LIGNES DIRECTRICES mVAM: DETECTER LES BIAIS DES OPERATEURS AVEC UN MODELE DE REGRESSION LINERAIRE

Les données sur les ménages collectées lors d'entretiens téléphoniques menés par des opérateurs peuvent être influencées par des comportements non intentionnels des opérateurs comme une mauvaise interprétation de la réponse de la personne interrogée ou une mauvaise façon de poser les questions. Cela peut conduire à une distorsion des données, connue aussi comme 'biais de l'opérateur'. MVAM utilise un modèle de régression linéaire comme technique pour identifier l'impact des opérateurs sur les données.

Exemple de comment faire une régression linéaire en utilisant le logiciel de statistique STATA pour prendre en compte l'effet dû à l'opérateur – Exemple de l'Irak.

Pour contrôler les différentes sources de biais notamment celui dû aux opérateurs, on a utilisé dans le cas de l'Irak les variables suivantes :

- 'Operateur' (Opr)
- '*Gouvernorat*' (ADM1),
- '*statut IDP*' (IDP) et
- '*type d'habitation'* (Habitation).

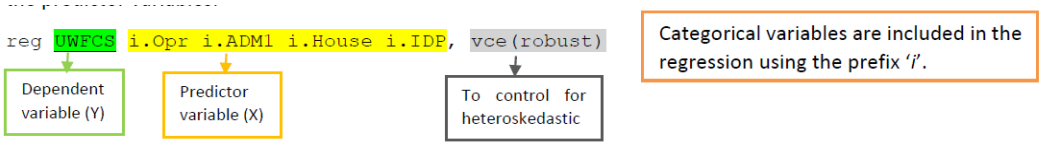
Initialement, d'autres variables comme l'*âge de la personne interrogée* et le *sexe du chef de ménage* avaient été inclus dans le modèle. Mais une analyse exploratoire et un ajustement du modèle initial ont indiqué que ces deux variables n'étaient pas des variables de prédiction significatives. Elles ont donc été éliminées du modèle.

La variable dépendante, '*FCS*', était la somme non-pondérée du compte des huit groupes alimentaires (score de consommation alimentaire non pondéré – UWFCS) car nous avons trouvé que cette diminution de variance permettait une différentiation par variables plus claire et donc des résultats plus clairs.

Cependant avant de faire la régression linéaire, nous avons défini les catégories de référence pour chaque variable qualitative suivant leur plus grande fréquence dans les données. Ainsi : pour *gouvernorat*=Baghdad, statut *IDP*='non IDP' et *Type d'habitation* = 'maison personnelle'

Le niveau de référence pour *operateurs* est Operateur.A parce qu'il.elle avait le UWFCS (score de consommation alimentaire non pondéré) moyen le plus proche de la moyenne de l'UWFCS de l'échantillon complet.

Dans Stata, la variable dépendante (UWFCS) est mise tout de suite après la commande de régression, suivie par les variables de prédiction:



TRADUCTION :
Dependent variable (Y) = Variable dépendante (Y)
Predictor Variable (X) = Variable de prédiction (X)
To control heteroskedastic = pour contrôler l'hétéroscedasticité

Categorical variables are included in the regression using the prefix 'i ' = Les variables catégorielles sont incluses dans la régression en utilisant le préfixe 'i'

Tableau 1: Résultats de la régression

	UWFCS Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Operator	**Baseline operator=Operator.A**					
Operator.B	-1.76149	.757631	-2.32	0.020	-3.247557	-0.2754238
Operator.C	-4.520824	.6407806	-7.06	0.000	-5.777692	-3.263955
Operator.D	1.457308	.7017217	2.08	0.038	0.0809062	2.833711
Operator.E	4.444539	.6606128	6.73	0.000	3.14877	5.740308
Operator.F	-.7736828	.6999972	-1.11	0.269	-2.146703	0.599337
Operator.G	.9491623	.7370538	1.29	0.198	-0.4965427	2.394867
ADM1	**Baseline ADM1: Baghdad**					
Anbar	-2.452756	1.022152	-2.40	0.017	-4.457671	-0.4478405
Babil	-.6120194	.8925465	-0.69	0.493	-2.362718	1.138679
Basrah	-.7146482	.9299657	-0.77	0.442	-2.538743	1.109447
Diyala	-1.376795	.9428552	-1.46	0.144	-3.226172	0.4725823
Duhok	.0270307	.8729092	0.03	0.975	-1.68515	1.739211
Erbil	-.3876416	.8071716	-0.48	0.631	-1.97088	1.195597
Kerbala	.3683594	.931366	0.40	0.693	-1.458482	2.195201
Kirkuk	-.9448069	.7902964	-1.20	0.232	-2.494945	0.6053315
Missan	-.8401069	1.265091	-0.66	0.507	-3.321538	1.641324
Muthanna	1.908981	1.186616	1.61	0.108	-0.4185251	4.236486
Najaf	.7378798	1.055241	0.70	0.484	-1.331938	2.807697
Ninewa	-2.277519	.7722409	-2.95	0.003	-3.792242	-0.7627957
Qadissiya	.0113681	1.072007	0.01	0.992	-2.091335	2.114071
Salah al-Din	-.6064122	.7360993	-0.82	0.410	-2.050245	0.8374205
Sulaymaniyah	-2.579634	.8465911	-3.05	0.002	-4.240192	-0.9190753
Thi-Qar	-1.252819	1.095065	-1.14	0.253	-3.40075	0.895112
Wassit	-2.811444	1.142835	-2.46	0.014	-5.053074	-0.5698134
kerbala	-18.63017	.7689348	-24.23	0.000	-20.13841	-17.12193
House	**Baseline House: Own home**					
Camp	-6.504216	1.35819	-4.79	0.000	-9.168258	-3.840175
Guest	-2.496454	.9953222	-2.51	0.012	-4.448744	-0.544165
Other	-4.840922	1.695378	-2.86	0.004	-8.166346	-1.515498
Rental	-3.376204	.4093596	-8.25	0.000	-4.179149	-2.57326
Unfinished_building	-5.087508	1.426357	-3.57	0.000	-7.885258	-2.289758
IDP	**Baseline IDP: Non IDP**					
YES IDP	-3.124104	.4936916	-6.33	0.000	-4.092463	-2.155746
_cons	45.54907	.6699399	67.99	0.000	44.235	46.86313

La p-value associée à un test à deux queues qui teste l'hypothèse selon laquelle chacun coefficient est différent de zéro. Pour l'éviter, la p-value doit être plus petite que 0,05 (vous pouvez également choisir 0,10).

Les résultats de la régression (Tableau 1) montrent que l'opérateur C et E ont un effet opérateur extrêmement significatif (valeur p: 0.000). Ils reportent un score de consommation alimentaire non pondéré respectivement inférieur et supérieur de 4,5 fois en comparaison avec l'opérateur A défini comme référence. Par ailleurs , il y a également d'autres opérateurs – opérateurs B et D – qui ont un impact significatif (négatif et positif respectivement) sur l'UWFCS. Lors de l'interprétation des résultats de la régression, également des effets autre que le biais introduit par les opérateurs doivent être pris en considération. Ainsi des diminutions significatives de la moyenne du UWFCS sont associées au fait d'être un ménage déplacé ou de vivre dans un camp/une location/des bâtiments non terminés ou d'être accueillis (gratuitement par quelqu'un ou dans une famille d'accueil). Nous avons aussi noté une diminution statistiquement significative parmi les personnes originaires des gouvernorats les plus directement affectés par le conflit.

Un des postulats principaux du modèle de régression (OLS) qui a un impact sur la validité de tous les tests (p, t et F) est que les résidus se comportent « normalement ». Les résidus (indiqués ici par la

lettre "`r`") sont la différence entre les valeurs observées (y) et les valeurs prédites. Si ces hypothèses ne sont pas confirmées, vous ne pouvez pas analyser vos données en utilisant une régression linéaire multiple car les résultats obtenus ne seront pas valides.

Après avoir fait une analyse de régression, dans STATA vous pouvez utiliser la commande *predict* pour créer des résultats :

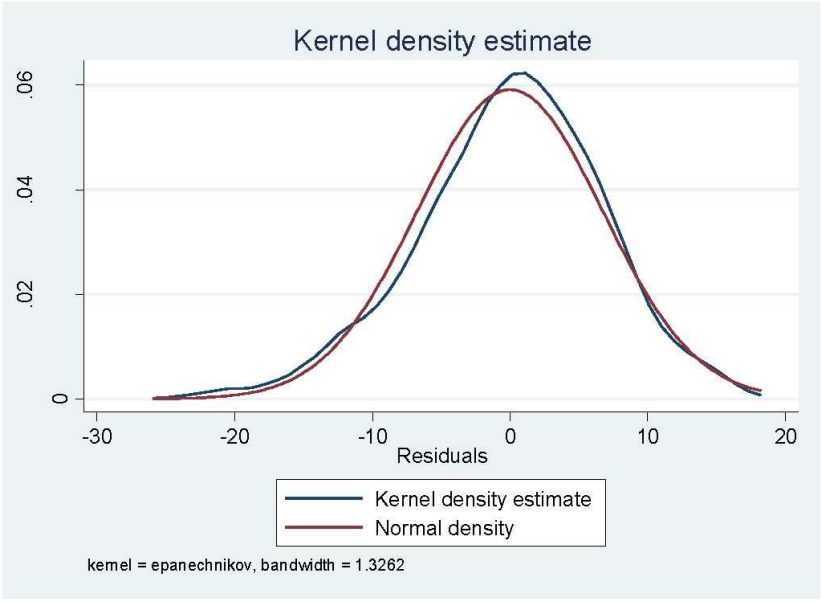
```
predict r, resid
```

et la commande kdensity pour créer le graphe de la densité estimée par noyaux (kernel density plot) avec l'option normale qui demande qu'une densité normale soit superposée sur le graphique. kdensity aide à vérifier la normalité dans les résidus.

Kdensity r, normal: →

L'option `normal`, superpose une distribution normale pour faire la comparaison.

Figure 1 : Densité par noyaux.



Le test graphique suggère que les résidus sont distribués normalement. Si les résidus ne suivent pas une distribution « normale », il faut que vous vérifiez si une variable a été omise, la spécification du modèle ou sa linéarité.

Annexe:

Liste des noms des variables:

- **ADM1_NAME**= zone administrative.
- **OPERATOR=** Nom des opérateurs.
- **House Type= Type de bâtiment de** résidence.
- **IDP_YN=** statut IDP.
- **Staples=** Nombre de jours où les personnes interrogées ont consommé des céréales, des racines et /ou des tubercules au cours des 7 derniers jours.
- **Veg=** Nombre de jours où les personnes interrogées ont consommé des légumes et/ou des feuilles au cours des 7 derniers jours.
- **Fruits=** Nombre de jours où les personnes interrogées ont consommé des fruits au cours des 7 derniers jours.
- **Protéins=** Nombre de jours où les personnes interrogées ont consommé des œufs, de la viande et /ou des produits de la mer comme plat principal au cours des 7 derniers jours.
- **Pulses=** Nombre de jours où les personnes interrogées ont consommé des légumes secs, des noix et /ou des graines au cours des 7 derniers jours.
- **Dairy=** Nombre de jours où les personnes interrogées ont consommé du lait (frais ou en poudre) et /ou d'autres produits laitiers au cours des 7 derniers jours.
- **Fats=** Nombre de jours où les personnes interrogées ont consommé de l'huile, des matières grasses et /ou du beurre au cours des 7 derniers jours.
- **Sugars=** Nombre de jours où les personnes interrogées ont consommé du sucre et des sucreries au cours des 7 derniers jours.
- **UWFCS=** score non pondéré de la consommation alimentaire.

Syntaxe STATA:

*****************Convertir une variable chaine en une variable numérique *****************

- **encoder ADM1_NAME, gen (ADM1)**
- **encoder Operator, gen(Opr)**
- **encoder HouseType, gen(House)**
- **encoder IDP_YN, gen(IDP)**

******************Créer le score de consommation alimentaire non pondéré ****************

- **gen UWFCS = Staples+Veg+Fruits+Proteins+Pulses+Dairy+Fats+Sugars**

*******************Modèle de régression non linéaire************************************

- **reg UWFCS i.Opr i.ADM1 i.House i.IDP, vce(robust)**

****************** Vérifier la normalité des résidus *************************************

- **predict r, resid** génère les résidus

Méthode Graphique

- **Kdensity r, normal** → produit le graphique de la densité et superpose la distribution normale.