Analyse de la participation au marché du travail et des déterminants du salaire : Application d'un modèle de sélection endogène et essais d'utilisation des techniques en grande dimension

GUSAROV
Nikita

Sous la direction de Michal Urdanivia

Université Grenoble Alpes
Faculté d'Economie et Gestion

Mémoire de Master 1 MIASHS
Parcours : Chargé d'études économiques et statistiques

Année universitaire 2018-2019

L'Université n'entend donner aucune approbation ni l'approbation aux opinions émises dans ce travail : ces opinions doivent être considérées comme propres à leur auteur.

Sommaire

Introduction

Ce mémoire se base principalement sur deux articles dont un est l'article méthodologique fondamental présenté par Heckman en 1983 et l'autre un *working paper* portant sur l'application de ces méthodes écrit par Aeberhardt et al. en 2007. Les deux traitent des problèmes de la présence de sélection endogène lors des études du marché de travail.

Nous utilisons aussi dans ce travail les œuvres de Belloni et al. sur les implémentations des méthodes de pénalisation pour la sélection des variables en présence des données de grande dimension, ainsi que le travail fondamental de Tibshirani publié en 1996 introduisant le terme de pénalisation *L1* (LASSO).

De cette façon ce travail a un double objectif. D'un côté nous essayerons de répliquer les études d'Aeberhardt sous une forme simplifiée sur les données de l'Institut National de la Statistique et des Etudes Economiques. D'autre, nous allons faire une application des modèles pénalisés plus complexes en nous plaçant dans une situation de traitement des données en haute dimension induite par introduction des relations non-linéaires, afin de mieux comprendre la loi décrivant la formation du salaire sur le marché.

Le travail est divisé en 3 parties, dont la première comprend les aspects méthodologiques et théoriques qui seront utilisés dans le corps de ce travail. La partie suivante présente l'échantillon étudié et la partie 3 regroupe tous les modèles, parmi lesquels le modèle de sélection endogène traditionnel et les modèles pénalisées.

1 La Méthodologie

Le comportement de différents modèles sera étudié dans la situation où la variable dépendante n'est observable que pour une certaine partie d'échantillon. Nous allons procéder à partir des outils les plus simples, comme l'estimation d'un modèle simple par la méthode des moindres carrés ordinaires, vers des techniques plus complexes comme le modèle de sélection endogène, le modèle de sélection construite avec la pénalisation aux deux étapes. Cela nous permettra de comparer les performances des modelés différents et d'observer comment le traitement des différents biais change nos perceptions des relations sous-jacentes. Dans cette section nous décrivons les principes théoriques pour chaque modèle.

1.1 Les Modèles de Sélection Endogène

Supposons qu'on s'intéresse à l'influence de X sur Y mais on n'observe Y que si une variable indicatrice $D = 1$. Ceci peut correspondre à plusieurs cas :

- Non-réponse à une enquête. L'échantillon utilisé est composé des seuls répondants ($D = 1$) à l'enquête ;
- Du fait du mode d'échantillonnage, on n'observe Y que lorsque $Y > 0$ (troncature).
- Auto-sélection : on observe le salaire d'un individu que lorsque ce dernier a choisi d'être actif.

Dans ce travail nous nous plaçons dans le cas d'un modèle de sélection généralisé. C'est-à-dire qu'on considère le modèle de sélection suivant [1]:

$$\begin{cases} Y = X\beta_0 + \varepsilon \\ D = \mathbb{I}\{Z\gamma_0 + \eta \geq 0\} \end{cases} \quad (1)$$

On suppose qu'on observe (D, X, Z) mais Y est connue seulement si $D = 1$. On pose

[1] $X = (x_{i,j})_{1\leq i\leq N, 1\leq j\leq p} \in M_{N,p}$ avec N le numero des observations et p le numero des regresseurs (y compris la constante).
$Z = (z_{i,j})_{1\leq i\leq N, 1\leq j\leq m} \in M_{N,m}$ avec N le numero des observations et m le numero des regresseurs (y compris la constante).
$Y = (y_i)_{1\leq i\leq N} \in M_{N,1}$ et $D = (d_i)_{1\leq i\leq N} \in M_{N,1}$ avec $d_i \in \{0,1\}$

que ε et η sont a priori corrélés. Un tel modèle est dit modèle de sélection endogène (ou un modèle Tobit de type II).

Les hypothèses sur les erreurs pour ce modèle sont suivantes :

1. (ε, η) sont indépendants de (X, Z) ;
2. $\eta \sim \mathcal{N}(0,1)$;
3. $E(\varepsilon|\eta) = \delta_0 \eta$.

Les hypothèses 2 et 3 sont satisfaites lorsque η est gaussienne mais sont plus faibles en général.

On a alors :

$$E(Y|X, Z, \eta) = X\beta_0 + E(\varepsilon|X, Z, \eta) = X\beta_0 + E(\varepsilon|\eta) = X\beta_0 + \delta_0 \eta \quad (2)$$

Comme Y n'est pas observable que lorsque $D = 1$, on obtient[2] :

$$E(Y|X, Z, D = 1) = X\beta_0 + \delta_0 \lambda(Z\gamma_0) \quad (3)$$

Ici γ_0 est identifié par l'équation modélisant une variable binaire (Probit ou Logit) :

$$D = \mathbb{I}\{Z\gamma_0 + \eta \geq 0\} \quad (4)$$

A son tours, δ_0 et β_0 sont identifiés par la régression de Y sur $(X, \lambda(Z\gamma_0))$ conditionnellement sur $D = 1$. Où $\lambda(Z\gamma_0)$ est le ratio de Mills inverse (Green, 2002) déterminé comme :

$$\lambda(Z\gamma_0) = {f(Z\gamma_0)}\Big/{\mathrm{F}(Z\gamma_0)} \quad (5)$$

Alors, pour estimer ce modèle on peut suivre la procédure à deux étapes proposée par Heckman en 1979 :

1. Estimer $\widehat{\gamma_0}$ en construisant un modèle binaire pour les relations D, Z ;
2. Estimer $\widehat{\beta_0}$ et $\widehat{\delta_0}$ en régressant Y sur $X, \lambda(Z\widehat{\gamma_0})$.

Il existe d'autres méthodes pour résoudre ce problème, parmi lesquels :

- L'estimation par maximum de vraisemblance ;
- Autres techniques semiparametriques, décrites et étudiés par plusieurs auteurs, parmi lesquels nous pouvons citer : Pagan A. et Ullah A. (1999), Li Q. et Racine J.S. (2007).

Nous pouvons obtenir les effets marginaux moyens pour le modèle de sélection endogène en combinant les effets marginaux obtenus pour les deux étapes de modélisation en utilisant la logique suivante.

On détermine l'effet marginal moyen de la variable x_j sur y comme :

$$\frac{\partial E[Y \mid X = x,\ D = 1]}{\partial x_j}, \forall \mathrm{j} \in \{1, \ldots, \mathrm{p}\} \quad (6)$$

Sachant que :

$$E[Y \mid X = x, D = 1] = X\hat{\beta} + \delta\lambda(Z\hat{\gamma}),\quad Z\hat{\gamma}: D = 1 \quad (7)$$

On obtient[3] :

$$\frac{\partial E[Y \mid X = x, D = 1]}{\partial x_j} = E\left[\frac{\partial(X\hat{\beta})}{\partial x_j} + \frac{\partial(\delta\lambda(Z\hat{\gamma}))}{\partial x_j}\right] \quad (8)$$

Ou :

$$\lambda(Z\hat{\gamma}) = {f(Z\hat{\gamma})}\Big/{\mathrm{F}(Z\hat{\gamma})} \quad (9)$$

Avec une dérivée démontrée par Woodbridge (2005):

[2] $E(Y|X, Z, D = 1) = E[E(Y|X, Z, D = 1, \eta)|X, Z, D = 1] = E[E(Y|X, Z, D = 1, \eta)|X, Z, \eta \geq -Z\gamma_0] =$
$= E[X\beta_0 + \delta_0 \eta|X, Z, \eta \geq -Z\gamma_0] = X\beta_0 + \delta_0\lambda(Z\gamma_0)$

[3] Ici pour simplifier l'écriture on pose que $X = Z$ (ce qui signifie que $x_j = z_j$), mais les formules tiennent pour les autres cas aussi

$$\frac{\partial \lambda(x)}{\partial x} = -\lambda(x)\big(x + \lambda(x)\big) = -x\,\lambda(x) - \lambda^2(x) \quad (10)$$

Sachant cette dérivée et en appliquant les règles basiques de dérivation on obtient :

$$\frac{\partial E[Y \mid X = x, D = 1]}{\partial x_j} = \hat{\beta}_j - E[\delta\lambda(Z\,\hat{\gamma})\,(Z\,\hat{\gamma} + \lambda(Z\,\hat{\gamma}))\,(\frac{\partial(Z\,\hat{\gamma})}{\partial x_j})] \quad (11)$$

Où :

$$E\left[\frac{\partial(Z\,\hat{\gamma})}{\partial x_j}\right] = E\left[\frac{\partial(Z\,\hat{\gamma})}{\partial z_j}\right] = \hat{\gamma}_j \quad (12)$$

En supposant que les variables x_j et x_i (ainsi que z_j, z_i) sont indépendantes et $x_j \neq f(x_i)$, on obtient l'effet marginal pour le modèle complet, lequel pour une variable x_j est :

$$m_{x_j} = \frac{\partial E[Y \mid X = x, D = 1]}{\partial x_j} = \widehat{\beta_j} - \delta\widehat{\gamma_j} E\big[\lambda(Z\,\hat{\gamma})\big(Z\,\hat{\gamma} + \lambda(Z\,\hat{\gamma})\big)\big] \quad (13)$$

1.2 Les Modèles Pénalisés en Grande Dimension

Pour être claire nous décrivons d'abord les concepts de base d'un modèle LASSO simple pour après introduire ces concepts pour un modèle binaire. Dans la dernière partie de cette section nous proposons un aperçu sur les méthodes d'inférence pour les modèles de ce type figurant dans la littérature ainsi que les idées clés qu'on implémentera dans notre travail.

1.2.1 LASSO (Least Absolute Shrinkage and Selection Operator)

Le LASSO est un modèle pénalisé minimisant la somme des carrés des résidus sous contrainte que la somme des valeurs absolues des coefficients est inférieure à une constante. Par sa nature cette contrainte a une tendance de rendre nuls certains arguments, réduisant de cette façon la dimensionnalité du modèle la rendant plus interprétable. La méthode combine les meilleures traites des modèles pénalisés (la régression Ridge) et les modèles de sélection des variables, elle à la fois :

- Augmente la précision de prédiction par réduisant certains coefficients à 0 ;
- Facilite l'interprétation du modèle.

Dans ce travail les représentations des modèles proposés par R. Tibshirani et puis développés A. Belloni (en collaboration avec V. Chernozhukov) vont être utilisés.

Pour une régression linéaire de N observations avec p variables on peut écrire l'équation[4] :

$$y_i = \sum_{j=1}^{p} x_{i,j} b_j + \varepsilon_i, \quad \forall i \in 1, \dots, N \quad (14)$$

Ou on peut réécrire le modèle sous la forme matricielle comme[5] :

$$Y = X\beta + \varepsilon \quad (15)$$

L'estimateur de LASSO pour ce modèle peut être défini sous la forme présentée par Belloni et al. (2014), laquelle est plus compréhensible que celle de Tibshirani (1996) :

$$\hat{\beta} = \arg\min_b \Big(\sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{i,j} b_j)^2 + \lambda \sum_{j=1}^{p} |b_j|\gamma_j\Big), \quad \lambda > 0 \quad (16)$$

[4] Ici $x_{i,1} = const, \quad \forall i \in 1, \dots, N$

[5] $Y = \begin{pmatrix} y_1 \\ \dots \\ y_N \end{pmatrix}, \quad \beta = \begin{pmatrix} b_1 \\ \dots \\ b_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_N \end{pmatrix}, \quad X \in M_{N,p}$

Le LASSO est un cas particulier d'un modèle pénalisé par un terme de pénalisation étant :

$$\lambda \sum_{j=1}^{p} |b_j| \gamma_j \tag{17}$$

Dans ce cas λ représente le niveau de pénalité général de façon que plus λ est grande, moins des coefficients non-nuls aura le modèle final, et pour $\lambda = 0$ un modèle linéaire simple est obtenu. Quand à γ_j elle joue le rôle de la pénalité respectif a b_j (il est populaire de définir γ_j en fonction de la variance).

Pour faciliter la résolution une autre formulation du modèle est proposée, facilitant la solution au niveau du calcul. Il est proposé de réécrire les termes b_j comme[6] :

$$b_j = b_j^+ - b_j^- \text{ avec } b_j^+ \geq 0 \; ; \; b_j^- \geq 0 \tag{18}$$

Cela permet de transformer le programme d'optimisation avec p variables sous 2^p contraintes dans un programme d'optimisation ne contenant que $2p$ variables et $(2p + 1)$ contraintes.

$$\hat{\beta} = \underset{b_j^+,\ b_j^-}{\arg\min} \left(\sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{i,j}(b_j^+ - b_j^-))^2 + \lambda \sum_{j=1}^{p} (b_j^+ - b_j^-)\gamma_j \right), \quad \lambda > 0 \tag{19}$$

1.2.2 Régression Logistique Pénalisé

Maintenant nous allons présenter la possibilité d'incorporer les idées de pénalisation dans les modèles binaires. Le cas le plus simple est un modèle Logit pénalisé, car elle pose le moins des problèmes pour le calcul. Le problème de régression logistique pénalisée peut être reformulé sous la forme d'un problème de résolution d'un programme quadratique. Dans cette partie nous suivons le raisonnement présenté par Arthur Charpentier (Pr. A Montréal et Ecole Polytechnique)[7], des explications plus détaillées peuvent être trouvées dans le travail de Lee et Algamal (2015).

$$\log(\mathcal{L}) = \frac{1}{n} \sum_{i=1}^{n} y_i(\beta_0 + x_i^T \beta) - log[1 + \exp(\beta_0 + x_i^T \beta)] \tag{20}$$

Comme la fonction est concave c'est possible de le reformuler en utilisant l'approximation quadratique de log-vraisemblance.

$$\log(\mathcal{L}) \approx \log(\mathcal{L})' = \frac{1}{n} \sum_{i=1}^{n} \omega_i [z_i + (\beta_0 + x_i^T \beta)]^2 \tag{21}$$

Où z_i est :

$$z_i = (\beta_0 + x_i^T \beta) + \frac{y_i + p_i}{p_i[1 - p_i]} \tag{22}$$

Avec p_i étant la prédiction :

$$p_i = \frac{\exp(\beta_0 + x_i^T \beta)}{1 + \exp(\beta_0 + x_i^T \beta)} \tag{23}$$

Et ω_i représente les poids :

$$\omega_i = p_i[1 - p_i] \tag{24}$$

Ce que nous donne un problème de MCO pénalisé.

---

[6] Lasso, (17/06/2009), Machinelearning.ru, site web

[7] « Classification from scratch, penalized LASSO logistic », partie 5/8, publié le 04/06/2018, mode d'accès : https://freakonometrics.hypotheses.org/52894

1.2.3 L'Inférence Pour les Modèles Pénalisés

Il existe plusieurs stratégies d'inférence pour les modèles pénalisés. Parmi les stratégies les plus connues nous pouvons énumérer :

- *Bootstrap*, suivant les idées de Belloni et al. (2018) qui dans leurs articles proposent d'utiliser *block multiplier procedure* or *nonoverlapping block boostrap* pour étudier les coefficients obtenus, démontrant son efficacité par une application sur les modèles temporelles et spatiales ;
- Réestimation du modèle sélectionné par LASSO par MCO simple, l'efficience de cette technique est montrée également par Belloni et al. (2013) ;
- Inférence post-sélection conditionnelle, étudiée par Lee et al. (2016) ;
- LASSO Bayessien, proposée par Park et Castella (2008) ;
- *Data splitting*, qui est utilisée pour la génération des *p*-valeurs. Cette méthode est illustrée par Wippert et al. (2017) dans leur étude.

Dans notre travail nous allons suivre les idées proposées par Belloni et al. (2013). Une réestimation des modèles basés sur les variables sélectionnées par LASSO nous permettra d'obtenir de résultats plus précis avec des estimateurs de la variance plus proches à la réalité, sans faire trop d'effort et sans faire appel à des techniques complexes, telles que Bootstrap.

Nous allons implémenter un algorithme de cross-sélection pour améliorer la précision de la sélection. L'algorithme de cross-sélection sera effectué pour plusieurs valeurs de λ (regularisant la puissance de la pénalisation) et nous permettra de choisir un modèle suffisamment creux (*sparse model*) pour notre cas. Nous choisissons λ suivant l'évolution des moyennes d'erreurs sur plusieurs *folds* afin de ne pas tomber dans une situation de sur-apprentissage.

Il faut noter que selon les hypothèses de Lee et al. (2016) le choix des variables risque d'être biaisé parce que LASSO peut sélectionner des variables différents conditionnels sur les jeux des données. Toutefois il existe des travaux qui s'opposent à cette supposition. Pour la référence on donne le travail de Zhao et al. (2017).

Nous automatisons la procédure pour le logiciel Python avec utilisation des fonctions du package '*SciKit-Learn*'. Le code est présent dans les annexes.

2 Les Données

Les données sont issues de l'Enquête Emploi Continue 2012 (EEC 2012) effectuée par l'Institut national de la statistique et des études économiques (INSEE). Cette enquête donne un aperçu du marché du travail de manière structurelle et conjoncturelle. C'est la seule source fournissant une mesure des concepts d'activité, de chômage, d'emploi et d'inactivité tels qu'ils sont définis par le Bureau international du travail (BIT) selon les indications d'INSEE. De plus, elle s'inscrit dans le cadre des enquêtes "Forces de travail" défini au niveau européen ("*Labour Force Survey*"). Cela permettra à répliquer les résultats obtenus en futur et les tester sur les autres données similaires.

Ces données incorporent l'information sur l'emploi, le chômage, la formation, l'origine sociale, la situation un an auparavant, et la situation principale mensuelle sur les douze derniers mois.

2.1 La Composition D'Echantillon

La formation d'échantillon suit en général la méthodologie d'Aeberhardt et al. (2007). L'échantillon dans notre cas comprenne les femmes françaises de 15 à 60 ans dont les deux parents ont été français à la naissance. Nous excluons les femmes pour lesquelles la

citoyenneté à la naissance d'au moins un des parents est inconnue.

A cette étape il faut commenter pourquoi nous avons choisi de travailler sur les données d'année 2012. La tranche d'âge est fixée de telle façon pour ne pas faire face à des problèmes du traitement des retraitées et préretraités, l'âge de retrait étant fixé à 60 ans et 8 mois pour la génération de l'année 1952 avant la mise en place des réformes qui ont modifié ce chiffre. La structure d'EEC 2012 ne permettant pas d'identifier les retraités parmi les autres inactives d'une façon précise ne nous laisse pas d'autres solutions simples répondant aux besoins de notre étude. Ainsi nous faisons appel à la loi du 9 novembre 2010 reportant l'âge de la retraite de 60 ans à 62 ans pour les femmes vers 2018, cela nous permet de limiter notre échantillon d'une manière suffisamment précise et conforme aux objectifs de notre étude. Ce faisant, nous risquons d'inclure dans notre échantillon les préretraités, ainsi qu'exclure les femmes qui continuent à travailler après 60. Toutefois, nous pouvons supposer que leurs parties dans l'échantillon sont non-significatives et vont être expliqués par le terme d'erreur.

Comme nous modélisons les salaires et envisageons à construire un modèle pénalisé notre échantillon final comprend les données sur les salariées et les inactives, à l'exception des étudiantes et des retraitées. Ce choix est contestable en raison de l'endogénéité potentielle de la décision quant à la durée des études et à la participation aux régimes de préretraite, mais il est suffisamment pertinent dans le contexte de notre étude.

Dans l'enquête, il est demandé aux individus de l'échantillon d'être renseignés sur leur situation professionnelle cela nous permet d'identifier parmi celles qui n'ont pas travaillé en 2012 les étudiantes au sens de CSP.

Nous excluons de notre analyse aussi les femmes qui ne reçoivent que des compensations non salariales (elles ne représentent qu'une très petite partie de la population). De plus, nous excluons de l'échantillon les femmes qui sont actives, ayant un emploi, mais qui n'ont pas renseigné son salaire (autrement nous risquons de sous-estimer les effets parce que la moyenne de la variable dépendante sera biaisée vers zéro). Enfin nous éliminons de notre échantillon les femmes qui n'ont pas indiqué son activé au sens du BIC.

Le code de Julia et du Python contenant la construction des critères logiques pour cette étape, ainsi que pour l'étape suivante est inclus dans l'annexe 1. Le logiciel Julia dans ce cas est utilisé pour un prétraitement des données originales et une réduction de dimensions. Après, le logiciel Python nous sert à identifier les individus nous intéressant par des critères logiques décrits dans cette partie de travail ainsi que dans la partie suivante.

2.2 La Participation aux Marché du Travail : la Véritable Inactivité et le Chômage

L'enquête EEC de 2012 fournit des informations précises sur la situation l'année d'entretien, aussi bien que pour quelques dernières années de la vie professionnelle des individus enquêtés. Afin de distinguer les femmes, qui sont en recherche d'emplois (les chômeurs), de celles qui sont vraiment inactives et sont exclues du marché du travail, nous allons nous tenir toujours à la méthode décrite dans le travail effectué par Aeberhardt et al. (2007). Cela est fait purement pour pouvoir observer les statistiques descriptives et mieux comprendre la situation sur le marché de travail. Nous n'allons pas utiliser cette distinction entre les véritables inactives et les chômeurs lors de la modélisation.

Parmi les difficultés identifiées par les auteurs nous avons le fait qu'il est difficile de trouver, parmi les femmes qui n'ont pas travaillé l'année d'enquête en se déclarant comme les chômeurs, celles qui étaient effectivement au chômage.

Premièrement, nous distinguons les femmes qui ont travaillé en 2012 de celles qui n'ont pas travaillé. Parmi celles qui ne l'ont pas fait, nous vérifions si elles ont déjà travaillé auparavant. Parmi celles qui n'ont jamais travaillé, nous ne retenons que les chômeurs qui n'étaient pas étudiants en 2012. Parmi celles qui avaient un emploi dans le passé, certaines

l'ont quitté il y a moins de cinq ans et d'autres il y a plus de cinq ans. Pour ces dernières, nous n'avons que très peu d'informations et nous considérons comme chômeurs ceux qui étaient au chômage au moment de l'interview fait dans le cadre de l'enquête. Pour celles dont le dernier emploi a été occupé au cours des cinq dernières années, nous avons plus d'informations, y compris leur situation actuelle et la raison pour laquelle ils ont quitté leur dernier emploi. Nous considérons comme chômeurs celles qui étaient au chômage lorsqu'elles ont quitté leur dernier emploi et étaient toujours au chômage au moment de l'entretien. Quelques femmes se déclarent au chômage juste après avoir quitté leur dernier emploi mais sont en dehors du marché du travail (retraitées, rentrées aux études ou à l'université ou sont inactives) au moment de l'entretien. De plus, parmi celles qui se déclarent au chômage, certaines ont quitté leur emploi pour des raisons de santé ou de famille, c'est-à-dire pour une autre raison que la mise à pied, le licenciement ou la résiliation du contrat de travail temporaire. Dans ce cas, nous ne savons pas si ces personnes ont participé au marché du travail en 2011 et nous les excluons du groupe des chômeurs.

Aeberhardt et al. (2007) indiquent aussi un problème possible de sous-estimation du nombre de chômeurs en en plaçant certaines dans le groupe des inactives.

2.3 Les Statistiques Descriptives

Avant de procéder avec les estimations nous définissons les abréviations que nous utiliserons ultérieurement dans notre travail. Les variables correspondent aux traits et caractéristiques des individus. Il faut noter, que nous utilisons des variables composées par des ensembles des caractéristiques aussi dans les étapes initiales de notre travail (avant de passer aux modèles pénalisés ou nous allons redéfinir les variables explicatives).

D'abord nous allons observer des statistiques descriptives pour l'échantillon complet (le salaire pour les femmes exclues du marché de travail est considéré comme zéro). Dans le Tableau 1 les caractéristiques principales sont présentées, telles que les moyennes et la variance ainsi que les descriptives des variables qui suivent la notation utilisée par Aeberhardt et al. (2007) dans leurs études. L'échantillon comprend 12285 observations :

Tableau 1. Les abréviations utilisées dans ce travail ainsi que les statistiques descriptives pour l'échantillon complet.

Abréviation	Caractéristique	Moyenne	Variance
WAGE	Salaire perçue par heure	7.535114	215.504663
EXPER	Expérience de travail en année	22.230606	165.095872
AG50	< 20 ans	0.013431	0.013252
AG51	20-29 ans	0.169719	0.140926
AG52	30-39 ans	0.226781	0.175365
AG53	40-49 ans	0.264469	0.194541
AG54	50-60 ans	0.315507	0.215980
DDIPL1	Diplôme supérieur à baccalauréat + 2 ans	0.129508	0.112745
DDIPL3	Baccalauréat + 2 ans	0.162312	0.135978
DDIPL4	Baccalauréat ou brevet professionnel	0.197314	0.158394
DDIPL5	CAP, BEP ou autre	0.247538	0.186278
DDIPL6	Brevet des collèges	0.085877	0.078509
DDIPL7	Aucun diplôme	0.177452	0.145975
HOUS1	Femme célibataire sans enfants	0.222304	0.172899
HOUS2	Femme célibataire avec enfants	0.308751	0.213441
HOUS3	Femme avec un conjoint qui travaille, avec enfants	0.165649	0.138221

HOUS4	Femme avec un conjoint qui travaille, sans enfants	0.137403	0.118533
HOUS5	Femme avec un conjoint sans travail, avec enfants	0.071551	0.066437
HOUS6	Femme avec un conjoint sans travail, sans enfants	0.094343	0.085449
RESID1	Pas en ZUS, pas en région parisienne	0.013350	0.013172
RESID2	Pas en ZUS, en région parisienne	0.001709	0.001707
RESID3	En ZUS, pas en région parisienne	0.875783	0.108796
RESID4	En ZUS, en région parisienne	0.109158	0.097250

Il est facile à voir que comme les variables composées correspondent à des parties de la population inégales il peut être intéressant de les décomposer en plusieurs variables. Dans le cas que nous étudions actuellement les femmes résidant hors ZUS sont sous-représentées aussi bien que les femmes qui cohabitent avec un conjoint sans travail et les mineures de 20 ans.

Maintenant pour supporter notre supposition que les femmes actives et inactives ont des différentes caractéristiques qui définissent leur choix d'être présente sur le marché du travail nous passons à la comparaison des femmes actives sur le marché de travail contre les femmes exclues du marché, dont on ne connaît pas le salaire. Les résultats des tests sont donnés dans le Tableau 2 ci-dessous[8] :

Tableau 2. La comparaison des moyennes pour les femmes actives et inactives.

Abréviation	Femmes Actives	Femmes Inactives	Tests	P-valeurs
WAGE	11.756272	0	nan	nan
EXPER	21.959741	22.714124	-2.98	0.00***
AG50	0.002413	0.033099	-11.16	0.00***
AG51	0.154940	0.196101	-5.69	0.00***
AG52	0.243205	0.197461	5.94	0.00***
AG53	0.304293	0.193380	14.06	0.00***
AG54	0.290577	0.360009	-7.84	0.00***
DDIPL1	0.153035	0.087509	11.14	0.00***
DDIPL3	0.194437	0.104965	13.94	0.00***
DDIPL4	0.213233	0.168896	6.08	0.00***
DDIPL5	0.244221	0.253457	-1.13	0.26
DDIPL6	0.074422	0.106325	-5.80	0.00***
DDIPL7	0.120650	0.278848	-20.58	0.00***
HOUS1	0.205232	0.252777	-5.96	0.00***
HOUS2	0.309754	0.306960	0.32	0.75
HOUS3	0.258319	0.000227	52.26	0.00***
HOUS4	0.214249	0.000227	46.23	0.00***
HOUS5	0.005969	0.188619	-30.67	0.00***
HOUS6	0.006477	0.251190	-37.12	0.00***
RESID1	0.007366	0.024031	-6.67	0.00***
RESID2	0.001778	0.001587	0.25	0.80
RESID3	0.873381	0.880073	-1.09	0.28
RESID4	0.117475	0.094310	4.06	0.00***

[8] Les statistiques descriptives complètes sont présentées dans l'annexe 2.

Il y a plus des jeunes femmes âgées ainsi que des jeunes femmes parmi celles qui ne travaillent pas. On observe qu'il y a plus des femmes avec un niveau d'études bas (n'ayant aucun diplôme ou brevet des collèges). Selon les résultats des tests d'égalité des moyennes nous ne pouvons rejeter l'hypothèse d'égalité des dernières que pour quatre variables décrivant la composition des ménages et les lieux de résidence. Cela nous permet de supposer que ces variables sont peu déterminatives quand il s'agit de la décision d'une femme d'entrer sur le marché de travail.

Il faut noter, que suivant les tests nous ne pouvons pas rejeter l'hypothèse qu'en total les sous-échantillons sont identiques et qu'il n'y a pas des différences significatives entre les femmes actives et inactives. Quand même les résultats des tests ne sont pas une preuve suffisante pour tenir à cette hypothèse. Dans des sections suivantes nous allons observer le rôle joué par le ratio de Mills inverse représentant le terme de correction, s'il sera important dans le modèle construit améliorant les estimations nous pourrions constater qu'il existe des différences suffisantes pour affecter notre modèle.

Nous nous intéressons aussi dans la composition du sous-échantillon des femmes inactives au sens large. Comme nous avons décrit dans la section précédente c'est possible de diviser les femmes considérées comme inactives par notre modèle en deux parties. C'est-à-dire, nous sommes capables de séparer les femmes strictement inactives, qui ne se manifestent pas sur le marché de travail, des femmes étant dans la situation de chômage. Les résultats sont regroupés dans le Tableau 3 plus bas[9] :

Tableau 3. Comparaison des femmes strictement inactives contre les chômeurs pour le sous-échantillon des femmes inactives.

Abréviation	Femmes Strictement Inactives	Femmes en Chômage	Tests	P-valeurs
EXPER	24.574112	16.402390	16.84	0.00***
AG50	0.022601	0.068725	-5.50	0.00***
AG51	0.169064	0.287849	-7.58	0.00***
AG52	0.189316	0.225100	-2.42	0.02**
AG53	0.193425	0.193227	0.01	0.99
AG54	0.410919	0.187251	14.98	0.00***
DDIPL1	0.084825	0.096614	-1.12	0.26
DDIPL3	0.104491	0.106574	-0.19	0.85
DDIPL4	0.161139	0.195219	-2.43	0.02**
DDIPL5	0.249486	0.266932	-1.10	0.27
DDIPL6	0.105078	0.110558	-0.49	0.62
DDIPL7	0.294981	0.224104	4.63	0.00***
HOUS1	0.232463	0.321713	-5.43	0.00***
HOUS2	0.275022	0.415339	-8.09	0.00***
HOUS3	0.000294	0.000000	1.00	0.32
HOUS4	0.000294	0.000000	1.00	0.32
HOUS5	0.204285	0.135458	5.37	0.00***
HOUS6	0.287643	0.127490	12.24	0.00***
RESID1	0.026123	0.016932	1.87	0.06*
RESID2	0.001761	0.000996	0.62	0.53
RESID3	0.875550	0.895418	-1.77	0.08*
RESID4	0.096566	0.086653	0.97	0.33

[9] Les statistiques descriptives complètes sont présentées dans l'annexe 2.

Il est assez évident que les sous-échantillons sont assez différents, parce que la partie des femmes complètement inactives est représentée par des femmes âgées, tandis que les femmes en chômage sont en plus jeunes.

Enfin il faut préciser que nous allons utiliser le logarithme de salaire horaire comme la variable dépendante et pas le salaire en euros tout simplement parce que la distribution de salaire horaire ne suit pas la loi normale et notre estimateur sera biaisé autrement.

3. La Modélisation

Maintenant nous procédons avec la modélisation et la comparaison des résultats obtenus.

Notre individu de référence sera représenté par une femme français travaillant à temps complet, ne résidant pas ni en ZUS, ni en région Parisien. Nous supposons aussi que cette femme n'a aucun diplôma et qu'elle réside seule sans enfants.

3.1 Le Modèle Simple

On commence par la modélisation en utilisant les Moindres Carrées Ordinaires (MCO) de l'échantillon restreinte. Ici nous utilisons seulement les données sur les femmes classées comme actives pour lesquelles le salaire horaire est observable. Les résultats obtenus par les MCO et présents sur le Tableau 4, seront nos résultats de référence pour le reste du mémoire :

Tableau 4. Les résultats d'estimation de l'échantillon réduit par les MCO.

Variable	Coefficient	Erreur standard	P-valeur
Constante	1.7132	0.044	0.000***
RESID1	Reference		
RESID2	0.0231	0.068	0.735
RESID3	0.0229	0.041	0.576
RESID4	0.1934	0.043	0.000***
DDIPL1	0.6596	0.017	0.000***
DDIPL3	0.4566	0.014	0.000***
DDIPL4	0.2806	0.014	0.000***
DDIPL5	0.1361	0.013	0.000***
DDIPL6	0.1342	0.016	0.000***
DDIPL7	Reference		
EXPER	0.0215	0.001	0.000***
EXPERSQ	-0.0003	2.8e-05	0.000***

Bien que le modèle est généralement correct (F-statistique est de 252.3), les résultats obtenus laissent à désirer parce que le modèle n'explique que une petite fraction de la variance du salaire (29.3% suivant la valeur de R-*squared*). Cet effet est normal en sciences sociales et l'économie en particulier pour des raisons d'impossibilité d'incorporer toutes les variables explicatives dans le modèle à cause de complexité des relations qu'on tente à étudier. Quand même il existe de moyens pour essayer à dépasser cette limitation, par exemple, c'est possible d'utiliser telles techniques de sélection algorithmique des régresseurs comme les modèles pénalisé, ce qu'on va essayer de faire dans la sous-section 4.3.

Parmi des autres problèmes de cette modèle nous pouvons identifier le fait qu'avec les résultats du test de Jacques-Bera nous ne pouvons pas retenir l'hypothèse de la normalité des résidus, ce qui pourra nous créer des complications pour des étapes suivants. Ce problème peut avoir lieu à cause d'une misspecification du modèle, bien que en sciences-sociales un tel

comportement des résidus n'est pas rare, comme c'est presque impossible de construire un modèle complet et nous ne pouvons que observer quelques effets des certaines variables.

Sachant que les résultats obtenus par le modèle de sélection seront assez proches des résultats obtenus lors de cet étape, nous pouvons déjà observer les trends principaux déterminant les relations entre le salaire et les caractéristiques des femmes.

Par exemple, nous observons que les variables décrivant le type de la résidence des individus en question sont peu significatives et qu'il existe seulement la distinction entre les femmes habitant en ZUS à Paris et toutes les autres femmes de notre échantillon. Ce fait peut être expliqué par ce que notre échantillon n'est pas balancé et que presque 90% des individus dans notre échantillon habitent en ZUS hors la région parisienne.

Le niveau des études à un effet strictement positive sur le montant de salaire perçu, ce que s'explique par une qualification plus haute des individus ayant un diplôme. Par un mécanisme identique l'expérience des individus sur le marché de travail a aussi un effet positif sur le salaire, bien que dans ce cas la relation soit non-linéaire. L'effet marginal de ce facteur décroit avec l'âge. C'est suivant cette hypothèse que nous n'avons pas directement intégré l'âge des femmes en question dans notre modèle à cet étape afin d'éviter les problèmes de la multicollinearité. Probablement il sera sensible d'intégrer d'autres variables lié à l'âge dans notre modèle quand nous allons traiter le problème en étudiant de relations non-linéaires sous-jacentes par des moyens d'utilisation des modèles pénalisés.

Nous estimons aussi le modèle pour l'échantillon complet en supposant que salaire est zéro pour les femmes exclues du marché. Les résultats sont présentés dans l'annexe 3 parce que nous savons par défaut que ces résultats sont biaisés suivant le raisonnement de Heckman (1979) et il n'y a aucun intérêt de les présenter à ce stade du travail.

3.2 Le Modèle de Sélection Endogène

Maintenant, quand nous avons vérifié les traits principaux ayant un effet sur le salaire perçu, nous allons réestimer notre modèle en utilisant des outils de sélection généralisée afin de corriger les résultats. Cela nous permettra d'éviter le biais apporté par le fait que nous utilisons seulement une partie de notre échantillon y excluant les personnes en recherche d'emploi. Comme c'était déjà précisé, nous allons estimer notre modèle de sélection endogène à deux étapes, pour pouvoir après comparer les résultats avec ceux obtenu pour le modèle avec pénalisation sur deux étapes.

3.2.1 Etape Probit (Etape de Sélection)

Nous estimons maintenant la première partie du modèle de sélection endogène suivant la procédure à deux étapes proposée par Heckman (1979). C'est un modèle binaire de type Probit lequel décrit la probabilité pour une femme d'avoir un emploi. Les variables sociodémographiques (vivre en couple, avoir des enfants, que le conjoint travaille ou non) se sont prouvées d'être des régresseurs appropriés, car leur impact sur le fait d'avoir un emploi doit être théoriquement significatif.

Tableau 5. Les résultats d'estimation du modèle Probit décrivant l'activité.

Variable	Coefficient	Erreur standard	P-valeur
Constante	-1.4921	0.131	0.000***
HOUS1	Reference		
HOUS2	0.2443	0.037	0.000***
HOUS3	2.9804	0.322	0.000***
HOUS4	3.2647	0.290	0.000***

HOUS5	-2.1027	0.077	0.000***
HOUS6	-1.9003	0.079	0.000***
RESID1	Reference		
RESID2	0.5821	0.408	0.154
RESID3	0.4839	0.120	0.000***
RESID4	0.4853	0.128	0.000***
DDIPL1	1.0372	0.062	0.000***
DDIPL3	1.0343	0.058	0.000***
DDIPL4	0.7448	0.052	0.000***
DDIPL5	0.5042	0.049	0.000***
DDIPL6	0.3276	0.063	0.000***
DDIPL7	Reference		
EXPER	0.0776	0.005	0.000***
EXPERSQ	-0.0016	0.000	0.000***

Par la nature du modèle Probit nous préférons à ne pas commenter les coefficients obtenus, mais passer directement aux effets marginaux moyens du modèle pour avoir plus de précision. Passons maintenant à l'étude des effets marginaux obtenus pour l'étape de sélection :

Tableau 6. Les effets marginaux pour le modèle Probit.

Variable	dy/dx	Erreur standard	P-valeur
HOUS1	Reference		
HOUS2	0.0489	0.007	0.000***
HOUS3	0.5966	0.064	0.000***
HOUS4	0.6535	0.057	0.000***
HOUS5	-0.4209	0.014	0.000***
HOUS6	-0.3804	0.014	0.000***
RESID1	Reference		
RESID2	0.1165	0.082	0.154
RESID3	0.0969	0.024	0.000***
RESID4	0.0972	0.026	0.000***
DDIPL1	0.2076	0.012	0.000***
DDIPL3	0.2070	0.011	0.000***
DDIPL4	0.1491	0.010	0.000***
DDIPL5	0.1009	0.010	0.000***
DDIPL6	0.0656	0.013	0.000***
DDIPL7	Reference		
EXPER	0.0155	0.001	0.000***
EXPERSQ	-0.0003	2.29e-05	0.000***

Nous remarquons que le fait d'avoir un conjoint sans emploi avec enfant (HOUS5) ou sans enfant (HOUS6) a un impact négatif sur la probabilité d'être en emploi. Le reste des variables ont un impact positif sur la probabilité d'avoir un emploi pour les femmes.

D'après les résultats que nous obtenons, il se pourrait qu'il y ait un effet incitatif du fait que le conjoint soit en emploi. Effectivement, nous remarquons que les variables ou les conjoint(e)s sont actifs ont un impact positif sur la probabilité d'être en emploi.

De plus, le fait d'avoir un enfant réduit la probabilité d'être en emploi pour les femmes ayant un conjoint et l'augmentent pour les femmes célibataires (bien que les coefficients sont

peu représentatifs et nous ne pouvons affirmer que cette supposition est absolument correcte).

Par ailleurs, le fait d'habiter en région parisienne mais pas dans une ZUS, impacte positivement les chances d'être ébauchée. Il faut aussi indiquer un résultat inacceptable même au seuil de 15% pour la variable RESID2, ce que peut s'expliquer par une représentation insuffisante de cette variable dans notre échantillon.

Ensuite, le fait d'avoir un diplôma exerce un impact positif sur la probabilité d'emploi. L'influence augmente avec le niveau de diplôme. D'ailleurs, plus le diplôme acquit a nécessité de longues études, plus est la probabilité d'emploi des personnes qui sont diplômées avec des études longues.

L'expérience se comporte exactement comme dans le modèle précédent exerçant un impact positif décroissant sur la probabilité d'être active pour une femme, ce que peut être expliqué par sa corrélation avec l'âge.

3.2.2 Etape MCO

Maintenant nous pouvons incorporer les résultats obtenus dans le modèle visant à expliquer les variations de salaire. Nous utilisons le ratio de Mills inversé (IMR ou imr) comme un des régresseurs, ce terme va représenter le terme de correction. Les résultats sont regroupés dans le Tableau 7 :

Tableau 7. Les résultats de deuxième étape du modèle de sélection endogène.

Variable	Coefficient	Erreur standard	P-valeur
Constante	1.7059	0.046	0.000***
RESID1	Reference		
RESID2	0.0251	0.068	0.713
RESID3	0.0250	0.041	0.544
RESID4	0.1953	0.043	0.000***
DDIPL1	0.6617	0.018	0.000***
DDIPL3	0.4586	0.015	0.000***
DDIPL4	0.2820	0.014	0.000***
DDIPL5	0.1372	0.013	0.000***
DDIPL6	0.1347	0.016	0.000***
DDIPL7	Reference		
EXPER	0.0217	0.001	0.000***
EXPERSQ	-0.0003	2.86e-05	0.000***
Ratio de Mills Inversé	0.0054	0.010	0.591

Nous remarquons que, comme nous l'avons déjà supposé, les résultats obtenus sont assez proches à ceux, obtenus lors d'estimation d'un échantillon incomplet par les MCO sans étape de sélection. Effectivement, tous les commentaires faits pour le modèle précèdent s'applique dans ce cas aussi. Ce qui nous intéresse maintenant ce sont les effets marginaux et les différences entre les coefficients des deux modèles.

3.2.3 Les Effets Marginaux et la Comparaison des Résultats entre les Modèles

Le tableau suivant (Tableau 8) nous permet de comparer les effets marginaux moyens des modèles construits auparavant :

Tableau 8. Comparaison des différents modèles.

Variable	MCO biaisé	MCO	Sélection endogène

R2	0.136	0.293	0.293
F-Statistique	199.8	252.3	229.6
Jarque-Bera	979.148	99166.533	99171.120
Echantillon	12285	7874	12285
	Coefficients		
Constante	-0.1625	1.7132	1.7059
DDIPL1	1.1182	0.6596	0.6616
DDIPL3	0.9633	0.4566	0.4585
DDIPL4	0.6784	0.2806	0.2820
DDIPL5	0.4357	0.1361	0.1371
DDIPL6	0.3068	0.1342	0.1347
DDIPL7	Reference		
EXPER	0.0684	0.0215	0.0212
EXPERSQ	-0.0013	-0.0003	0.0000
HOUS1	Reference		
HOUS2	nan	nan	-0.0000
HOUS3	nan	nan	-0.0002
HOUS4	nan	nan	-0.0002
HOUS5	nan	nan	0.0000
HOUS6	nan	nan	0.0000
RESID1	Reference		
RESID2	0.5591	0.0231	0.0251
RESID3	0.4566	0.0229	0.0249
RESID4	0.5886	0.1934	0.1953
Ratio de Mills Inversé	nan	nan	0.0054

Nous pouvons voir que le niveau d'étude (le type de diplôma en possession d'individu) reste relativement intacte par la correction. Nous pouvons dire que pour cette variable l'estimation par MCO simple sur un échantillon réduit a donné des résultats acceptables.

Cette observation s'applique aussi à telle variables comme le type de la résidence ou le fait d'avoir un travail à temps partiel.

Le terme affecté le plus par la correction (relativement aux changements de la valeur) c'est l'expérience (EXPER). Il est évident que le modèle simple a largement surestimé l'impact de l'expérience sur le salaire.

De plus, nous pouvons observer que les résultats obtenus par les MCO sur un échantillon réduit s'approche dans sa précision aux résultats obtenus par un modèle de sélection et leurs coefficient se diffère assez peu. Toutefois les deux modèles surpassent dans sa précision les estimations fait par les MCO sur l'échantillon complet (si on considère le salaire de femmes exclues du marché d'être zéro).

3.3 Les Modèles Pénalisés et la Sélection Endogène

Maintenant nous pouvons passer à l'application des techniques de pénalisation sur un modèle à haute dimension. La justification d'utilisation de ces méthodes est donnée dans la sous-section 4.3.1 et en sous-section 4.3.2 le modèle est estimé.

3.3.1 Les Variables et Non-linéarité

Comme nous avons vu précédemment les résultats obtenus avec l'ensemble des

variables proposés par les chercheurs dans des études antérieurs ne sont pas suffisamment représentatives (Aeberhardt et al. 2007, 2010). Bien que toutes ces variables aient un impact sur le salaire, nous avons des raisons pour douter la façon dont les variables agrégés sont composé. De plus, nous pouvons supposer que le salaire est déterminé par ces variables d'une façon non-linéaire, bien que nous ne savons pas exactement, quelle fonction décrit la relation étudiée. Afin d'inférer cette fonction, nous nous adressons à des méthodes de sélection du modèle par pénalisation (par le terme de pénalisation L1 – LASSO) en nous mettant dans un cas d'étude des régressions en grande dimension. Pour identifier la bonne fonction décrivant la relation étudiée, nous allons commencer par la création d'un ensemble des variables décrivant toutes les relations non-linéaires possibles.

Sur la Figure 1 la matrice de corrélation générée dans le Python est présentée. Le glossaire complet des variables se trouve dans l'annexe 4.

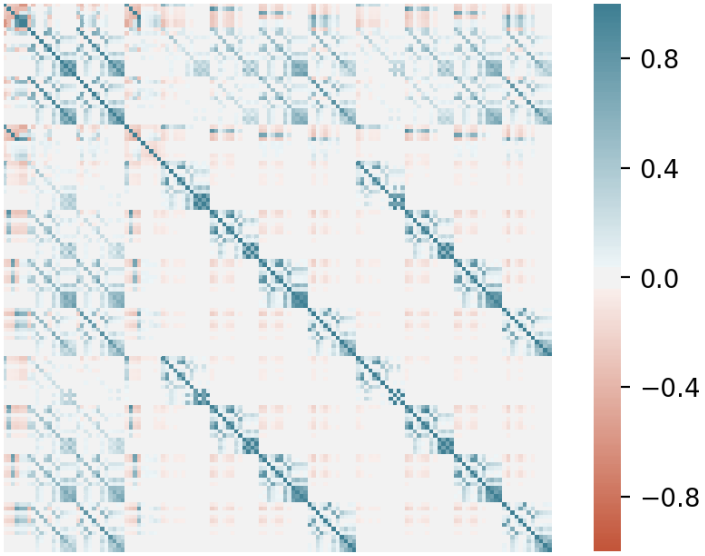

Figure 1. La matrice de corrélation pour des variables des modèles de grande dimension.

Comme c'est possible à voir à partir de cette matrice de corrélation il y a plusieurs variables interconnectées, ce qui s'explique par le fait qu'ils ont été créés comme des indices conditionnels regroupant plusieurs caractéristiques.

Dans tout un autre modèle il serait impossible d'utiliser cet ensemble de redresseurs, mais les techniques de pénalisation nous permettent de faire estimation directement sur ce liste des variables et d'y identifier celles qui sont les plus pertinentes étant donnée l'échantillon. L'estimateur qu'on utilisera tend d'égaliser les coefficients des variables peu importantes à 0. C'est à dire, avec un terme de pénalisation suffisamment importante nous allons tomber sur un modèle sans multicollinearité présente et sans des problèmes d'identification pour les réestimer après avec des techniques de sélection endogène utilisés en haut.

3.3.2 L'Application

Dans cette partie une extension d'étude classique est présentée. Nous allons choisir les variables explicatives automatiquement par un algorithme faisant appel à de régressions pénalisées.

Pour obtenir des choix des variables suffisamment pertinents nous allons implémenter des techniques de validation croisée. Le jeu des données est divisé en 5 *folds* (les échantillons

d'estimation de 80%, et les sous-échantillons de 20% pour les tests).

Etant limitée dans la puissance computationnelle nous allons tester un nombre des λ limité (dans ce travail nous utiliserons seulement 10 λ, distribués uniformement sur l'intervalle de $1e^{-4}$ à $1e^{+4}$ ce qui est un intervalle traditionnelle en fonctions du Python). Pour le critère de sélection de bon niveau de la pénalisation nous posons que le rapport d'erreurs moyennes obtenues pour l'ensemble des *folds* ne doit pas être inférieur à 3. Cette critère est testé avec un algorithme itérative (le code est présente dans les annexes).

3.3.2.1 LASSO Logistique

Comme il n'existe pas d'implémentation de Probit pénalisé dans le logiciel Python nous allons utiliser un Logit pénalisé à sa place. Cela est fait sur l'hypothèse que les différences entre ces deux modèles résident dans la forme de la fonction de répartition des résidus, qui sont suffisamment similaires : les deux sont centrées au zéro, ce qui nous va donner des choix des redresseurs pratiquement identique. Parce que nous n'allons pas faire l'analyser les estimateurs obtenus directement, mais utiliser ce modèles purement pour le choix des variables explicatives nous pouvons tolérer cette différence. Il faut quand même noter qu'il soit intéressant d'effectuer une analyse identique sur Probit-pénalisé pour confirmer l'hypothèse du choix des variables identiques.

Pour commencer nous présentons l'évolution de terme d'erreur moyen pour différentes λ sur la Figure 2, le code utilisé pour générer cette figure se trouve dans l'annexe 5 :

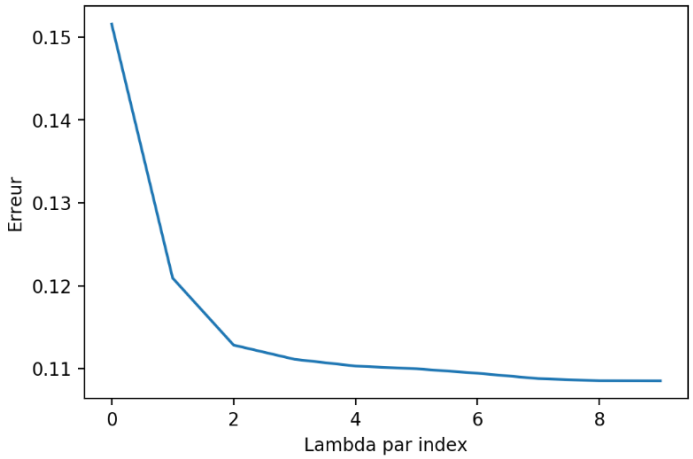

Figure 2. L'évolution d'erreur moyenne en dépendance de lambda pour le Logit pénalisé.

Notre algorithme choisi $\lambda = 0.005994$ (index = 2) pour éviter le sur-apprentissage. Cette valeur va résulter dans le choix d'un plus petit ensemble des variables pour la réestimation, ceux qui permettra de commenter les résultats.
Sur l'équation de sélection nous identifions l'ensemble des variables suivantes :

Tableau 9. Les variables et leurs coefficients captés par Logit pénalisé.

Variable	Coefficient
EXPER2	0.009648
EXPER1COUPLE	-0.111049
EXPER1NOZUS	-0.014255
EXPER1PARIS	0.001050
EXPER1COUPLECJACT	0.473121
EXPER1COUPLEENF	-0.037597

EXPER1AG51	0.020001
EXPER1AG52	0.003651
EXPER1AG54	-0.018200
EXPER1DDIPL1	0.059307
EXPER1DDIPL3	0.051706
EXPER1DDIPL4	0.038234
EXPER1DDIPL5	0.023835
EXPER1DDIPL6	0.016838
EXPER1AG51ENF	-0.070973
EXPER1AG52ENF	-0.013040
EXPER1AG52COUPLE	-0.000505
EXPER1AG52COUPLEENF	-0.035464
EXPER1AG53ENF	0.002607
EXPER1AG54ENF	0.003932

Nous pouvons constater, que les recruteurs perçoivent les caractéristiques personnelles d'un individu à travers un prisme d'expérience, de manière que chaque caractéristique est vue comme un avantage ou une inconvenance en dépendance des années d'expérience.

Nous n'allons pas étudier les coefficients obtenus pour cette modèle parce qu'ils sont biaisés vers zéro et il nous reste encore de les réestimer par un modèle traditionnel simple. De plus nous ne présentons pas les données la variance des coefficients obtenus ce qu'est impossible à faire sans utilisation des techniques avancées telles que Bootstrap.

Enfin nous illustrons que la matrice de covariance obtenu pour ces variables justifie l'utilisation de cet ensemble dans la modélisation, bien qu'on risque toujours de faire face au problème de la présence de multicollinearité :

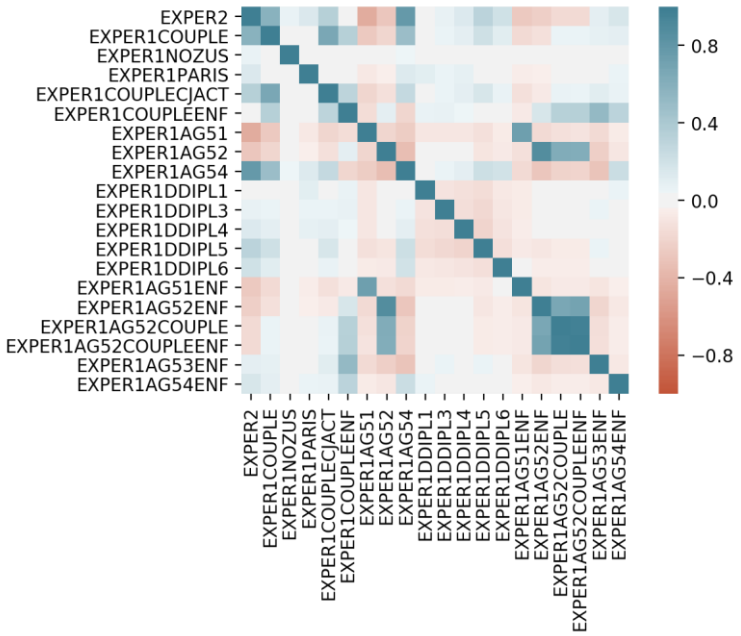

Figure 3. La matrice de corrélation des variables choisies.

Nous supposons quand même que en étudiant un nombre des λ plus grand il est possible d'obtenir des résultats encore plus précis, avec un moindre nombre des régresseurs.

3.3.2.2 LASSO MCO

Suivant la logique de présentation du Logit pénalisé nous passons à l'étude de l'équation principale par des méthodes de pénalisation. Pour commencer nous présentons l'évolution de terme d'erreur moyen pour différentes λ (l'algorithme utilisé est aussi présenté dans l'annexe 5 de ce travail) :

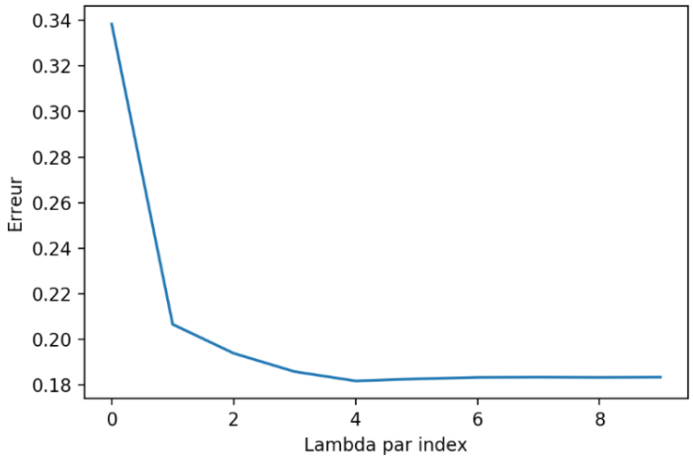

Figure 4. L'évolution de l'erreur moyenne en dépendance de lambda pour une régression pénalisée.

L'algorithme choisi $\lambda = 0.044293$ (index = 3) pour éviter le sur-apprentissage. Dans le cas de sélection des variables pour le modèle de salaire nous ne rencontrons pas des problèmes techniques. Sur l'équation décrivant la formation du salaire nous identifions un set des variables suivantes :

Tableau 10. Les variables et leurs coefficients captés par le modèle pénalisé.

Variable	Coefficient
EXPER1ENF	-0.000176
EXPER1PARIS	0.006042
EXPER1COUPLEENF	0.000343
EXPER1AG54	0.000306
EXPER1DDIPL1	0.025627
EXPER1DDIPL3	0.014069
EXPER1DDIPL4	0.006746
EXPER1DDIPL5	0.001279
EXPER1DDIPL6	0.001570
EXPER1AG54COUPLE	-0.000358

Maintenant nous pouvons observer que pour le salaire aussi le facteur principal le déterminant est l'expérience qui joue le rôle d'un multiplicateur des compétences acquises.

Enfin nous illustrons que la matrice de covariance obtenus pour ces variables justifie l'utilisation de cet ensemble dans la modélisation :

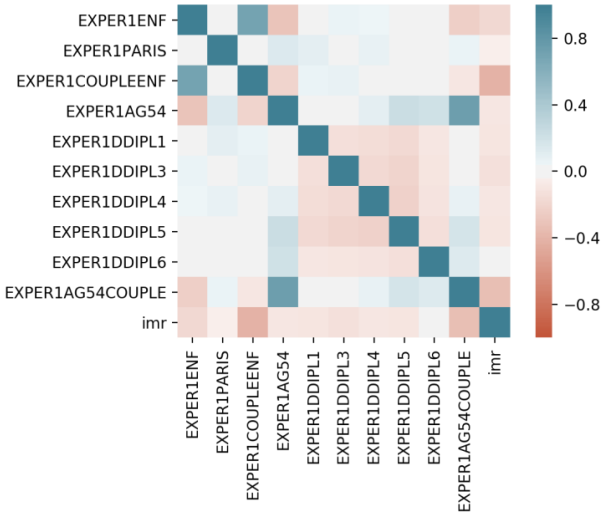

Figure 5. La matrice de corrélation des variables choisies.

3.3.2.3 Etape de Sélection Post-LASSO

Maintenant, ayant obtenus nos liste des régresseurs pour les deux étapes du modèle de sélection endogène, nous pouvons procéder est estimer les relations qui nous intéresse par des méthodes classiques. Cela nous permettra d'obtenir des coefficients moins biaisée que ceux obtenus pour les modèles de LASSO ainsi que effectuer la correction pour la présence de sélection endogène. On commence par le modèle Probit dans lequel nous introduisons l'ensemble des variables retenu par le LASSO Logistique.

Tableau 11. Les résultats de réestimation de l'équation de sélection pour les variables choisies.

Variable	Coefficient	Erreur standard	P-valeur
Constante	0.0342	0.040	0.395
EXPER2	0.0029	0.002	0.235
EXPER1COUPLE	-0.0607	0.002	0.000***
EXPER1NOZUS	-0.0135	0.005	0.004***
EXPER1PARIS	0.0011	0.002	0.613
EXPER1COUPLECJACT	0.5983	0.018	0.000***
EXPER1COUPLEENF	-0.0245	0.005	0.000***
EXPER1AG51	0.0376	0.009	0.000***
EXPER1AG52	0.0135	0.005	0.007***
EXPER1AG54	-0.0102	0.002	0.000***
EXPER1DDIPL1	0.0346	0.003	0.000***
EXPER1DDIPL3	0.0306	0.002	0.000***
EXPER1DDIPL4	0.0243	0.002	0.000***
EXPER1DDIPL5	0.0157	0.002	0.000***
EXPER1DDIPL6	0.0126	0.002	0.000***
EXPER1AG51ENF	-0.0897	0.010	0.000***
EXPER1AG52ENF	-0.0185	0.005	0.000***
EXPER1AG52COUPLE	-0.0215	0.023	0.352
EXPER1AG52COUPLEENF	-0.0396	0.025	0.120
EXPER1AG53ENF	0.0037	0.003	0.160
EXPER1AG54ENF	0.0075	0.003	0.018**

On obtient les effets marginaux moyens pour ce modèle, faut noter, que ce sont des effets marginaux moyens des variables utilisé directement dans le modèle et nous ne pouvons pas observer les effets marginaux des composantes sans des manipulations élaborés sur les données :

Tableau 12. Les effets marginaux pour le modèle de sélection pour les variables choisies.

Variable	dy/dx	Erreur standard	P-valeur
EXPER2	0.0006	0.001	0.395
EXPER1COUPLE	-0.0128	0.000	0.235
EXPER1NOZUS	-0.0028	0.001	0.000***
EXPER1PARIS	0.0002	0.000	0.004***
EXPER1COUPLECJACT	0.1260	0.004	0.613
EXPER1COUPLEENF	-0.0052	0.001	0.000***
EXPER1AG51	0.0079	0.002	0.000***
EXPER1AG52	0.0028	0.001	0.000***
EXPER1AG54	-0.0021	0.000	0.007***
EXPER1DDIPL1	0.0073	0.001	0.000***
EXPER1DDIPL3	0.0064	0.001	0.000***
EXPER1DDIPL4	0.0051	0.000	0.000***
EXPER1DDIPL5	0.0033	0.000	0.000***
EXPER1DDIPL6	0.0027	0.000	0.000***
EXPER1AG51ENF	-0.0189	0.002	0.000***
EXPER1AG52ENF	-0.0039	0.001	0.000***
EXPER1AG52COUPLE	-0.0045	0.005	0.000***
EXPER1AG52COUPLEENF	-0.0083	0.005	0.352
EXPER1AG53ENF	0.0008	0.001	0.120
EXPER1AG54ENF	0.0016	0.001	0.160

Comme nous avons déjà indiqué, on peut supposer que les recruteurs voient les caractéristiques personnelles d'un individu à travers un prisme d'expérience, et chaque caractéristique est vue comme un avantage ou une inconvenance en fonction des années d'expérience.

Les femmes sont plus susceptible de ne se présenter sur le marché de travail quand elles ont entre 20 et 29 ans ce que s'explique par une forte probabilité qu'elles sont en recherche d'un emplois qui les conviennent ou qu'elles restent hors du marché de travail pour des raisons familiales. Le fait d'avoir un conjoint (être en couple) augmente aussi la probabilité que la femme reste hors du travail, comme elle aura un soutien financier.

C'est aussi remarquable que la valeur d'un niveau d'études élevé s'augmente avec l'expérience.

3.3.2.4 Etape MCO Post-LASSO

Par analogie avec l'étape précèdent nous allons commencer par un aperçue des variables utilisés dans cet étape. Sur la matrice de corrélation nous observons l'absence de multicollinearité facilement identifiable (cela n'implique pas, qu'il n'y a pas absolument que les régresseurs ne sont pas entreliés entre eux).

Evidemment, comme nous modélisons des relations non-linéaires les effets marginaux observés ne sont pas représentatives. La situation est identique à celle d'étape précèdent. Dans le Tableau 13 nous regroupons les coefficients du modèle réestimé :

Tableau 13. Les résultats de réestimation de l'équation de salaire pour les variables choisies.

Variable	Coefficients	Erreur standard	P-valeur
Constante	2.1992	0.010	0.000***
EXPER1ENF	-0.0017	0.000	0.000***
EXPER1PARIS	0.0064	0.001	0.000***
EXPER1COUPLEENF	0.0011	0.001	0.062
EXPER1AG54	0.0006	0.000	0.092
EXPER1DDIPL1	0.0275	0.001	0.000***
EXPER1DDIPL3	0.0155	0.001	0.000***
EXPER1DDIPL4	0.0081	0.000	0.000***
EXPER1DDIPL5	0.0024	0.000	0.000***
EXPER1DDIPL6	0.0031	0.000	0.000***
EXPER1AG54COUPLE	-0.0018	0.000	0.000***
Ratio de Mills Inversé	-0.0421	0.011	0.000***

Pour ce modèle on observe R-*squared* de 0.276 ce qui nous indique que la part de la variance expliquée est assez proche à notre modèle traditionnelle de départ. Bien que ce modèle nous offre un aperçu des véritables relations entre le salaire et les caractéristiques le déterminant, nous ne pouvons pas d'observer les effets de chaque caractéristique séparément, comme nous avons fait au début de ce travail. Ici nous pouvons voir que les caractéristiques séparées n'affectent pas directement le niveau de salaire perçu, mais leurs effets sont conditionnels sur l'expérience d'individu.

Conclusion

Dans ce mémoire nous avons effectué une étude du marché du travail des femmes en France. Nous avons commencé par suivre la méthodologie des études traditionnelles dans ce domaine telles que les travaux de Heckman (1979) et d'Aeberhardt et al. (2007, 2010) qui ont investigué le marché de travail et ses inégalités en présence de sélection endogène. La réplication des approches techniques utilisées dans ces articles nous a permis de confirmer les suppositions faites par les auteurs sur les mécanismes affectant la formation de salaire.

De plus, dans ce travail nous avons effectué une modélisation expérimentale du marché de travail des femmes françaises en introduisant plusieurs sources de non-linéarité de la formation de leurs salaires. Pour pouvoir estimer ce modèle des technique contemporaines de traitement des données de grande dimension ont été utilisées, telles que la sélection des variables par des modèles pénalisés. Les résultats obtenus apportent un nouvel aperçu sur le mécanisme qui entrelie le salaire et les caractéristiques personnelles, comme nous trouvons des justifications pour le rôle dont les non-linéarités jouent dans ce processus. Quand même les résultats obtenus risquent d'être biaisés et il reste encore à effectuer un analyse théorique profond afin de confirmer et justifié complètement la méthodologie implémentée.

Ce travail ouvre plusieurs pistes pour futur recherches, parmi lesquelles se trouvent à la fois des questions théoriques et pratique. Surtout il peut être intéressant de comparer les résultats obtenus pour les femmes avec les résultats pour les hommes sur le marché de travail en France s'approchant ainsi encore plus à la méthodologie d'Aeberhardt et al. (2007, 2010). Ou cela peut prouver intéressant d'investiguer plus le côté technique de cette étude en vérifiant la pertinence des résultats obtenus par des modèles pénalisées.

Bibliographie

1. Aeberhardt R. et al., (2007), Wages and Employement of French Workers with African Origin, *IZA DP*, No.2898 ;
2. Aeberhardt R. et al., (2010), L'emploi et les salaires des enfants d'immigrés, *Economie et Statistique*, (433-434), pp.31-46 ;
3. Belloni et al., (2014), Hight-dimentional methods and inference on structural and treatment effects, *Journal of economic perspectives*, (28)2, pp.1-23 ;
4. Belloni et Chernozhukov, (2013), Least squares after model selection in high-dimensional sparse models, *Bernoulli*, (19)2, pp.521-547 ;
5. Charpenter A. (04/06/2018), Classification from scratch, penalized LASSO Logistic 5/8, *site web* [https://freakonometrics.hypotheses.org/52894], consulté le 25/03/2019 ;
6. Chernozhukov et al., (2018), LASSO-Driven Inference in Time and Space, *CEMMAP WP*, CWP36/18 ;
7. Green W.H. [2002], *Econometric analysis*, Pearson Education, 5-ème éd., 802p. ;
8. Hastie T., Tibshirani R. et Wainwright M. [2015], *Statistical Learning with sparsity : The Lasso and Generalizations*, Chapman and Hall/CRC, 367p. ;
9. Heckman, (1979), Sample Selection Bias as a Specification Error, *Econometrica*, (47)1, pp.153-161 ;
10. Lee et al., (2016), Exact post-selection inference, with application to the lasso, *The Annals of Statistics*, (44)3, pp.907-927 ;
11. Lee et Algamal, (2015), Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification, *Computers in Biology and Medicine*, (67), pp.136-145 ;
12. Li Q. et Racine J.S. [2007], *Nonparametric Econometrics Theory and Practice*, Princeton University Press, 768p. ;
13. Pagan A. et Ullah A. [1999], *Nonparametric Econometrics*, Cambridge University Press, 444p. ;
14. Park et Casella, (2008), The Bayesian Lasso, *Journal of the American Statistical Assosiation*, (103)482, pp.681-686 ;
15. Tibshirani R. (2015), Recent Advances in Post-Selection Statistical Inference, *Stanford University Lectures* [http://statweb.stanford.edu/~tibs/ftp/nips2015.pdf] ;
16. Tibshirani, (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, (58)1, pp.267-288 ;
17. Toomet, (2008), Sample Selection Models in R: Package sampleSelection, *Journal of statistical software*, (27)7, pp.1-23 ;
18. Wippert et al., (2017), Variable Selection and Inference in a follow-up Study on Back Pain, *University of Munchen Technical Report*, No.211 ;
19. Zhao et al. (2017), In defense of indefensible: A very naïve approach to high-dimensional inference, *arXiv* : https://arxiv.org/abs/1705.05543 ;
20. L'Enquête Emploi Continue 2012, INSEE, les données transmis par le responsable pédagogique ;
21. Enquête emploi en continu, Résumé, (dernière version de 15/07/2019), INSEE, *site web* [https://www.insee.fr/fr/metadonnees/source/serie/s1223/], consulté le 07/05/2019 ;
22. Lasso, (17/06/2009), *Machinelearning.ru*, *site web* [http://www.machinelearning.ru/wiki/index.php?title=%D0%9B%D0%B0%D1%81%D1%81%D0%BE], consulté le 10/02/2019 ;

Annexes

Annexe 1 : Extraits du code utilisé pour la génération d'échantillon

1.1 Le prétraitement des données avec le logiciel Julia

Dans cette section nous regroupons les parties du code sur Julia, qui a servi pour réduire la dimensionnalité du jeu des données originale, distribué par l'INSEE :

```
# Chargement des packages
using DataFrames
using CSV
using DataFramesMeta

# Sélection des variables nous intéressantes parmi plus des 300 disponibles
data = @select(eemploi2012, :ACTEU, :SEXE, :DDIPL,
                :FORDAT, :MATRI, :TYPMEN5, :ZUS,
                :REG, :AG, :AGQ, :AGEQ, :AGE, :AG5, :NBENF3, :NBENF6,
                :NBENF18, :SALRED, :NBHEUR, :NATPERC,
                :NATMERC, :PAIPERC, :PAIMERC, :RGI, :IDENT,
                :NOI, :TRIM, :DIP, :DIP11, :CONTRA, :CSER,
                :CSPM, :CSPP, :ANNEE, :CHPUB, :NAFG4N,
                :ANCENTR4, :TPP, :ACTEU6CJ,
                :ADFDAP, :ADEBEN, :ACTEU6, :INSCAC)

# Vérification des variables manquantes et choix de l'étape d'enquête relevant
data1 = @where(data, :RGI .== 1,
                :AG .!== missing,
                :AG5 .!== missing,
                :SEXE .!== missing,
                :NBENF18 .!== missing,
                :FORDAT .!== missing,
                :DDIPL .!== missing,
                :DIP11 .!== missing,
                :DIP .!== missing,
                :CSPP .!== missing,
                :CSPM .!== missing)
```

1.2 Les critères de composition d'échantillon sous Python

Les noms des variables dans le jeu de données initial est identique à celui du glossaire de l'EEC 2012, les données ont été prétraitées dans le logiciel *Julia* afin de réduire la dimensionnalité :

```
# Identification des français
critfr = (((b.PAIPERC == 1) & (b.PAIMERC == 1)) &
            ((b.NATPERC == 1) & (b.NATMERC == 1)))

# Identification de femmes
critfem = (b.SEXE == 2)

# Identification d'actives
critwork = (b.ACTEU == 1)

# Identification de ceux, qui ont indiqué son salaire
salafich = ((b.SALRED > 0) & (b.SALRED.notna()) & (b.NBHEUR > 0) &
(b.NBHEUR.notna()))
```

```
# Critères pour identifier les chômeurs
# Non-étudiantes
critstud = ((b.ACTEU == 2) & b.ADEBEN.isna() & (b.ACTEU6 != 5))
# Les femmes en recherche d'emplois avec le dernier emploi avant 2007
critquit = ((b.ACTEU == 2) & (b.ADFDAP < 2007))
# Les femmes en recherches d'emplois avec le dernier emploi après 2007
critunemp = ((b.ACTEU == 2) & (b.ADFDAP >= 2007) &
                      (b.INSCAC == b.ADFDAP))
# Concatenation
critchom = (critstud | critquit | critunemp) & (b.ACTEU6 != 5)

# Identification des individus exclue de marché
critinact = ((b.ACTEU == 3) | (b.ACTEU == 2)) & (b.ACTEU6 != 5)

# Obtention d'échantillon entier
crit = ((critwork & salafich) | (critinact) | (critchom)) & (b.AG < 61)
```

Annexe 2 : Statistiques descriptives

2.1 Les statistiques descriptives complets pour la comparaison des femmes actives contre inactives

	Femmes Actives		Femmes Inactives	
Taille d'échantillon	7874		4411	
Abréviation	Moyenne	Variance	Moyenne	Variance
WAGE	11.756272	286.613928	0	0
EXPER	21.959741	143.838465	22.714124	202.718483
AG50	0.002413	0.002407	0.033099	0.032011
AG51	0.154940	0.130950	0.196101	0.157681
AG52	0.243205	0.184080	0.197461	0.158506
AG53	0.304293	0.211726	0.193380	0.156020
AG54	0.290577	0.206168	0.360009	0.230455
DDIPL1	0.153035	0.129632	0.087509	0.079869
DDIPL3	0.194437	0.156651	0.104965	0.093969
DDIPL4	0.213233	0.167786	0.168896	0.140402
DDIPL5	0.244221	0.184601	0.253457	0.189260
DDIPL6	0.074422	0.068892	0.106325	0.095042
DDIPL7	0.120650	0.106107	0.278848	0.201138
HOUS1	0.205232	0.163133	0.252777	0.188924
HOUS2	0.309754	0.213833	0.306960	0.212784
HOUS3	0.258319	0.191614	0.000227	0.000227
HOUS4	0.214249	0.168368	0.000227	0.000227
HOUS5	0.005969	0.005934	0.188619	0.153077
HOUS6	0.006477	0.006436	0.251190	0.188136
RESID1	0.007366	0.007313	0.024031	0.023459
RESID2	0.001778	0.001775	0.001587	0.001585
RESID3	0.873381	0.110601	0.880073	0.105569
RESID4	0.117475	0.103688	0.094310	0.085435

2.2 : Les statistiques descriptives pour la comparaison des femmes vraiment inactives contre les chômeurs

	Femmes Strictement Inactives		Femmes en Chômage	
Taille d'échantillon	3407		1004	
Abréviation	Moyenne	Variance	Moyenne	Variance
EXPER	24.574112	194.507644	16.402390	179.173913
AG50	0.022601	0.022096	0.068725	0.064066
AG51	0.169064	0.140522	0.287849	0.205196
AG52	0.189316	0.153521	0.225100	0.174604
AG53	0.193425	0.156058	0.193227	0.156046
AG54	0.410919	0.242136	0.187251	0.152340
DDIPL1	0.084825	0.077653	0.096614	0.087366
DDIPL3	0.104491	0.093600	0.106574	0.095311
DDIPL4	0.161139	0.135213	0.195219	0.157265
DDIPL5	0.249486	0.187298	0.266932	0.195875
DDIPL6	0.105078	0.094064	0.110558	0.098433
DDIPL7	0.294981	0.208028	0.224104	0.174055
HOUS1	0.232463	0.178476	0.321713	0.218431
HOUS2	0.275022	0.199443	0.415339	0.243075
HOUS3	0.000294	0.000294	0.000000	0.000000
HOUS4	0.000294	0.000294	0.000000	0.000000
HOUS5	0.204285	0.162601	0.135458	0.117226
HOUS6	0.287643	0.204965	0.127490	0.111347
RESID1	0.026123	0.025448	0.016932	0.016662
RESID2	0.001761	0.001758	0.000996	0.000996
RESID3	0.875550	0.108994	0.895418	0.093738
RESID4	0.096566	0.087267	0.086653	0.079223

Annexe 3 : Estimation de l'échantillon complet par les Moindres Carrés Ordinaires

Cette modèle est surement biaisée, alors nous la présentons que dans les annexes ainsi que sur l'étape de comparaison des performances des différents modèles.

Variable	Coefficient	Erreur standard	P-valeur
Constante	-0.1625	0.082	0.048**
RESID2	0.5591	0.208	0.007***
RESID3	0.4566	0.079	0.000***
RESID4	0.5886	0.085	0.000***
DDIPL1	1.1182	0.041	0.000***
DDIPL3	0.9633	0.036	0.000***
DDIPL4	0.6784	0.034	0.000***
DDIPL5	0.4357	0.031	0.000***
DDIPL6	0.3068	0.041	0.000***
EXPER	0.0684	0.003	0.000***
EXPERSQ	-0.0013	6.79e-05	0.000***

Nous pouvons facilement voir dans quelle mesure les coefficients sont biaisés par rapport à des modèles.

Annexe 4 : La liste des variables pour les modèles en grande dimension

4.1 Les idées méthodologiques

Dans cette partie d'annexe nous tentons de présenter la façon dont les variables sont génère d'une manière synthétique et claire pour ne pas encombrer ce travail par énumération des toutes les variables.
Il faut commencer par dire que dans cette partie notre individus de référence et représenté par une femme sans aucun diplôma qui a moins de 20 ans, sans enfants, résidant sans conjoint dans une zone urbaine sensible, mais non en région parisien.

Abréviation	Caractéristique
EXPER1	Expérience (présenté avant comme EXPER)
EXPER2	Expérience au carrée (présenté avant comme EXPERSQ)
ENF	Femme avec des enfants
COUPLE	Femme vivant avec un conjoint (conjoint inactive comme référence)
NOZUS	Femme résidant hors ZUS
PARIS	Femme habitant en région Parisien
COUPLECJACT	Femme habitant avec un conjoint active
AG51	20-29 ans
AG52	30-39 ans
AG53	40-49 ans
AG54	50-60 ans
DDIPL1	Diplôme supérieur à baccalauréat + 2 ans
DDIPL3	Baccalauréat + 2 ans
DDIPL4	Baccalauréat ou brevet professionnel
DDIPL5	CAP, BEP ou autre
DDIPL6	Brevet des collèges

Le reste des régresseurs est obtenu par une intermultiplication de ces variables afin d'avoir des relations non-linéaires complexes. Par exemple :

Abréviation	Caractéristique
COUPLEENFNOZUSPARIS	Femme en couple avec des enfants résidant en ZUS à Paris

L'ensemble finale contienne 141 variables.

4.1 La liste complète des variables

AG51	EXPER1COUPLENOZUS	AG52PARIS	AG54COUPLENOZUSPARIS	EXPER1AG53PARIS
AG52	EXPER1COUPLEPARIS	AG52COUPLECJACT	AG54ENFNOZUSPARIS	EXPER1AG53COUPLECJACT
AG53	EXPER1NOZUSPARIS	AG52COUPLEENF	AG54COUPLEENFNOZUSPARIS	EXPER1AG53COUPLEENF
AG54	EXPER1COUPLENOZUSPARIS	AG52COUPLENOZUS	EXPER1AG51ENF	EXPER1AG53COUPLENOZUS
AG55	EXPER1ENFNOZUSPARIS	AG52COUPLEPARIS	EXPER1AG51COUPLE	EXPER1AG53COUPLEPARIS
DDIPL1	EXPER1COUPLEENFNOZUSPARIS	AG52NOZUSPARIS	EXPER1AG51NOZUS	EXPER1AG53NOZUSPARIS
DDIPL3	EXPER1AG51	AG52COUPLENOZUSPARIS	EXPER1AG51PARIS	EXPER1AG53COUPLENOZUSPARIS
DDIPL4	EXPER1AG52	AG52ENFNOZUSPARIS	EXPER1AG51COUPLECJACT	EXPER1AG53ENFNOZUSPARIS
DDIPL5	EXPER1AG53	AG52COUPLEENFNOZUSPARIS	EXPER1AG51COUPLEENF	EXPER1AG53COUPLEENFNOZUSPARIS
DDIPL6	EXPER1AG54	AG53ENF	EXPER1AG51COUPLENOZUS	EXPER1AG54ENF
EXPER1	EXPER1DDIPL1	AG53COUPLE	EXPER1AG51COUPLEPARIS	EXPER1AG54COUPLE
EXPER2	EXPER1DDIPL3	AG53NOZUS	EXPER1AG51NOZUSPARIS	EXPER1AG54NOZUS
ENF	EXPER1DDIPL4	AG53PARIS	EXPER1AG51COUPLENOZUSPARIS	EXPER1AG54PARIS
COUPLE	EXPER1DDIPL5	AG53COUPLECJACT	EXPER1AG51ENFNOZUSPARIS	EXPER1AG54COUPLECJACT
NOZUS	EXPER1DDIPL6	AG53COUPLEENF	EXPER1AG51COUPLEENFNOZUSPARIS	EXPER1AG54COUPLEENF
PARIS	AG51ENF	AG53COUPLENOZUS	EXPER1AG52ENF	EXPER1AG54COUPLENOZUS
COUPLECJACT	AG51COUPLE	AG53COUPLEPARIS	EXPER1AG52COUPLE	EXPER1AG54COUPLEPARIS
COUPLEENF	AG51NOZUS	AG53NOZUSPARIS	EXPER1AG52NOZUS	EXPER1AG54NOZUSPARIS
COUPLENOZUS	AG51PARIS	AG53COUPLENOZUSPARIS	EXPER1AG52PARIS	EXPER1AG54COUPLENOZUSPARIS
COUPLEPARIS	AG51COUPLECJACT	AG53ENFNOZUSPARIS	EXPER1AG52COUPLECJACT	EXPER1AG54ENFNOZUSPARIS
NOZUSPARIS	AG51COUPLEENF	AG53COUPLEENFNOZUSPARIS	EXPER1AG52COUPLEENF	EXPER1AG54COUPLEENFNOZUSPARIS
COUPLENOZUSPARIS	AG51COUPLENOZUS	AG54ENF	EXPER1AG52COUPLENOZUS	
ENFNOZUSPARIS	AG51COUPLEPARIS	AG54COUPLE	EXPER1AG52COUPLEPARIS	
COUPLEENFNOZUSPARIS	AG51NOZUSPARIS	AG54NOZUS	EXPER1AG52NOZUSPARIS	
EXPER1ENF	AG51COUPLENOZUSPARIS	AG54PARIS	EXPER1AG52COUPLENOZUSPARIS	
EXPER1COUPLE	AG51ENFNOZUSPARIS	AG54COUPLECJACT	EXPER1AG52ENFNOZUSPARIS	
EXPER1NOZUS	AG51COUPLEENFNOZUSPARIS	AG54COUPLEENF	EXPER1AG52COUPLEENFNOZUSPARIS	
EXPER1PARIS	AG52ENF	AG54COUPLENOZUS	EXPER1AG53ENF	
EXPER1COUPLECJACT	AG52COUPLE	AG54COUPLEPARIS	EXPER1AG53COUPLE	
EXPER1COUPLEENF	AG52NOZUS	AG54NOZUSPARIS	EXPER1AG53NOZUS	

Annexe 5 : Extraits du code avec les algorithmes utilisés

5.1 L'algorithme pour le modèle de sélection endogène à deux étapes :

```
# Pour f  étant le jeu des données étudié
# Pour 'varia' étant une liste des noms des régresseurs

# Création des objets pour la régression Probit
D1 = f['ACTIV']
Z1 = f[varia].astype(float)
Z1 = sm.add_constant(Z1)

# Régression Probit
probit_reg1 = sm.Probit(D1, Z1,  missing = 'drop').fit() # Probit

# Extraction des valeurs générées par Probit
f['D_hat'] = probit_reg1.fittedvalues
# Création d'une variable correspondante au ratio de Mills inversé
f['imr'] = stats.norm.pdf(f['D_hat'])/stats.norm.cdf(f['D_hat'])

# Selection de sous-échantillon des femmes actives
f2 = f[(f['ACTIV'] == 1)]

# Création des objets pour la régression linéaire simple
Y1 = f2['LWAGE']
X1 = f2[variables]
X1 = sm.add_constant(X1) # Constante

# Régression MCO
lin_reg1 = sm.OLS(Y1, X1, missing = 'drop').fit(cov_type='HC0')
```

5.2 L'algorithme du choix de λ pour le modèle Logit :

```
# Le cas de LASSO logistique
# Le choix des paramettres
n = 10
k = 3

# La regression logistique penalisée avec validation croisée
logit_reg3 = LogisticRegressionCV(Cs = n, cv = 5, penalty = 'l1',
                                  solver = 'liblinear', max_iter = 150,
                                  n_jobs = -1)
logit_reg3.fit(Z1, D1)

# L'algorithm du choix de lambda
error = 1-logit_reg3.scores_[1].mean(axis=0)
for i in range(1, n-1):
    s1 = (error[i-1] - error[i])
    s2 = (error[i] - error[i+1])
    if (k > s1/s2):
        alpha = logit_reg3.Cs_[i]
        break

# Réestimation du modèle
logit_reg3 = LogisticRegression(penalty = 'l1', C = alpha,
                                solver = 'liblinear', max_iter = 250).fit(Z1, D1)
```

5.3 L'algorithme du choix de λ pour le modèle pénalisé simple :

```
# Le cas de LASSO standard
lin_reg3 = LassoCV(eps=1e-4, n_alphas = n, cv = 5, max_iter = 200, n_jobs = -1)
lin_reg3.fit(X1, Y1)

# L'algorithme du choix de lambda
error = lin_reg3.mse_path_.mean(axis=1)
for i in range(1, n-1):
    s1 = (error[i-1] - error[i])
    s2 = (error[i] - error[i+1])
    if (k > s1/s2):
        alpha = lin_reg3.alphas_[i]
        break

# Réestimation du modèle
lin_reg3 = Lasso(alpha = alpha, max_iter = 250).fit(X1, Y1)
```