SARTORI
Romain
M1DS2E

Devoir économétrie n°3

Exercise 3 (30 points) Effet des variables continues et discrètes dans des modèles LPM, Logit et Probit

1. *Reprendre les données des exercices 1 et 2 et ajouter la variable lrd0, correspondant au logarithme du revenu d'inactivité (revenu perçu si la personne ne travaille pas).*

   ➔ Script SAS

```
/* Chargement d'un nouvelle base de donnée contenant mident et lrd0*/
data newtable2;
set mydata (keep=mident lrd0);
run;

/* Fusion des tables obtenus dans l'exercice précédent et newtable que nous venons de
créer*/
data merged_data3;
merge newtable2 merged_p2reg;
by mident;
run;
```

2. *Régresser (indépendamment) par MCO l'activité principale (actifp) et l'activité secondaire (actifs) sur les variables sexe et lrd0. Interpréter et commenter les résultats. Idem avec femme et lrd0, puis avec homme et lrd0. Commenter les similitudes et différences.*

Nous commençons par régresser l'activité principale par les différentes indicatrices de sexe et par lrd0.

➔ Script SAS

```
/*Différentes regression de actifp sur les indicatrices de sexe et lrd0*/
/*Regression sur lrd0 et des deux sexes*/
proc reg data=merged_data3;
  model actifp = sexeN lrd0;
run;
```

Variable dépendante : actifp

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	154.54008	77.27004	528.90	<.0001
Erreur	10545	1540.56534	0.14609		
Total sommes corrigées	10547	1695.10542			

Root MSE	0.38222	R carré	0.0912
Moyenne dépendante	0.79882	R car. ajust.	0.0910
Coeff Var	47.84818		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	3.48277	0.08340	41.76	<.0001
sexeN	1	0.06835	0.00793	8.62	<.0001
lrd0	1	-0.26606	0.00818	-32.52	<.0001

```
/*Regression sur lrd0 et sexe_masculin*/
proc reg data=merged_data3;
 model actifp = sexe_homme lrd0;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifp

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	154.54008	77.27004	528.90	<.0001
Erreur	10545	1540.56534	0.14609		
Total sommes corrigées	10547	1695.10542			

Root MSE	0.38222	R carré	0.0912
Moyenne dépendante	0.79882	R car. ajust.	0.0910
Coeff Var	47.84818		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	3.61947	0.08689	41.66	<.0001
sexe_homme	1	-0.06835	0.00793	-8.62	<.0001
lrd0	1	-0.26606	0.00818	-32.52	<.0001

```
/*Regression sur lrd0 et sexe_feminin*/
proc reg data=merged_data3;
 model actifp = sexe_femme lrd0;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifp

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	154.54008	77.27004	528.90	<.0001
Erreur	10545	1540.56534	0.14609		
Total sommes corrigées	10547	1695.10542			

Root MSE	0.38222	R carré	0.0912
Moyenne dépendante	0.79882	R car. ajust.	0.0910
Coeff Var	47.84818		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	3.55112	0.08479	41.88	<.0001
sexe_femme	1	0.06835	0.00793	8.62	<.0001
lrd0	1	-0.26606	0.00818	-32.52	<.0001

Ensuite, nous nous régresserons l'activité secondaires par ses mêmes variables.

➔ Script SAS

```
/*Différentes regression de actifs sur les indicatrices de sexe et lrd0*/
/*Regression sur lrd0 et des deux sexes*/
proc reg data=merged_data3;
 model actifs = sexeN lrd0;
run;
```

Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	2.97787	1.48894	49.14	<.0001
Erreur	10545	319.50933	0.03030		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17407	R carré	0.0092
Moyenne dépendante	0.03157	R car. ajust.	0.0090
Coeff Var	551.37157		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	0.36649	0.03798	9.65	<.0001
sexeN	1	-0.01201	0.00361	-3.33	0.0009
lrd0	1	-0.03006	0.00373	-8.07	<.0001

```
/*Regression sur lrd0 et sexe_masculin*/
proc reg data=merged_data3;
  model actifs = sexe_homme lrd0;
run;
```

La procedure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	2.97787	1.48894	49.14	<.0001
Erreur	10545	319.50933	0.03030		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17407	R carré	0.0092
Moyenne dépendante	0.03157	R car. ajust.	0.0090
Coeff Var	551.37157		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	0.34246	0.03957	8.65	<.0001
sexe_homme	1	0.01201	0.00361	3.33	0.0009
lrd0	1	-0.03006	0.00373	-8.07	<.0001

```
/*Regression sur lrd0 et sexe_feminin*/
proc reg data=merged_data3;
 model actifs = sexe_femme lrd0;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	2	2.97787	1.48894	49.14	<.0001
Erreur	10545	319.50933	0.03030		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17407	R carré	0.0092
Moyenne dépendante	0.03157	R car. ajust.	0.0090
Coeff Var	551.37157		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	0.35447	0.03862	9.18	<.0001
sexe_femme	1	-0.01201	0.00361	-3.33	0.0009
lrd0	1	-0.03006	0.00373	-8.07	<.0001

Commentaires sur les modèles sur actifp :

Tous les modèles étudiant la variation de l'activité principale (variable actifp) ont comme point commun un R^2 similaire. Ainsi, pour tous nos modèles, le sexe et le revenu d'inactivité expliquent pour 9,12 % des variations de l'activité principale.

Les trois modèles présentent un coefficient du logarithme du salaire d'inactivité significatif et négatif. Cela indique qu'une hausse du logarithme du revenu d'inactivité est associée à une diminution de l'activité principale. Plus précisément, une augmentation de 1 dans le logarithme du revenu d'inactivité (lrd0) entraîne une baisse d'environ 23.33% de la probabilité d'activité principale, à sexe donné. Ce résultat est intuitif puisqu'un revenu plus élevé en inactivité produit une incitation à ne pas travailler, ou en tout cas, n'incite pas à travailler.

Les coefficients sexe_homme et sexe_femme sont significatifs et respectivement négatifs et positifs. A revenu d'inactivité identique, un homme a environ 6% plus de chances qu'une femme d'avoir une activité principale.Commentaire sur les modèles sur actifs :

Ici aussi, les trois modèles ont un R^2 similaire (de 0.0092). Pour tous nos modèles, le sexe et le revenu d'inactivité expliquent pour 0.92% des variations de l'activité secondaire, ce qui est très peu ce qui remet en cause la qualité du modèle.

Dans les trois modèles, le coefficient du logarithme du salaire d'inactivité est négatif et significatif. L'intuition de ce résultat pourrait être qu'une hausse du revenu d'inactivité permettrait au individus de dégager d'avantage de temps pour une activité secondaire. On pourrait interpréter les coefficients de la première regression comme suit : une augmentation de 1 dans le logarithme du revenu d'inactivité (lrd0) entraîne une baisse d'environ 25.95% de la probabilité d'activité secondaire, à sexe donné.

Les coefficients sexe_homme et sexe_femme sont tout les deux significatifs et positifs.

3. *Reprendre la question avec un modèle Logit puis avec un modèle Probit. Commenter les similitudes et différences par rapport aux MCO.*

Nous commencerons par estimer tous les modèles (sur actifs et actifp) en probit par le biais de la procédure probit.

➔ Script SAS

```
/* Reprise des modèles en probit*/
/* Regression sur actifs*/
proc probit data=merged_data3;
 model actifs(event='1') = sexeN lrd0;
run;
proc probit data=merged_data3;
 model actifs(event='1') = sexe_homme lrd0;
run;
proc probit data=merged_data3;
 model actifs(event='1') = sexe_femme lrd0;
run;
```

```
/* Regression sur actifp*/
proc probit data=merged_data3;
 model actifp(event='1') = sexeN lrd0;
run;
proc probit data=merged_data3;
 model actifp(event='1') = sexe_homme lrd0;
run;
proc probit data=merged_data3;
 model actifp(event='1') = sexe_femme lrd0;
run;
```

Ensuite, on estime tous les modèles (sur actifs et actifp) en logit par le biais de la procédure logistic.

```
/* Reprise des modèles en logit*/
/* Regression sur actifs*/
proc logistic data=merged_data3;
 model actifs(event='1') = sexeN lrd0;
run;
proc logistic data=merged_data3;
 model actifs(event='1') = sexe_homme lrd0;
run;
proc logistic data=merged_data3;
 model actifs(event='1') = sexe_femme lrd0;
run;
/* Regression sur actifp*/
proc logistic data=merged_data3;
 model actifp(event='1') = sexeN lrd0;
run;
proc logistic data=merged_data3;
 model actifp(event='1') = sexe_homme lrd0;
run;
proc logistic data=merged_data3;
 model actifp(event='1') = sexe_femme lrd0;
run;
```

Comparaison de la méthode linéaire (par MCO) et les méthodes non-linéaires (par Probit et Logit).

Tout d'abord, les modèles sont semblables dans la mesure où les variables explicatives présentent des p-values significatives et que les signes des coefficients sont similaires entre les modèles. En termes de différences, on constate des différences en termes d'intervalles de confiances de 95% avec de légères variations.

Les coefficients ne sont pas interprétables directement dans les modèles binaires (à la différence des MCO). On se servira de l'analyse des signes pour observer des résultats analogues aux résultats du modèle linéaire.

Les analyses logit et probit offrent des rapports de cotes ainsi que des effets de Type III, fournissant ainsi des détails supplémentaires sur l'importance et l'impact des variables indépendantes incluses dans le modèle.

Dans les régressions logit, probit et MCO, l'impact de lrd0 est constaté comme positif et significatif, suggérant ainsi qu'une augmentation de lrd0 est associée à une probabilité accrue d'être actif.

De même, dans les régressions logistique, probit et MCO, l'effet de sexe_homme est observé comme négatif et significatif, ce qui implique que les hommes ont statistiquement moins de chances d'être actifs par rapport aux femmes.

Les rapports de cotes (pour la régression logistique) et les effets marginaux (pour la régression probit) ne sont pas directement comparables aux coefficients de la régression MCO. Cependant, ils facilitent une interprétation plus intuitive des effets des variables indépendantes sur la probabilité d'être actif.

Les régressions logistique et probit présentent des résultats similaires en ce qui concerne la direction et la signification des effets de lrd0 et sexe_homme sur la probabilité d'être actif. Cependant, les coefficients, les intervalles de confiance et les statistiques d'ajustement du modèle peuvent varier en raison des différences dans les fonctions de lien et les hypothèses de distribution des erreurs. Bien que les coefficients de la régression MCO ne soient pas directement comparables à ceux des régressions logistique et probit, ils indiquent également des effets similaires en termes de direction et de signification.

4. *(papier-crayon) Exprimer, pour chaque modèle, la probabilité prédite d'activité principale et d'activité secondaire pour un homme et pour une femme, pour un log-revenu d'inactivité de 9.5, 10, 11, 11.5 et 12, puis calculer ces différentes probabilités et représentez-les sur un même graphique. Calculer la différence de probabilité d'activité entre homme et femme pour chacun de ces niveaux de revenu. Calculer l'effet d'une augmentation de 1% du revenu d'inactivité sur la probabilité d'activité dans chacun des modèles pour chacun de ces niveaux de revenu. Conclure sur les possibilités d'interpréter la valeur des coefficients des modèles binaires.*

Régression Logit

-> Avec homme

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 9,5 * (-0,5092))) = 0,98947

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 10 * (-0,5092))) = 0,94289

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 11 * (-0,5092))) = 0,9916

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 11,5 * (-0,5092))) = 0,98958

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 12 * (-0,5092))) = 0,99257

-> Avec femme

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 9,5)) = 0,93645

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 10)) = 0,96130

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 11)) = 0,98603

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 11,5)) = 0,99163

P (actifs = 1) = 1/(1 + exp(-(-3,2343 + 1,0447 + 12)) = 0,99504

Modèle Probit Homme

$0,0096 \cdot \Phi(-3,09 + 0,48 \cdot \ln\delta + 0,0454)$

$\ln\delta = 9,5$ => 0,000938

$\ln\delta = 10$ => 0,0004298

$\ln\delta = 11$ => 0,0004026

$\ln\delta = 11,5$ => 0,0003916

$\ln\delta = 12$ => 0,0003802

Modèle Probit femme

$0,0096 \cdot \Phi(3,09 + 0,48 \cdot \ln\delta)$

$\ln\delta = 9,5$ => 0,0009896

$\ln\delta = 10$ => 0,0004446

$\ln\delta = 11$ => 0,0004096

$\ln\delta = 11,5$ => 0,0003996

$\ln\delta = 12$ => 0,0003856

Modèle logit - expression générale avec homme :

$0,0096 * \exp(-3,18 + 0,49 \cdot \ln\delta + 0,0494)/(1 + \exp(-3,18 + 0,49 \cdot \ln\delta + 0,0494))^2$

(-> on remplace par les différents niveaux de revenu

$\ln\delta = 9,5$ => 0,0009541

$\ln\delta = 10$ => 0,0004468

$\ln\delta = 11$ => 0,0004226

$\ln\delta = 11,5$ => 0,0004113

$\ln\delta = 12$ => 0,0004000

Modèle logit femme

$0,0096 * \exp(-3,18 + 0,49 \cdot \ln\delta)/(1 + \exp(-3,18 + 0,49 \cdot \ln\delta))^2$

$\ln\delta = 9,5$ => 0,0009581

$\ln\delta = 10$ => 0,0004652

$\ln\delta = 11$ => 0,0004285

$\ln\delta = 11,5$ => 0,0004173

$\ln\delta = 12$ => 0,0004052

Modèle MCO => Femme et homme = 0,0096 pour tous les revenus

Conclure sur les possibilités d'interpréter la valeur des coefficients des modèles binaires.

En utilisant les dérivées et les variations de probabilité précédemment calculées, nous sommes en mesure de quantifier l'impact d'une variation du revenu d'inactivité sur la probabilité d'activité et de comparer cet impact entre les sexes. L'examen des résultats des différents modèles révèle que l'effet d'une augmentation du revenu d'inactivité sur la probabilité d'activité est généralement plus faible dans les modèles logit et probit que dans le modèle MCO. Cela indique que, pour les modèles binaires, une augmentation de 1% du revenu d'inactivité entraîne une variation de la probabilité d'activité moins prononcée par rapport au modèle MCO. Aussi, on observe une différence entre les probabilités d'activité entre les sexes. Le modèle MCO semble démontrer que les probabilités d'activités des hommes sont plus élevées que celles des femmes tandis que les modèles binaires semblent montrer l'inverse. Or, selon l'Insee en 2020, parmi les 15-64 ans, 68 % des femmes et 75 % des hommes participent au marché du travail. On pourrait alors être tenter d'avancer que le modèle des MCO offre des résultats plus proches de la réalité.

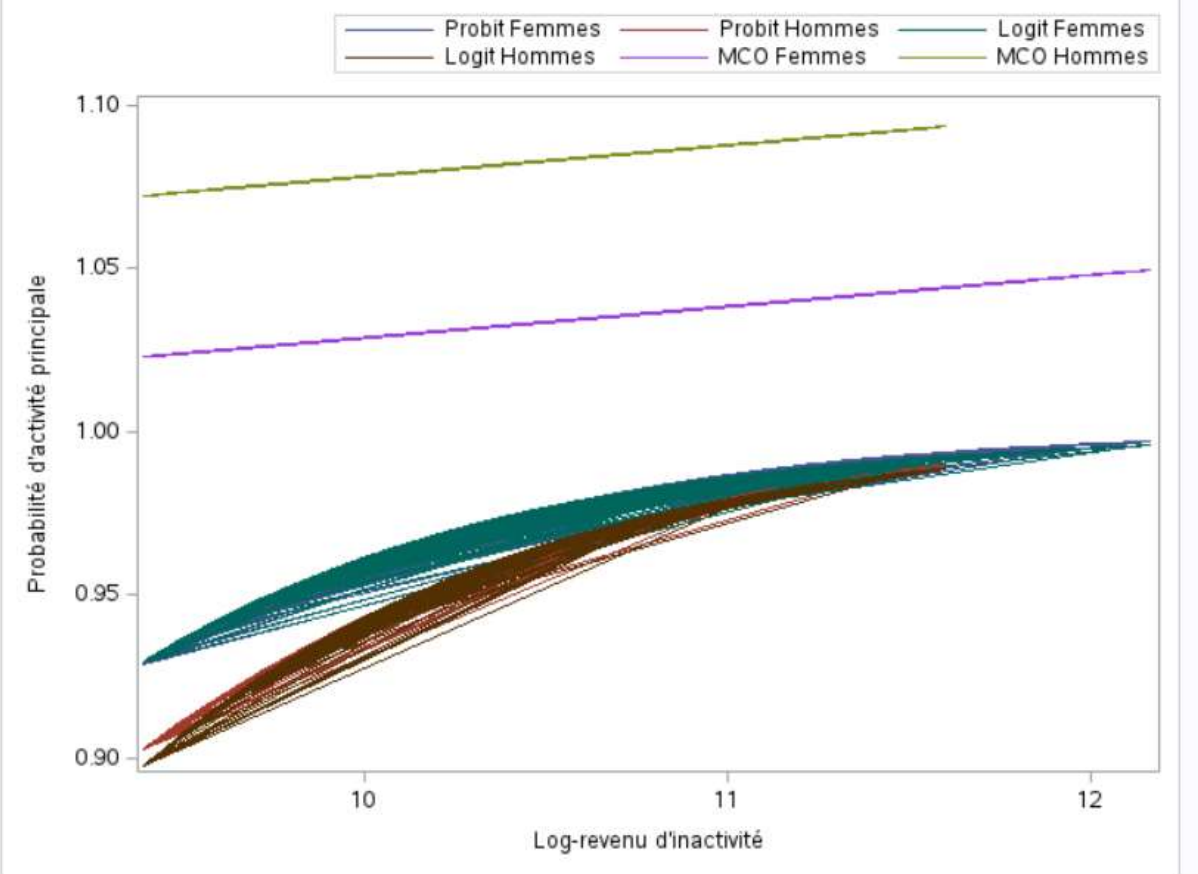
Les modèles binaires ne permettent pas une interprétation directe des résultats, car ils mesurent la variation de la probabilité de succès pour un changement d'une unité dans la variable explicative, mais cette variation est non linéaire et dépendante du niveau des autres variables indépendantes dans le modèle. En revanche, ces modèles permettent de retranscrire les impacts des variables explicatives sur la probabilité d'activité.

5. (SAS à nouveau) Calculer ces différentes probabilités pour chaque ligne de l'échantillon et tracer sur un même graphique toutes les courbes de probabilité d'activité principale en fonction du log-revenu d'inactivité. Idem pour la probabilité d'activité secondaire. Vérifier la cohérence avec la question précédente.

```
/* Création des variables de probabilité d'activité principale et secondaire */
/*Attention il faut tout run en même temps*/
data probabilities;
set merged_data3;
/* Probit */
if sexe='Femme' then do;
p_main_probit_f = cdf('NORMAL', -2.9105 + 0.4661 * lrd0);
p_second_probit_f = 1 - p_main_probit_f;
end;
else do;
p_main_probit_m = cdf('NORMAL', -2.9105 + 0.4661 * lrd0 - 0.1692);
p_second_probit_m = 1 - p_main_probit_m;
end;
/* Logit */
if sexe='Femme' then do;
p_main_logit_f = 1 / (1 + exp(-(-7.2343 + 1.0447 * lrd0)));
p_second_logit_f = 1 - p_main_logit_f;
end;
else do;
p_main_logit_m = 1 / (1 + exp(-(-7.2343 + 1.0447 * lrd0 - 0.4072)));
```

```
p_second_logit_m = 1 - p_main_logit_m;
end;
/* MCO */
if sexe='Femme' then do;
p_main_mco_f = 0.9328 + 0.0096 * lrd0;
p_second_mco_f = 1 - p_main_mco_f;
end;
else do;
p_main_mco_m = 0.9328 + 0.0096 * lrd0 + 0.0494;
p_second_mco_m = 1 - p_main_mco_m;
end;
run;

/* Formation du graphique*/
proc sgplot data=probabilities;
   series x=lrd0 y=p_main_probit_f / legendlabel='Probit Femmes';
   series x=lrd0 y=p_main_probit_m / legendlabel='Probit Hommes';
   series x=lrd0 y=p_main_logit_f / legendlabel='Logit Femmes';
   series x=lrd0 y=p_main_logit_m / legendlabel='Logit Hommes';
   series x=lrd0 y=p_main_mco_f / legendlabel='MCO Femmes';
   series x=lrd0 y=p_main_mco_m / legendlabel='MCO Hommes';
   xaxis label='Log-revenu d''inactivité';
   yaxis label='Probabilité d''activité principale';
   keylegend / location=outside position=topright;
run;
```



Les différentes courbes reflètent les différents modèles comme l'indique la légende au-dessus du graphique. Nous avons au total 6 courbes, les deux courbes de MCO en haut et le courbe probit et logit en bas. Il est difficile de distinguer les différentes courbes se trouvant sur la partie inférieure du graphique, je ne suis pas parvenu à régler ce problème. Aussi, courbes des modèles par MCO ont des probabilités supérieures à 1 ce qui reflète une erreur de modélisation.

Sinon, toutes les courbes sont croissantes traduisant qu'une augmentation du log-revenu d'inactivité est associée à une augmentation de la probabilité d'activité principale. On peut aussi observer des différences dans l'ordonnée à l'origine des sexes pour chaque modèle. Cela reflète ce que nous avions discuté précédemment. À savoir, qu'à log revenu d'inactivité donnée dans les MCO, les hommes ont une probabilité d'activité supérieure aux femmes. La relation inverse est présente dans les modèles binaires : à log revenu d'inactivité donnée, les femmes ont une probabilité d'activité supérieure aux hommes.

Enfin, on observe que les courbes probit/logit convergent vers 1 et semblent avoir une pente semblable. Ces résultats suggèrent que les divers modèles prévoient des probabilités d'activité similaires pour des niveaux équivalents de log-revenu d'inactivité.