Statistiques et économétrie appliquées Printemps 2019

Projet B

Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment

Groupe 69

Liam SVOBODA, Gaëtan MERMINOD, Gianmarco MURGIDA, Danica KOSTIC

1. Nettoyage de la base de données

Lignes 4 à 64 du Logfile.

2. Statistiques descriptives

2.1 Lignes 68 à 101 du Logfile et Fichier Excel 2.1.

	Groupe de contrôle			CCT			CCT + Formation			p-value
	Moyenne	Ecart-type	Observations	Moyenne	Ecart-type	Observations	Moyenne	Ecart-type	Observations	(CCT vs CCT + formation)
Variables										
Mère vit dans le foyer	0.9560117	0.2051693	1023	0.9637681	0.1869634	966	0.9636872	0.1871546	1074	0.9925
Nombre d'années d'éducation de la mère	4.211253	2.982941	942	4.275299	3.124042	919	4.02629	2.82054	1027	0.0652
Nombre d'années d'éducation du père	3.880517	3.393664	929	3.876836	3.24239	885	3.700694	3.131954	1009	0.24
Homme chef de famille	0.8523949	0.3548816	1023	0.8509317	0.3563403	966	0.8631285	0.3438718	1074	0.434
Taille du ménage	6.419355	2.842764	1023	6.236025	2.629089	966	6.456238	3.016643	1074	0.0805
Masculin	0.4868035	0.5000703	1023	0.5041408	0.5002418	966	0.5037244	0.5002191	1074	0.985
Age (en mois) au premier transfert	61.9912	30.99149	1023	60.95756	31.13353	966	61.27654	31.72565	1074	0.8182
Poids de naissance	6.813193	1.527763	683	6.70436	1.560191	634	6.844865	1.754179	674	0.1166
Accès à l'eau courante	0.0928641	0.2903839	1023	0.1180124*	0.3227899	966	0.1173184*	0.3219493	1074	0.96
Consommation de nourriture par tête	3022.756	4191.761	1023	3027.385	4560.298	966	2850.817	2585.665	1074	0.3003

2.2 Lignes 104 à 129 du Logfile.

2.3. Discutez les résultats de vos statistiques descriptives et partagez vos conclusions sur la validité de la méthode. N'oubliez pas d'utiliser les termes économétriques pour répondre à ces questions !

Les résultats obtenus montrent une significativité au seuil de 10% pour la variable "Accès à l'eau courante" (s3awater_access_hh_05) lorsqu'on compare les groupes CCT et CCT+Formation au groupe de contrôle. Il existe des différences significatives au seuil de 10% pour les variables "Nombre d'années d'éducation de la mère" (ed_mom) et " Taille du ménage" (s1hhsize_05) lorsqu'on compare la moyenne des groupes CCT et CCT+Formation. Toutes les significativités énumérées ne sont en réalité pas très fortes et ne sont par conséquent pas très utiles. L'allocation aléatoire de traitement est sûrement ce qui explique ces différences entre les groupes. En effet c'est le fruit du hasard que les moyennes divergent à un seuil significatif. Toutefois, ces différences montrent qu'en pratique, les allocations aléatoires ne distribuent pas toujours les caractéristiques de façon égale entre les groupes. Ce concept sera développé au point 3.1.4.

Concernant la méthode RCT, elle est valide pour plusieurs raisons :

1. Une famille attribuée au groupe de contrôle, ne pourra pas bénéficier des avantages des groupe de traitement.
2. Les caractéristiques endogènes seront, en moyenne, distribuées de manière égale à travers les groupes.
3. Il est aussi dit qu'il y a un taux d'attrition faible durant les trois périodes.

Un taux d'attrition faible nous permet d'éliminer la crainte qu'une partie des sujets étudiés aient quitté l'expérience où n'aient pas respecté les directives, comme par exemple de dépenser l'argent reçu dans les groupes CCT et CCT+Formation autrement que pour l'éducation des enfants.

3. Analyse d'impact

3.1 Lignes 135 à 164 du Logfile et Fichier Excel 3.1.

	1		2		3		
Variable dépendante	CCT	CCT + Formation	CCT	CCT + Formation	CCT	CCT + Formation	P-value test CCT=CCTtraining
Score au test de langue (standardisé)	0.0467 (0.0497)	0.0793 (0.0486)	0.0260 (0.0804)	0.184** (0.0796)	0.0260 (0.0981)	0.184* (0.101)	0.0384
Score au test de mémoire (standardisé)	0.0587 (0.0556)	0.134** (0.0543)	0.0303 (0.0495)	0.0849* (0.0484)	0.0303 (0.0534)	0.0849 (0.0583)	0.3338
Score au test de mémoire associative (standardisé)	0.0715 (0.0549)	0.122** (0.0537)	0.0621 (0.0492)	0.106** (0.0480)	0.0621 (0.0606)	0.106* (0.0629)	0.3307
Score au test de compétences interpersonnelles (standardisé)	0.0629 (0.0560)	0.0521 (0.0549)	-0.0456 (-0.0793)	-0.0728 (-0.0787)	-0.0456 (-0.0944)	-0.0728 (0.0666)	0.7456
Score au test de motricité globale (standardisé)	0.0986 (0.0673)	0.0748 (0.0659)	0.0821 (0.105)	0.0852 (0.103)	0.0821 (0.0866)	0.0852 (0.100)	0.9609
Score au test de motricité fine (standardisé)	0.0590 (0.0479)	0.140*** (0.0468)	0.0438 (0.0751)	0.129* (0.0751)	0.0438 (0.0877)	0.129 (0.0892)	0.2695

3.1.1. Pourquoi n'est-il en théorie pas nécessaire de rajouter des variables de contrôle dans le cas d'un RCT (Randomized Controlled Trial) ?

En théorie, si on effectue une sélection aléatoire, les propriétés non-observées des individus sont distribuées en moyenne dans les mêmes proportions dans les différents groupes. Les variables de traitement seront donc indépendantes des variables omises (càd : $Cov(x_i,w_i)=0$).

3.1.2. Quel(s) changement(s) constatez-vous à l'ajout de variables de contrôle et pourquoi ? Si vous observez un changement, à quelle(s) variable(s) est-il dû ?

En général, l'ajout des variables de contrôles fait varier les coefficients des variables de traitement. Le coefficient de la variable CCT pour les différentes régressions n'est pas significatif ni avant, ni après l'ajout des variables de contrôles.
Concernant les coefficients de la variable de traitement CCT+Formation, sa significativité augmente quand la variable bweight est ajoutée à la régression du score de test de langue. Pour les régressions des scores des tests de mémoire, de mémoire associative et de motricité fine, le coefficient de la variable CCT+Formation est significatif avant et après l'ajout des variables de contrôle. Pour les deux derniers tests (comportement interpersonnel & motricité globale), le coefficient n'est pas significatif ni avant, ni après l'ajout des variables de contrôle.
Nous arrivons à la conclusion que le RCT a créé des groupes plutôt similaires visible grâce aux faibles changements de significativité des coefficients. Nous avons aussi remarqué une forte augmentation de la valeur du R^2 lors de l'ajout de la variable i.age_transfer à nos 4 premières régressions, elle a par conséquent un fort effet de prédiction.

3.1.3. QUESTION BONUS : Pourquoi n'ajoutons-nous que des variables récoltées avant l'expérience (2005) et non des variables récoltées après le traitement (2008) comme variables de contrôle ?

Si l'on utilise les variables de 2008, l'effet attribué aux groupes de traitement sera faussé car les variables de contrôle auraient potentiellement été affectées par le traitement réalisé avant 2008. De plus certaines variables ont changé durant les 3 ans. Or les variables d'un essai contrôlé randomisé doivent rester constantes durant le

traitement. Les sujets de cette étude sont des familles ; les enfants grandissent, certains naissent, des familles se séparent et d'autres changements peuvent se produire.

3.1.4. Citez deux raisons pour lesquelles on ajoute généralement tout de même des variables de contrôle ?

On ajoute les variables de contrôle de toute façon pour assurer la validité interne de l'étude. En effet, il se peut que la sélection n'était pas parfaitement aléatoire ou que par malchance, les caractéristiques non-observées aient été distribuées de façon non-homogène comme c'est le cas dans cette étude. Dans notre étude, le poids moyen des enfants (bweight) n'était pas égal entre les groupes.
Une autre raison, qui est surtout valable pour les expériences qui portent sur des longues périodes, est que les variables de contrôle peuvent fluctuer et être influencées par les traitements. Ceci aurait pour effet des compositions parfois totalement différentes entre les groupes.
On constate en ajoutant les variables de contrôle, que notre modèle permet d'effectuer des prédictions de plus haute qualité, en d'autres termes notre R^2 augmente.

3.1.5. Imaginez que vous observiez les résultats ci-dessous pour la variable Accès à l'eau courante au test d'équilibre de l'enquête initiale, et que nous ne contrôlions pas pour cette variable. L'effet des CCT sur la variable Score au test de langue (standardisé)serait-il estimé correctement en estimant la régression de la colonne 3 ? Si non, serait-il surestimé ou sous-estimé ? Référez-vous à la formule du Biais de Variable Omise pour répondre à cette question.

Non il ne serait pas estimé correctement. En effet, on aura :

$$\beta_{1s}=\beta_1+ \beta_{2*}\text{cov}(CCT_i;AàE_i)/\text{var}(CCT_i)$$

Avec:
β_{1s}:Coefficient de corrélation de la variable binaire CCT avec le test d'équilibre de l'enquête sans l'inclusion de la variable Accès à l'eau (AàE).
β_1:Coefficient de corrélation de la variable binaire CCT avec le test d'équilibre de l'enquête avec l'inclusion de la variable Accès à l'eau (AàE).
β_2: Coefficient de la variable Accès à l'eau (AàE).

Il est donc apparent dans le tableau que $\text{cov}(CCT_i;AàE_i)<0$ parce que la moyenne de AàE est plus faible dans le groupe CCT que dans le groupe de contrôle et ce, de manière significative au seuil de α=5%. De plus, on constate aussi que β_2 est positif, c'est à dire que l'accès à l'eau influence positivement le score au test ce qui est intuitif.

Ainsi,

$$\beta_{1s}<\beta_1$$

Ce qui signifie que le biais de variable omise est négatif. Et donc que l'effet de CCT est faussement amoindri par l'omission de la variable AàE.

3.1.6. Qu'est-ce qu'un cluster et en quoi est-il important d'en tenir compte lors des analyses économétriques ?

Si les observations pour un individu ne sont pas indépendantes alors :

$$\text{cov}(\epsilon_{it};\epsilon_{it+1}) \neq 0$$

Il se pourrait que l'individu s'améliore dans le temps plus il joue à un jeu par exemple.

Il est important de tenir compte de ce fait car c'est une violation de l'hypothèses que les observations sont indépendantes. Cette hypothèse est importante pour l'estimation avec OLS. Lorsqu'on tient compte de ce phénomène, les coefficients ne changent pas mais les erreurs-types changent et deviennent robustes. Dans notre cas, cela permet de contrôler pour les caractéristiques liées à la géographie.

3.1.7. Quels changements observez-vous dans les résultats ?

Nous pouvons constater que les valeurs des coefficients ne changent pas car le clustering n'impacte pas le calcul de ces derniers. Néanmoins, nous remarquons maintenant que les erreurs-types sont robustes car le clustering corrige pour l'hétéroscédasticité. Cette modification a un impact sur la significativité de certains coefficients. C'est le cas ici car on constate que les variables dans la colonne CCT+Formation deviennent moins significatives dans la 3^{e} colonne qu'elles ne l'étaient dans la 2^{e} colonne.

3.2 Peut-on conclure à un impact causal des programmes de CCT et de CCT + Formation et si oui lequel (n'oubliez pas de commenter les coefficients obtenus)? Peut-on conclure à un impact causal différent du programme de CCT et du programme de CCT + Formation. Si oui lequel est le plus efficace?

À présent, nous avons distribué les sujets de l'expérience de façon aléatoire et apporté des variables de contrôle au modèle puisque la distribution aléatoire a mené à des différences significatives entre les trois groupes pour certaines caractéristiques. De plus, nous avons pris en compte le clustering ce qui nous a permis d'avoir les erreurs-types robustes nécessaires en présence d'hétéroscédasticité. Grâce à ces méthodes et opérations, nous pouvons supposer que la validité interne de l'expérience est assurée pour la 3^{e} colonne du tableau.

Le tableau montre des coefficients de corrélation faiblement significatifs pour le groupe CCT+Formation associés aux scores des tests de langue et de mémoire associative et ce, au seuil de 10% (c.f. Colonne 3 du tableau 3.1). Nous pouvons affirmer un effet causal entre recevoir un CCT+Formation et la performance aux deux tests précédents. Toutefois, nous ne pouvons pas rejeter l'hypothèse H_0 selon laquelle les coefficients sont nuls pour CCT vis-à-vis tous les tests ainsi que pour CCT+Formation vis-à-vis des trois tests qui ne sont pas mentionnés ci-dessus à savoir, le score au test de mémoire, le score au test de compétences interpersonnelles et le score au test de motricité globale.

Tous les résultats obtenus avec les régressions sont plus élevés pour le groupe CCT+Formation que ceux de CCT. Néanmoins, l'hypothèse d'égalité des moyennes pour le score du test de langue entre CCT et CCT+Formation est la seule à pouvoir être rejetée au seuil de 5%. Pour les coefficients des autres tests, les coefficients du traitement CCT et CCT+Formation sont considérés comme égaux.

Nous pouvons conclure qu'il y a un avantage à bénéficier du traitement CCT+Formation.

4. Hétérogénéité

4.1 Lignes 346 à 258 du Logfile et Fichier Excel 4.1.

Variables indépendantes	Filles	Garçons
Score au test de langue (standardisé)	0.1412176 (0.1413398)	-0.059653 (0.1182698)
Score au test de motricité globale (standardisé)	0.2584311 (0.1664311)	-0.0569286 (0.066742)

4.1.1. Interprétez les coefficients.

À première vue, les coefficients sont négatifs pour les garçons et positifs pour les filles. Toutefois, les résultats ne sont pas significativement différents de 0. Cela n'est pas surprenant car nous avons remarqué en établissant le tableau du point 3.1 que le traitement CCT n'influence aucun des tests de manière significative.

4.1.2. Peut-on conclure que l'effet du programme est différent en fonction du sexe ? Écrivez l'hypothèse à tester et testez manuellement dans Stata si la différence entre les deux coefficients est significative. Supposez que la covariance entre les deux est nulle.

Pour le langage :
H_{0L}: cct_z_language_08M = cct_z_language_08F
H_{1L}: cct_z_language_08M ≠ cct_z_language_08F

Pour la motricité :
H_{0M}: cct_z_grmotor_08M = cct_z_grmotor_08F
H_{1M}: cct_z_grmotor_08M ≠ cct_z_grmotor_08F

Nous constatons des statistiques de test z de -1.0899405 et -1.7586927 pour le test de langue et de motricité respectivement.

En p value nous trouvons ensuite :

- 0.2757394 qui n'est donc pas significatif, même à 10%
- 0.0786297 qui est significatif au seuil de 10%

En conclusion, on ne rejette pas H_{0L}, l'hypothèse selon laquelle il n'y a pas de différence entre le coefficient de CCT pour les garçons et les filles par rapport au test de langue. Toutefois, on peut bel et bien rejeter H_{0M}, l'hypothèse qui dit qu'il n'y a pas de différence entre les sexes pour le coefficient de CCT pour le test de motricité. En prenant en compte les informations du point 4.1.1, on déduit que le sexe de l'individu impacte le coefficient de CCT sur le test de langue mais pas au test de motricité.

4.2 Lignes 422 à 430 du Logfile.

4.2.1

Variables explicatives	Score au test de langue (standardisé)	Score au test de motricité globale (standardisé)
CCT	0.0981421 (0.1260663)	0.1613566 (0.1457456)
CCT_male	-0.1387365 (0.1434097)	-0.1528045 (0.1426947)
male	0.0600858 (0.0796235)	0.1082158 (0.1054516)

4.2.2. Interprétez chacun des coefficients. Peut-on conclure que l'effet du programme est différent en fonction du sexe?

La version simplifiée de nos régression ressemble à:

$$y_i = \alpha + \beta_0 CCT_i + \beta_1 male_i + \beta_2 CCT\_male_i$$

La variable male prend la valeur 0 si l'individu est une femme et 1 s'il est un homme.

Pour une femme, l'équation est: $y_f = \alpha + \beta_0$

Seul la variable CCT reste car les deux autres sont égales à 0 car la variable male a pris la valeur 0.

Pour un individu de sexe masculin, la variable binaire male est égale à 1.

$$y_m = \alpha + \beta_0 + \beta_1 + \beta_2$$

Le coefficient β_1 est le résultat qu'aurait en plus un garçon comparé à une femme si aucun des deux appartenait au groupe CCT.
Le coefficient β_2 est l'effet supplémentaire sur le score si l'individu est un homme appartenant au groupe CCT en comparaison à une femme du même groupe. Dans notre cas, le fait d'être un homme appartenant au groupe CCT a un effet négatif sur le score aux tests de langue et de motricité globale.
Tous les résultats affichés dans le tableau ci-dessus ne sont pas significativement différent de zéro.

4.3. Expliquez quelle est la différence théorique entre ces deux stratégies. Dans quel cas doit-on utiliser l'une ou l'autre ?

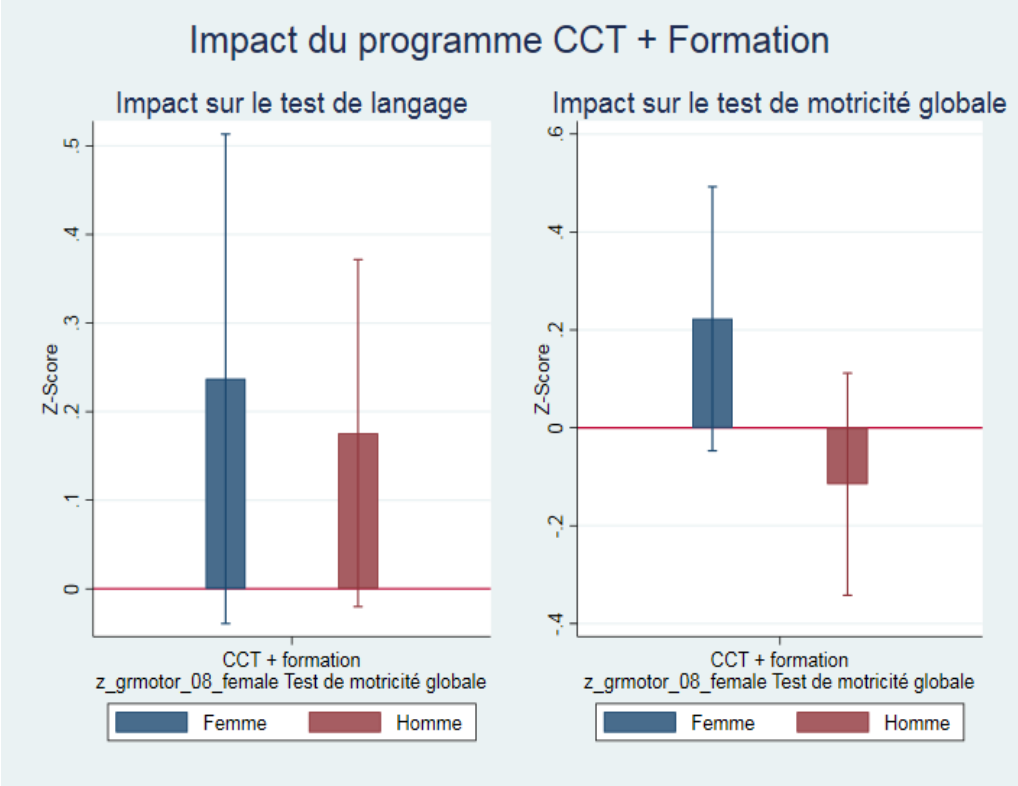
Avec la première stratégie, on ne distingue pas l'impact du sexe par rapport à celui du traitement CCT sur les scores des tests. Ce qui signifie que le coefficient qui résulte de la régression nous donne l'effet du traitement et celui du sexe confondu.

Avec la deuxième méthode, on distingue d'une part l'effet du sexe, l'effet du traitement et finalement l'effet d'interaction entre le sexe et le traitement.

On observe dans ce cas que l'effet d'interaction est statistiquement nulle.

Pour connaître l'effet d'un sexe, la première méthode est la meilleure. La seconde permet de directement comparer les effets du traitement sur les hommes par rapport à son effet sur les femmes.

5. Graphiques



5.3. Comment s'interprète votre graphique ?

Les extrémités des rectangles de couleur qui ne touchent pas la ligne du zéro représentent la valeur des coefficients. À savoir ; 0.2369756, 0.1757119, 0.2227674, -0.1152404 si on les énumère de gauche à droite.
Pour analyser ces coefficients nous utiliserons le fait que les moustaches délimitent les intervalles de confiance à 95%.
On constate donc que toutes les moustaches traversent la ligne rouge du zéro ce qui signifie qu'aucun des coefficients sont significativement différents de 0 au seuil de α=5%. Cela confirme ce que l'on trouve lorsqu'on observe les p-values obtenues à partir des régressions. On remarque pour le test de langage que les coefficients pour les garçons et les filles sont uniquement significatifs au seuil de α=10% et ceux du test de motricité ne le sont même pas au seuil de α=10%.
Intéressons-nous à présent à la configuration des moustaches du graphique de droite. Bien que les deux coefficients ne soient pas différents de 0 au seuil de 5%, on peut rejeter l'hypothèse de l'égalité des coefficients au seuil de α=5%. L'effet du traitement CCT+Training sur le score du test de motricité des garçons est donc significativement différent de son effet sur celui des filles au seuil de significativité α=5%. On ne peut toutefois pas en dire de même pour le score au test de langage.