**MEMORIA: Ganancias y gastos de familias filipinas**

Artem Vartanov y Mario Baldocchi

Titulos a tamaño 14, texto tamaño 12

**Descripción de los datos**

Los [datos](https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure) a estudiar reflejan las ganancias y gastos de diversas familias filipinas, además de ciertos aspectos importantes a la hora de reforzar nuestra idea sobre el poder adquisitivo de una familia concreta.

Estos datos se podrían en diversos ámbitos laborales, pero nosotros supondremos que nuestro cliente es un banco, al que le interesa poder saber cuándo conceder un crédito o denegarlo.

Entendemos que se pueden sacar el certificado sobre la renta, pero para ahorrar el tiempo, teniendo en cuenta sus gastos que vemos en sus transacciones de la cuenta de nuestro banco(privacidad de datos?, eso no existe en Filipinas), queremos estimar su salario.

La variable que pretendemos estimar o la variable respuesta es *Total.Household.Income*, que representa la ganancia familiar anual medida en pesos filipinos (alrededor de 0.016 EUR a la hora de redactar la memoria).

**Preprocesado**

Al inicio, realizamos la conversión de todas las columnas que sean categóricas a factor con as.factor <lineas 20-36>. Esto se hace para posteriormente quedarnos solo con columnas numéricas.

Seguimos analizando. Vemos que hay solo dos columnas que tienen valores N/A - *Household.Head.Occupation* y *Household.Head.Class.of.Worker,* que representan la ocupación del miembro con ganancias principales de la familia y su tipo de relación con el empleador. Tenemos muchas opciones, entre las más sencillas sería o bien eliminar las filas que tengan algún N/A, que serían 3355 filas. O bien, eliminar estas dos columnas, porque al analizar el dataframe sin estas columnas nos quedan 0 filas con N/A. En este caso decidimos quitar estas columnas, pero las podremos tener en cuenta más adelante para la búsqueda de outliers, pues al ser dos variables categóricas nunca estarán dentro del modelo de regresión (a no ser que las convirtamos a dummy, pero serían demasiadas variables, pues tienen muchas categorías cada una).

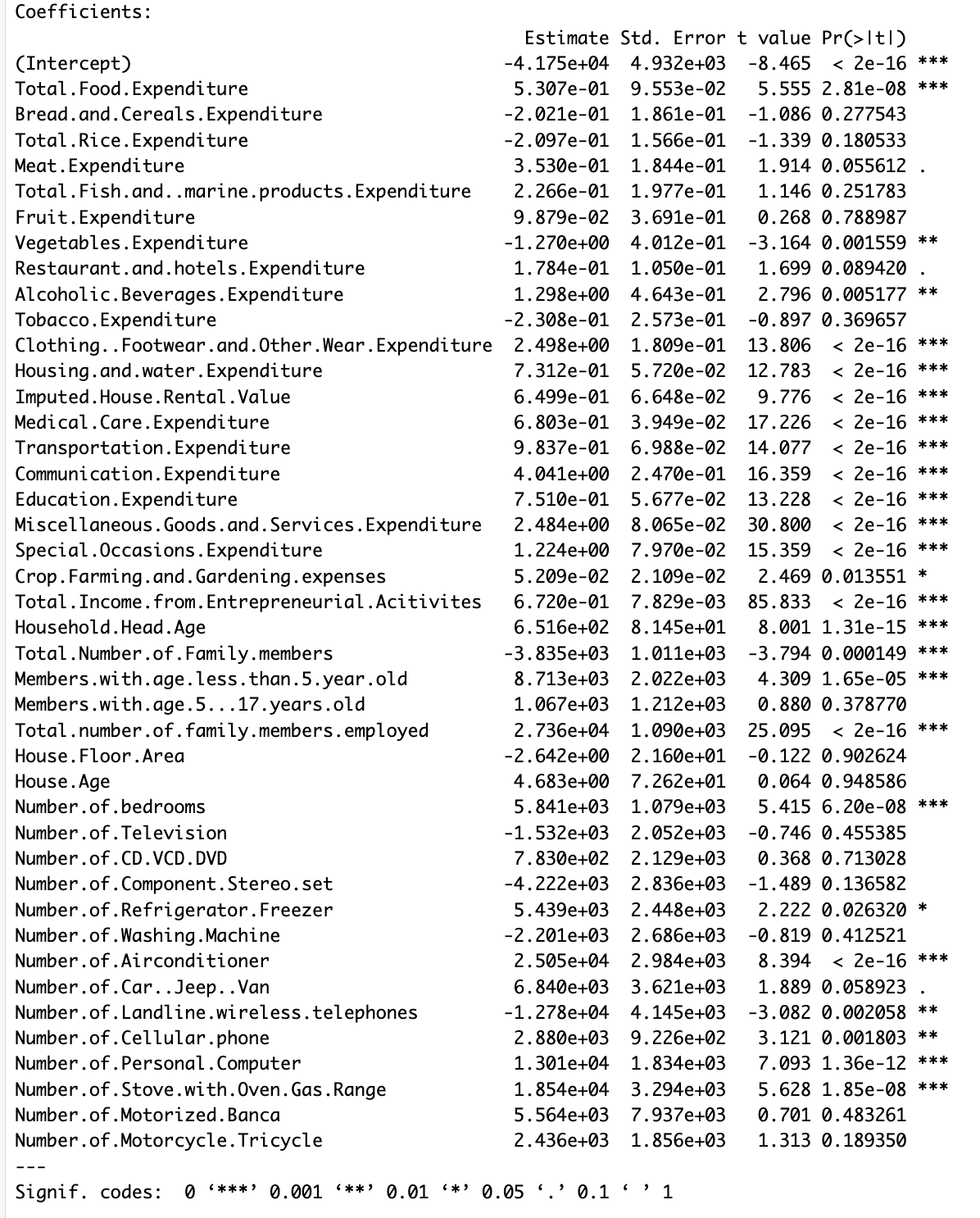
**Correlaciones**

A continuación pasamos a analizar las variables correlacionadas con la variable respuesta. Pero antes de ello, primero quitamos las columnas no numericas.

\*\*\*FALTA TEXTO\*\*\*

**Regresión lineal múltiple completa (RLM\_FULL)**

Generamos el modelo de regresión múltiple que usará todas las variable numéricas y todo el conjunto de datos. En la *Figura 1* aparece el summary del modelo completo. Vemos que no todas las variables tienen el p-valor < 0.05 (valor prefijado para rechazar hipótesis nulas). Esto quiere decir que no podemos estar seguros de que el coeficiente de esa variable no sea 0, en otras palabras, que debería estar en el modelo esa variable.

Figura 1: Summary de RLM\_FULL

Cabe destacar que el R2 ajustado es 0.7981 que mide la cantidad de variabilidad explicada por el modelo, teniendo en cuenta la penalización por introducir más variables en el modelo(usamos este criterio para poder contrastarlo con modelos posteriores, pues tendrán diferente número de variables). En este caso, como tenemos 42 predictores, la penalización es bastante alta.

Después de calcular la raíz cuadrada del error del modelo nos sale 132910.2, lo cual parece ser un error crítico.