Utiliser l'économétrie : conduire une modélisation économétrique

Fiches méthodes & outils dans l'accompagnement des porteurs de projet du Fonds de dotation Paris 2024 à l'évaluation d'impact

Dans le cadre du dispositif d'accompagnement des porteurs de projet du Fonds de dotation Paris 2024, un accompagnement dédié sur la mesure d'impact est proposé.

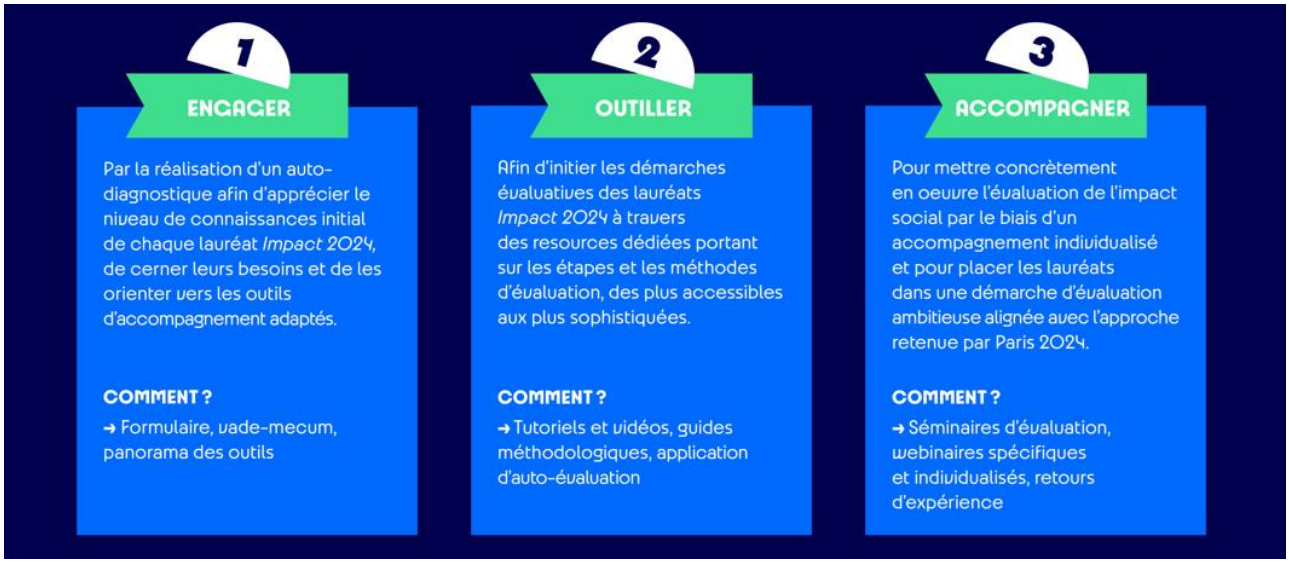
Pour votre organisation, les objectifs sont les suivants :

- Apporter des éléments de connaissance sur les impacts de vos projets
- Communiquer en interne et en externe
- Améliorer vos projets
- Essaimer

Pour Paris 2024, les évaluations d'impact permettront de :

- Contribuer à l'évaluation de la stratégie Impact & Héritage
- Apporter de la lisibilité sur la valeur ajoutée de vos projets
- Léguer un héritage méthodologique

Différents niveaux d'accompagnement sont proposés concernant la mesure d'impact :



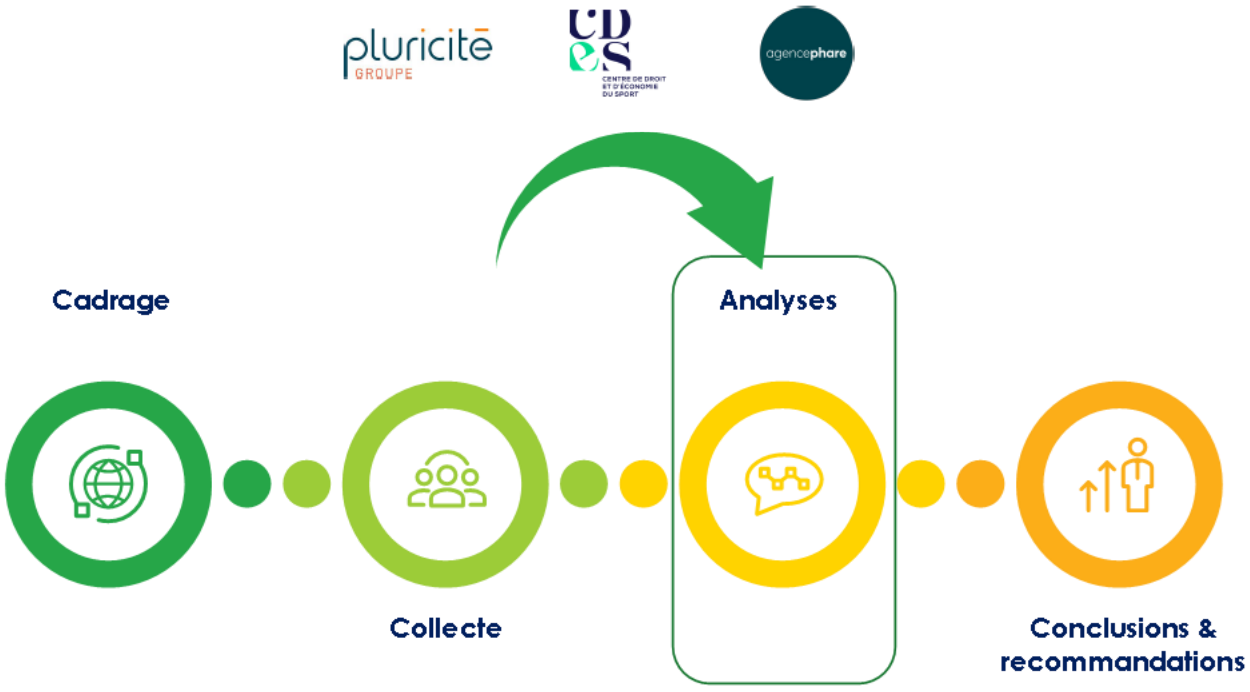
Les documents d'application s'inscrivent dans une logique d'outillage, visant à sécuriser les démarches d'évaluation d'impact des porteurs de projet à travers des vade-mecum portant sur les méthodes d'évaluation et les étapes et les outils à mettre en place. Avec ces documents, il s'agit de favoriser l'acquisition des notions clefs, la compréhension des tenants et aboutissants de la mesure d'impact et l'appropriation de solutions concrètes pour engager le travail, l'organiser – le baliser.

Retrouvez les documents d'application sur les sujets suivants :

Fiches outils	Fiches processus
⊕ Organiser un benchmark ⊗ Conduire des entretiens ⊗ Conduire une étude de cas ⊗ Utiliser la facilitation graphique ⊗ Utiliser l'infographie ⊗ Utiliser la datavisualisation ⊗ Utiliser les personae ⊗ Animer un focus groups évaluatif ⊗ Utiliser l'observation participante ⊗ Mobiliser la méthode des scénarios ⊗ Elaborer un référentiel d'évaluation ⊗ Construire un diagramme logique d'impact ⊗ Mener des enquêtes bénéficiaires ⊗ **Utiliser l'économétrie**	⊗ Bâtir un protocole d'évaluation ⊗ Formuler ses questions évaluatives ⊗ Opter pour l'évaluation participative ⊗ Vérifier l'évaluabilité du projet ⊗ Formuler des recommandations ⊗ Communiquer et rendre utile la démarche ⊗ Faire un contrôle qualité de son rapport

Retrouvez les documents d'application et d'autres outils sur la plateforme :

https://accompagnementimpact2024.org/



1 L'essentiel, en un coup d'œil

La modélisation économétrique, c'est quoi ?

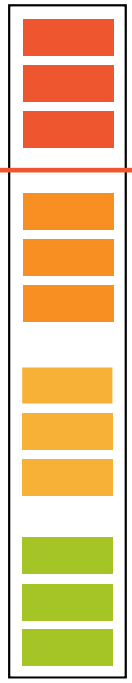
L'économétrie est un outil d'analyse quantitative, permettant de vérifier l'existence de certaines relations entre des phénomènes et de mesurer concrètement ces relations sur la base d'observations de faits réels. L'économétrie se base donc sur les mathématiques et les statistiques afin d'identifier des relations entre différentes variables. Mais tout comme vous n'avez pas besoin d'être un pro en microprocesseur pour utiliser un ordinateur, il n'est pas foncièrement nécessaire d'être un crack en math pour pouvoir se servir de l'économétrie comme outil d'analyse.

La démarche économétrique consiste donc à représenter à l'aide d'équations le comportement d'un phénomène observé et à estimer les coefficients des équations en recourant à l'historique du phénomène et ceci dans le but de le comprendre, de l'expliquer, de le reproduire et de le prévoir.

Concrètement cela permet de comprendre l'impact « toutes choses étant égales par ailleurs » de plusieurs variables explicatives sur une variable à expliquer (qu'elle soit quantitative, qualitative binaire, qualitative ordonnée, comme par exemple le fait de prendre une licence sportive ou pas, d'avoir de meilleurs résultats scolaires, de trouver un job…). On parle d'analyse « toutes choses étant égales par ailleurs » (ou *Ceteris paribus* en latin pour briller en société) lorsqu'on laisse de côté un certain nombre de paramètres d'une situation donnée pour n'en étudier qu'un seul à la fois : ainsi, dans un modèle théorique l'influence de la variation d'une quantité (la variable explicative) sur une autre (la variable expliquée) est examinée à l'exclusion de tout autre facteur (exemple, l'âge, le niveau de formation, le fait d'habiter dans un quartier prioritaire…).

L'évaluation économétrique de l'impact d'un projet a pour objectif de quantifier les effets du projet sur un résultat (état de santé, salaire, emploi, niveau de compétences, etc.) mesuré sur une population. L'exercice est particulier : les bénéficiaires possèdent des caractéristiques difficiles à mesurer (motivation, processus de sélection complexe, situation antérieure, …). Une simple comparaison de la situation des bénéficiaires avec celle des non-bénéficiaires ne permet donc pas en général de séparer les effets propres du programme de ceux qui résultent de ces caractéristiques. Dès lors, le principal défi de l'évaluation est d'identifier le contrefactuel, c'est-à-dire la situation, fondamentalement inobservable, qui aurait prévalu en l'absence du projet. Pour répondre à ce problème, on utilise l'économétrie.

Niveau de complexité de l'outil !



COMPLEXE	L'outil / la méthode nécessite un niveau d'expertise et d'expérience relativement important et suppose un fort engagement de ressources (en temps, etc.).
ELABORÉ	L'outil / la méthode requiert un niveau d'expertise et d'expérience élevé qui pourra être compensé par un fort niveau d'engagement.
INTERMÉDIAIRE	L'outil / la méthode implique une relative exigence technique et implique un engagement de ressources (temps passé...) assez raisonnables.
ACCESSIBLE	L'outil / la méthode peut se mettre en place relativement facilement, sans appeler un niveau d'expertise et / ou d'expérience dédié.

Atouts

- Un outil d'analyse quantitative puissant
- Un outil qui permet de convertir les propositions qualitatives (comme « la relation entre deux variables ou plus est positive ») en propositions quantitatives
- La capacité à déterminer des conclusions tirées à partir des données et d'un ensemble d'hypothèses...

Limites

- Le piège central de l'économétrie ? La différence entre causalité et de corrélation.
- Les résultats sont à interpréter avec précaution : des variables peuvent avoir été oubliées ; la corrélation observée peut être accidentelle
- Poser un modèle économétrique c'est-à-dire le modèle statistique ou mathématique qui représente la relation entre deux ou plusieurs variables est souvent complexe et susceptible d'erreurs
- Il est nécessaire de disposer d'un jeu de données suffisamment important. En effet, plus on dispose d'un nombre important d'observations plus le modèle sera fiable et les prédictions satisfaisantes

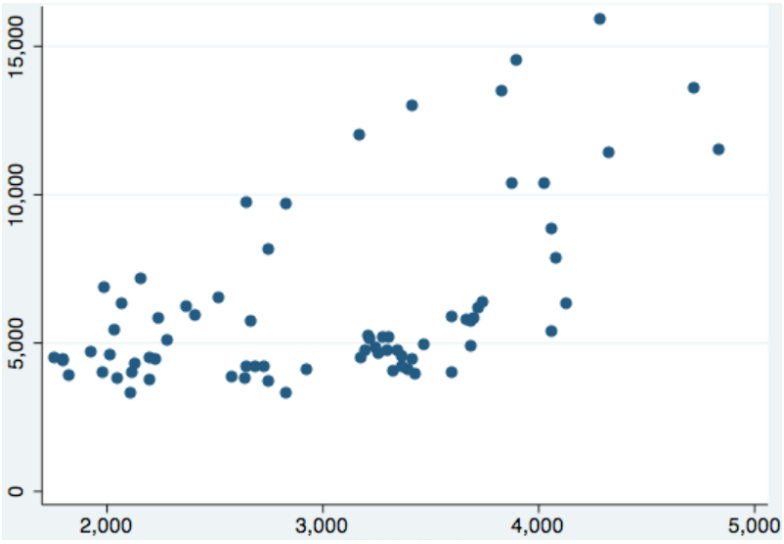
2 Un modèle simple à deux variables

Commençons donc avec un modèle simple à deux variables. Vous souhaitez tester s'il existe une relation linéaire entre les variables y et x. Pour le poser sous la forme d'une équation, votre but est d'estimer les coefficients de l'équation ci-dessous (avec un terme d'erreur mais qui n'est pas présenté dans cette fiche pour plus de simplicité).

$$y = \beta_0 + \beta_1 x$$

Le modèle ci-dessus est un modèle du niveau mathématique de classe de seconde, le fameux y = f(x).

En réalisant un graphique entre les 2 variables (par exemple l'âge d'un bénéficiaire de votre projet et le montant de son précédent salaire avant d'intégrer le projet) pour chacun des bénéficiaires de votre projet, cela donne ça (chaque point correspondant à un demandeur d'emploi):



Source : www.captaineconomics.fr

Il semble bien qu'il y ait une relation croissante entre l'âge et son salaire. Graphiquement on remarque en effet une tendance : plus l'âge est élevé, plus son salaire est élevé (en moyenne). C'est bien beau tout cela, mais si vous êtes évaluateur du projet et que vous allez voir votre directeur en lui disant cela, il va vous dire "tu es bien sympa mon petit, mais il me faut des chiffres précis et être sûr que cette relation est significative et n'est pas le fruit du hasard".

Et là, deux solutions s'offrent à vous. (1) La technique de l'autruche : vous laissez traîner le problème et inventez des chiffres au hasard en faisant des calculs à l'arrache sur Excel. (2) Vous ouvrez un logiciel économétrique et faites une étude simple de régression linéaire.

Et aussi incroyable que cela puisse paraître, la seconde solution est beaucoup plus simple et rapide que la première. Il suffit d'un seul calcul dans un logiciel spécialisé (type Stata, SAS, Sphinx...) pour avoir l'ensemble des réponses aux questions de votre directeur. C'est là qu'intervient la magie de la "régression linéaire et de la méthode des moindres carrés ordinaire". Sans rentrer dans les détails, le but de cette technique est de tracer une droite sur le graphique précédent, tel que l'écart (au carré) entre les points du nuage et votre droite de régression soit le plus faible possible.

Vous pouvez le faire avec une règle et un crayon de bois, et tâtonner en calculant pour chaque point l'écart entre les points et votre droite. Ou bien utiliser une fonction automatique sur un logiciel économétrique. Dans Stata, la commande "regress price weight" indique que vous souhaitez faire une régression linéaire entre votre variable y (âge) et votre variable x (salaire). Ensuite il faut savoir lire les résultats de ces tableaux pétris de chiffres.

Pour rappel des mathématiques de classe de seconde, le coefficient va correspondre à l'ordonnée à l'origine et votre coefficient Béta 1 à la pente de la droite de régression. Sur le tableau ci-dessous qui consiste à regarder le lien entre le poids d'une voiture et son prix de vente, vous avez ces deux informations : Béta 0 correspondant au coefficient de la constante (_cons) et bleu, soit -6,70 et Béta 1 correspondant au coefficient devant le poids (weight) en rouge, soit 2,05. Votre équation initiale peut donc s'écrire: " price = -6,70 + 2,05 * weight "

Source	SS	df	MS
Model	184233937	1	184233937
Residual	450831459	72	6261548.04
Total	635065396	73	8699525.97

Number of obs	= 74
F(1, 72)	= 29.42
Prob > F	= 0.0000
R-squared	= 0.2901
Adj R-squared	= 0.2802
Root MSE	= 2502.3

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	2.044063	.3768341	5.42	0.000	1.292857	2.795268
_cons	-6.707353	1174.43	-0.01	0.995	-2347.89	2334.475

Source : www.captaineconomics.fr

En moyenne donc, et en se basant sur les données de cet échantillon, une voiture de 2000 livres (lbs) coûtera environ "-6,70 + 2,05 * 2000 = 4000 dollars).

Mais cette relation est-elle significative ? Pour cela, il convient de regarder le "t-stat" du tableau de régression (en vert), et de comparer la valeur de ce "t'stat" (dans notre cas 5,42) à des valeurs déterminées statistiquement. Il est possible de ne pas rentrer dans les calculs, et uniquement de se baser sur cette règle "si le t-stat est supérieur à 1,96 en valeur absolue, alors la variable est significative". Ici c'est le cas, donc vous pouvez aller voir votre directeur en lui disant "eh chef, une voiture qui pèse 1 livre de plus qu'une autre coûtera en moyenne 2,05 dollars de plus, et la relation est significative". Si vous voulez même faire le savant, vous pouvez lui dire "[...] et cette relation est significative avec un intervalle de confiance de 95%", la valeur de t-stat de 1,96 correspondant à cet intervalle de confiance.

3 Un modèle avec plusieurs variables explicative

Nous allons continuer avec notre modèle qui essaye d'expliquer le prix d'une voiture (notre variable dépendante "Y") en fonction de variables explicatives (X1, X2, D1).

La variable X1 correspond comme dans l'article précédent au poids d'une voiture. La variable X2 représente la consommation en carburant de la voiture et la variable D1 est une variable indicatrice, prenant la valeur 0 si la voiture est une voiture domestique (donc américaine dans notre exemple) ou 1 si la voiture est une voiture étrangère. Notre modèle ressemble donc à cela :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D_1 + \varepsilon$$

Le dernier terme correspond au terme d'erreur, qui représente la déviation entre ce que le modèle prédit et la réalité. Comme précédemment notre but ici va être de déterminer (1) les variables significatives, c'est à dire voir si les différents coefficients sont différents de 0, (2) la valeur de la constante alpha et des différents coefficients "beta" qui permettent de minimiser l'erreur entre notre droite de régression linéaire estimée et les valeurs réelles de Y et enfin (3) la précision de notre modèle, en utilisant, entre autre, le "R-squared". Le raisonnement est le même qu'avec seulement une variable.

```
. regress price weight mpg foreign      Source: CaptainEconomics.fr
```

Source	SS	df	MS
Model	317252881	3	105750960
Residual	317812515	70	4540178.78
Total	635065396	73	8699525.97

Number of obs	= 74
F(3, 70)	= 23.29
Prob > F	= 0.0000
R-squared	= 0.4996
Adj R-squared	= 0.4781
Root MSE	= 2130.8

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	3.464706	.630749	5.49	0.000	2.206717	4.722695
mpg	21.8536	74.22114	0.29	0.769	-126.1758	169.883
foreign	3673.06	683.9783	5.37	0.000	2308.909	5037.212
_cons	-5853.696	3376.987	-1.73	0.087	-12588.88	881.4934

Les étapes à suivre sont les suivantes :

Etape 1: Tester la significativité des variables. Pour cela, il suffit de regarder le "t-stat" (t) ou bien la P-value (P>?t?), et comparer ces valeurs à des "valeurs seuils". Pour faire simple, une variable est significative avec un intervalle de confiance de 95% si son t-stat est supérieur à 1,96 en valeur absolue, ou bien si sa P-value est inférieure à 0,05. Dans notre exemple, on voit que la variable "mpg", qui correspond à la consommation en essence de la voiture n'est pas significative (t-stat trop faible en valeur absolue et P-value trop forte). De plus, l'intervalle de confiance à 95%, allant de -126.17 à 169.99 comprend la valeur 0. Il est donc impossible de rejeter l'hypothèse Béta2 = 0.

Les deux autres variables "weight" et "foreign" sont significatives (t-stat de 5,49 et 5,37 donc supérieur à la valeur seuil de 1,96). De plus, l'intervalle de confiance ne comprend pas la valeur 0. Pour Béta1 par exemple, l'intervalle de confiance permet de dire "je suis sûr à 95% que la valeur de Béta1 se trouve entre 2,20 et 4,72. Le coefficient (=3.467 pour Béta1 par exemple) correspond exactement au milieu de l'intervalle de confiance de la variable.

Mais on fait quoi maintenant qu'on a trouvé que la variable "consommation de la voiture" n'est pas significative? Et bien on relance la régression, mais en supprimant la variable. En effet, les résultats de la régression peuvent être modifiés par l'inclusion de variables non significatives, et il est donc préférable d'analyser le résultat d'une régression finale contenant uniquement des variables significatives. Voici donc le résultat de notre nouvelle régression.

```
. regress price weight foreign
```

Source: CaptainEconomics.fr

Source	SS	df	MS
Model	316859273	2	158429637
Residual	318206123	71	4481776.38
Total	635065396	73	8699525.97

Number of obs	= 74
F(2, 71)	= 35.35
Prob > F	= 0.0000
R-squared	= 0.4989
Adj R-squared	= 0.4848
Root MSE	= 2117

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	3.320737	.3958784	8.39	0.000	2.531378	4.110096
foreign	3637.001	668.583	5.44	0.000	2303.885	4970.118
_cons	-4942.844	1345.591	-3.67	0.000	-7625.876	-2259.812

Etape 2 (nouvelle régression): C'est bon, nos deux variables sont significatives (t-stat > 1,96 en valeur absolue).

Etape 3: Étude des coefficients. La valeur estimée de Béta1 est égale à 3,32 et celle de Béta3 à 3637. Comment lire cela? Cela signifie que "toutes choses égales par ailleurs", une voiture pesant une livre (unité de masse américaine) de plus, coûtera en moyenne 3,32 $ de plus. Même raisonnement en ce qui concerne l'analyse du coefficient de notre variable indicatrice : "toutes choses égales par ailleurs", une voiture étrangère coûte en moyenne 3637 dollars de plus qu'une voiture américaine.

Etape 3: Mais quelle est la précision de notre modèle ? Pour cela, il est possible de regarder le "R-squared", qui mesure la proportion de la variance de Y (variable dépendante) qui est expliquée par la variation des toutes les variables explicatives. Le R-squared est par construction compris entre 0 et 1 ; plus on se rapproche de 1, plus le modèle est précis. Dans notre exemple, 49% de la variation de Y peut être expliquée par les variations de X1 et D1. En gros, c'est pas mal mais pas terrible non plus. Il manque en effet de nombreuses variables à notre

modèle pour que celui-ci permette d'estimer avec précision le prix d'une voiture en fonction de ses caractéristiques.

Il n'existe pas de valeur du R-squared à partir de laquelle le modèle peut-être considéré comme bon ou mauvais (cela dépend du modèle). Pour donner un ordre d'idée dans cette situation, un R-squared proche de 0,8 est signe d'un bon modèle, tandis que si votre R-square est proche de 0,2 , c'est pas la folie (peut-être pas mal de variables omises).

4 Dans quel cas utiliser la modélisation économétrique

4.1 Pourquoi ?!

- Lorsque l'on cherche à expliquer un phénomène observable afin de voir les caractéristiques qui l'influence
- Pour pouvoir faire des prédictions ou influer sur le futur de ce phénomène

4.2 A quel moment ?!

- Au début de l'évaluation, pour tester des hypothèses causales
- L'approche qualitative permettra ensuite d'explorer les causalités, de les expliquer finement

5 Exemples plus complexes

Ex1 : Une variable à expliquer quantitative :

Tentons d'expliquer les variations dans le poids des enfants à la naissance (la variable à expliquer qu'on note « poids » et qui est égale au poids en kilos de l'enfant à la naissance) en considérant le sexe (noter « sexe » qui vaut 1 si l'enfant est un garçon, 0 si c'est une fille), l'ordre de naissance (noté « rang » et qui est égale à 1 si 1er enfant,2 si 2eme enfant, …), le revenu familiale (noté « revenu ») et le nombre moyen de cigarettes fumés quotidiennement durant la grossesse (noté « cigarette »).

On considère alors le modèle suivant :

$$poids = \beta_0 + \beta_1 masculin + \beta_2 rang + \beta_3\, log(revenu) + \beta_4 cigarettes + \mu$$

Avec β_i les estimateurs du modèle à estimer, β_0 la constante à estimer et μ le terme d'erreur (une variable aléatoire qui résume tout ce que le modèle n'explique pas). Pour chaque variable décrite, on dispose des données empiriques pour N observations.

Le passage de la variable revenu en mode logarithmique va permettre de lisser la série et d'estimer l'élasticité du coefficient (si le revenu augmente de 1€ alors le poids de l'enfant va varier de X%). On va alors appliquer la méthode dite des Moindres Carrées Ordinaire (MCO) pour estimer les différents coefficients et vérifier la validité du modèle, en cherchant :

- Quel est le pouvoir explicatif du modèle ? Est-ce la liaison découverte entre la variable à expliquer et les variables explicatives est significative ? (c'est-à-dire transposable dans la population et non pas propre à l'échantillon observé)
- Quel est l'apport marginal de chaque variable explicative dans l'explication des valeurs de la valeur à expliquer ? (c'est-à-dire un paramètre est-il significativement différent de 0 ?)
- Quelle sont les propriétés (notamment la précision) des paramètres « β » obtenus ? (biais, variance)
- Quelle sera la qualité de la prédiction des valeurs de la variable à expliquer à partir des valeurs des variables explicatives ? (Intervalle de prédiction, fourchettes, …)

Ex2 : Une variable à expliquer qualitative :

En 2015, la Mission locale de Lyon a cherché à expliquer le taux de sortie positif (emploi, formation, contrat en alternance…) en fonction de différentes variables (sexe, niveau de qualification, nationalité, type d'hébergement, couverture Sociale, ayant droit / bénéficiaire, top Complémentaire Santé, RSA, RQTH, au moins un enfant, permis pour un véhicule motorisé, résidence en QPV/QVA, niveau de priorité du quartier, nombre de contact avec la ML) des jeunes qu'ils ont accompagnés durant l'année 2014.

Ici la variable à expliquer (la sortie) n'est plus numérique mais binaire, elle vaut donc 1 si le bénéficiaire est sorti positivement du dispositif et 0 sinon. Elle est définie dans la suite par la variable Y et l'ensemble des variables explicatives par la matrice X. Soit :

$$y_i = \begin{cases} 1 \; si \; l'individu \; i \; a \; connu \; une \; sortie \; positive \; de \; son \; accompagnement \\ 0 \; si \; l'individui \; n'a \; pas \; connu \; une \; sortie \; positive \; de \; son \; accompagnement \end{cases}$$

On va ainsi faire les hypothèses suivantes :

- Il existe une variable latente du modèle, inobservable, qui représente la probabilité de sortie du jeune n°i : $Y_i^* = X_i\beta + \mu_i$
- Il existe une certaine valeur seuil (noté « c ») au-delà de laquelle la proportion des { Yi = 1} l'emporte sur celle des { Yi = 0}

On a ainsi pu construire le modèle suivant :

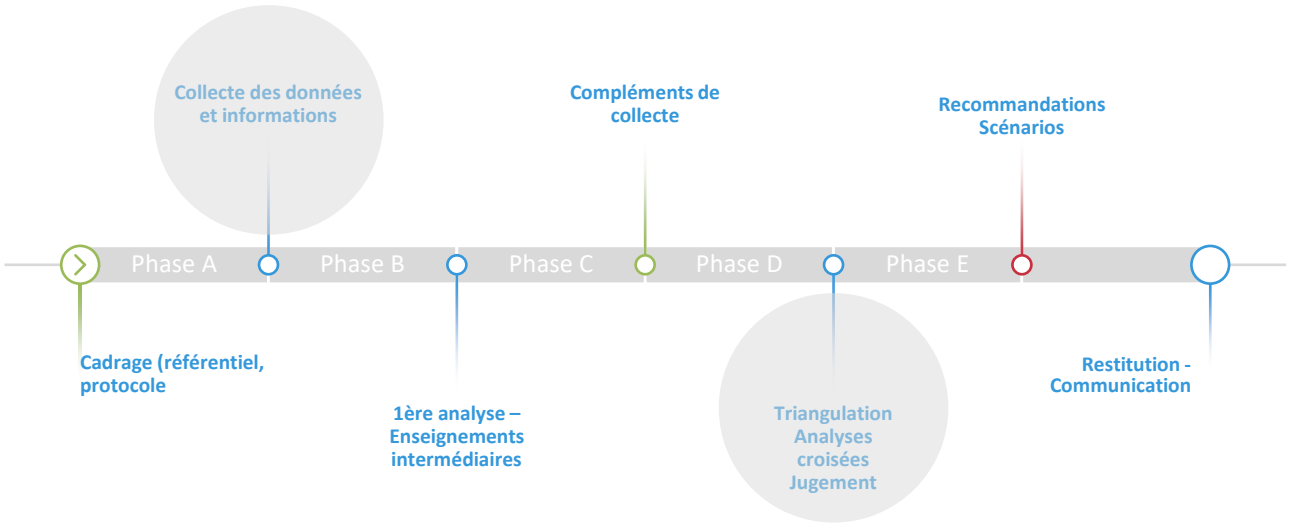
$$y_i = \begin{cases} 1 \; si \; Y_i^* > c \\ 0 \; si \; Y_i^* < c \end{cases}$$

Avec $Y_i^* = X_i\beta + \mu_i$

Ce modèle ne détermine pas exactement la réalisation de la sortie positive, mais fournit plutôt une mesure de la proportion d'observations pour lesquelles on constate une sortie positive.

Le modèle va donc dépendre de la distribution statistique de μ (le terme d'erreur). Si on suppose que celui-ci suit une loi logistique on appliquera un modèle logit binaire et si on celui-ci suit une loi normale on appliquera un modèle probit binaire.

6 Etapes de mise en œuvre de la modélisation économétrique



- **Etape 1 :** Définir la variable à expliquer et les variables explicatives à considérer, par exemple du côté des variables explicatives le fait d'avoir trouvé un travail, d'avoir pris une licence, d'avoir de meilleurs résultats scolaires, et de l'autre l'intégration dans le projet, la durée de l'accompagnement, les caractéristiques du bénéficiaire, etc….

- **Etape 2 :** Nettoyage des données : transformation en variables binaires ou dichotomiques des variables ouvertes (ex : 0 si homme, 1 si femme)
- **Etape 3 :** en fonction de la nature de la variable à expliquer et des données disponibles choisir le modèle le plus adapté. Le cas le plus simple reste un modèle MCO pour une variable quantitative et un modèle logit/probit pour les variables qualitatives **binaires** mais il existe une multitude de modèles : modèle linéaire, probit, logit, bivariés, multivariés, à variables qualitatives polytomiques, non-ordonnés, tobit, à variable dépendante limité (liste non exhaustive)
- **Etape 4 :** vérifier les hypothèses d'hétéroscédasticité et d'auto-correlation
- **Etape 5 :** vérifier le niveau de validité du modèle, noté R^2 (plus il s'approche de 1 plus le modèle est bon)
- **Etape 6 :** en fonction du modèle interprétation des résultats
- **Etape 7 :** calcul éventuel des projections

7 Pour en savoir plus

- http://eric.univ-lyon2.fr/~ricco/cours/cours/Generalites%20Econometrie.pdf
- https://halshs.archives-ouvertes.fr/cel-01261163/document
  http://eric.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression.pdf