# Guía VAM-m:
# Identificando el sesgo del operador mediante el modelo de regresión lineal

La recolección de datos del hogar mediante las llamadas telefónicas en tiempo real realizadas por los operadores pueden ser influenciadas por las conductas no intencionadas de los operadores, tales como: una mala interpretación de la respuesta del encuestado o una mala manera de plantear las preguntas por los operadores. Esto puede resultar en una distorsión de datos, también conocido como "sesgo del operador". El modelo de regresión lineal se utiliza en VAM-m como una técnica para identificar tales efectos del operador.

**Ejemplo de cómo llevar a cabo una regresión lineal utilizando el software estadístico STATA[1] explicando el efecto del operador – Ejemplo de Irak**

En el caso de Irak, se consideró establecer un conjunto de covariables para controlar las fuentes de sesgo de tales efectos del operador. Las cuales se conforman por:

- 'Operador' (Opr)
- '*Provincia*' (ADM1),
- 'Estado *PDI*[2]' (PDI) and
- '*Tipo de vivienda'* (Casa).

Inicialmente, otras variables como edad del encuestado y sexo del jefe del hogar se incluyeron en el modelo, pero el análisis explorativo y el modelo apropiado incial indicaron que esas dos variables no eran indicadores significativos, así que se eliminaron del modelo.

La variable dependiente Puntaje de Consumo Alimentario –PCA, era la suma no ponderada de ocho grupos alimentarios (Puntaje de Consumo Alimentario no ponderado –PCANP[3]), y encontramos que la varianza disminuida permitió una segregación más limpia por variables y por ende, resultados más claros.

Sin embargo, antes de implementar la regresión lineal, los niveles de línea base se definieron a continuación, representando el nivel más común para cada variable en el modelo: provincia=Bagdad, estado PDI="no PDI" y tipo de vivienda="casa propia". El nivel de línea base para operadores fue Operador.A con la media PCANP más cercana al promedio de PCANP como una muestra completa.

[1] Data Analysis and Statistical Software

[2] Personas Desplazadas Internamente

[3] Un-weighted food consumption score (UWFCS) – Puntaje de Consumo Alimentario no ponderado (PCANP)

En Stata, la variable dependiente (PCANP) se identifica inmediatamente después del comando de regresión, seguido de las variables predictoras:

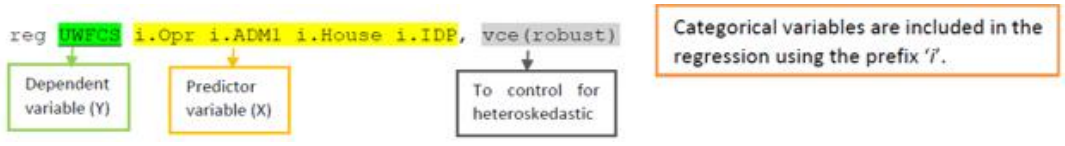



Tabla 1: Resultado de la regresión

| | 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 |

El resultado del valor p de dos colas comprueba la hipótesis que cada coeficiente es diferente de 0. Para rechazar esto, el valor de p tiene que ser menor de 0.05 (usted también puede seleccionar un valor alfa de 0.10)

Los resultados de la regresión (Tabla 1) comprueban que el operador C y E son muy significativos para el "efecto operador" (valor P: 0.000). En particular, se encuentran respectivamente por debajo y con un sobre registro de Puntaje de Consumo Alimentario no ponderado –PCANP de 4.5, en comparación al operador A definido como línea basal. Sin embargo, existen otros operadores – Operadores B y D – que tienen un impacto significativo (respectivamente negativo y positivo) en PCANP. Cuando se están interpretando los resultados de regresión, se debe considerar también otros factores además del sesgo del operador. De hecho, la disminuciones significativas en el promedio de PCANP se asociaron con los hogares de personas internamente desplazadas que viven en un campamento/casa rentada/edificios no terminados de construir o huéspedes (se hospeda gratis con alguien o con una familia anfitriona). También, observamos estadísticamente una disminución significativa entre los encuestados de provincias que habían sido directamente más afectados por el conflicto.

Uno de los principales supuestos del modelo de regresión (OLS)4 que impacta la validez de todas estas pruebas (p, t y F) es que los residuos se comportan "normales". Los residuos (indicados con la letra "r") son la diferencia entre los valores observados (Y) y los valores predichos. Si estos supuestos no son confirmados, usted no puede analizar los datos utilizando unas múltiples regresiones lineales, y no podrá obtener resultados válidos.

Después de ejecutar un análisis de regresión en STATA usted puede utilizar el comando predecir para crear residuales:

```
predict r, resid
```

y el comando *kdensity*[5] para producir un diagrama de densidad de núcleo (*kernel*) con la opción normal, solicitando que una densidad normal pueda ser superpuesta en el diagrama. *Kdensity* significa estimación de densidad de *kernel*, y le ayudará a revisar la normalidad en los residuos:

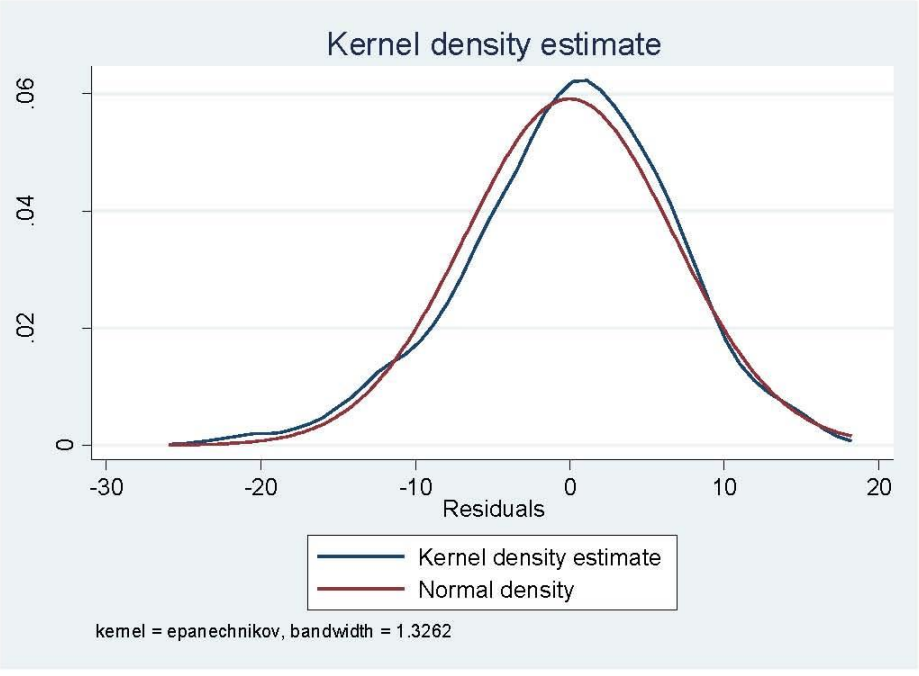
```
Kdensity r, normal
```

La opción normal, superpone una distribución normal para comparar.

[4] Ordinary Least Squares - mínimos cuadrados ordinarios (MCO) o mínimos cuadrados lineales

[5] Univariate Kernel density durity estimation

Figura 1: diagrama de densidad *kernel.*




La prueba gráfica sugiere que los residuos son distribuidos normalmente. Si los residuos no siguen un patrón “normal”, entonces debe verificar si una variable, un modelo específico o lineal se omitió.

**Annexo : Lista de nombre de variables:**

- **ADM1_NOMBRE=** área administrativa.
- **OPERADOR=**Nombre de los operadores.
- **Tipodevivienda=** Tipo de vivienda.
- **PDI_YN=** Estado de Personas desplazadas Internamente.
- **Cereales y tubérculos=** Número de días que los encuestados consumieron cereales incluyendo maiz, granos, arroz, pan, tubérculos/o raíces en los 7 días previos.
- **Veg=** Número de días que los encuestados consumieron vegetales y/o hojas en los 7 días previos.
- **Frutas=** Número de días que los encuestados consumieron frutas en los 7 días previos.
- **Proteínas=** Número de días que los encuestados consumieron huevos, carne, pescado y/u otro marisco como plato principal en los 7 días previos.
- **Legumbres=** Número de días que los encuestados consumieron legumbres, nueces y semillas en los 7 días previos.
- **Lácteos=** Número de días que los encuestados consumieron leche (polvo o fresca) y/u otro producto lácteo en los 7 días previos.
- **Grasas=** Número de días que los encuestados consumieron aceite, grasas y/o mantequilla en los 7 días previos.
- **Azúcares=** Número de días que los encuestados consumieron azúcar y/o dulces en los 7 días previos.
- **PCANP=** Puntaje de Consumo Alimentario no ponderado.

**Sintaxis de STATA:**

***************Transformando una variable de cadena en una variable numérica *******************

- **codificar ADM1_Nombre, gen(ADM1)**
- **codificar Operador, gen(Opr)**
- **codificar Tipodevivienda, gen(Casa)**
- **codificar PDI_YN, gen(PDI)**

*****************Creando un Puntaje de Consumo Alimentario no ponderado ********************

- **gen PCANP= Cereales+Veg+Frutas+Proteínas+Legumbres+Lácteos+Grasas+Azúcares**

*******************Modelo de regresión lineal*************************************************

- **reg PCANP i.Opr i.ADM1 i.Casa i.PDI, vce(robusto)**

****************** Verificando la normalidad de residuos**************************************

- **predecir r, residuo** → genera residuos

**Método gráfico**

- **Kdensity r, normal** → produce un diagrama de densidad kernel con una distribución normal superpuesta.