mVAM GUIDANCE:

DETECTING OPERATOR BIAS THROUGH A LINEAR REGRESSION MODEL

The household data collected through live telephone interviews conducted by operators might be influenced from some unintentional operator's behaviors such as a wrong interpretation of respondent answer or a wrong manner of posing questions by the operators. These can result in data distortion, also known as '*operator bias'*. The linear regression model is used in mVAM as a technique to identify such operator effects.

Example of how to carry out linear regression using the statistical software STATA for accounting for the operator effect - Iraq example

In the case of Iraq, controlling for sources of bias such as operator effects a set of covariates were considered. They consisted of:

- 'Operator' (Opr)
- '*Governorate*' (ADM1),
- '*IDP status*' (IDP) and
- '*Housing type'* (House).

Initially, other variables such as *age of the respondent* and *gender of the head of households* were included in the model. But, exploratory analysis and initial model fitting indicated that those two variables were not significant predictors, so they were removed from the model.

The dependent variable, '*FCS*', was the un-weighted sum of the eight food group counts (un-weighted food consumption score - UWFCS) as we found that the decreased variance permitted much cleaner segregation by variables and hence clearer results.

However, before the linear regression was implemented, baseline levels were defined as follows, representing the most commonly observed level for each variable in the model: *governorate*=Baghdad, *IDPstatus*='non IDP' and *HousingType* = 'own home'. The baseline level for *operators* was Operator.A with the UWFCS mean closest to the average UWFCS of the complete sample.

In Stata, the dependent variable (UWFCS) is listed immediately after the regress command followed by the predictor variables:

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

Dependent variable (Y)

Predictor variable (X)

To control for heteroskedastic

Categorical variables are included in the regression using the prefix '*i*'.

Table1: Output of the regression

	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

The 2 tail p-value test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (you could choose also an alpha of 0.10).

The results of the regression (Table 1) demonstrate that operator C and E have an extremely significant "operator effect" (p-value: 0.000). In particular, they are respectively under and over-reporting the un-weighted food consumption score by 4.5 in comparison to the operator A defined as baseline. However, there are a few other operators —operators B and D— that have significant impact (negative and positive respectively) on UWFCS. When interpreting results of the regression, also factors other than operator bias need to be considered. Indeed, significant decreases in average UWFCS were associated with households who were IDPs and live in a camp/rental house/unfinished buildings or guests (staying with someone else for free or with host family). We also noticed a statistically significant decrease amongst respondents from governorates that have been most directly affected by conflict.

One of the main assumption of the regression model (OLS) that impacts the validity of all tests (p, t and F) is that residuals behave 'normally'. Residuals (here indicated by the letter "`r`") are the difference between the observed values (Y) and the predicted values. If these assumptions are not satisfied, you cannot analyze your data using multiple linear regression as valid results will not be obtained.

After running a regression analysis, in STATA you can use the *predict* command to create residuals:

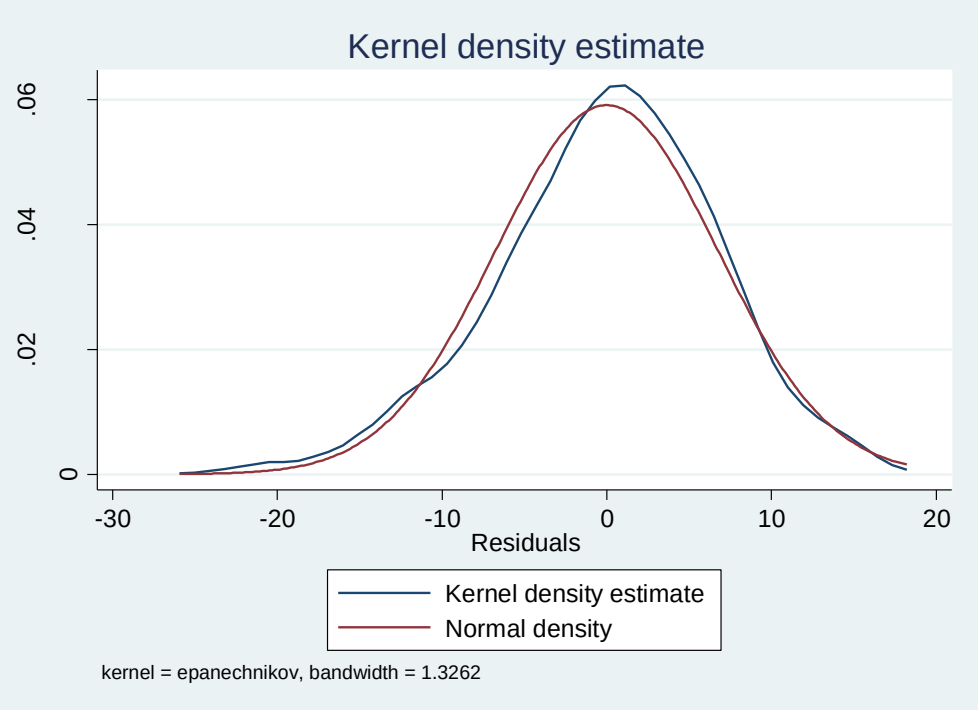
```
predict r, resid
```

and the kdensity command to produce a kernel density plot with the normal option requesting that a normal density be overlaid on the plot. kdensity stands for kernel density estimate and it will help to check for normality in the residuals:

```
Kdensity r, normal
```

The option `normal`, overlays a normal distribution to compare.

Figure1: Kernel density plot.



The graphical test suggests that residuals are normally distributed. If residuals do not follow a 'normal' pattern then you should check for omitted variable, model specification or linearity.

Annex:

List of Variable Names:

- **ADM1_NAME=** Administrative area.
- **OPERATOR=**Name of the operators.
- **HouseType=** Residence housing type.
- **IDP_YN=** IDP status.
- **Staples=** Number of days respondents consumed main staple starches, including cereals, grains, tubers and/or roots in the previous 7 days.
- **Veg=** Number of days respondents consumed vegetables and/or leaves in the previous 7 days
- **Fruits=** Number of days respondents consumed fruits in the previous 7 days.
- **Proteins=** Number of days respondents consumed eggs, meat, fish and/or other seafood as a main dish in the previous 7 days.
- **Pulses=** Number of days respondents consumed pulses, nuts, and/or seeds in the previous 7 days.
- **Dairy=** Number of days respondents consumed milk (powdered or fresh) and/or other dairy products in the previous 7 days.
- **Fats=** Number of days respondents consumed oil, fat and/or butter in the previous 7 days.
- **Sugars=** Number of days respondents consumed sugar and/or sweets in the previous 7 days.
- **UWFCS=** un-weighted food consumption score.

Stata Syntax:

*****************Converting a string variable to a numeric variable**********************

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

******************Creating the un-weighted food consumption score*********************

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

*******************Linear Regression Model**************************************

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

****************** Checking Normality of Residuals***********************************

- **predict r, resid** → generates residuals

Graphical Method

- **Kdensity r, normal** → produces kernel density plot with normal distribution overlaid.