M2 – Econométrie – Guillaume BOURGEOIS

Application : modèle avec variable dépendante binaire

Présentation du jeu de données :

Nous nous intéressons aux facteurs influençant la probabilité qu'une personne soit à la retraite

Notre variable d'intérêt est la variable « *retire* » qui est égale à 1 si l'individu est à la retraite et 0 sinon

Nous souhaitons étudier l'effet de 6 déterminants potentiels :

- *age*
- hstatusg *: égale à 1 si la personne est « en bonne santé » et 0 sinon*
- hhincome : revenu annuel du ménage (milliers $)
- educyear : nombre d'années d'études
- married : égale à 1 si l'individu est marié et 0 sinon
- hisp : égale à 1 si l'individu est d'origine hispanique et 0 sinon

Code pour importer les données dans R :

```
df0=read.csv(file="https://raw.githubusercontent.com/guillaume-bourgeois92/econometrie-M2/main/probit_insurance.csv")
```

Vous aurez besoin des packages suivants :

```
library(wooldridge)

library(mfx)

library(stargazer)

library(dplyr)

library(pastecs)
```

Vous pouvez vous aider du code R utilisé pour l'exemple utilisé dans les diapos de cours (lien)

Questions :

1) Donnez quelques statistiques descriptives de ce jeu de données : *stargazer ou stat.desc*

2) Faire une estimation par MCO du modèle et afficher les valeurs prédites (*predict(mpl) si vous avez enregistré les résultats de l'estimation dans un objet qui se nomme mpl)*. Discuter des limites de cette approche et interpréter les résultats :

```
mpl=lm(Y~x1+x2, data=…)
summary(mpl) # afficher les résultats
stargazer(mpl, type="text") # afficher les résultats
```

3) a) Faire une estimation « logit » du modèle : `logit=glm(retire ~ age + hstatusg+ hhincome + educyear + married + hisp, family=binomial(link= « logit »), data=…)`

b) Calculer les effets marginaux moyens : `margin_logit=logitmfx(retire ~ age + hstatusg+ hhincome + educyear + married + hisp, data=df0, atmean = F)`

c) Donner les effets marginaux des variables sur les log-odds (directement donnés par les coefficients estimés lors du probit) `logit$coefficients`

d) Donner les effets des variables sur les odds-ratios (calculer l'exponentielle des coefficients)

e) Calculer le pseudo-R^2, le R^2 de MacFadden et réaliser un test de significativité globale des coefficients → aidez vous du code utilisé pour réaliser l'exemple dans les diapos de cours

f) Commenter les résultats

4) Reprendre les questions 3a)b)d)e)f) pour le modèle probit (dans un modèle probit on peut pas faire les interprétations en terme de *log-odds* ou de *odds-ratios*)

5) Calculer la probabilité pour l'individu avec les caractéristiques suivantes d'avoir une assurance santé : retire=0, age=60, hstatusg=1, hhincome=45.3, educyear=11, married=1, hisp=0 : *créer un nouveau data.frame avec les caractéristiques de l'individu en respectant bien les noms des variables, puis predict(probit, data.frame.créé, type= "response", se.fit=T)*