

Le traitement de la non réponse endogène: une application à l'enquête post formation

Pauline Givord(), Alejandra Arbelaez (*),
Anne Bucher(*), Melchior Fosse(**)*

(Dares (** Insee)*

JMS-31 mars 2022

Outline

1. Introduction
2. Stratégie empirique
3. Application
4. Discussion

Outline

1. Introduction
2. Stratégie empirique
3. Application
4. Discussion

L'enquête Post-Formation

- Enquête réalisée depuis 2019 chaque trimestre auprès des stagiaires de la formation professionnelle
- Chaque vague couvre une cohorte de sortants, 6 à 9 mois après la sortie de formation (formation suivie, interrompue ou abandonnée)
- Entre 25 000 et 30 000 répondants chaque trimestre (2 000 par région FM hors Corse)
- L'enquête vise à rendre compte de la perception des stagiaires sur la formation qu'ils ont suivi, son déroulé, ainsi que sur ce qu'elle leur a apporté.

L'enquête Post-Formation - collecte

- Les données sont collectées par voie postale ou via internet - collecte en auto-administré
- Les taux de réponse sont faibles (environ 33%, avec tendance à la baisse)
- Test vague 3 : utilisation de relances téléphoniques pour les personnes ne répondant pas au questionnaire auto-administré - généralisé à partir de la vague 10 (dernière disponible)
- Quel redressement de la non réponse utiliser?
 - Soupçon de non réponse **endogène** - propension à répondre de la personne est corrélée avec les variables d'intérêt
 - Exemple : personnes qui ont abandonné la formation en cours moins intéressée à répondre
 - Méthode proposée par Castel et Sillard (2020) dans le cadre de l'enquête Epicov

Cette étude

- applique d'une méthode de redressement de la non réponse endogène dans l'enquête Post Formation
- discute mise en œuvre pratique et illustre hétérogénéité selon les variables d'intérêt

Outline

1. Introduction
2. Stratégie empirique
3. Application
4. Discussion

Stratégie empirique (I)

Modéliser simultanément y_i et r_i avec un modèle Heckman - cas d'une variable d'intérêt binaire

$$\left\{ \begin{array}{l} \text{(i)} \quad y_i^* = c^1 + \mathbf{z}_i \chi + \epsilon_i \\ \text{(ii)} \quad y_i = \mathbf{1}(y_i^* \geq 0) \\ \text{(iii)} \quad r_i^* = c^0 + \mathbf{z}_i \beta + \mathbf{w}_i \psi + \nu_i \\ \text{(iv)} \quad r_i = \mathbf{1}(r_i^* \geq 0) \end{array} \right. \quad (1)$$

- y_i^* : variable latente qui correspond à la variable d'intérêt y_i
- r_i^* : variable latente qui correspond au fait de répondre à l'enquête r_i
- z_i : variables disponibles dans la base d'échantillonnage
- w_i : instrument

Instrument : fait d'appartenir au sur-échantillon qui a bénéficié de relances téléphoniques

- ① non corrélé aux variables d'intérêt car tirages *indépendants* des deux échantillons
- ② effet significatif sur le fait de répondre à l'enquête

Estimations des indicateurs

- Deux méthodes:
 - Imputation pour les non répondants de la variable d'intérêt à partir du modèle (loi de Bernouilli de paramètre
$$\mathbb{P}(y_i = 1 | \mathbf{z}_i, \mathbf{w}_i, r_i = 0) = \frac{\Phi_2(-(c^0 + \mathbf{z}_i \beta + \mathbf{w}_i \psi, c^1 + \mathbf{z}_i \chi); -\rho)}{\Phi(c^1 + \mathbf{z}_i \chi)}$$
 - Répondération : estimateur de la probabilité de réponse r_i à partir du modèle (respectivement $\mathbb{P}(r_i = 1 | \mathbf{z}_i, \mathbf{w}_i, y_i = 1)$ et $\mathbb{P}(r_i = 1 | \mathbf{z}_i, \mathbf{w}_i, y_i = 0)$) et utilisation des poids $\pi_i \hat{r}_i$
- Un modèle différent doit être utilisé pour chaque variable d'intérêt

Outline

1. Introduction
2. Stratégie empirique
- 3. Application**
4. Discussion

Application - Enquête Post Formation

- Application aux 15 indicateurs principaux tirés de l'enquête (tableaux de bord)
 - **déroulé du parcours de formation** (délai avant entrée, test de compétences à l'entrée, abandon éventuel avant la fin, obtention d'une certification)
 - **modalités de la formation** (à distance, période en entreprise) et de son suivi (réorganisation de la vie personnelle nécessaire, perception d'une aide financière)
 - **indicateurs de satisfaction** (sentiment d'avoir reçu une information suffisante sur la formation, d'avoir été suffisamment accompagné, sentiment que la formation a été utile, contacts avec les employeurs, d'avoir reçu des opportunités d'emploi)
 - **Suite de la formation** (poursuite en formation, en emploi)
- Forme réduite : différences dans les moyennes dans les deux échantillons - si fort écart, soupçons de réponse endogène

Instrument : fait d'appartenir au sur-échantillon qui a bénéficié de relances téléphoniques

- ① non corrélé aux variables d'intérêt car tirages *indépendants* des deux échantillons
- ② effet significatif sur le fait de répondre à l'enquête

Condition 1- Première étape

Groupe initial	Echantillon principal		Echantillon téléphonique	
	Part (%)	Taux de réponse (%)	Part (%)	Taux de réponse (%)
<i>Vague 3</i>				
Ensemble	100	34,5	100	54,8
G1 : Adr. postale + Mail + Tél.	62,4	37,0	62,8	63,2
G2 : Adr. postale + Mail	18,1	33,1	18,5	34,1
G3 : Adr. postale + Tél.	11,7	30,2	11,1	61,8
G4 : Adresse postale	7,7	24,0	7,6	25,8
<i>Vague 10</i>				
Ensemble	100	31,9	100	44,9
G1 : Adr. postale + Mail + Tél.	65,6	34,1	66,1	51,1
G2 : Adr. postale + Mail	21,9	32,0	22,3	30,6
G3 : Adr. postale + Tél.	7,5	20,7	6,7	48,7
G4 : Adresse postale	5,0	18,7	4,9	20,5

Note: les groupes initiaux sont définis par le type de coordonnées (adresse postale, mail, numéro de téléphone mobile) disponible initialement dans la base de sondage.

Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).

Source: Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Estimation en forme réduite

Variable	Ech. principal (%)	Ech. téléphonique (%)	Diff.	rho	Imputation (%)
<u>Abandon</u>	7	8,6	-1,6	-0,23	12,1
<u>Formation à distance</u>	7	5	2,0	0,52	3,0
<u>Délais d'entrée en formation court</u>	72,7	77,1	-4,4	-0,25	80,4
Information suffisante	82,7	83,6	-0,9	-0,09	85,3
<u>Test de compétences</u>	61,4	57,4	4,0	0,22	51,9
Formation utile	85,7	85,8	-0,1	-0,03	86,3
<u>En emploi</u>	49,3	54,1	-4,8	-0,30	59,8

Notes : Les variables soulignées sont celles pour lesquelles le rho est significatif.

Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).

Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Outline

1. Introduction
2. Stratégie empirique
3. Application
4. Discussion

Discussion

- Repose sur extrapolation linéaire (deux points) : forte dépendance à la forme paramétrique
 - variables filtrées (conditionnelle) - biprobit pas forcément adapté
 - sous-populations - corrélation identique?
 - côté multidimensionnel de la non réponse (voir Dutz et al., 2022)
- Un modèle par variable : pb pour calculs de la précision (bootstrap)
- Pas d'identification distincte de la non réponse endogène et des effets de mode (réponse différente par téléphone)
 - Effets de mode - travaux en cours avec le DMS pour estimer simultanément effets de mode (administré vs collecte par téléphone)
 - trois échantillons: mixtes, auto-administrés seuls, téléphone seul
- Arbitrage coût/précision: tailles d'échantillons à discuter

Estimateur d'Heckman par repondération après troncature avec le modèle propre et le modèle d'abandon

Variable	Vague 3		Vague 10	
	Modèle propre (%)	Modèle abandon (%)	Modèle propre (%)	Modèle abandon (%)
<i>Déroulé du parcours de formation</i>				
Abandon	11,4	11,4	10,7	10,7
Obtention d'une certification	53,1	69,0	84,1	85,5
Délais d'entrée en formation court	80,1	82,1	88,7	84,5
Test de compétences	52,2	72,1	43,1	65,3
<i>Modalités de la formation et de son suivi</i>				
Période en entreprise	52,8	54,8	34,7	39,1
Formation à distance	3,1	10,7	39,7	52,3
Réorganisation de la vie personnelle	-	-	21,5	42,4
Aide financière	-	-	6,0	19,6
<i>Satisfaction</i>				
Information suffisante	85,1	89,2	81,2	86,0
Accompagnement suffisant	37,5	52,1	30,8	47,1
Contacts avec des employeurs	23,8	36,2	15,6	25,5
Opportunités d'emploi	48,4	63,7	42,6	51,4
Formation utile	86,4	91,1	81,2	87,2
<i>Suites de la formation</i>				
Poursuite en formation	10,8	17,5	16,0	13,5
En emploi	60,8	63,0	71,8	58,4

Notes: Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).

Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).

Source: Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Bibliographie

- CASTELL, L. et SILLARD, P. (2021). Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman. Rapport technique M2021-02, Institut National de la Statistique et des Études Économiques.
- DUTZ, D., HUITFELDT, I., LACOUTURE, S., MOGSTAD, M., TORGOVITSKY, A. et van DIJK, W. (2021). Selection in surveys. Working Paper 29549, National Bureau of Economic Research.

Merci!

Présentation de l'enquête Post-Formation

Table: Présentation de l'enquête Post-Formation

Vague	Sortants	Champ	Collecte	Taux de sondage (effectifs échantillons)	Taux de réponse (effectifs enquêtés)
1	T4 2018	France entière	T3 2019	43% (75 000)	35% (26 000)
2	T1 2019	France entière	T4 2019	51% (90 000)	33% (30 000)
3	T2 2019	France entière	T1 2020	37% (90 000)	38% (34 000)
4	T3 2019	France métro.	T2 2020	45% (75 000)	36% (27 000)
5	T4 2019	France métro. S2 2019 DOM	T3 2020	42% (90 000)	35% (32 000)
6	T1 2020	France métro.	T4 2020	47% (71 500)	37% (27 000)
7	T2 2020	France métro. S1 2020 DOM	T1 2021	48% (86 000)	37% (32 000)
8	T3 2020	France métro.	T2 2021	41% (71 500)	32% (23 000)
9	T4 2020	France métro. S2 2020 DOM	T3 2021	35% (91 500)	28% (26 000)
10	T1 2021	France métro.	T4 2021	29% (71 500)	34% (24 000)

Source: Dares.

Description des principales variables

Variable indicatrice	Définition
<i>Déroulé du parcours de formation</i>	
Abandon	Le sortant a abandonné sa formation avant la fin
Obtention d'une certification*	Le sortant a obtenu en intégralité ou en partie une certification
Délais d'entrée en formation court	Le sortant a déclaré un délai de 3 mois ou moins entre les démarches et le début de la formation
Test de compétences	Le sortant a fait une évaluation des compétences (test de positionnement) avant le début de la formation
<i>Modalités de la formation et de son suivi</i>	
Période en entreprise	Le sortant a déclaré qu'au moins une partie de la formation se déroulait en entreprise
Formation à distance	Le sortant a déclaré que la formation se déroulait au moins en partie à distance
Réorganisation de la vie personnelle	Le sortant a déclaré avoir du réorganiser sa vie personnelle pour suivre la formation
Aide financière*	Le sortant a déclaré avoir reçu une aide financière pour réorganiser sa vie personnelle
<i>Satisfaction</i>	
Information suffisante	Le sortant considère avoir été bien informé sur le contenu de la formation
Accompagnement suffisant	Le sortant considère avoir été accompagné pendant la formation pour préparer sa recherche d'emploi
Contacts avec des employeurs	Le sortant a déclaré avoir été mis en contact par l'organisme de formation avec des employeurs potentiels
Opportunités d'emploi	Le sortant a déclaré avoir obtenu des opportunités d'emploi grâce à la formation
Formation utile	Le sortant considère que la formation était utile
<i>Suites de la formation</i>	
Poursuite en formation	Le sortant a déclaré être de nouveau en formation à la date d'enquête
En emploi	Le sortant a déclaré être, à la date d'enquête, travailleur indépendant, salarié d'une entreprise, salarié d'un ou de particulier(s), fonctionnaire ou avoir une promesse d'embauche.

Notes: Variables indicatrices reconstruite à partir des réponses à l'enquête post formation. *Ces variables sont conditionnelles.

Source: Dares, exploitation des vagues 3 et 10 de l'enquête post formation.

Moyenne par repondération sur observables des principales variables

Variable	Vague 3		Vague 10	
	Echantillon principal (%)	Echantillon téléphonique (%)	Echantillon principal (%)	Echantillon téléphonique (%)
<i>Déroulé du parcours de formation</i>				
Abandon	7,0	8,6	8,3	8,9
Obtention d'une certification	85,4	83,8	81,6	81,7
Délais d'entrée en formation court	72,7	77,1	80,1	83,9
Test de compétences	61,4	57,4	59,4	56,3
<i>Modalités de la formation et de son suivi</i>				
Période en entreprise	42,2	46,5	32,8	33,5
Formation à distance	7,0	5,0	46,0	44,1
Réorganisation de la vie personnelle	-	-	36,7	31,9
Aide financière	-	-	16,0	14,0
<i>Satisfaction</i>				
Information suffisante	82,7	83,6	82,5	82,4
Accompagnement suffisant	41,4	40,3	40,9	38,2
Contacts avec des employeurs	25,3	25,0	20,8	20,0
Opportunités d'emploi	51,1	50,3	44,7	44,4
Formation utile	85,7	85,8	83,9	83,6
<i>Suites de la formation</i>				
Poursuite en formation	11,3	10,7	10,5	12,0
En emploi	49,3	54,1	50,7	56,7

Notes : Variables indicatrices reconstruite à partir des réponses à l'enquête post formation. La moyenne classique correspondent ici au valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).

Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).

Source : Dares, exploitation des vagues 3 et 10 de l'enquête post formation.

Moyenne par repondération sur observables et redressement sur inobservables des principales variables (I)

Variable	Moyenne classique (%)	rho	Valeur p	Repondération après troncature (%)
<i>Vague 3</i>				
<i>Déroulé du parcours de formation</i>				
Abandon	7,4	-0,23	0,00	11,4
Obtention d'une certification	85,0	0,17	0,01	79,3
Délais d'entrée en formation court	73,8	-0,25	0,00	80,1
Test de compétences	60,4	0,22	0,00	52,2
<i>Modalités de la formation et de son suivi</i>				
Période en entreprise	43,2	-0,25	0,00	52,5
Formation à distance	6,5	0,52	0,00	3,1
Réorganisation de la vie personnelle	-	-	-	-
Aide financière	-	-	-	-
<i>Satisfaction</i>				
Information suffisante	83,0	-0,09	0,10	85,1
Accompagnement suffisant	41,1	0,12	0,01	36,8
Contacts avec des employeurs	25,2	0,04	0,41	23,8
Opportunités d'emploi	50,9	0,06	0,22	48,4
Formation utile	85,7	-0,03	0,57	86,4
<i>Suites de la formation</i>				
Poursuite en formation	11,2	0,03	0,68	10,8
En emploi	50,5	-0,30	0,00	60,8
<i>Vague 10</i>				

Notes: Variables indicatrices reconstruites à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Le rho correspond ici au coefficient de corrélation entre les résidus de l'équation d'outcome et de sélection. Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).

Champ : sortants de formation de France entière au T2 2019 (vague 3)

Source: Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation

Moyenne par repondération sur observables et redressement sur inobservables des principales variables (II)

Variable	Moyenne classique (%)	rho	Valeur p	Repondération après troncature (%)
<i>Vague 10</i>				
<i>Déroulé du parcours de formation</i>				
Abandon	8,4	-0,12	0,31	10,6
Obtention d'une certification	81,6	-0,11	0,33	84,1
Délais d'entrée en formation court	80,9	-0,37	0,00	88,7
Test de compétences	58,8	0,37	0,00	43,3
<i>Modalités de la formation et de son suivi</i>				
Période en entreprise	32,9	-0,06	0,52	35,0
Formation à distance	45,6	0,16	0,06	39,2
Réorganisation de la vie personnelle	35,8	0,43	0,00	21,5
Aide financière	15,6	0,44	0,03	6,0
<i>Satisfaction</i>				
Information suffisante	82,5	0,04	0,68	81,2
Accompagnement suffisant	40,3	0,24	0,00	31,0
Contacts avec des employeurs	20,6	0,18	0,07	15,8
Opportunités d'emploi	44,6	0,04	0,66	42,9
Formation utile	83,8	0,09	0,35	81,2
<i>Suites de la formation</i>				
Poursuite en formation	10,8	-0,23	0,02	16,0
En emploi	51,9	-0,53	0,00	71,8

Notes: Variables indicatrices reconstruites à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Le rho correspond ici au coefficient de corrélation entre les résidus de l'équation d'outcome et de sélection.

Champ : sortants de formation de France métropolitaine au T1 2021 (vague 10).

Source: Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Caractéristiques initiales des deux échantillons(I)

Table: Caractéristiques initiales des deux échantillons

Variable	Vague 3				Vague 10			
	Echantillon principal		Echantillon téléphonique		Echantillon principal		Echantillon téléphonique	
	Effectifs	%	Effectifs	%	Effectifs	%	Effectifs	%
Ensemble	74498	100%	14902	100%	60736	100%	10721	100%
<i>Par sexe</i>								
Femmes	35596	47,8	7070	47,4	28059	46,2	5007	46,7
Hommes	38902	52,2	7833	52,6	32677	53,8	5714	53,3
<i>Par âge</i>								
Moins de 25 ans	20527	27,6	4095	27,5	14414	23,7	2537	23,7
25-50 ans	44854	60,2	8984	60,3	37826	62,3	6685	62,4
Plus de 50 ans	9117	12,2	1824	12,2	8496	14,0	1499	14,0
<i>Par niveau de diplôme</i>								
Infra BAC (ou niveau BAC non obtenu)	42681	57,3	8546	57,3	37201	61,3	6575	61,3
Diplôme niveau BAC ou plus	26826	36,0	5363	36,0	21400	35,2	3783	35,3
Niveau non renseigné	4991	6,7	994	6,7	2135	3,5	363	3,4
<i>Par objectif du stage</i>								
Création d'entreprise	2226	3,0	437	2,9	2340	3,9	411	3,8
Formations qualifiantes / professionnalisantes	39419	52,9	7881	52,9	37680	62,0	6654	62,1
Formations préparatoires	8864	11,9	1779	11,9	11870	19,5	2105	19,6
Adaptation au poste (AFPR-POEI)	6524	8,8	1304	8,7	3989	6,6	700	6,5
Objectif non renseigné	17465	23,4	3502	23,5	4857	8,0	851	7,9

Lecture : L'ensemble des caractéristiques initiales des deux échantillons sont bien identiques, ce qui est conforme au fait que la sélection dans l'un ou l'autre des échantillons est aléatoire.

Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).

Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Caractéristiques initiales des deux échantillons(II)

Table: Caractéristiques initiales des deux échantillons

Variable	Vague 3				Vague 10			
	Echantillon principal		Echantillon téléphonique		Echantillon principal		Echantillon téléphonique	
	Effectifs	%	Effectifs	%	Effectifs	%	Effectifs	%
<i>Par région de résidence</i>								
Auvergne Rhone Alpes	5484	7,4	1094	7,3	4984	8,2	882	8,2
Bourgogne Franche Comte	5484	7,4	1096	7,4	4980	8,2	881	8,2
Bretagne	5460	7,3	1091	7,3	4976	8,2	872	8,1
Centre Val De Loire	5473	7,3	1099	7,4	4976	8,2	871	8,1
Corse	1146	1,5	231	1,6	981	1,6	174	1,6
Grand Est	5477	7,4	1095	7,3	4981	8,2	879	8,2
Guadeloupe	1056	1,4	204	1,4	-	-	-	-
Guyane	646	0,9	125	0,8	-	-	-	-
Hauts De France	5479	7,4	1094	7,3	4984	8,2	878	8,2
Ile De France	5483	7,4	1106	7,4	4975	8,2	885	8,3
La Reunion	3815	5,1	762	5,1	-	-	-	-
Martinique	1753	2,4	351	2,4	-	-	-	-
Mayotte	458	0,6	86	0,6	-	-	-	-
Normandie	5443	7,3	1094	7,3	4980	8,2	878	8,2
Nouvelle Aquitaine	5481	7,4	1100	7,4	4981	8,2	879	8,2
Occitanie	5483	7,4	1100	7,4	4978	8,2	887	8,3
Pays De La Loire	5450	7,3	1085	7,3	4981	8,2	881	8,2
Provence Alpes Cote D Azur	5427	7,3	1090	7,3	4979	8,2	874	8,2
<i>Par groupe de disponibilité des coordonnées</i>								
Gr. 1 : Adr. postale + Mail + Tél.	46519	62,4	9356	62,8	39848	65,6	7087	66,1
Gr. 2 : Adr. postale + Mail	13475	18,1	2754	18,5	13284	21,9	2395	22,3
Gr. 3 : Adr. postale + Tél.	8735	11,7	1657	11,1	4558	7,5	716	6,7
Gr. 4 : Adresse postale	5769	7,7	1135	7,6	3046	5,0	523	4,9

Lecture : L'ensemble des caractéristiques initiales des deux échantillons sont bien identiques, ce qui est conforme au fait que la sélection dans l'un ou l'autre des échantillons est aléatoire.

Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).

Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.