Enoncés des Travaux Dirigés

Econométrie des Variables Qualitatives

Professeur de CM :
Xavier Joutard

Chargées de TD :
Kenza Elass - kenza.elass@univ-amu.fr
Bertille Picard - bertille.picard@univ-amu.fr

Licence 3 et Magistère 1
Année 2020-2021
Semestre 2 - Marseille

Organisation du cours

Modalités d'évaluation des Travaux Dirigés :

- Évaluation de connaissance durant le semestre
- Application finale sur ordinateur avec projet à rendre

Organisation des séances de Travaux Dirigés :

- 8 séances de 1h30
- Si possible la dernière séance se déroulera en salle informatique

ATTENTION : consulter le matériel sur `https://ametice.univ-amu.fr` et l'emploi du temps en ligne. Si un lien est nécessaire pour les cours, vous le trouverez sur la page du cours.

Documents de référence :

- J. Stock, M. Watson, Principes d'Économétrie, traduction J. Trabelsi, Pearson
- Introductory Econometrics : A Modern Approach, Jeffrey Wooldridge, si possible une édition récente

Sommaire

Chapitre 1 : Variables dépendantes binaires et modèle de probabilité linéaire

Ce chapitre se concentre sur des rappels des notions utiles au cours (modèle de régression linéaire et moindres carrés ordinaire (MCO).

1.1 - Quelles sont les hypotheses des MCO ? Pour chaque hypothèse, fournissez un exemple pour lequel l'hypothèse est validée et un exemple pour lequel I'hypothèse est violée.

1.2 - Considérons le modèle de régression linéaire $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Supposons connue la valeur de $\beta_0 = 0$. Déduire l'estimateur des MCO de β_1.

b. Pour une régression linéaire fournissant un $\beta_1 = 0$, montrez que $R^2 = 0$.

c. Une régression linéaire avec un $R^2 = 0$ implique-t-elle automatiquement $\beta_1 = 0$?

d. Expliquez la différence entre β_0 et β_1.

e. Quelle est la différence entre Y_i et les valeurs prédites $\hat{Y}_i$?

f. Quelle est la différence entre l'erreur de la régression u_i et le terme résiduel $\hat{u}_i$?

g. Montrez que I'hypothèse des moindres carrés, $E(u_i \mid X_i) = 0$ implique que : $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$

h. Montrez que l'estimateur des MCO de β_0 et β_1 est sans biais.

1.3

a. Montrez que le coefficient de détermination R^2 de la régression de Y sur X est le carré de la corrélation empirique entre X et Y.

b. Montrez que le R^2 de la régression de Y sur X est le même que le R^2 de la régression de X sur Y.

c. Montrez que $\beta_1 = r_{XY}(S_Y \mid S_X)$, où r_{XY} est la corrélation empirique entre X et Y, avec Sx et Sy correspondant aux écarts-types empiriques de X et de Y respectivement.

1.4 - Supposons qu'un chercheur se propose d'estimer la régression du score des tests sur l'effectif des classes (CS). Les résultats des estimations pour 100 classes sont

donnés par

$$\text{ScoreTest } = 520,4 - 5,82 \times CS, \quad R^2 = 0,08$$

a. Déterminez la valeur attendue du score des tests pour une classe de 22 élèves.

b. Supposons que la classe comportait 19 élèves l'année dernière et 23 élèves cette année. Quelle est la variation attendue du score des tests induite par cette variation de l'effectif de la classe ?

c. Pour un échantillon de 100 classes, quel est le score des tests moyen associé à une taille moyenne des classes de $21,4$?

d. Quelle est la part de la variance du score des tests expliquée par le modèle ?

1.5 - Considérons un échantillon aléatoire composé de salariés diplômés de l'université, âgés entre 25 et 26 ans. L'estimation, à partir de cet échantillon, de la régression du salaire hebdomadaire moyen (SHEM, mesuré en) par rapport à I'âge (mesuré en années) fournit

$$\text{SHEM} = 696,7 + 9,6 \times \hat{\text{Age}}, \quad R^2 = 0,023, \quad SER = 624,1$$

a. Explicitez la signification des valeurs des paramètres estimés, 696,7 et 9,6 .

b. L'erreur-type de la régression (SER) est de 624,1 . Quelle est l'unité de mesure de la SER (, années, sans unité) ?

c. Le R^2 de la régression est égal à $0,023$. Quelle est son unité de mesure (, années, sans unité) ?

d. A partir des résultats de l'estimation, calculez le salaire d'un cadre âgé de 25 ans ; de 45 ans.

e. D'après ce que vous savez sur la distribution des salaires, pensez-vous que l'hypothèse de normalité de la distribution du terme d'erreur de cette régression est plausible ? Pensez-vous que la distribution est symétrique ? Quelle est la valeur minimale des salaires ? Est-elle compatible avec la distribution normale ?

Variables dépendantes binaires et modèle de probabilité linéaire

1. 6 - Un de vos amis utilise des données individuelles pour étudier les facteurs

incitant les étudiants à fumer. En particulier, il cherche à déterminer l'impact du niveau d'éducation sur la décision de fumer.

a. Peut-on dire qu'être fumeur est une variable continue ? Expliquez.

b. Comment interpréteriez-vous les paramètres de cette régression ?

c. Supposons que le modèle de probabilité linéaire fournit une valeur estimée de Y égale à $1,15$. Expliquez pourquoi ce résultat est aberrant.

1.7 - Pour chaque exemple de variable étudiée, quelle est la nature de la variable (quantitatif : continu/discret, qualitatif : nominal/ordonné) ?
- L'age, le département de naissance, le nombre de frères et soeurs, être fumeur, le prix d'une action

1.8 - Considérons un modèle de probabilité linéaire, $Y_i = \beta_0 + \beta_1 X_i + u_i$, où $P(Y_i = 1 \mid X_i) = \beta_0 + \beta_1 X_i$
a. Montrez que $E(u_i \mid X_i) = 0$.

b. Montrez que $\text{var}(u_i \mid X_i) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]$.

c. Le terme résiduel u_i est-il hétéroscédastique ? Expliquez.

Chapitre 2 : les modèles probit et logit

Les exercices 2.1 à 2.5 sont basés sur le scénario suivant :
Aux Etats-Unis, quatre cent candidats au permis de conduire ont été aléatoirement sélectionnés et répartis en deux groupes : ceux qui I'ont obtenu (Pass $_i = 1$) et ceux qui ont échoué (Pass $_i = 0$). Les données supplémentaires portent sur leur sexe (Homme $_i = 1$ si le candidat est de sexe masculin, et $= 0$ sinon) et sur lent expérience en matière de conduite (Expérience, en années). Le tableau suivant résume les différents modèles estimés.

Variable dépendante : *Pass*							
	Probit	Logit	Probabilité linéaire	Probit	Logit	Probabilité linéaire	Probit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Expérience	0,031 (0,009)	0,040 (0,016)	0,006 (0,002)				0,041 (0,156)
Homme				- 0,333 (0,161)	- 0,622 (0,303)	- 0,071 (0,034)	- 0,174 (0,259)
Homme × *Expérience*							- 0,015 (0,019)
Constante	0,712 (0,126)	1,059 (0,221)	1,774 (0,034)	1,282 (0,124)	2,197 (0,242)	0,900 (0,022)	0,806 (0,200)

2.1 - Utilisez les résultats de la colonne (1) pour répondre aux questions suivantes.

a. La probabilité de réussir l'examen du permis de conduire dépend-elle de l'expérience ? Expliquez.

b. Construisez I'intervalle de confiance à 95% pour la réponse à la question (a).

c. Quelle est la probabilité, pour un homme avec 10 ans d'expérience de conduite, d'obtenir le permis de conduire ?

d. Quelle est la probabilité, pour un homme sans aucune expérience de conduite, d'obtenir le permis de conduire ?

2.2-

a. Répondre aux questions a., c. et d. de l'exercice 2.1 pour le modèle logit donné en

colonne (2).

b. Estimez les probabilités de réussir l'examen pour des valeurs d'Expérience comprises entre 0 et 60 et en utilisant les résultats des modèles probit et logit donnés en colonnes (1) et (2). Les modèles probit et logit donnent-ils les mêmes prédictions ?

2.3- *a.* Répondre aux questions a., c. et d. de l'exercice 2.1 en utilisant les résultats de la colonne (3). *b.* Estimez les probabilités pour des valeurs d' Expérience comprises entre 0 et 60 en utilisant les résultats de la colonne (3). Pensez-vous que, pour cette étude, le modèle de probabilité linéaire soit approprié ? Pourquoi ?

2.4 - Utilisez les résultats des colonnes (4) − (6).

a. Proposez une importante variable omise susceptible de biaiser les estimations des colonnes (1),(2) et (3). Quelle est sa nature, et comment peut-elle biaiser les résultats ?

b. Estimez la probabilité de réussir l'examen de permis de conduire pour un homme et pour une femme.

c. Les modèles des colonnes de (4) a (6) sont-ils différents ? Pourquoi ?

2.5 - Utilisez les résultats de la colonne (7).

a. Quelle est la probabilité, pour un homme avec 10 ans d'expérience de conduite, d'obtenir le permis de conduire ?

b. Quelle est la probabilité, pour une femme avec 2 ans d'expérience de conduite, d'obtenir le permis de conduire ?

c. L'effet de l'expérience sur la performance à l'examen est-il différent pour un homme ou pour une femme ? Expliquez.

2.6 - On utilise un modèle logit afin d'analyser la probabilité d'accéder à un emploi stable (CDI) dans les 4 ans après avoir quitté le système éducatif en France. Un échantillon de 17435 individus est utilisé, dont 58.9% arrive à trouver un emploi stable dans la période.

	Model 1	Model 2
Constant	0.69	0.683
	(0.27)	(0.27)
Years of education	0.167	0.166
	(0.009)	(0.009)
Female	-0.198	-0.197
	(0.04)	(0.03)
Months since leaving full-time	-0.058	-0.059
	(0.006)	(0.006)
Months spent unemployment	-0.124	-0.124
since leaving full-time education	(0.003)	(0.003)
Local area rate of unemployment	-0.05	-0.034
as a percentage	(0.01)	(0.008)
Paris Region	0.166	–
	(0.05)	
South West	0.212	–
	(0.05)	
East and Center	0.371	–
	(0.07)	
North East	-0.037	–
	(0.06)	
West and North West	-0.020	–
	(0.05)	
Correct Predictions of $y = 1$	9025	9045
Correct Predictions of $y = 0$	2853	2808
log likelihood	-10,560.9	-10,587.5

a. A partir du tableau ci-contre, expliquez et interprétez les résultats obtenus pour le Modèle 1

b. Quel est l'effet marginal d'une année d'éducation supplémentaire sur la probabilité de trouver un emploi stable ?

c. L'exclusion des variables muettes pour les régions produit les résultats dans la colonne « Modèle 2". Quelles conclusions tirez-vous concernant le r´le de ces variables muettes régionales (la référence est "Sud-est") ? Procéder à un test adéquat pour valider ce rôle (valeur critique à 5% du chi $^2, x_5^2 = 11,1$)

Chapitre 3 : Estimation et inférence des modèles probit et logit

3.1 - Pourquoi les paramètres des modèles probit et logit sont-ils estimés par la méthode du maximum de vraisemblance plutôt que par les MCO ?

3.2 - Déterminez la fonction de vraisemblance pour les modèles probit et le logit

3.3 - Supposons qu'une variable aléatoire Y admette trois modalités (1,2,3) soit la distribution de probabilité suivante : $P(Y = 1) = p, P(Y = 2) = q$ et $P(Y = 3) = 1 - p - q$. Un échantillon de taille n est aléatoirement tiré de cette distribution. Les variables aléatoires sont notées, $Y_1, Y_2, \ldots \ldots Y_n$

a. Déterminez la fonction de vraisemblance.

b. Déterminez l'expression de p et q estimés par le maximum de vraisemblance.

3.4 - On dispose d'une série d'informations sur le logarithme des salaires dont la distribution est celle d'une loi Normal $\mathrm{N}(0, \sigma^2)$

a. Si le logarithme des salaire suit une loi Normal, quelle est la loi de distribution des salaires ?

b. Construisez la fonction de vraisemblance du modèle.

3.5 - Pour quelle(s) raison(s) pourrait on dire que l'estimation par maximum de vraisemblance est moins robuste que l'estimation par moindres carrés ordinaires ?

3.6 - On considère le modèle logit caractérisé par

$$P(y_i = 1 \mid x_i) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

a. On cherche à déterminer si la variable x_i affecte le choix des individus y_i. Autrement dit, on veut tester si $\beta_2 = 0$. Pour cela proposez trois tests statistiques. Pour chacun des tests expliquez votre approche (hypothèse nulle, statistique de test, ...).

b. Donnez la contribution individuelle à la fonction de vraisemblance. Déterminez le vecteur score ainsi que la matrice d'information. Déterminez les statistiques de test pour chacun des tests.

c. Le score test estimé est de 9,95 et la Wald test est de 7,41 ; étant donné que la valeur critique à 5% du chi ^{2} est ${\chi_1}^2 = 3,84$ que pouvez vous en conclure ? La log-vraisemblance du modele est 85,57 alors qu'elle est de 12,84 dans le modèle contraint $\beta_2 = 0$, que pouvez-vous en dire ?

Chapitre 4 : Autres modèles de variables dépendantes limitées

4.1 - Un laboratoire pharmaceutique a mis au point un nouveau médicament. Pour un certain dosage, il cherche à apprécier l'importance des effets secondaires de ce médicament sur T patients ; quatre catégories d'effets sont recensées : aucun effet (1), effet léger (2), effet sévère (3), provoque la mort du patient (4). La gravité de la réaction dépend en fait du niveau d'allergie (non-observable) du patient i mesuré par un indicateur y_i^* mais déterminé par des caractéristiques individuelles observables x_1. Dans la mesure où plus cet indicateur est élevé, plus la probabilité que la réaction soit grave augmente, on peut définir trois valeurs-seuils : 0, a et b telles que :

$$y_1 = 1 \quad \text{si } y_1^* \leq 0$$

$$y_1 = 2 \quad \text{si } 0 < y_1^* \leq a$$

$$y_1 = 3 \quad \text{si } a < y_i* \leq b$$

$$y_1 = 4 \quad \text{si } b \leq y_1^*$$

Étant donné que ${y_1}^* = x_i'\beta +$ ui où les ui sont indépendants et identiquement distribués selon une loi normale centrée réduite.

a. Nommez ce modèle.

b. Écrivez la fonction de vraisemblance du modèle comme une fonction de β, a et *b.*

4.2 - On explique les dépenses d'investissement de N firmes (y_1^* pour $i = 1, \ldots, N$) par des variables x_i selon une régression linéaire avec terme d'erreur distribué suivant une loi Normal $N(0, \sigma^2)$. Pour cela, on est confronté à une double censure ; on dispose en effet des observations suivantes :

- pour toutes les firmes dont le niveau d'investissement réel est inférieur au seuil a_1, on ne peut observer les dépenses exactes mais seulement le seuil a
- pour toutes les firmes dont le niveau d.investissement réel est supérieur au seuil b_1, on ne peut observer les dépenses exactes mais seulement le seuil b_1
- En revanche, pour toutes les autres firmes ayant des dépenses d'investissement comprises entre ces deux bornes, on observe le niveau exact de ces dépenses.

a. Écrivez de manière formelle le modèle décrit ci-dessus.

b. Écrivez la fonction de vraisemblance correspondante à ce modèle.

c. Déterminez l'espérance conditionnelle suivante : $\mathrm{E}\left(\mathrm{y}_i \mid \mathrm{a}_1 < \mathrm{y}_1 < \mathrm{b}_1\right)$.

4.3 - On analyse les déterminants de la fréquentation des matchs de rugby dans les stades de France - le nombre de places vendues, y - en fonction des caractéristiques du match (adversaire, temps, jour, classement,...) regroupées dans le vecteur x selon un modèle de régression linéaire avec erreurs distribuées suivant une loi Normal :

$$y_i = X_i'\beta + u_i \quad \text{où } u_i \mid X_i \sim N\left(0, \sigma^2\right)$$

Notez que l'unité d'observation correspond à un match, c'est à dire à une rencontre dans un stade à une date donnée. Il arrive fréquemment que les matchs sont "à guichet fermé" et remplissent tout le stade. Dans ce cas, certaines demandes de spectateurs ne pourront être satisfaites. On notera m_1 la capacité d.accueil maximum du stade le jour de match.

a. En utilisant une variable latente, écrivez de manière formelle le modèle décrit ci dessus. De quel type de modèle s'agit-il ?

b. Supposons que vous ne reteniez que l'information binaire sur chaque match indiquant si "le match est à guichet fermé" ou si "le match n'est pas à guichet fermé". Comparez cette information à celle que vous exploitez dans la question précédente. Quelle modélisation allez vous appliquer dans ce cas ?

4.4 - Quels modèles utiliseriez-vous pour

a. Une étude expliquant le nombre de minutes par mois qu'une personne consacre à l'utilisation de son téléphone portable ?

b. Une étude expliquant les notes (de A à F) à l'examen d'Introduction à l'Économie ?

c. Une étude expliquant le choix des consommateurs entre trois différentes marques de soda : Orangina, Perrier ou Schweppes ?

d. Une étude sur le nombre de téléphones portables par famille ?