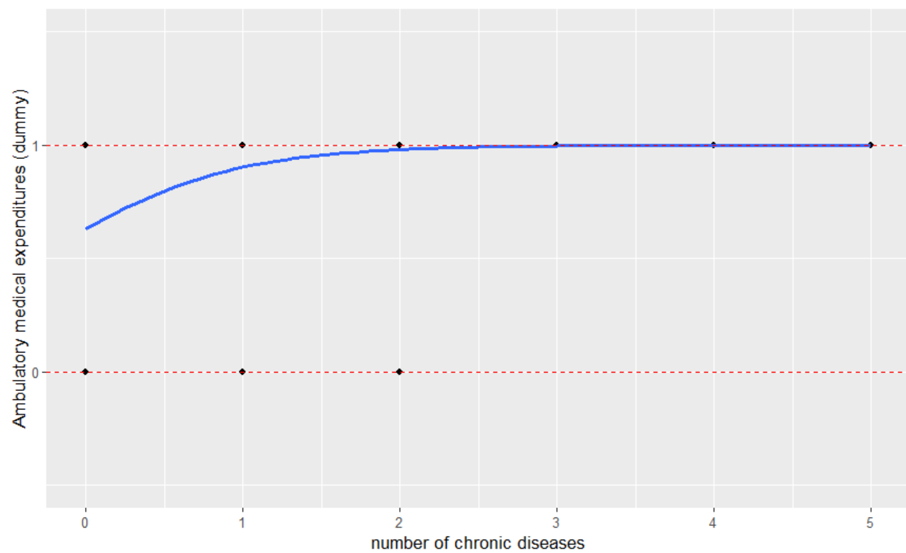
$$F(x\beta) = \frac{exp(x\beta)}{1 + exp(x\beta)}$$

as in probit all coefficients are statistically different from zero and equal sine,

Table 3:

	Dependent variable:
	dambexp
totchr	1.644^{***}
	(0.155)
age	0.243^{***}
	(0.051)
educ	0.176^{***}
	(0.022)
Constant	-2.712^{***}
	(0.355)
Observations	1,893
Log Likelihood	−961.005
Akaike Inf. Crit.	1,930.010
Note:	*p<0.1; **p<0.05; ***p<0.01

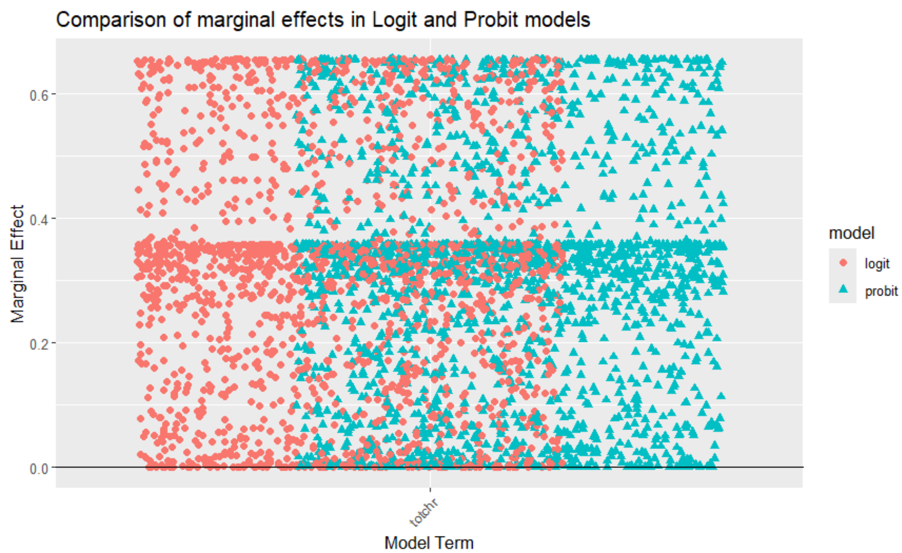
in this framework logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable, then in this case for every one unit change in *totchr*, the log odds of expenditure (versus expenditure zero) increases by 1.644. as in the probit case predictors define a probability measure since is constrained between 0 and 1.



the shape is similar if we compare it with probit model.
if we compare all the marginal effect (LPM, Logit and Probit), remember that in logit and probit model $E[y_i|x_i] = Pr[y_i = 1|x_i]$ represent the probability of success, in logit the marginal effect is:

$$\frac{\partial E[y_i|x_i]}{\partial x_{ij}} = F_i(1 - F_i)\beta_j$$

then in average the marginal effect is greater in logit model (0.35) with respect to probit model (0.25), in the LPM model the avarage marginal effect is simply the coefficient since is linear. in reality we do not interpret the effective result of AME in logit but only the direction, if we want to have a magnitude of the variation we need to calculate the MEM.



as we can see in the picture if we compare the marginal effects between probit and logit we can argue that logit can fit better the low level of *totchr* and probit work better for high level of it.