SARTORI

Romain

M1 DS2E

Devoir n°2 économétrie

Exercise 2 (30 points) Estimation d'un modèle binaire par régression linéaire (Linear Probability Model, ou LPM).

1) *Reprendre la table créée à partir de Panel95light.csv dans l'exercice 1. Télécharger la table PanelEuropeen95.csv. Avec le logiciel de votre choix extraire les variables mident, mois, actif (indicatrice díactivité le mois considéré), actifp (activité principale), actifs (activité secondaire ou épisodique) et agea (âge mesuré en années). Importer sous SAS la table contenant mident, mois, actif et agea, puis la fusionner avec la table de l'exercice 1, selon les variables mident et mois (ce couple constitue l'identifiant unique dans ce panel). Cette méthode pourra être utilisée tout au long de l'année pour enrichir la table au fur et à mesure. La table obtenue constitue la table de travail pour l'exercice.*

⇨ *Voir le code SAS*

2) Régresser (MCO) actifs sur les différentes indicatrices d'éducation.

Pour appliquer les MCO sur nos différentes indicatrices on utilise la proc REG avec les données que l'on vient de fusionner. Voici quelques informations sur le résultats de cette régression :

- Le nombre d'observations et d'observations utilisées sont les mêmes, au nombre de 10 548.
- Le Root MSE qui mesure l'écart entre les valeurs prédites par le modèles et les valeurs observée est relativement faible : 0.17346. Cela peut indiqué une certaine fiabilité du modèle avec une bonne capacité de prédiction.
- Les valeurs du R^2 et du R^2 ajusté sont respectivement de 0.0163 et de 0.0159. On note que le R^2 est légèrement supérieur à son équivalent ajusté ce qui peut donner un indice sur une éventuelle corrélation élevée entre les variables indépendantes du modèle.
- Le coefficient de variation est de 549.4543 ce qui peut refléter une forte variabilité par rapport à la moyenne.
- La valeur F est de 43.67 et la p-value est <.0001. Cela reflète le fait que le modèle explique une partie significative de la variation de la variable dépendante.
- Les variables d'étude sont toute significatives avec une p-value <.0001 pour chacune d'entre elles.
- Les coefficients des variables d'études sont négatifs reflétant une relation inverse avec la variable dépendante.

Ces informations nous alternent sur un problème de multicolinéarité que nous avons déjà pu observer dans nos modèles précédents. Le modèle est toutefois significatifs mais malheureusement explique une faible variation de la variable dépendante.

3) *Ajouter à la table de travail la probabilité d'activité secondaire ou épisodique prédite par ce modèle, notée p1reg, ainsi qu'un estimateur de la variance du résidu, notée sigma1reg. Représenter graphiquement la distribution de p1reg et la commenter.*

⇨ Voir le code SAS

Explication de la démarche :

- On utilise la procédure data pour calculer les estimateurs de résidus

- La procédure means la moyenne de l'estimateur des résidus établi en amont avec pour résultat 0.0300750. On peut utiliser cette valeur pour estimer la variance des erreurs du modèle, plus la valeur est faible plus l'ajustement du modèle est bon.

- L'étape data « _null_ » créer une étape data temporaire qui utilise la fonction CALL SYMPUT pour assigner la valeur de "residu_sq_mean" (la moyenne des carrés des résidus) à une macro-variable nommée 'sigma1reg'.

- La proc sql permet de créer une nouvelle table.

- Enfin, on utilise la proc univariate sur cette nouvelle table pour analyser la distribution des valeurs prédites de p1reg sur « actif ». On obtient les informations suivantes :
    - Moyenne : 0.03157, Écart-type : 0.022324
    - Tous les test de Goodness of fit à savoir Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling ont respectivement des p-value de <0.01, <0.005 et <0.005
    - L'analyse des quantiles (voir tableau en annexe) nous montre de grands écarts entre les valeurs au extrémités (10% et en dessous et 95% et au-delà) remettant en cause l'aspect normal de la distribution.
    - Enfin, la représentation graphique en annexe propose que la distribution ne suit pas l'aspect d'une loi normale.

Sachant que la distribution ne suit pas une loi normale, il est important de prendre des précautions supplémentaires dans l'analyse statistique pour éviter des conclusions erronées ou biaisées.

4) *Régresser actifs sur les différentes indicatrices d'éducation par MCP et commenter les différences avec la régression précédente*

Régresser en utilisant la méthode des moindres carrés pondérés améliore la démarche par les MCO sur les points suivants :

- Prise en compte de l'hétéroscedasticité
- La méthode des MCP peut être utilisée pour pondérer les observations en fonction de la structure de corrélation des données, ce qui améliore l'efficacité des estimations des coefficients.
- La méthode des MCP peut être appliquée dans des cas de non IID pour tenir compte de la structure de dépendance ou de l'hétérogénéité des variances entre les observations.

Comparons à présent les résultats entre nos deux méthodes de régression.

- Comme pour l'approche avec MCO, tous les coefficients sont négatifs et significatifs de part leurs faible p-value.
- Le R^2 est très légèrement supérieur avec la méthode des MCP. Le modèle explique donc un peu mieux la variation de la variable dépendante dans l'approche MCP que dans les MCO.
- Le RMSE est légèrement plus élevé dans l'approche MCP (0.1903>0.17346) ce qui indique que la mesure l'écart entre les valeurs prédites par le modèles et les valeurs observée est légèrement plus élevée que dans l'approche MCO
- Les MCO et MCP offrent des résultats similaires, toutes fois les MCP explique un peu mieux la variation de la variable dépendante « actif » ce qui pourrait le rendre de facto plus intéressant pour notre étude.

5) *Créer les variables agea2 = agea2 et agea3 = agea3.*

⇨ *Voir le code SAS*

6) *Régresser actifs sur agea, agea2, agea3 et les différentes indicatrices d'éducation, et l'indicatrice de sexe féminin.*

Pour cette régression sur les indicatrices *agea2 = agea2 et agea3 = agea3, nous gardons l'entiereté de notre échantillon soit 10548 observations. Dévrivons les résultats sur la qualité du modèle :*

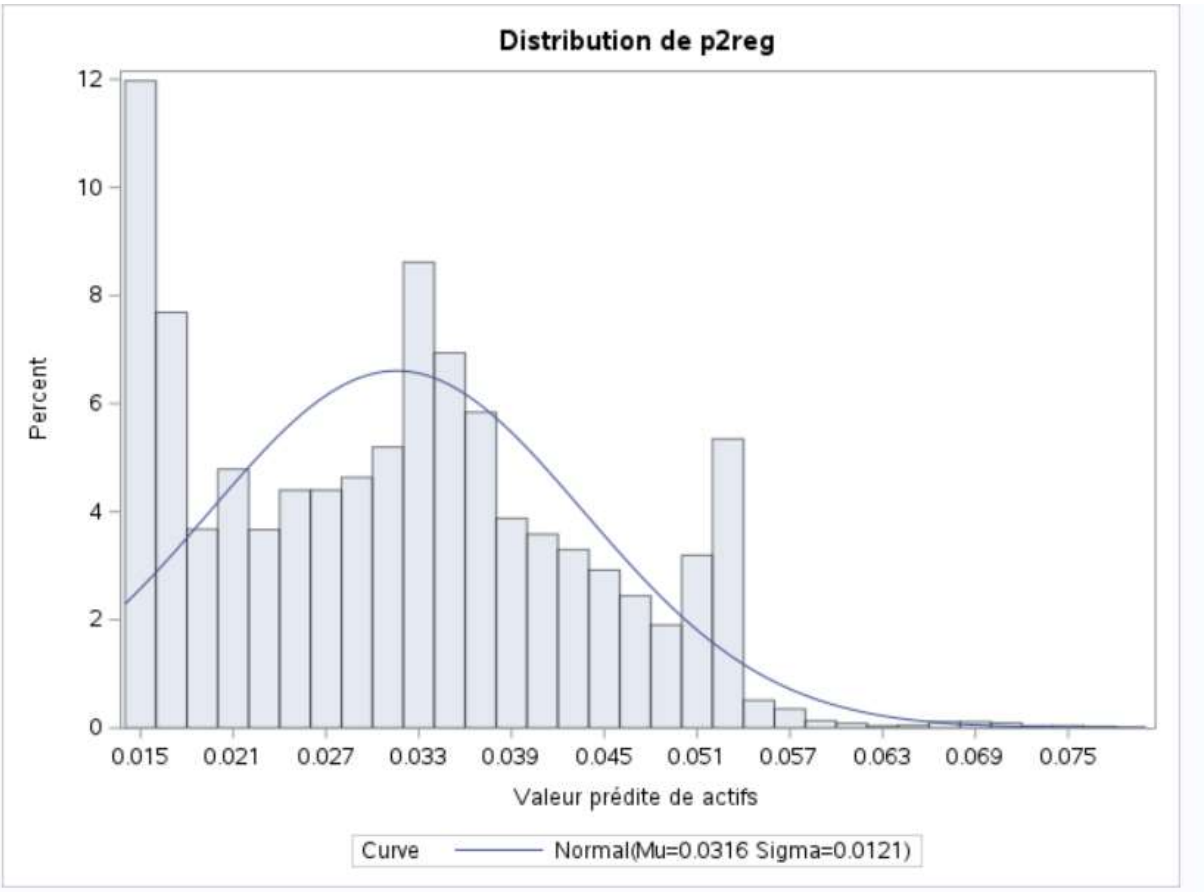
- *Le R^2 et le R^2 ajusté 0.0048 et de 0.0044 respectivement.* On note que le R^2 est légèrement supérieur à son équivalent ajusté ce qui peut donner un indice sur une éventuelle corrélation élevée entre les variables indépendantes du modèle.
- Le Root MSE qui mesure l'écart entre les valeurs prédites par le modèles et les valeurs observée est relativement faible : 0.17448. Cela peut indiquer une certaine fiabilité du modèle avec une bonne capacité de prédiction. On notera que c'est le moins élevé de nos dernières régressions.
- *La valeur F est de 12,63 et la p-value inférieure à 0,0001. Cela souligne qu'une partie significative du modèle pour expliquer la variance de la variable dépendante.*

Listing des résultats des coefficients et p-values des variables indicatrices.

	Intercept	*Agea*	*Agea 2*	*Agea3*
Coefficient	0.37166	-0.02526	0.00060982	- 0.00000465
p-value	*p < 0,0004*	*p<0.0023*	*p<0.0040*	*p<0.0077*
Relation		*Négative*	*Quadratique*	*Cubique*

Dans l'ensemble, les résultats du modèle indiquent que l'âge et le sexe ont un impact significatif sur la variable "actifs". Cependant, la valeur relativement faible du coefficient de détermination (R^2) suggère que d'autres facteurs pourraient également jouer un rôle dans l'explication de la variation de la variable "actifs". Il serait donc judicieux de prendre en compte ces autres facteurs dans de futurs modèles pour une explication plus complète.

7) *Ajouter à la table de travail la probabilité d'activité secondaire ou épisodique prédite par ce nouveau modèle, notée p2reg. Représenter graphiquement la distribution de p2reg et la commenter.*



Description et commentaire :

- La distribution est asymétrique (Skewness = 0.36819377), les valeurs de p2reg sont concentrées à gauche vers les plus faibles valeurs.
- La moyenne et la médiane sont relativement proches l'une de l'autre (0.031570 vs 0.031864). Cela peut s'expliquer par l'asymétrie positive.
- Les tests de normalité Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling indiquent que la distribution de p2reg ne suit pas une distribution normale avec p-value inférieur à 0.01.

La distribution n'a pas l'air de suivre une loi normale, elle est asymétrique avec des valeurs concentrées vers les plus faibles. On aperçoit trois point dans lesquels les valeurs se concentrent au minimum, autour de la moyenne et autour de 0.055 qui allonge la courbe vers la droite.

8) *Peut-on appliquer la méthode des MCP ici ? Expliquer les différences par rapport au modèle 1.*

Les tests de Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling et leurs p-value associé faible supposent une distribution non normale et donc propice à des problèmes d'hétéroscédasticités ce que les méthodes MCP peuvent corriger.

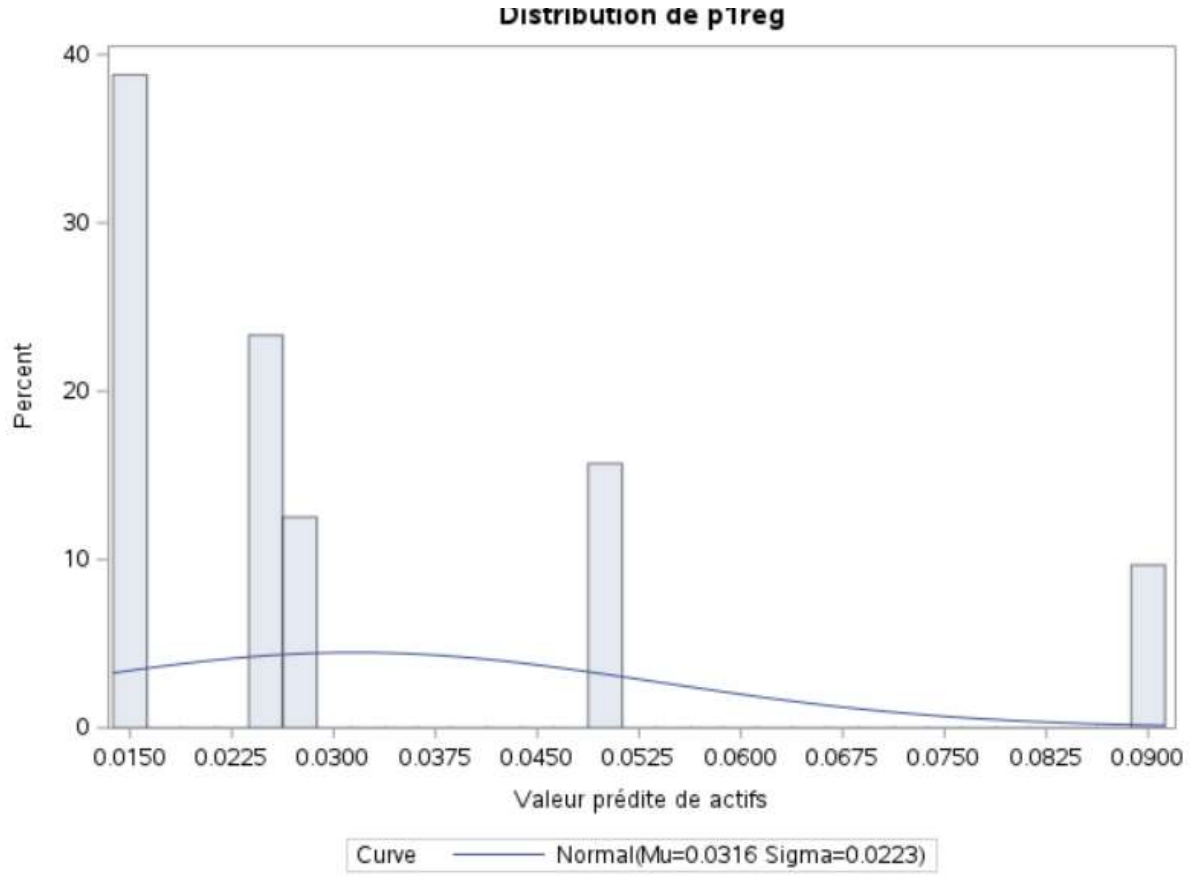
Pour utiliser la méthode des MCP, il faut choisir des variables à pondérer. Or, contrairement au modèle 1 nous n'avons ici qu'une variable indicatrice à la place des 4 du modèle 1 ce qui rend la répartition de la pondération plus compliquée.

9) *Comparer les résultats obtenus en remplaçant l'indicatrice de sexe féminin par l'indicatrice de sexe masculin, puis par la variable sexeN. Commenter précisément chacune des différences entre ces trois modèles.*

Comparaison : Ici, la moyenne et la médiane est équivalente à la régression effectuée avec l'indicatrice femme. Tous les résultats sont semblable entre les modèles avec les deux indicatrices. Toutefois ce résultat est peu probable est relève assurément d'une limite du modèle ou plus probablement encore une erreur de programmation de ma part.

Cependant, si cela n'est pas une erreur, le variable actif serait plus sensible aux différences d'âge qu'à la différence d'indicatrice homme ou femme. Cependant, l'exact similitude entre les modèles semblent plutôt refléter une erreur de modélisation.

Annexe :


Distribution de p1reg
Percent
40
30
20
10
0
0.0150
0.0225
0.0300
0.0375
0.0450
0.0525
0.0600
0.0675
0.0750
0.0825
0.0900
Valeur prédite de actifs
Curve
Normal(Mu=0.0316 Sigma=0.0223)