Modélisation de la demande de transport

4B

économétrie des choix discrets

Fabien Leurent
ENPC / LVMT

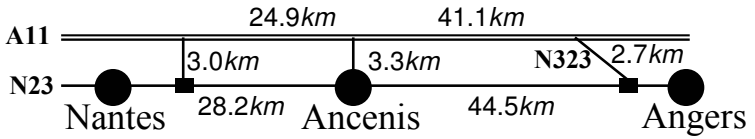
Plan

- Observer les choix
  - Préférences révélées : RP (Revealed Preferences)
  - Préférences déclarées : SP (Stated Preferences)
- Grouper les observations
- Estimer un modèle de choix
  - Se ramener à une régression linéaire : cas d'une distribution d'arbitrages prix-temps
  - La méthode du maximum de vraisemblance : application au modèle logit
  - *Régression linéaire pour choix discrets*
- Conclusion

Préférences révélées

- Observer, dans quel objectif ?
  - Paramètres de choix, pour transfert et prédiction
- Observation de situations réelles, qui révèlent les choix
- Situation de choix
  - Quelles options ? Comment les identifier ?
  - Décrire les options
  - Qui choisit ?

Exemple : choix d'itinéraire routier entre Angers et Nantes



O-D	Itinéraires	Prix €	Temps h	VL obs	*VL sim*
Nantes-	Route gratuite N23	10.0	0.967	1500	*1522*
Angers	Autoroute A11	13.2	0.617	3500	*3477*
Nantes-	Route gratuite N23	3.88	0.400	2600	*2621*
Ancenis	Autoroute A11	5.17	0.300	1300	*1279*
Angers-	Route gratuite N23	6.11	0.567	500	*456*
Ancenis	Autoroute A11	7.44	0.417	450	*494*

Exemple : choix modal avec une enquête-ménages de déplacement

- Éventail d'options, pour chaque déplacement
  - Automobile
  - TC
  - Éventuellement autres modes
- Option choisie : une par déplacement
- Options alternatives non choisies
  - calculées au moyen d'un codage des réseaux et d'une recherche d'itinéraire

Principes

- Identifier les décideurs *u*
  - Groupe d'appartenance ?
  - Caractères $Y_{u_\ell}^k$ (descriptifs de l'état)
  - Paramètres β_u (comportement)
- Pour chaque décideur *u*, ensemble de choix
  - Caractères X_m^k de chaque option *m* : ex. prix, temps, nb ruptures de charge
  - Pour indiquer l'option choisie $m^*(u)$: une variable binaire $y_u(m)$ de valeur 1 pour l'option $m^*(u)$, ou 0 si $m \neq m^*(u)$

Conditions d'application des RP

- Existence des options
- Caractérisation des décideurs (observation)
- Caractérisation des options (codage)
- Réalisation des enquêtes sur le choix : précautions
  - Niveau d'information des décideurs sur les options non choisies ?
  - Attention aux contraintes qui peuvent entraîner des captivités (Cf modèle Satchmo, L Hivert INRETS 1989)

Préférences déclarées

- *En anglais : Stated Preferences (SP)*
- Objectif
  - Situations virtuelles
  - Connaissance approfondie du comportement individuel
- Situations
  - Existantes : variation de l'offre (modification de l'infra, changement de fréquence ou de confort)
  - Hypothétiques : option de type nouveau, ex. nouveau mode de transport
- Méthode
  - Par questionnement d'individus, avec des questions ajustées au répondant pour en retirer autant d'information que possible

Exemple : enquête USAP 1992

Scénario à cinq options abstraites (Temps, Prix)
Réponses des motifs privés, ou professionnels, pour un déplacement à 150 km en heure creuse

Option	Temps (h)	Prix (€)	*CG €*	Privés	Prof
1	3	15	*43,7*	135	13
2	2,5	18	*41,2*	60	22
3	2	23	*41,2*	135	32
4	1,75	26	*41,9*	70	24
5	1,25	34	*45*	65	56

Applications

- Depuis la fin des années 1970
  - Voyageurs : choix modal en urbain (métro léger-Orlyval, tramway) et en interurbain (train)
  - Fret : choix modal des chargeurs
  - Choix de localisation des ménages et des entreprises
- Répétition de scénarios pour un répondant
  - Variations 'intra-individu' : l'APT est-il stable ?
- Enquête de consentement à payer
  - Combien voulez-vous payer pour avoir tel changement ?

Méthode des SP, 1/2

- Concevoir les scénarios de choix
  - Différencier les options par leurs caractères
  - Ouvrir l'éventail des possibilités : attention aux interpolations
  - Fixer le type de réponse : une seule option, ou des proportions, ou un classement des options
- Énoncer les scénarios de manière
  - concrète, réaliste, pour bien « mettre en situation de choix »
  - compréhensible ! Surtout pour mode nouveau
- Ajuster les scénarios à chaque répondant
  - En fonction des caractères socio-démographiques, des habitudes de choix, et des réponses aux questions précédentes
  - Intérêt des enquêtes assistées par ordinateur

Méthode des SP, 2/2

- Identifier les décideurs
  - Segmentation selon des critères à fixer
  - Plan d'enquête : carré latin, carré gréco-latin etc
- Taille d'enquête
  - Quelques dizaines à quelques centaines
- Coût
  - Conception
  - Durée d'administration individuelle
  - Nb d'interviews
  - Administration par téléphone ou en face-à-face

Grouper les observations

- Démarche commune aux RP et aux SP
- Les modes abstraits : une simplification nécessaire
  - Options décrites de manière abstraite par les caractères
  - ces options abstraites sont supposées identiques pour des décideurs, malgré les aléas du réel (itinéraires…)
- Les segments de demande
  - Contrainte statistique : constituer des groupes de décideurs avec un effectif suffisant >= 30
  - Segmentation selon des caractères observés directement

Simulation versus Estimation

- Deux perspectives bien distinctes pour un modèle

$$Y = F(\Theta, X)$$

- En simulation
  - On connaît *F* et Θ
  - On se donne l'input *X*
  - On en déduit l'output *Y*
- En estimation (paramétrique)
  - On connaît *F* mais pas Θ
  - On a des observations conjointes (X_i, Y_i)
  - On confronte ces observations à *F* pour estimer Θ
- *Estimation non-paramétrique*
  - *Non seulement Θ, mais encore F est à déterminer*

Estimer un modèle de choix

- Mettre en forme les observations
  - Méthode agrégée : par groupes d'individus
  - Méthode désagrégée : données individuelles
- Estimateurs
  - Moindres carrés : régression linéaire dans cas simples
  - Maximum de vraisemblance : formuler les probabilités des options, par segment de demande et par scénario
- Parfois : reformuler pour se ramener à un modèle statistique simple

Concurrence Train / Avion

Observations par scénario *s* : modes 1 = Train, 2 = Avion, caractères le temps T_m, le prix P_m, et un effectif enquêté n_m

s	P_{tr}	P_{av}	T_{tr}	T_{av}	n_{tr}	n_{av}
1	1,5	15,0	3,7	2,3	67	23
2	2,0	12,0	5,9	2,5	37	53
3	2,4	23,0	9,9	2,1	28	62
4	2,3	18,0	8,1	1,9	21	69
5	2,0	15,2	3,0	2,0	66	24
6	2,7	21,0	8,5	2,1	21	69
7	1,8	17,0	4,6	2,2	47	43
8	1,4	20,0	3,8	1,8	59	31

Modèle prix-temps : position statistique

- Estimer A la fonction de répartition des APT
- Hypothèse d'une distribution log-normale

$$A(x) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$$

- Par scénario *s*
  - APT de coupure $\alpha_s^* = \frac{P_{2s} - P_{1s}}{T_{1s} - T_{2s}}$
  - Si $P_{1s} < P_{2s}$ alors $A(\alpha_s^*)$ est la proportion du mode 1, observée par la proportion empirique

$$\hat{p}_{1s} = \frac{n_{1s}}{n_{1s} + n_{2s}}$$

Reformulation

- Relation observé-modélisé

$$\hat{p}_{1s} = A(\alpha_s^*)$$

$$\hat{p}_{1s} = \Phi\left(\frac{\ln(\alpha_s^*)-\mu}{\sigma}\right)$$

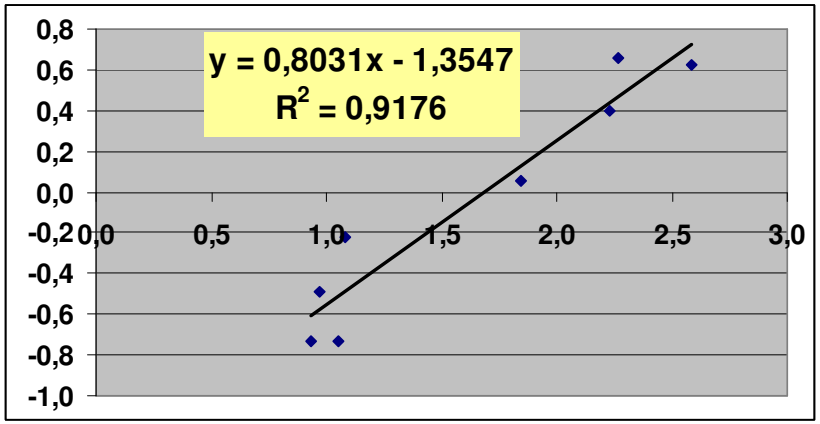
$$\Phi^{-1}(\hat{p}_{1s}) = \frac{\ln(\alpha_s^*)}{\sigma} - \frac{\mu}{\sigma}$$

- Modèle de régression linéaire *Y = a X + b*

Traitement des données

s	$\hat{p}_{1s}$	$\Phi^{-1}(\hat{p}_{1s})$	α_s^*	$\ln(\alpha_s^*)$
1	0,74	0,66	9,63	2,26
2	0,41	-0,22	2,94	1,08
3	0,31	-0,49	2,64	0,97
4	0,23	-0,73	2,53	0,93
5	0,73	0,62	13,26	2,58
6	0,23	-0,73	2,85	1,05
7	0,52	0,06	6,32	1,84
8	0,66	0,40	9,31	2,23

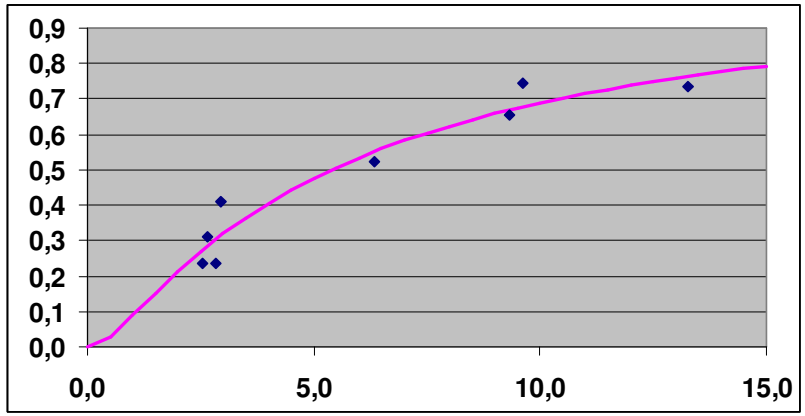
Résultats, 1/2



$\sigma = 1.25 \rightarrow M\_APT = 11.7 = \exp(\mu + \sigma^2/2)$

$\mu = 1.69 \quad S\_APT = 28.6 = M.\sqrt{\exp(\sigma^2) - 1}$

Résultats, 2/2



Commentaires

- Effectifs $n_s = n_{1s} + n_{2s}$
  - Par scénario pour observer la répartition modale
- Erreur d'observation ?
  - Fluctuation d'échantillonnage
  - Effets des tailles de groupe ?
  - Solution = pondérer les observations => moindres carrés pondérés
- Incertitude d'estimation
  - SE*a*, SE*b*, cov(*a*,*b*)
  - Propagation sur μ, σ et cov(μ,σ), et sur M_APT, S_APT

L'estimation par maximum de vraisemblance

- La fonction de vraisemblance
  - Soit une VA *X* dont la distribution de probabilité dépend d'un paramètre Θ, avec une fonction de densité f(Θ, *X*)
  - Si on ne connaît pas Θ, l'observation de valeurs x_i de *X* apporte de l'information sur Θ : précisément, l'observation {*X* = *x*} donne à une valeur Θ du paramètre une **vraisemblance *L*(Θ | *x*) = f(Θ, *x*)**

  + se généralise à un ensemble d'observations
- Estimateur du maximum de vraisemblance : choisir Θ qui maximise *L*
- Propriétés statistiques
  - Plus la taille est grande moins le biais est grand
  - Variance minimale : statistiquement efficace
  - Imprécision d'estimation → $\mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j] \approx -[\frac{\partial^2 \ln(L)}{\partial\theta_i \partial\theta_j}]^{-1}$
  - Simplifier un modèle

Vraisemblance et choix discret

- Modèle multinomial simple
  - Problème d'estimer des proportions π_m
  - On observe des $m^*(i)$ pour des individus *i*
  - Vraisemblance (si *indépendance*) $L(\pi|m^*) = \prod_{i\in I} \pi_{m^*(i)}$

$$L(\pi|m^*) = \prod_{i\in I}\prod_{m\in M} \pi_m^{y_m(i)}$$
$$= \prod_{m\in M} \pi_m^{I_m} \quad \text{avec } I_m \text{ nb d'occurrences de } m$$

- Modèle de choix discret
  - Problème d'estimer des proportions $\pi_m(\Theta, X)$
  - Vraisemblance $L(\Theta|m^*) = \prod_{i\in I} \pi_{m^*(i)}(\Theta, X_i)$
    $= \prod_{i\in I}\prod_{m\in M} \pi_m(\Theta, X_i)^{y_i(m)}$
  - Regrouper les termes par option ? Possible de grouper selon les valeurs des X_i donc selon les segments

Application et exemple

- Traitement numérique : avec la log-vraisemblance
$$\Lambda(\Theta) \equiv \ln L(\Theta) = \sum_i \ln \pi_{m^*(i)}$$
- Modèle logit : concurrence entre TC et automobile
  - Deux facteurs de choix : le prix et le temps
  - Un paramètre à estimer : l'arbitrage prix-temps
  - TC : $V_{TC} = \theta\, T_{TC} + P_{TC}$: on retient P_{TC} = 1 €
  - VP : $V_{VP} = \theta\, T_{VP} + P_{VP}$: on retient P_{VP} = 0.5 €
- Observations :

Individu	T_{TC}	T_{VP}	Choix
1	20'	30'	TC
2	35'	10'	VP
3	60'	20'	VP

EMV pour une distribution d'APT

  - *Quand on dispose d'un logiciel d'estimation non linéaire, on peut traiter par maximum de vraisemblance les problèmes pour lesquels un traitement par régression existe aussi*
  - *Revenons au cas de la distribution d'APT, déjà traité*
- Fonction de vraisemblance
$$L(\mu,\sigma) = \prod_s A_s^{n_{1s}} (1-A_s)^{n_{2s}} \quad \text{avec} \quad A_s = A(\alpha_s^*, \mu, \sigma)$$
- L'EMV est un estimateur de nature désagrégée
  - Même si la formule de L est agrégée selon les scénarios *s*
  - Dans la formule de *L*, on a regroupé les observations individuelles qui partagent les mêmes scénarios et choix
  - Chaque individu intervient dans le produit, précisément dans un scénario *s* et dans le terme de son option choisie

Modèle logit additif

- Formule additive par segment *u*, scénario *s*, option *m*:
$$U_{sm}(\Theta_u) = V_{sm}(\Theta_u) + \varepsilon_{sm}$$
$$V_{sm} \leftarrow \begin{cases} Y_s^\ell \text{ caractères du décideur} \\ X_{sm}^k \text{ caractères de l'option} \\ \theta_u^k, \theta_u^\ell \text{ paramètres du décideur} \end{cases}$$
- Forme linéaire (entre autres)
$$V_{sm} = \theta_m^0 + [\sum_k \theta_m^u . X_{sm}^k] + [\sum_\ell \theta_u^\ell . Y_u^\ell]$$
- Reste linéaire envers Θ quand attributs composites
- Intérêt pour le maximum de vraisemblance
  - La fonction de log-vraisemblance $\Lambda = \ln L$ est concave
  - La probabilité $\pi_{sm}(u)$ ne peut être nulle

Spécification d'une utilité logit

- Identifiabilité ?
  - On ne peut estimer à la fois les Θ_u et le θ_ε des erreurs ε, donc on fixe le θ_ε à 1 (plus simple) ou un des θ_u^k à 1
  - Pour évaluer l'arbitrage entre un facteur quelconque et un prix, il faut prendre le ratio des coefficients θ^k / θ_P
- Constantes modales θ_m^0
  - Pour *M*-1 options sur *M*, on spécifie un paramètre de 'constante modale'
  - Ce paramètre prend en compte les attributs non mesurés
  - Dans l'estimation, il permet de reproduire exactement la répartition obervée
- Caractères fixes du client
  - Ex. Y_u^ℓ : revenu, motorisation, abonnement TC
  - Appartenance à un groupe (tranche d'âge, sexe, profession): codée par variable binaire (dummy) Y_u^ℓ = 1 si oui ou 0 sinon

Commentaires sur la spécification

- Effets d'agrégation
  - La formule linéaire dissocie les Y^{ℓ} et les X^{k}
  - Souvent on prend un coef θ^{k} de Y^{k} identique pour tous les clients : alors c'est un regroupement, une agrégation
  - Ex. Y^{k} = durée de marche : le coefficient pourrait dépendre de la tranche d'âge !
- Cas important : l'appartenance à un groupe
  - le traitement par une variable binaire (dummy) permet de reproduire la répartition pour ce groupe
  - cela risque de « forcer » plutôt que de refléter l'effet réel (qui serait de particulariser les paramètres pour ce groupe)
  - L'appartenance simultanée à plusieurs groupes, est le plus souvent modélisée par la superposition des appartenances, et non comme leur croisement

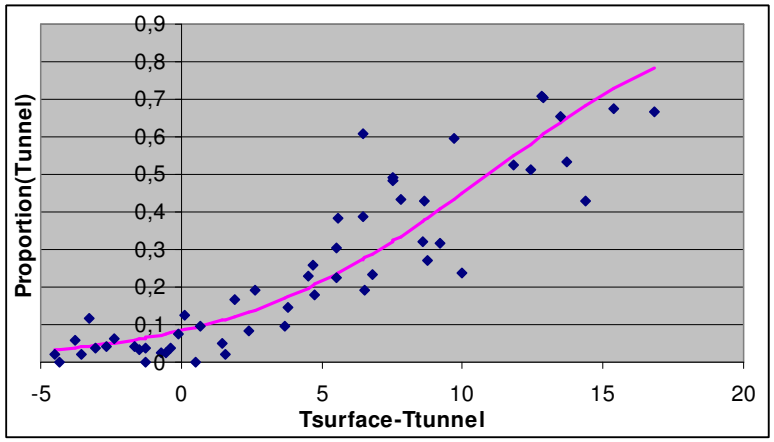
Exemples

- Tunnel à péage Prado-Carénage à Marseille
  - en concurrence avec réseau routier gratuit
  - Caractères Prix, Temps, Distance
- Liaison rapide Roissy-Place de l'Etoile
  - Métro rapide direct
  - Concurrent des TC plus lents, automobile et taxi

Exemple (1/2)

- Tunnel Prado-Carénage
  - Caractères et paramètres
  - Prix du tunnel <– coef = constante modale θ_0
  - T_{tunnel}, $T_{surface}$ <– coef = $\theta_{APT} = \alpha . \theta_0$/Péage
  - D_{tunnel}, $D_{surface}$ <– coef = $\theta_{dist} = \beta . \theta_0$/Péage
  - Groupes par tranche de ΔTemps et de ΔDistance
- Résultats du modèle logit « pur »
  - α = 9.2 €/h
  - β = 0.02 €/km

Exemple (2/2)



Simplifier un modèle

- Pourquoi simplifier ?
  - Si la valeur d'un paramètre ne diffère pas significativement de 0
  - Garder les paramètres significatifs : expliciter la partie utile
- Comment simplifier
  - Le coefficient *t* de Student : $t_\theta = \theta/SE_\theta$. Si $|t_\theta|<2$ alors la valeur 0 est dans $[\theta-2SE_\theta, \theta+2SE_\theta]$ intervalle de confiance à 95%
  - Mieux vaut utiliser la vraisemblance

Simplifier le modèle grâce à la théorie de la vraisemblance

- Principe
  - Soit θ' un sous-ensemble de *N'* paramètres extrait de θ en fixant les *N-N'* autres paramètres à 0
  - La variable $x = 2[\Lambda(\theta_{ML})-\Lambda(\theta'_{ML})]$ est distribué $\chi^2_{N-N'}$
  - Donc l'hypothèse « θ= θ' U 0 » a une probabilité critique

$$\Pr\{ \chi^2_{N-N'} > x \}$$

- Exemple pour Prado-Carénage
  - Le modèle à APT unique a une Λ^* = -1 599.3
  - Le modèle logit à APT distribué a une Λ^* = -1 587.8
  - la probabilité de l'hypothèse « $\sigma_{APT} = 0$ » est < 10^{-4}

Moindres carrés (1/2)

- Probabilité de l'option 1 : $p_X = F(\Theta.X)$
  $$F^{-1}(p_X) = \Theta.X$$
- Proportion observée $\hat{p}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} y_i$
- n_x individus avec $X = x$, $y_i \in \{0,1\}$
- En moyenne $\hat{p}_x = p_x$
- Approximation $F^{-1}(\hat{p}_x) \approx \Theta.x + \frac{\hat{p}_x - p_x}{F'(F^{-1}(\hat{p}_x))}$

Moindres carrés (2/2)

- Modèle $F^{-1}(\hat{p}_x) \approx \Theta.x + \eta_x$
  $$\eta_x = (\hat{p}_x - p_x)/F'(F^{-1}(p_x))$$
- Erreur
  - Moyenne nulle
  - Variance $\mathrm{var}[\eta_x] = \frac{p_x.(1-p_x)}{n_x F'^2(F^{-1}(p_x))}$
- Estimateur des moindres carrés pondérés
  $$\hat{\Theta} = [\frac{x.x^t}{\mathrm{var}[\eta_x]}]^{-1}[\sum_x \frac{F^{-1}(\hat{p}_x)}{\mathrm{var}[\eta_x]}x]$$
  - Il est consistant et approx. normal : asymptotiquement, la matrice de covariance est $[x.x^t / \mathrm{var}(\eta_x)]^{-1}$

Récapitulation sur le modèle logit

- Spécification flexible
- Mais dont les hypothèses méritent discussion
- Estimation par maximum de vraisemblance
- Éventuelles simplifications
- Estimation par moindres carrés

Conclusion

- Interdépendance de la spécification et de l'estimation
- La forme de l'observation dépend des spécifications 'maîtrisées'