¿Cómo citar?: Montero Granados. R (2016): Modelos de regresión lineal múltiple. *Documentos de Trabajo en Economía Aplicada*. Universidad de Granada. España.

Modelos de regresión lineal múltiple

Roberto Montero Granados
Departamento de Economía Aplicada
Universidad de Granada

Resumen

La regresión lineal múltiple trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variables independientes. En este tipo de modelos es importante testar la heterocedasticidad, la multicolinealidad y la especificación. En este curso trataremos de introducirnos en el mundo de la modelización, con creación de dummies, configurando un individuo de referencia, factores de ponderación, variables de interacción, interrelación, etc. Es particularmente importante entender lo que se está haciendo en cada momento porque estos principios sirven para prácticamente todos los modelos que se emprendan a continuación y después, con modelos más complejos y menos intuitivos, serán más difíciles de comprender.

Keywords: Regresión lineal, Stata,

Índice

1. Introducción

Dicen los que estudian neurociencia que la inteligencia humana se configura mediante la relación. Que las neuronas, relacionando axones construyen ideas simples y que estos grupos relacionándose entre sí pueden construir ideas más complejas. Que las relaciones entre fenómenos y su aparente causalidad nos hacen incluso más felices porque creemos entender mejor el mundo y así podemos aprovechar mejor sus beneficios y protegernos mejor de sus peligros. Antiguamente relacionar conceptos era algo fácil e intuitivo (una flor + una mariposa = una fruta fresca; una nube + algo de viento = agua también fresca) pero entonces unos señores (Bernoulli, Gauss, Laplace…) pensaron que aquello podía complicarse algo más y, para regocijo de los profesores de matemáticas, inventaron la probabilidad y la estadística.

Hay magnitudes con comportamiento relativamente constante es decir que cada vez que se miden el resultado es el mismo (el tamaño de un folio, el peso de un coche, la distancia a las estrellas…) o cuyo movimiento es fijo o constantemente acelerado (los objetos al caer, las estrellas otra vez…) y cuando queremos comparar dos o más lo que se hace son proporciones o fórmulas que pueden ser más o menos exactas (por ejemplo, el número π que ya tiene 10 billones de decimales, ¡sabe Dios para qué!). Estas son magnitudes deterministas.

Otras magnitudes, en cambio, son absolutamente diferentes cada vez que las observamos (un individuo puede medir más por la mañana que por la tarde, cada pieza de fruta tiene un peso distinto, cada año la producción de la fábrica es distinta…). Esto puede ser debido a que nos equivoquemos al medirlas, a que sean tan grandes que no podamos medirlas enteras o a que estén sometidas a algún tipo de influencia que no controlamos. Hay muchas variables físicas que no son deterministas (el ruido, la posición de un electrón, la generación de ondas) pero casi todas las variables sociales de una población son así (la edad, la renta, las tendencias políticas). A estas magnitudes las denominamos estocásticas.

La correlación es una medida de la similitud de la variabilidad de dos magnitudes estocásticas (que, como varían, las denominamos variables). La ventaja de

la correlación como medida de asociación es la sencillez de cálculo y su inconveniente es que, a veces, no es suficiente para comprender la relación entre ambas.

Si abrimos la base "satisfacción" y calculamos el coeficiente de correlación entre edad y estasalud…

```
. pwcorr edad estasalud
```

	edad	estasa~d
edad	1.0000	
estasalud	0.3808	1.0000

El índice de correlación es 0.3808, lo que significa que un cambio en edad suele representar un cambio, en el mismo sentido, en estasalud, pero ¿Es un índice alto o bajo? ¿Suficiente o insuficiente? El primer problema que tenemos es que para interpretar, siquiera superficialmente, el coeficiente tenemos que saber cómo están medidas las variables.

La regresión es una técnica estadística que consiste en calcular dicha similitud en forma de función matemática. Esta función nos ofrece mucha más información sobre dicha relación. Por ejemplo, el modelo más sencillo: la regresión lineal simple, ya nos informa de las siguientes magnitudes: la magnitud de la correlación; el incremento marginal, el valor de una de ellas cuando la otra es cero y si dicha relación puede considerarse significativa o fuerte (distinta de una relación normal) o no significativa o débil (similar a una relación normal)

Si abrimos la base "satisfacción" y calculamos la regresión lineal entre las mismas variables del ejemplo anterior: edad y estasalud…

```
. mean edad

Mean estimation                   Number of obs    =      7,747

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        edad |   47.63986   .2029268      47.24207    48.03765
--------------------------------------------------------------
```

```
. generate edad_recode= edad-47
(3 missing values generated)
. regress estasalud edad_recode

      Source |       SS           df       MS      Number of obs   =     7,725
-------------+----------------------------------   F(1, 7723)      =   1309.49
       Model |  694.014765         1  694.014765   Prob > F        =    0.0000
    Residual |  4093.09669     7,723  .529987918   R-squared       =    0.1450
-------------+----------------------------------   Adj R-squared   =    0.1449
       Total |  4787.11146     7,724  .619771033   Root MSE        =      .728

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 edad_recode |    .016782   .0004638    36.19   0.000     .0158729    .0176911
       _cons |   2.120836   .0082883   255.88   0.000     2.104589    2.137083
------------------------------------------------------------------------------
```

El resultado nos informa de la correlación entre ambas: la varianza común es de un 14.5% (raíz (R2)= coeficiente de correlación. Raíz(.1450)= .3808); de la magnitud de dicha correlación: un incremento de un año de edad implica un incremento de .016 en el estado de salud; de la esperanza de vida esperada para un individuo: en nuestro caso debido a la transformación de edad_recode = edad- 47, implica que un individuo de 47 años tiene una esperanza de estasalud de 2.12; y de que dichas magnitudes son todas significativamente distintas de lo normal: la significación de R2 se puede medir con la F y la significación de las magnitudes se mide con su error estándar y el p-valor de todas es superior a 0.000.

Existen muchas técnicas de regresión en función del tipo de variables y de la forma funcional supuesta entre ellas. Las más elementales (aunque las más potentes en el sentido de que se puede obtener más información) son las lineales. La regresión lineal supone que la relación entre dos variables tiene una forma lineal (o linealizable mediante alguna transformación de las variables). La regresión lineal tiene una versión "simple" que empareja dos variables, pero esta suele ser insuficiente para entender fenómenos mínimamente complejos en la que influyen más de dos variables, esta versión es la "múltiple". En el modelo de regresión lineal múltiple suponemos que más de una variable tiene influencia o está correlacionada con el valor de una tercera variable. Por ejemplo en el peso de una persona pueden influir edad, género y estatura, en la renta pueden influir trabajo, capital físico, conocimientos, etc. En el modelo de regresión lineal múltiple esperamos que los sucesos tengan una forma funcional como

$$y_j = b_o + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j$$

donde *y* es la variable endógena, *x* las variables exógenas, *u* los residuos y *b* los coeficientes estimados del efecto marginal entre cada *x* e *y*.

2. Tipos de variables

En regresión se trabaja con variables. Y lo que se hace es calcular siempre promedios (promedio de una variable, de una correlación, de una tendencia, de una función, de un ratio; promedios de variables estandarizadas, estudentizadas o refuncionalizadas con cualquier otra función) y su desviación típica (la desviación típica es una denominación que se reserva a la muestra y a la población, cuando se refiere a un parámetro estimado - la media, la tendencia u otro - se suele denominar error estándar). Una vez calculados ambos se interpretan conjuntamente (si son altos, bajos, en términos absolutos o lo que es más usual, en relación a algo como por ejemplo una distribución normal). El 95% de lo que hacen los estadísticos es eso. Lo que hace tremendamente gruesos y desagradables los libros de estadística es que las variables tienen formas muy distintas y su cálculo, aunque similar es ligeramente distinto, por lo que hay que rehacer casi todos los conceptos exprofeso para cada tipo de variables. Los libreros y los vendedores de software explotan constantemente este fenómeno y andan siempre incentivando a los estadísticos a introducir modelos nuevos, algunos completamente inútiles, pero que obligan a los usuarios a actualizar sus librerías (físicas y virtuales)

A pesar que hay muchos tipos de variables los estadísticos se divierten poniendo a un mismo tipo de variable nombres distintos, para confundir a los estudiantes. Así en la función:

$$y_j = b_o + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j$$

y es una variables que puede denominarse alternativamente como endógena, dependiente, regresando, explicada o variable respuesta, entre otros. *x* son unas variables que puede denominarse: exógena, independiente, regresor o explicativa. Aunque todo el mundo evita hablar de causalidad (porque correlación no es, en

absoluto, prueba de causalidad), todo el modelo parece indicar un sentido de los efectos desde las variables *x* hacia la variable *y*. de forma que el valor de esta última parece formarse a partir de los valores o la influencia de los valores las primeras.

En regresión lineal múltiple sólo suele haber una variable endógena y puede haber varias variables exógenas. Es decir se individualiza el fenómeno observado. También puede darse el caso de la existencia de varias variables endógenas, pero su solución es difícil por lo que no es el caso general.

Dichas variables (tanto endógenas como exógenas) pueden adoptar dos formas generales:

- Continuas: Las variables continuas son aquellas que llenan el espacio. Son números reales (que pueden tener o no decimales) y servirán incluso cuando su rango no sea desde $-\infty$ hasta $+\infty$. Suelen ser variables cuantitativas (como el peso o la edad) pero también pueden ser consideradas continuas variables cualitativas cuando pueden ordenarse y tienen un número no bajo de elementos (se dice que con más de siete elementos puede considerarse cuantitativa. Ejemplos son el número de escalones de una escalera, habitaciones de una vivienda, árboles plantados…) Todas las variables de recuento (números enteros) siempre que su rango sea alto (más de 7 elementos) podrían considerarse como continuas. Dentro de las variables continuas tienen especial relevancia las conocidas como porcentajes. Estos ratios pueden considerarse variables continuas normales cuando se mueven en un rango central relativamente amplio pero deben considerarse de forma especial cuando se mueven cerca de sus extremos porque sus tasas de crecimiento se ven constreñidas al intervalo [0; 1]

- Discretas: Las variables discretas son aquellas que se mueven "a saltos". Además de las variables de recuento suelen ser factores cualitativos que indican alguna característica del individuo (como el género, color, idioma…) Si las características son sólo dos se suelen llamar dicotómicas (género, bebedor…). Si son más de dos se suelen denominar simplemente factor.

El tipo de variable es más importante si afecta a la variable endógena (porque nos obligará a utilizar uno u otro modelo de regresión) pero no es tan importante si

afecta a la variable exógena. No obstante a las variables exógenas factor y ordenadas también se les puede extraes más contenido informativo si se las transforma en dummies.

En el primer caso tenemos que para cada tipo de variable hay un modelo de regresión completamente distinto:

Tipo de Variable	**Modelo**
Continua	Lineal
Dicotómica	Logit o probit
Recuento	Poison o Binomial
Factor ordenado	Logit o probit Ordenada
Factor	Logit o probit Multinomial
Porcentaje	Regresión fraccional

Sin embargo en el caso de las variables exógenas la distinta forma sólo exigirá una distinta interpretación de forma que basta con saber cómo están codificadas para interpretar los coeficientes estimados y demás parámetros. Quizá el caso con el que debemos tener especial precaución sea el de los factores porque su introducción directamente en el modelo no puede interpretarse de una forma lógica. Por ejemplo si se introduce la variable color de pelo en un modelo y el resultado es 3 ¿qué puede significar? Pues absolutamente nada.

Dentro del modelo lineal es interesante distinguir entre varios tipos de variable dependiente, porque puede condicionar el tipo de modelo de regresión:

- Libre: Se tienen datos de una muestra que abarca toda la posible medida de la variable que esta puede tener en la población.
- Censurada: Faltan datos en la muestra de la variable en alguna zona de la que sí hay datos en la población (el estado de salud en una encuesta que sólo recoge a mayores de 16 y hasta 85 años). A veces se pueden ajustar modelos para regresión censurada (Modelos Tobit)
- Truncada: Faltan datos en la muestra, o estos dan un salto a partir de un momento. Por ejemplo está prohibido trabajar por menos del salario mínimo de forma que quien no cobra el salario mínimo dice cobrar cero. Esto hace que la

regresión se trunque. A veces se pueden ajustar modelos para regresión truncada.

- Doble Valla: Surge como evolución del modelo truncado donde parece que hay dos fenómenos que cuantifican una relación: una primera que determina el acceso o no al fenómeno y una segunda que determina el grado con el que se accede. Por ejemplo para obtener unas determinadas calificaciones en la Universidad primero hay que acceder: la probabilidad condicionada de acceso sería la primera valla y la probabilidad condicionada de sacar buenas calificaciones la segunda.

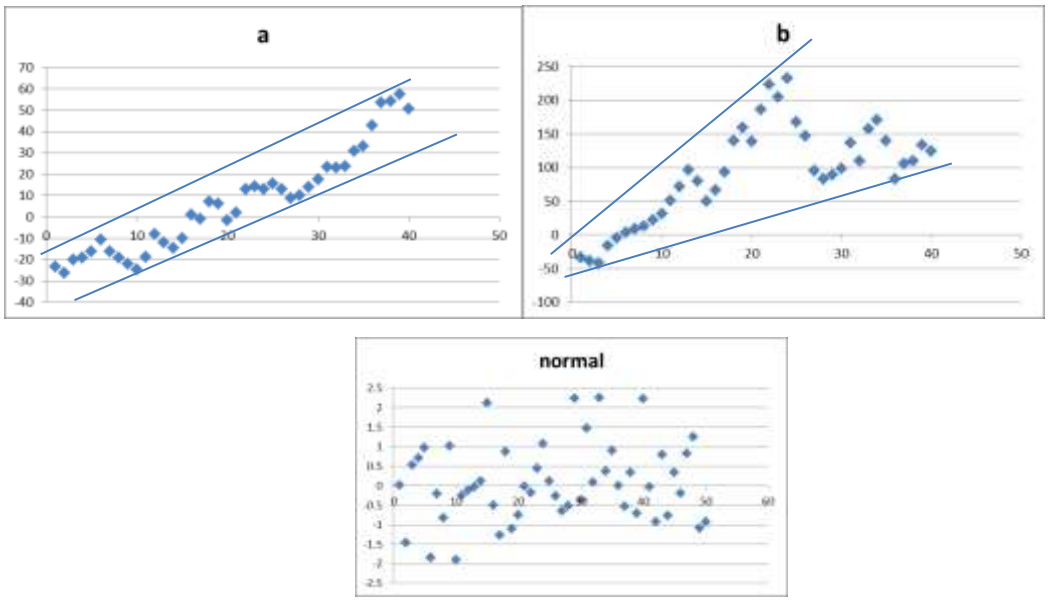
3. Hipótesis y Estimación

Para que los resultados de la regresión sean "confiables" (confiable es una forma coloquial de referirse a: insesgados, es decir que sus resultados sean parecidos a los reales; y óptimos, es decir que su varianza sea mínima) es necesario que:

a) La relación entre las variables sea lineal. Ser lineal no significa que forzosamente tenga que ser una línea recta sino también que pueda ser lineal con alguna transformación.

b) Las perturbaciones (es decir los efectos provocados aleatoriamente o por variables no incluidas en el modelo) deben ser: de media cero, homocedásticas y no autocorrelacionadas. Se suelen resumir estos bajo la denominación de "esfericidad" de los residuos.

Por ejemplo, si los siguientes gráficos son los residuos de tres modelos el modelo a no tiene residuos de media cero (aunque parece homocedásticos), los residuos del modelo b son, además, heterocedásticos y ambos parecen muy autocorrelacionados (un valor parece depender del valor anterior). Solo el modelo "normal" tiene unos residuos centrados en 0 (E(u) = 0), no parecen abrirse o cerrarse (son homocedásticos) y no tienen tendencia (no están autocorrelacionados)

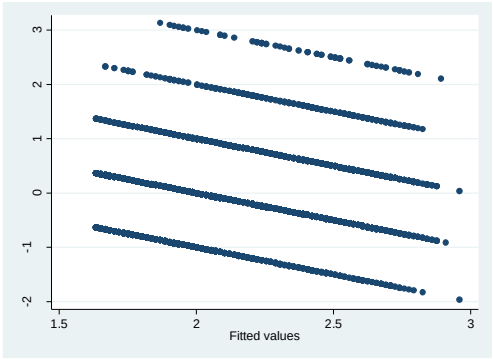


Aunque digan que una imagen vale más que mil palabras, en estadística esto no siempre se cumple. El análisis gráfico casi nunca es tan explícito en la vida real. En el caso de la base de Satisfacción si graficamos los residuos de una regresión cualquiera.

. regress estasalud edad

(output omitido)

. rvfplot



El resultado no es siempre igual a la teoría. No es evidente que no haya tendencia, ni qué pueden implicar las distintas bandas. Aunque parece no se puede afirmar rotundamente si hay o no esfericidad.

El software dice que tiene dos formas de estimación de una regresión lineal. Un primero por mínimos cuadrados ordinarios (MCO), que consiste en resolver la ecuación:

$$\hat{b} = (X'X)^{-1}X'y$$

Donde $\hat{b}$ es el vector de estimación de los coeficientes, *y* es el vector de la dependientes *X* es la matriz de variables dependientes y *X'* es la traspuesta de *X*

Y un segundo mediante máxima verosimilitud (ML). Que consiste en maximizar la ecuación:

$$\ln L(Y) = -\frac{n}{2}\ln(2\pi^2) - \frac{n}{2}\ln(2\sigma^2) - \frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}$$

Derivando respecto de β y de σ e igualando a 0. Ambos procedimiento llevan a los mismos resultados (la práctica totalidad de las veces) pero hay ocasiones en que, por repugnantes e inexpugnables problemas matemáticos, no es posible resolver el sistema por MCO y debe resolverse por máxima verosimilitud.

La ventaja de resolver por MCO es que obtenemos medidas de ajuste confiables (R^2 y $\overline{R}^2$). Por ML no podemos obtener (en la mayoría de los casos) una R^2 pero podemos obtener otras medidas de ajuste como el Criterio de información bayesiano (BIC) y el de Akaique (AIC) (en ambos casos cuanto más pequeños mejor).

4. Modelando

Si en regresión lineal simple se dice que se necesitan al menos 30 datos para que el teorema central del límite entre en vigor y las estimaciones sean consistentes, en regresión múltiple necesitaremos además un número mínimo de casos en función de las variables a introducir. Se dice que, además de los 30 casos general se necesitan un mínimo de 10 casos por variable adicional (Si *k* es el número de variables independientes el mínimo sería de *k+2* y algunos autores sugieren necesario *k·20*)

Aunque no es normal hay que tener en cuenta que un exceso de variables independientes puede hacer subir artificialmente el R^2 pero también reducir la significación estadística de las variables significativas.

Para comprobar el efecto de la adición incontrolada de variables en el coeficiente de determinación, en la base de datos "satisfacción" se han introducido sucesivamente variables en el orden de la lista. Desde el primero:

```
. regress estasalud edad

      Source |       SS           df       MS      Number of obs   =     7,725
-------------+----------------------------------   F(1, 7723)      =   1309.49
       Model |  694.014765         1  694.014765   Prob > F        =    0.0000
    Residual |  4093.09669     7,723  .529987918   R-squared       =    0.1450
-------------+----------------------------------   Adj R-squared   =    0.1449
       Total |  4787.11146     7,724  .619771033   Root MSE        =      .728

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |    .016782   .0004638    36.19   0.000     .0158729    .0176911
       _cons |   1.332081    .023597    56.45   0.000     1.285824    1.378337
------------------------------------------------------------------------------
```

Hasta

```
. regress estasalud edad genero est_recode ocup_1d nac_esp frec_prim_publ frec_urg_publ
frec_esp_pub frec_hosp_pub izq_der

      Source |       SS           df       MS      Number of obs   =     5,165
-------------+----------------------------------   F(10, 5154)     =    164.47
       Model |  721.581673        10  72.1581673   Prob > F        =    0.0000
    Residual |   2261.2102     5,154   .43872918   R-squared       =    0.2419
-------------+----------------------------------   Adj R-squared   =    0.2404
       Total |  2982.79187     5,164  .577612678   Root MSE        =    .66237

------------------------------------------------------------------------------
     estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
          edad |   .0129958   .0006109    21.27   0.000     .0117981    .0141934
        genero |  -.0254716   .0186058    -1.37   0.171    -.0619468    .0110036
    est_recode |   -.069245   .0105595    -6.56   0.000    -.0899461   -.0485438
       ocup_1d |   .0084289   .0043011     1.96   0.050    -3.15e-06    .0168609
       nac_esp |  -.0606391   .0360727    -1.68   0.093    -.1313569    .0100787
frec_prim_publ |   .0250182    .002548     9.82   0.000     .0200229    .0300134
 frec_urg_publ |   .0423709   .0081534     5.20   0.000     .0263866    .0583551
  frec_esp_pub |    .073405   .0061831    11.87   0.000     .0612835    .0855266
 frec_hosp_pub |   .0555632   .0239583     2.32   0.020     .0085947    .1025317
       izq_der |  -.0114292   .0050795    -2.25   0.024    -.0213872   -.0014713
         _cons |   1.564264   .0735283    21.27   0.000     1.420118    1.708411
------------------------------------------------------------------------------
```

Es decir introduciendo más variables hemos conseguido incrementar el R2 desde el 14.5% hasta el 24.19%. Aunque en el camino algunas variables que originariamente eran significativas han perdido dicha condición.

Antes de obtener resultados hay que preparar los datos de que dispongamos. Los datos se suelen obtener de muy diversas fuentes y con codificaciones a veces inverosímiles. Pero el delicado programa estadístico, para no indigestarse, necesita que los cocinemos un poco. Normalmente tendremos que jugar con datos ausentes, datos irracionales (en el sentido coloquial no matemático), re-escalar variables, linealizarlas, etc. Para la preparación de los datos se puede disponer de unas reglas elementales generales pero probablemente sea el momento en que la experiencia y el genio del modelador tengan más cabida e importancia.

Datos ausentes

Hemos de saber que el programa desprecia cualquier individuo a quién falte uno de los valores de las variables incluidas en el análisis. En datos micro o macro esto puede ser radicalmente importante (puede haber países o regiones enteras sin una variable y eso excluye a dicha región del análisis) pero también puede serlo en casos de encuestas. Hay preguntas que, por comprometidas o por que no son bien comprendidas, no son cumplimentadas por muchos encuestados. Entonces, a pesar de su interés estadístico hemos de pensar si su inclusión en el modelo trae más ventajas que inconvenientes ya que si se incluye dicha variable se pierden todas las encuestas de los individuos que no contestaron dicha pregunta. Algunos autores estiman los datos faltantes también mediante regresión (del resto de información de la base de datos o incluso de información externa) Hacer esto para tener más datos puede afectar a los supuestos del modelo y, por lo tanto, a la confiabilidad de los resultados por lo que es una cuestión que el investigador debe ponderar.

En la base de datos satisfacción podemos estar interesados en los condicionantes del voto. Podemos estimar la regresión de edad, nivel de estudios, estado de salud y ocupación sobre si votó en las últimas elecciones generales. El resultado es:

```
. regress voto edad est_recode estasalud ocup_1d

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(4, 7356)      =    144.12
       Model |  107.779478         4  26.9448695   Prob > F        =    0.0000
```

```
    Residual |  1375.28233     7,356  .186960622   R-squared       =    0.0727
-------------+----------------------------------   Adj R-squared   =    0.0722
       Total |  1483.06181     7,360  .201502964   Root MSE        =    .43239

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0072301   .0003331    21.70   0.000     .0065771    .0078831
  est_recode |   .0351379   .0056993     6.17   0.000     .0239656    .0463102
   estasalud |   -.018707   .0070907    -2.64   0.008    -.0326067   -.0048072
     ocup_1d |  -.0131749   .0023377    -5.64   0.000    -.0177576   -.0085923
       _cons |   .4119321   .0331784    12.42   0.000      .346893    .4769712
------------------------------------------------------------------------------
```

Todas las variables son significativas y el número de observaciones válidas es de 7361. Si ahora ampliamos el número de variables con la autovaloración del individuo entre izquierdas y derechas el resultado es:

```
. regress voto edad est_recode estasalud ocup_1d izq_der

      Source |       SS           df       MS      Number of obs   =     5,165
-------------+----------------------------------   F(5, 5159)      =     64.90
       Model |  47.2833933         5  9.45667867   Prob > F        =    0.0000
    Residual |  751.724545     5,159   .14571129   R-squared       =    0.0592
-------------+----------------------------------   Adj R-squared   =    0.0583
       Total |  799.007938     5,164  .154726557   Root MSE        =    .38172

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |    .005927   .0003614    16.40   0.000     .0052185    .0066356
  est_recode |   .0144313   .0060582     2.38   0.017     .0025546    .0263079
   estasalud |  -.0163668   .0075913    -2.16   0.031     -.031249   -.0014847
     ocup_1d |  -.0083462   .0024622    -3.39   0.001    -.0131732   -.0035192
     izq_der |    .002176   .0029235     0.74   0.457    -.0035552    .0079073
       _cons |    .563934   .0383841    14.69   0.000     .4886849    .6391831
------------------------------------------------------------------------------
```

O sea que se reduce el número de observaciones a 5165 y, en realidad la variable no es ni tan siquiera significativa luego descartarla podría ser una opción correcta.

Valores improbables o imposibles.

En ocasiones hay individuos con valores extraordinarios (altura superior a 2.5m; rentas brutalmente inmorales, …) otras veces sucede que se ha producido un error de transcripción (donde dije 20 pensando en otra cosa, escribí 200); también pueden ocurrir errores informáticos (yo sufro un problema con mi PC con la importación de bases de datos porque lo tengo configurado para usar el "." como símbolo decimal y puede ser que interprete 158.200 como 158.2) En estos casos es necesario repasarlos (En stata un

sumarize suele ser suficiente). En caso de duda con los datos es necesario rastrear y evaluar individuo por individuo.

```
. sum edad

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        edad |      7,747    47.63986    17.86102         18         97
```

En este caso todos los parámetros parecen correctos. Sería sospechoso, por ejemplo encontrar que el máximo es superior a 200 años (si es que hablamos de humanos no bíblicos) o que aparezcan edades por debajo de 18 años que es el límite inferior de la encuesta.

Un outlayer (un valor improbable o imposible) con una base de datos con pocos grados de libertad (poca diferencia entre el número de observaciones y variables) puede confundir mucho las estimaciones.

Creación de variables dummies

Una variable dummy es una variable dicotómica (0; 1) que hemos construido expresamente con algún propósito informativo. Por ejemplo y aunque sólo sea un tema de terminología, el género de un individuo no es una dummy sino una variable discreta dicotómica (con dos opciones), para ser una dummy tiene que ser un artificio construido con algún propósito interpretativo.

El propósito más usual para construir dummies es el de obtener información sobre la influencia de cada escalón de una variable factor ordinal o de un factor puro. Un factor ordinal se puede introducir como tal en un modelo (un factor puro no se puede de ninguna manera no porque el programa no pueda estimarlo sino porque no tiene interpretación alguna) pero su interpretación será más potente si lo dummyficamos (no existe esa palabra, creo, pero la utilizaremos de todas formas).

Cuando se introducen factores dummyficados hay que tener en cuenta que hemos de excluir una dummy del análisis, de lo contrario el mismo programa la excluye porque estaría perfectamente colineada con el resto. (si *d1*, *d2* y *d3* son las dummies de *d* cualquiera de ellas se puede expresar como combinación de las otras 2. Por ejemplo $d1_j=1- d2_j- d3_j$). Y también hay que tener en cuenta que la interpretación de los $\hat{\beta}$ serían cuanto cambia la variable dependiente cuando el individuo pasa de la variable excluida a la dummy estimada.

Esta interpretación provoca que en el caso de los factores dummyficados es posible que haya variables dummies de las que se obtengan estimaciones significativas y otras que no. En el caso de ser variables normales se dice que suele ser prudente excluir a las variables no significativas pero en el caso de las dummies hemos de incluir todo el paquete siempre que, al menos una dummy, sea significativa.

Si con la base "satisfacción" estimamos el modelo de edad y est_recode sobre estasalud los resultados muestran que ambas variables están correlacionadas. La edad está positivamente correlacionada con estasalud y el nivel de estudios lo está inversamente. Dada la codificación de edad implica que un año más de edad implica 0.013 peor estasalud (que está codificada de 1 a 5) y cada grado más en estudios implica una -0.10 mejor estasalud. Pero ¿Qué es un grado más en estudios? ¿es igual pasar de primaria a secundaria que de grado a posgrado? Para ello creamos dummies (también aprovecharemos para crear una variable edad respecto a su promedio y así aumentar la potencia de la información)

```
. regress estasalud edad est_recode

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(2, 7358)      =    644.34
       Model |  651.950926         2  325.975463   Prob > F        =    0.0000
    Residual |  3722.42891     7,358  .505902272   R-squared       =    0.1490
-------------+----------------------------------   Adj R-squared   =    0.1488
       Total |  4374.37984     7,360  .594345087   Root MSE        =    .71127

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0134507   .0005175    25.99   0.000     .0124362    .0144652
  est_recode |  -.1032535   .0081198   -12.72   0.000    -.1191706   -.0873364
       _cons |   1.716913   .0369765    46.43   0.000     1.644429    1.789398
------------------------------------------------------------------------------

. generate edad_recode = edad - 47
(3 missing values generated)
```

```
. tabulate est_recode, generate (est_)

 est_recode |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        166        2.25        2.25
          1 |      1,681       22.77       25.02
          2 |      2,438       33.02       58.04
          3 |      1,758       23.81       81.85
          4 |      1,249       16.92       98.77
          5 |         91        1.23      100.00
------------+-----------------------------------
      Total |      7,383      100.00

. regress estasalud edad_recode est_1 est_2 est_4 est_5 est_6

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(6, 7354)      =    215.26
       Model |  653.490243         6   108.91504   Prob > F        =    0.0000
    Residual |   3720.8896     7,354  .505968126   R-squared       =    0.1494
-------------+----------------------------------   Adj R-squared   =    0.1487
       Total |  4374.37984     7,360  .594345087   Root MSE        =    .71131

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 edad_recode |   .0133599   .0005612    23.80   0.000     .0122597    .0144601
       est_1 |   .2990886   .0593292     5.04   0.000     .1827863    .4153909
       est_2 |   .1063492   .0251728     4.22   0.000     .0570032    .1556952
       est_4 |  -.0853054   .0223086    -3.82   0.000    -.1290368   -.0415741
       est_5 |  -.1985412    .024811    -8.00   0.000    -.2471778   -.1499045
       est_6 |  -.3067945   .0759593    -4.04   0.000    -.4556965   -.1578925
       _cons |   2.134103   .0147502   144.68   0.000     2.105189    2.163018
------------------------------------------------------------------------------
```

El resultado final es ahora mucho más interesante, muestra el efecto de la edad (que es igual que antes de 0.103 por año) y el de cada grado. Pero la especial configuración del factor "estudios" hace que podamos afirmar que el nivel de estudios 1 (analfabeto) incrementa la mala salud en 0.30 respecto a los de secundaria (la dummy excluida). También los que solo tienen estudios de primaria tienen peor salud (0.11) que los que tienen secundaria. En cambio los que disponen de un nivel de estudios superior a primaria tienen un mejor autovaloración del estado de salud en -0.09, -0.20, -0.31 si tienen estudios de bachiller, grado o posgrado respectivamente. Además la especial configuración de las variables hace que, cuando todas sean 0 tenemos a un individuo de 47 años con estudios de secundaria. Ese sería nuestro individuo de referencia y la esperanza de su autovaloración del estado de salud es de 2.13

La "linealización"

El modelo lineal exige que la relación entre dependiente e independientes sea lineal. Sin embargo, en ocasiones observamos fenómenos que no tienen este carácter pero que

pueden linealizarse (probablemente esa palabra tampoco exista) con relativa facilidad. Los procedimientos más usuales para linearizar variables son:

En el caso de factores ordinales: Creación de dummies de forma que aunque la relación del factor con la dependiente no sea lineal, las de cada dummy con la dependiente sí lo será por construcción porque sólo hay un escalón entre la referencia y cada dummy (Aunque probablemente nunca nos haga falta, llamaremos a esta operación la dummyficación linearizadora)

En el caso de variables cuantitativas (sean números Reales, Enteros o Naturales) las operaciones más usuales son: tomar logaritmos o crear polinomios.

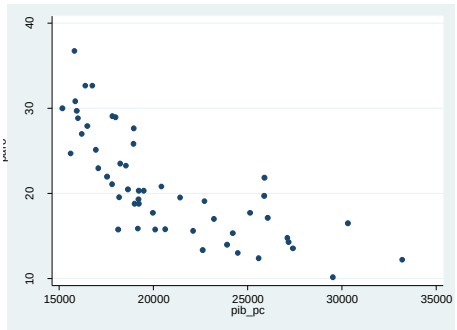
Se suelen tomar logaritmos cuando se dispone de datos cuantificados en unidades monetarias, sobre todo si son datos de diversos momentos del tiempo, pero también si son cortes transversales. El motivo es que los datos en unidades monetarias tienden a la acumulación en el tiempo (la inflación, la producción, los salarios…) y pueden llevar a crecimientos que parecen exponenciales cuando son lineales.

También puede ser recomendable tomar logaritmos cuando la variabilidad de la variable sea muy alta (por ejemplo se introduzcan en la misma base de datos población de países como EEUU o China y otros como Andorra o Liechtenstein). Los logaritmos homogenizan un poco la base de datos y hace que sus estimaciones sean más robustas.

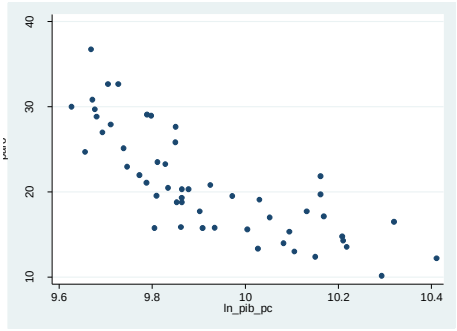
Cuando se trabaja con la variable transformada en logaritmos (por su facilidad matemática se utiliza, usualmente, el logaritmo natural) lo único que hay que hacer es tener cuidado con la interpretación de los $\hat{\beta}$ ya que no se corresponderá con el incremento de *y* ante una unidad de *x* sino ante una unidad de *ln(x)*. Pero esto no es un inconveniente sino incluso una virtud ya que, dadas las propiedades de los logaritmos, pueden interpretarse en el sentido de elasticidades (si se regresa *ln(x)* sobre *ln(y)*) como el cambio porcentual en *y* cuando cambia un *1% x*, o en el sentido de semielasticidades (si se regresa *ln(x)* sobre *y*) como el cambio en *y* de un cambio de un *1%* en *x*.

En la base "provincias", supongamos que queremos conocer el cambio en el paro (paro) en función del PIB per cápita provincial (pib_pc). Una primera regresión, directa entre muestra que dicha influencia es significativa pero el gráfico parece indicar cierta no linealidad. Si se construye la variable ln_pib_pc = ln(pib_pc) y se vuelve a correr la regresión la significación y el ajuste aumentan. El gráfico muestra que se ha corregido algo la no linealidad

```
. regress paro pib_pc

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     67.16
       Model |  1149.98372         1  1149.98372   Prob > F        =    0.0000
    Residual |   856.20034        50  17.1240068   R-squared       =    0.5732
-------------+----------------------------------   Adj R-squared   =    0.5647
       Total |  2006.18406        51  39.3369423   Root MSE        =    4.1381

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pib_pc |  -.0010924   .0001333    -8.19   0.000    -.0013601   -.0008246
       _cons |   43.58039   2.826607    15.42   0.000     37.90298    49.2578
------------------------------------------------------------------------------

. twoway (scatter paro pib_pc)
```




```
. generate ln_pib_pc = ln(pib_pc)

. regress paro ln_pib_pc

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     82.52
       Model |  1249.25134         1  1249.25134   Prob > F        =    0.0000
    Residual |  756.932716        50  15.1386543   R-squared       =    0.6227
-------------+----------------------------------   Adj R-squared   =    0.6152
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.8908

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ln_pib_pc |  -24.94791   2.746331    -9.08   0.000    -30.46407   -19.43174
       _cons |   268.4069   27.25166     9.85   0.000     213.6704    323.1435
------------------------------------------------------------------------------

. twoway (scatter paro ln_pib_pc)
```



La construcción de polinomios sigue una filosofía similar. Se trata de introducir un polinomio de la variable en lugar de la variable misma. De forma que se consiga la linealidad de la relación. Es decir en lugar de introducir la variable *x* o además de introducir la variable *x* como regresor introducimos x^2; x^3, etc.

En Stata hay varias formas de tratar con variantes polinomiales de una variable. Las tres más usuales son: creación de variables; introducción de polinomios; y regresión polinomial. La primera consiste en crear manualmente potencias de la variable original y en introducirlas en el modelo como regresores (normalmente basta con incluir una potencia cuadrática). La segunda consiste en dar directamente sobre la regresión la orden para que Stata haga las dos cosas automáticamente. La ventaja de esta segunda opción es que podemos calcular los efectos marginales también automáticamente (aunque en realidad tampoco es difícil hacerlo manualmente). La tercera opción consiste en utilizar la regresión polinomial que consiste en dejar que sea Stata quién busque además las potencias más adecuadas para el polinomio a construir.

Con la base de datos "provincias" estudiamos la relación entre el desempleo y el pib_pc de la provincia. La regresión inicial ya es buena pero queremos mejorar la liearidad con potencias de pib_pc. Para ello, en la primera alternativa creamos pib_pc y tras introducirla se observa una mejora el ajuste. En la segunda alternativa damos directamente la orden para que Stata cree e introduzca la variable. El resultado es igual al anterior.

```
. regress paro pib_pc

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     67.16
       Model |  1149.98372         1  1149.98372   Prob > F        =    0.0000
```

```
    Residual |   856.20034        50  17.1240068   R-squared       =    0.5732
-------------+----------------------------------   Adj R-squared   =    0.5647
       Total |  2006.18406        51  39.3369423   Root MSE        =    4.1381

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pib_pc |  -.0010924   .0001333    -8.19   0.000    -.0013601   -.0008246
       _cons |   43.58039   2.826607    15.42   0.000     37.90298     49.2578
------------------------------------------------------------------------------

. generate pib_pc2 = pib_pc^2

. regress paro pib_pc pib_pc2

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     52.48
       Model |  1367.68608         2  683.843039   Prob > F        =    0.0000
    Residual |  638.497977        49   13.030571   R-squared       =    0.6817
-------------+----------------------------------   Adj R-squared   =    0.6687
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.6098

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pib_pc |   -.005638   .0011182    -5.04   0.000     -.007885   -.0033909
     pib_pc2 |   1.00e-07   2.45e-08     4.09   0.000     5.10e-08    1.50e-07
       _cons |   92.85591   12.30497     7.55   0.000     68.12814    117.5837
------------------------------------------------------------------------------

. regress paro c.pib_pc##c.pib_pc

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     52.48
       Model |  1367.68603         2  683.843013   Prob > F        =    0.0000
    Residual |  638.498029        49   13.030572   R-squared       =    0.6817
-------------+----------------------------------   Adj R-squared   =    0.6687
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.6098

------------------------------------------------------------------------------
             paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
           pib_pc |  -.0056379   .0011182    -5.04   0.000     -.007885   -.0033909
                  |
c.pib_pc#c.pib_pc |   1.00e-07   2.45e-08     4.09   0.000     5.10e-08    1.50e-07
                  |
            _cons |    92.8559   12.30497     7.55   0.000     68.12814    117.5837
------------------------------------------------------------------------------
```

La ventaja de la segunda opción es que nos permite calcular el efecto marginal y la predicción puntual con sus intervalos de confianza también automáticamente. Por ejemplo el efecto marginal y la E(paro) cuando el PIB pc de la provincia es de 20.500€ (el promedio en la base es 20.763€ pero elegimos 20.500€ por redondear un poco)

```
. margins, dydx(pib_pc)

Average marginal effects                        Number of obs     =         52
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : pib_pc

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pib_pc |  -.0014722   .0001489    -9.89   0.000    -.0017714   -.0011731
------------------------------------------------------------------------------
```

```
. margins, at(pib_pc=(20500))

Adjusted predictions                              Number of obs     =         52
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : pib_pc          =       20500

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    19.4347   .6597508    29.46   0.000     18.10888    20.76052
------------------------------------------------------------------------------
```

Aunque los valores de los parámetros se pueden calcular también fácilmente mediante sencillas operaciones. En nuestro caso $y = \beta_o + \beta_2 x + \beta_2 x^2$

Efecto marginal $= \frac{\partial y}{\partial x} = \beta_o + 2\beta_o x$ *= -0.0056+2·1e-7·20500 = - 0.0015*

Esperanza media = 92.86 – 0.056·20500 + 1e-7·20500^2*= 19.5*

Pero el ordenador también nos ofrece significación e intervalos de confianza que nosotros no podemos calcular tan fácilmente.

Regresión de polinomios fraccionales

La tercera opción es utilizar una nueva regresión que nos ofrece Stata que denomina regresión de polinomios fraccionales. Esta opción consiste en que es Stata quien, también automáticamente calcula los polinomios que mejor ajuste consiguen. En nuestro caso le indicamos al programa que busque hasta dos polinomios. El software busca entre 44 modelos y encuentra que los dos mejores candidatos son x^{-2} y, por otro lado $x^{-2}+ x^{-1}$. Una vez hallados crea las correspondientes variables y ejecuta la regresión. Sin embargo cuando ejecuta la regresión lineal el segundo polinomio no es significativo (es decir el que nosotros digamos que queremos dos polinomios y que haya encontrado los dos mejores no quiere decir que haya dos significativos). Por ello repetimos el análisis con sólo un polinomio. El ajuste es algo mejor que en los modelos anteriores, pasando la $\bar{R}^2$ del 66.87% al 68.42%. El resultado obviamente es el mismo que si manualmente hubiésemos creado la variable X-2 y la hubiésemos introducido

manualmente en el modelo pero este procedimiento nos muestra el mejor ajuste entre bastantes potencias.

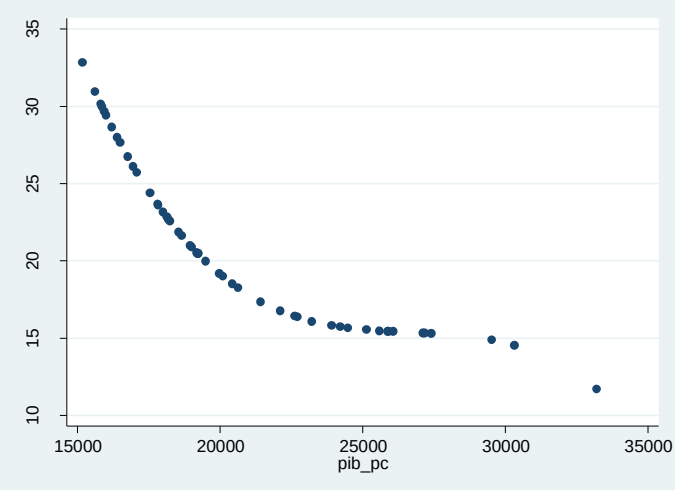
```
. fp <pib_pc>, replace : regress paro <pib_pc>
(fitting 44 models)
(....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%)

Fractional polynomial comparisons:
------------------------------------------------------------------------------
      pib_pc |   df    Deviance  Res. s.d.   Dev. dif.   P(*)   Powers
-------------+----------------------------------------------------------------
     omitted |    0    337.512      6.272     64.619    0.000
      linear |    1    293.235      4.138     20.342    0.000   1
       m = 1 |    2    276.544      3.525      3.650    0.192   -2
       m = 2 |    4    272.894      3.438      0.000       --   -2 -1
------------------------------------------------------------------------------
(*) P = sig. level of model with m = 2 based on F with 47 denominator dof.

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     60.39
       Model |  1427.17364         2  713.586819   Prob > F        =    0.0000
    Residual |  579.010418        49  11.8165391   R-squared       =    0.7114
-------------+----------------------------------   Adj R-squared   =    0.6996
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.4375

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pib_pc_1 |   1.62e+10   5.62e+09     2.89   0.006     4.93e+09    2.75e+10
    pib_pc_2 |   -1039971   550928.4    -1.89   0.065     -2147104    67160.66
       _cons |   30.91398   13.15244     2.35   0.023     4.483169    57.34479
------------------------------------------------------------------------------

. fp <pib_pc>, fp(-2) replace : regress paro <pib_pc>
-> regress paro pib_pc_1

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =    111.50
       Model |  1385.06775         1  1385.06775   Prob > F        =    0.0000
    Residual |  621.116301        50   12.422326   R-squared       =    0.6904
-------------+----------------------------------   Adj R-squared   =    0.6842
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.5245

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pib_pc_1 |   5.66e+09   5.36e+08    10.56   0.000     4.58e+09    6.73e+09
       _cons |   6.234912   1.472243     4.23   0.000     3.277824    9.191999
------------------------------------------------------------------------------
```

Si pedimos al programa un ajuste con 3 polinomios el programa propone: x^{-2}; $x^{-2}+ x^{-1}$; y $x^2+\ln(x)x^2+\ln(x)^2x^2$ los tres resultan significativos y el ajuste sube ($\overline{R}^2$ = 70.13%) lo que ahora surge es un problema de interpretación de dichos coeficientes. Para intentar intuir algo razonable estimamos los valores previstos por el modelo y los graficamos en función del pib_pc.

```
. fp <pib_pc>, dimension(3) replace : regress paro <pib_pc>
(fitting 164 models)
(....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%)

Fractional polynomial comparisons:
------------------------------------------------------------------------------
      pib_pc |   df    Deviance  Res. s.d.   Dev. dif.   P(*)   Powers
-------------+----------------------------------------------------------------
```

```
     omitted |     0    337.512       6.272     65.977    0.000
      linear |     1    293.235       4.138     21.699    0.001   1
       m = 1 |     2    276.544       3.525      5.008    0.214   -2
       m = 2 |     4    272.894       3.438      1.358    0.276   -2 -1
       m = 3 |     5    271.536       3.428      0.000       --   2 2 2
-------------------------------------------------------------------------------
(*) P = sig. level of model with m = 3 based on F with 46 denominator dof.

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(3, 48)        =     40.90
       Model |  1442.09809         3  480.699363   Prob > F        =    0.0000
    Residual |  564.085965        48  11.7517909   R-squared       =    0.7188
-------------+----------------------------------   Adj R-squared   =    0.7013
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.4281

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pib_pc_1 |  -.0000511   .0000171    -2.98   0.004    -.0000855   -.0000167
    pib_pc_2 |   9.54e-06   3.24e-06     2.95   0.005     3.03e-06     .000016
    pib_pc_3 |  -4.46e-07   1.53e-07    -2.91   0.005    -7.54e-07   -1.38e-07
       _cons |    175.078   39.91103     4.39   0.000     94.83153    255.3245
------------------------------------------------------------------------------
```

```
. fp predict paro_est, fp

. twoway (scatter paro_est pib_pc)
```



Regresión Cox-Box

Otro modelo que se utiliza cuando la relación no es lineal es el modelo Cox-Box. Este utiliza una transformación de las variables para conseguir la linealización concretamente se construye una batería de nuevas variables que dependen de λ que denotamos como Z(λ) y cuyo valor es

$$Z(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & si\ \lambda \neq 0 \\ \ln(y) & si\ \lambda = 0 \end{cases}$$

El software permite hacer la transformación Cox-Box sólo a la variable dependiente, sólo a la variables independientes, a todas (ambos lados) con el λ igual, o a todas (ambos lados) con parámetros diferentes λ para las independientes y θ para la dependiente. Dicen los defensores que este modelo reduce al máximo la posible heterogeneidad y no normalidad de los residuos.

Una ventaja respecto a la regresión fraccional es que en la regresión Cox-Box se pueden modificar varias variables independientes mientras que el modelo fraccional las debe estimar una a una.

Con la base "satisfacción" estudiamos la relación de edad y estudios sobre la renta familiar. Para que no exista confusión con edades en las que el individuo no está emancipado truncaremos el modelo sobre 30 años. Estimamos dos modelos: uno sencillo lineal, para comparación, y otro Cox-Box. En el caso del modelo Cox-Box hemos debido generar manualmente la variable edad2 como el cuadrado de la edad. Además hemos debido truncar la variable est_recode a un número mayor que cero (la Cox-Box, como calcula ln(y) y eleva y^{λ} no admite valores 0). Otra opción hubiese sido recodificar esta variable sumándole una unidad).

De los distintos modelos posibles se ha optado por el que realiza la transformación a ambos lados de la regresión estimando los dos parámetros λ para las independientes y θ para la dependiente.

```
. regress rentafam c.edad##c.edad est_recode if edad >30

      Source |       SS           df       MS      Number of obs   =     4,411
-------------+----------------------------------   F(3, 4407)      =    436.40
       Model |  2785.93774         3  928.645913   Prob > F        =    0.0000
    Residual |  9378.04412     4,407  2.12798823   R-squared       =    0.2290
-------------+----------------------------------   Adj R-squared   =    0.2285
       Total |  12163.9819     4,410  2.75827253   Root MSE        =    1.4588

------------------------------------------------------------------------------
     rentafam |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         edad |   .0580353   .0111453     5.21   0.000     .0361849    .0798857
              |
c.edad#c.edad |  -.0005091   .0000993    -5.13   0.000    -.0007037   -.0003145
              |
   est_recode |   .6912437   .0217104    31.84   0.000     .6486804    .7338071
        _cons |   1.379555   .3021103     4.57   0.000     .7872674    1.971843
------------------------------------------------------------------------------

. generate edad2 = (edad^2)
```

```
(3 missing values generated)

. boxcox rentafam edad edad2 est_recode if est_recode>0 & edad>30, model(theta)
Fitting comparison model

Iteration 0:   log likelihood = -8245.0196
Iteration 1:   log likelihood = -8190.6199
Iteration 2:   log likelihood = -8190.3497
Iteration 3:   log likelihood = -8190.3497

Fitting full model

Iteration 0:   log likelihood = -7715.0722
Iteration 1:   log likelihood = -7676.0376
Iteration 2:   log likelihood = -7675.9149
Iteration 3:   log likelihood = -7675.9149

                                                Number of obs   =      4,281
                                                LR chi2(4)      =    1028.87
Log likelihood = -7675.9149                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
    rentafam |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /lambda |   1.641999   .1449672    11.33   0.000     1.357868    1.926129
      /theta |   .7816837   .0298325    26.20   0.000     .7232131    .8401543
------------------------------------------------------------------------------

Estimates of scale-variant parameters
------------------------------
             |      Coef.
-------------+--------------
Notrans      |
       _cons |   1.809667
-------------+--------------
Trans        |
        edad |    .002309
       edad2 |  -1.53e-06
  est_recode |   .2761177
-------------+--------------
      /sigma |   1.063921
------------------------------

-------------------------------------------------------------------
   Test               Restricted
    H0:             log likelihood       chi2       Prob > chi2
-------------------------------------------------------------------
theta=lambda = -1     -9869.7622      4387.69           0.000
theta=lambda =  0     -8099.2734       846.72           0.000
theta=lambda =  1     -7715.0722        78.31           0.000
-------------------------------------------------------------------
```

El resultado de la Regresión Cox-Box muestra la estimación más ajustada para λ y θ, así como los parámetros de las independientes para dichos valores. Los valores, aunque difieren un poco no cambian el signo de la regresión lineal. Como puede observarse el modelo Cox-Box no reporta errores estándar luego no se puede hacer inferencia. El investigador debe ponderar los pros y los contras.

Configurar el individuo de referencia

Esta operación no mejora el ajuste ni la significación ni la magnitud de los coeficientes estimados pero es muy útil para que los resultados sean más inteligibles y

la potencia de su interpretación mayor. Consiste en preparar todas las variables que estemos interesados en introducir en el modelo rescaldándolas mediante combinaciones lineales, mediante estandarización o mediante creación de dummies para que cuando sean 0 coincidan con un individuo que, como mínimo exista y mejor si coincide con alguna característica medianamente interesante como que sea un individuo promedio de la población.

Si se hacen estas operaciones entonces, además, la probabilidad condicionada de la variable y quedará recogida en la constante del modelo (*E(y|x) = cte*) con lo que ganamos un parámetro estimado interesante más.

Por ejemplo en la base "provincias" intentaremos obtener evidencia sobre los condicionantes del paro provincial. Para ello regresamos pib_pc y nacim1000 sobre paro. Los resultados muestran que ambas variables son significativas y una estimación de su influencia. El ajuste es muy bueno ($\bar{R}^2$ = 67.73%). Sin embargo la constante (34.69) no nos dice nada. Si queremos obligar a hablar a la constante lo que se puede hacer es configurar un individuo de referencia (por ejemplo una provincia con PIB per cápita de 20500 y con nacimientos promedio de 9 nacimientos por 1000 habitantes (valores ambos cercanos a los promedios de 20763 y 8.86 respectivamente) tendría una tasa de paro de 21.32%. Con este cambio el resto de parámetros (ajuste, estimaciones, pvalor…) no cambian en absoluto.

```
. regress paro pib_pc nacim1000

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     54.52
       Model |   1384.1945         2  692.097249   Prob > F        =    0.0000
    Residual |  621.989557        49  12.6936644   R-squared       =    0.6900
-------------+----------------------------------   Adj R-squared   =    0.6773
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.5628

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pib_pc |  -.0010918   .0001148    -9.51   0.000    -.0013224   -.0008612
   nacim1000 |     1.0018   .2332229     4.30   0.000     .5331213    1.470479
       _cons |   34.69033   3.194685    10.86   0.000     28.27037    41.11029
------------------------------------------------------------------------------

. generate pib_pc_prom = pib_pc - 20500

. generate nacim1000_prom = nacim1000 - 9

. regress paro pib_pc_prom nacim1000_prom

      Source |       SS           df       MS      Number of obs   =        52
```

```
-------------+----------------------------------   F(2, 49)        =     54.52
       Model |   1384.1945         2  692.097249   Prob > F        =    0.0000
    Residual |  621.989557        49  12.6936644   R-squared       =    0.6900
-------------+----------------------------------   Adj R-squared   =    0.6773
       Total |  2006.18406        51  39.3369423   Root MSE        =    3.5628

------------------------------------------------------------------------------
          paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
   pib_pc_prom |  -.0010918   .0001148    -9.51   0.000    -.0013224   -.0008612
nacim1000_prom |     1.0018   .2332229     4.30   0.000     .5331213    1.470479
         _cons |   21.32477   .4960391    42.99   0.000     20.32795     22.3216
------------------------------------------------------------------------------
```

5. Resultados

Estamos observando que existen diversas formas de plantear los modelos de regresión. Si sólo existiese un método no habría discusión alguna sobre su formulación, no habría dudas sobre el conocimiento de las variables económicas y no habría cursos de econometría. Nuestro objetivo será entonces decir algo, si podemos, pero minimizando la posibilidad de equivocarnos.

Una persona puede equivocarse o mentir de dos formas: Diciendo que algo es cierto cuando no lo es (error tipo I) y diciendo que algo no es cierto cuando sí que lo es (error tipo II). En econometría, y en Ciencia (con mayúscula) en general tratamos de evitar al máximo posible (es decir minimizar) el error tipo I, de forma que sólo digamos que algo es cierto cuando estemos absolutamente seguros de que no es cierto lo contrario.

Denominamos α como la probabilidad de cometer un error tipo I es decir la probabilidad de decir que algo se cumple cuando no se cumple. Cuanta más pequeña sea α casi siempre será mejor porque estamos más seguros de lo que estamos diciendo. α también se llama nivel de significación (su complementaria: 1- α, que cuanto más grande mejor, se suele llamar nivel de confianza).

$$P(\text{aceptar algo} | \text{ algo es falso}) = \alpha$$

Escogemos minimizar α como medida del error, mejor que el error tipo II por dos razones: a) porque es más fácil de calcular estadísticamente y b) porque parece que es más importante estar más seguro de lo que se afirma que de lo que se rechaza. (¿Qué es más trágico, equivocarse al introducir un medicamento que no cura o equivocarse

rechazando un medicamento que sí cura? Las dos cosas son importantes: si uno minimiza el primer error significa que si digo que cura es que cura con una confianza 1-α (puede colarse α veces algún medicamento inútil). Si minimizo el segundo es que si digo que no cura es que no cura con cierta probabilidad (puede rechazarse un medicamento bueno). Luego parece, aunque es una cuestión discutible, que si hay que equivocarse, es mejor equivocarse en un sentido que en el contrario.

En la práctica para formular dicha probabilidad se plantea la realización de pruebas o test. Los test son un cálculo que hacemos con nuestros resultados, de la que se obtiene un resultado (parámetro) que, como sabemos que se distribuye según una determinada distribución (normal, t, F, etc.) podemos afirmar si el resultado de nuestra cuenta entra dentro de la zona de rechazo o de aceptación. Para equivocarnos lo menos posible ponemos unas zonas de aceptación muy pequeñas y una zonas de rechazo muy grandes (así cuando aceptemos algo es porque estamos bastante seguros) normalmente la zona de aceptación es del 1% o del 5% y su complementaria la de rechazo.

Para hacerlo todavía más útil casi todos los test se plantean como una aceptación o rechazo de que algo sea 0 (por eso se suele llamar hipótesis nula Ho=0). Por ejemplo que la recta de regresión estimada tenga tangente 0 (Ho: $\hat{\beta} = 0$), que dos parámetros sean iguales (Ho: $\widehat{\beta_1} - \widehat{\beta_2} = 0$), que una función sea normal (Ho: *F(Y|X)-N(Y|X)=0*); que la correlación entre un valor y él mismo retardado sea 0 (Ho: $corr(x_t; x_{t-n}) = 0$) y así sucesivamente. De hecho cuando un test no dice cuál es la Hipótesis nula es casi seguro que la Ho es que el parámetro estimado sea cero. Y al contrario si por algún capricho del destino una Ho no es que el parámetro sea cero (por ejemplo algún test de normalidad) entonces el programa nos lo recordará y nosotros tendremos que recordarlo en nuestro trabajo.

Tipos de errores

En una investigación estamos sometidos a errores constantes. Podemos equivocarnos en el tipo de muestreo (que no sea aleatorio, que no esté correctamente estratificado…), Podemos equivocarnos al medir las variables (porque los individuos mientan, porque los entrevistadores copien mal las respuestas, porque no se entiendan

las preguntas…). Pero supongamos que minimizamos dichos errores o que, al menos estamos razonablemente seguros de que el muestreo has sido correcto (o que tenemos a toda la población como en el caso de las variables macroeconómicas) y que los errores de medida se compensan unos con otros (es decir, el entrevistador torpe se equivoca con errores de suma cero). Ambos errores se reducen con la ampliación de la muestra, de forma que, si no estamos seguros de cualquiera de ellos bastaría con incrementar la muestra. Si las poblaciones son normales y grandes (más de 100.000 individuos), hay una relación cuadrática inversa entre el error *e* y el tamaño de la muestra necesaria *n* en la forma:

$$n = \frac{Z^2 p(1-p)}{e^2}$$

donde Z es el nivel de significación deseado (normalmente Z = 0.95 ó 0.99) y p la probabilidad esperada (en caso de duda p =0.5).

Pero los errores más difíciles de tratar son las perturbaciones estocásticas que sufre cualquier variable que se precie de ello. De hecho sólo estaremos razonablemente seguros de que nuestro modelo es razonablemente bueno cuando reduzcamos las perturbaciones a una masa informe de puntos. Dicho de otra forma, mientras las perturbaciones no explicadas por nuestro modelo tengan alguna forma definida puede significar que nuestro modelo adolece de alguna falta de variable explicativa o de algún defecto estructural que deberíamos corregir antes de presentar los resultados.

Para comprobarlo tenemos dos baterías de pruebas: gráficos y test numéricos.

En los test gráfico tendremos que observar si existe cierta esfericidad o si, por el contrario se aprecian tendencias entre ***cualquiera*** de las variables del modelo y los residuos. El problema de los test gráficos es que los residuos tienen unidades y, por lo tanto puede engañarnos la vista.

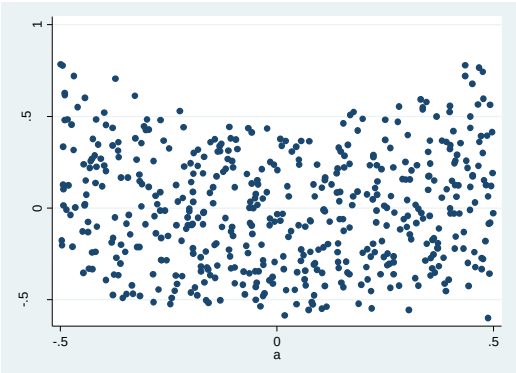
En los test numéricos trataremos de descartar la presencia de errores de especificación, de heterocedasticidad, etc.

Por ejemplo si abrimos la base "artificial" y estimamos una regresión de a y b sobre y, los resultados parecen buenos, tanto en significación como en ajuste. Sin embargo los gráficos parecen sugerir cierta correlación de los residuos con y y con a (no tanto en b). Cuando se produce correlación de los residuos con y puede ser un problema de especificación o de variables omitidas. Cuando se produce correlación entre los residuos y alguna variable independiente puede ser un problema también de especificación o variables omitidas pero también un problema de la temida ¡endogeneidad!.

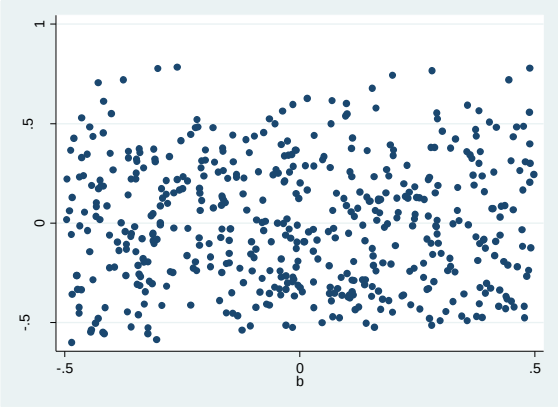
Además de la regresión le hemos solicitado a Stata que se reporten los coeficientes estandarizados (beta). Esto no es necesario pero por ver cuál de las dos variables tiene un mayor peso en la varianza de la variable dependiente.

```
. regress y a b, beta

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(2, 497)       =    401.41
       Model |  79.5929409         2  39.7964705   Prob > F        =    0.0000
    Residual |  49.2735182       497  .099141888   R-squared       =    0.6176
-------------+----------------------------------   Adj R-squared   =    0.6161
       Total |  128.866459       499  .258249417   Root MSE        =    .31487

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
           a |   1.038869   .0484181    21.46   0.000                .5952589
           b |   .9430685    .049771    18.95   0.000                .5256779
       _cons |   .0682061   .0141062     4.84   0.000                      .
------------------------------------------------------------------------------
```

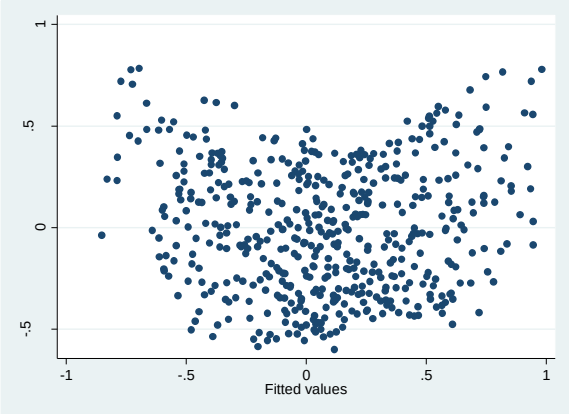
```
. rvpplot a
```




```
. rvpplot b
```




```
. rvfplot
```



Dado que no nos quedamos conformes practicamos estimamos los residuos y hacemos dos pruebas previas la de correlación con las variables del modelo y la de normalidad. Curiosamente no mide correlación entre los residuos y la variable a. Pero sí entre los residuos y la dependiente. La prueba de normalidad nos obliga a rechazar la hipótesis nula de normalidad (pero este test es muy exigente en realidad no se suele exigir porque no sale correcto casi nunca)

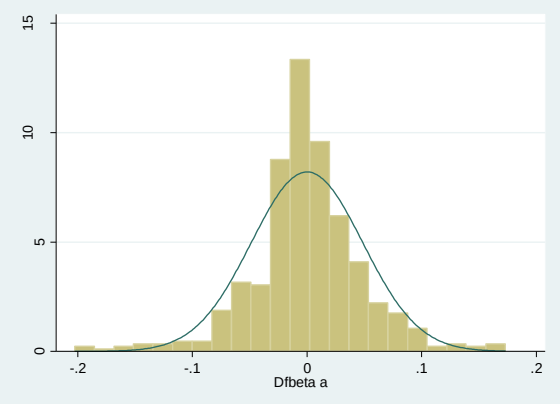
```
. predict residuos, residuals

. correlate residuos y a b
(obs=500)

             | residuos        y        a        b
-------------+------------------------------------
    residuos |   1.0000
           y |   0.6184   1.0000
           a |  -0.0000   0.5843   1.0000
           b |   0.0000   0.5133  -0.0208   1.0000


. sktest residuos

                    Skewness/Kurtosis tests for Normality
                                                      ------ joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
```

```
-------------+--------------------------------------------------------------
    residuos |        500     0.0571        0.0000       33.92         0.0000
```

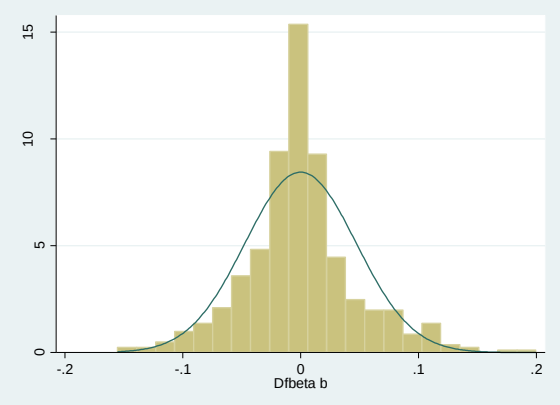
Como seguimos algo preocupados (incluso algo más si cabe) seguimos realizamos los tests correspondientes. Por orden el heterogeneidad (hetttest), el de multicolinealidad (vif), el de especificación o variables omitidas (ovtest; test de observaciones relevantes (DFbetas). No se detecta heterocedasticidad (¡bien!) porque se acepta la Ho. No se detectga multicolinealidad (¡bien!) porque ninguna variable tiene un VIF>10 (algunos autores hablan incluso de >30) y la media no es mucho más grande de 1. Atención porque falla el test de variables omitidas. Se rechaza la Ho de especificación correcta. El test dfbeta nos genera dos variables (una por independiente) y nos arroja, para cada observación, cuantas veces se reduciría el error si se omite dicha variable. Algunos autores dicen que hay que plantearse la exclusión cuando $|dfbeta|>2/\sqrt{n}$*, en nuestro caso como n=500 el límite estaría en 0.08. Otros autores dicen que sólo es preocupante cuando |dfbeta|>1. En nuestro caso no es preocupante porque, además, no hay valores raros.*

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      =     0.02
         Prob > chi2  =   0.8874

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
           a |      1.00    0.999566
           b |      1.00    0.999566
-------------+----------------------
    Mean VIF |      1.00


. estat ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                  F(3, 494) =     27.04
                   Prob > F =      0.0000



. dfbeta
                      _dfbeta_1: dfbeta(a)
                      _dfbeta_2: dfbeta(b)

. summarize _dfbeta_1 _dfbeta_2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   _dfbeta_1 |        500   -7.10e-06    .0486284  -.2026569   .1732518
   _dfbeta_2 |        500    9.83e-06    .0472394   -.155364   .1997676
```

```
. histogram _dfbeta_1, normal
(bin=22, start=-.20265689, width=.01708676)
```




```
. histogram _dfbeta_2, normal
(bin=22, start=-.15536402, width=.01614235)
```



La solución a los problemas de los residuos pasa por controlar la Heterocedasticidad, la Multicolinealidad y la especificación del modelo.

a) Heterocedasticidad.

Si no es muy acusada no es importante. Si es muy acusada será necesario utilizar estimadores robustos. Stata proporciona estimadores robustos de casi todos los modelos. Los estimadores robustos no cambian ni el ajuste ni los parámetros $\hat{\beta}$. Solo aumenta el error estándar estimado y, por tanto aumenta el p-valor por lo que corremos el riesgo de que una variable que fuese significativa deje de serlo.

Los problemas de heterocedasticidad tambien pueden ser debidos a la omisión de alguna variable relevante por lo que el uso de estimadores robustos solo se recomienda cuando el tests Reset de Ramsey (estat ovtest) no sea significativo, es decir no indique que hay variables omitidas.

b) Multicolinealidad.

La colinealidad no sólo es normal sino que es esperable y deseable. Es imposible que unas variables que explican y son explicadas por un fenómeno sean tan completamente independientes que no estén correlacionadas en algún grado. El problema surge cuando hay, como mínimo, dos variables muy, muy, muy correlacionadas, entonces sucede que una de ellas le "roba" la correlación al resto haciendo que las demás aparezcan como no significativas o incluso significativas con un signo distinto al esperado. Esto es normal, por ejemplo en el caso de la renta, la edad y el nivel educativo. Lo que hay que hacer en estos casos es sacrificar una de ellas y quedarnos con la variable que tenga más sentido interpretativo.

c) Error de especificación.

El error de especificación se refiere a que falta por incluir alguna interacción o alguna variable en forma polinómica. El test consiste en regresar a la variable dependiente con potencias de ella misma por lo que las variables omitidas deben ser potencias o interacciones de las variables dependientes. Lamentablemente el test no nos ofrece pistas sobre las variables díscolas por lo que se impone utilizar la lógica y como último recurso, claro está, la prueba y error.

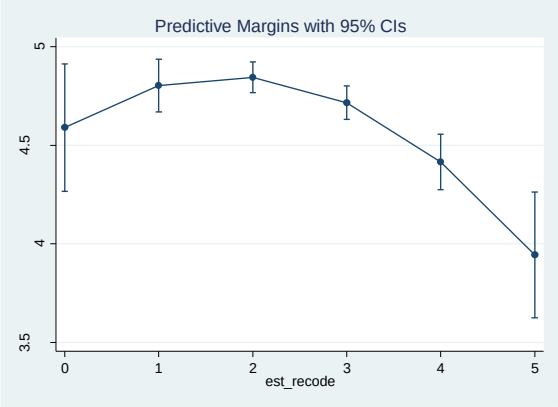
En la base de datos "satisfacción" vamos a estudiar los condicionantes del voto. La primera regresión, aunque el ajuste es bajo, muestra resultados relativamente significativos. Sin embargo los test de variables omitidas y de heterocedasticidad dicen que hay problemas. El test de inflación de la varianza también es correcto $VIF_i<30$ *y media* $VIF\approx 1$.

```
. regress izq_der edad est_recode estasalud ocup_1d nac_esp

      Source |       SS           df       MS      Number of obs   =     5,165
-------------+----------------------------------   F(5, 5159)      =     25.49
       Model |  420.853127         5  84.1706253   Prob > F        =    0.0000
    Residual |  17036.6524     5,159   3.3023168   R-squared       =    0.0241
-------------+----------------------------------   Adj R-squared   =    0.0232
       Total |  17457.5055     5,164  3.38061687   Root MSE        =    1.8172

------------------------------------------------------------------------------
     izq_der |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0125577   .0017199     7.30   0.000     .0091859    .0159295
  est_recode |  -.1496953   .0288776    -5.18   0.000    -.2063075    -.093083
   estasalud |  -.1088377   .0361265    -3.01   0.003    -.1796609   -.0380145
     ocup_1d |  -.0579368   .0117543    -4.93   0.000    -.0809802   -.0348934
     nac_esp |  -.1906452   .0989072    -1.93   0.054    -.3845452    .0032547
       _cons |   5.278402   .1952128    27.04   0.000     4.895702    5.661102
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of izq_der
       Ho:  model has no omitted variables
                F(3, 5156) =      4.24
                  Prob > F =      0.0053

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of izq_der

         chi2(1)      =     5.89
         Prob > chi2  =   0.0152

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
  est_recode |      1.60    0.625868
     ocup_1d |      1.35    0.738589
        edad |      1.33    0.749768
   estasalud |      1.18    0.848290
     nac_esp |      1.03    0.970659
-------------+----------------------
    Mean VIF |      1.30
```

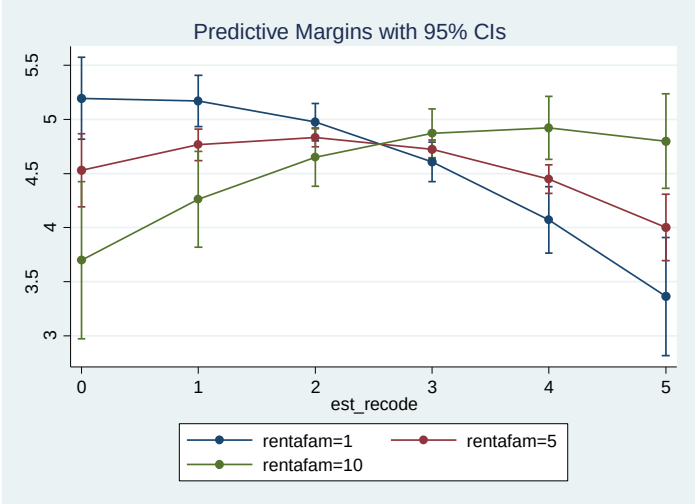
Tras un poco de estudio del modelo y de la teoría económica (y bastante de prueba y error) proponemos este otro modelo, cuyo ajuste continúa siendo bajo, pero ya estamos seguros de que no existe error de especificación. También desaparece la heterocedasticidad. Lo que aparece ahora es un pequeño problema de colinealidad entre renta y estudios. La multicolinealidad de los términos de interacción es normal (son el producto de dos variables en el modelo) y no es relevante. Lo difícil ahora es interpretar el sentido de variables como el nivel de estudios, que está en estado, cuadrático y como interacción con renta familiar. Pero como hemos utilizado los automatismos de Stata para introducir las variables entonces podemos pedirle a Stata que nos haga un análisis marginal que resulta interesante

```
. regress izq_der rentafam estasalud ocup_1d nac_esp c.rentafam#c.est_recode
c.est_recode##c.est_recode c.edad##c.edad
```

```
      Source |       SS           df       MS      Number of obs   =     3,971
-------------+----------------------------------   F(9, 3961)      =     13.77
       Model |   416.26945         9  46.2521612   Prob > F        =    0.0000
    Residual |  13301.5779     3,961  3.35813631   R-squared       =    0.0303
-------------+----------------------------------   Adj R-squared   =    0.0281
       Total |  13717.8474     3,970  3.45537718   Root MSE        =    1.8325

------------------------------------------------------------------------------
                 izq_der |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
                rentafam |  -.1663013   .0508508    -3.27   0.001    -.2659974   -.0666052
               estasalud |  -.1136832   .0414981    -2.74   0.006    -.1950428   -.0323236
                 ocup_1d |  -.0489874   .0142198    -3.45   0.001    -.0768663   -.0211085
                 nac_esp |  -.3267636   .1124485    -2.91   0.004    -.5472261   -.1063011
                         |
  c.rentafam#c.est_recode |   .0651922   .0174587     3.73   0.000     .0309633    .0994212
                         |
              est_recode |  -.0043596   .1282763    -0.03   0.973    -.2558534    .2471342
                         |
c.est_recode#c.est_recode |  -.0854879   .0254971    -3.35   0.001    -.1354766   -.0354992
                         |
                    edad |  -.0323012   .0100137    -3.23   0.001    -.0519336   -.0126689
                         |
           c.edad#c.edad |    .000453   .0000995     4.55   0.000     .0002579     .000648
                         |
                   _cons |   6.551325   .3545609    18.48   0.000     5.856186    7.246464
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of izq_der
       Ho:  model has no omitted variables
                 F(3, 3958) =      1.87
                   Prob > F =      0.1324

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of izq_der

         chi2(1)      =     2.71
         Prob > chi2  =   0.0995

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
    rentafam |      8.58    0.116525
   estasalud |      1.17    0.852384
     ocup_1d |      1.47    0.680204
     nac_esp |      1.06    0.946032
  c.rentafam#|
c.est_recode |     25.60    0.039057
  est_recode |     24.02    0.041629
          c. |
  est_recode#|
c.est_recode |     24.79    0.040338
        edad |     32.85    0.030442
      c.edad#|
      c.edad |     34.12    0.029305
-------------+----------------------
    Mean VIF |     17.07

. margins, at(est_recode=(0 1 2 3 4 5 )) plot

(output omitido)
```




```
. margins, at(est_recode=(0 1 2 3 4 5) rentafam=(1  5 10)) plot
(output omitido)
```



El resultado es interesante porque no sólo muestra una relación no lineal entre el nivel de estudios y la ideología sino que muestra una interacción curiosísima con la renta. De forma que parece que los que tienen más estudios se hacen bastante de izquierdas cuando no disponen de un nivel de renta familiar alto.

El ajuste del modelo

¿Cuánto de buena es tu idea del modelo? Preguntar eso es de evidente mal gusto, pero lo peor es que podemos medirlo fácilmente. Resulta que la variable dependiente tiene una variabilidad total (que al cuadrado llamamos suma del cuadrado Total -SCT), resulta que los residuos tienen otra variabilidad restante (que al cuadrado llamamos

suma del cuadrado de los residuos - SCR), luego tu modelo ajusta la diferencia. Se puede construir un índice que llamamos coeficiente de determinación como:

$$R^2 = 1 - \frac{SCR}{SCT}$$

Que es un valor que oscila entre 0 (pésimo ajuste) y 1 (máximo ajuste). Ojo un ajuste superior a 0.8 también debe preocuparnos porque igual nos hemos equivocado en algo como haber puesto variables proxi o combinaciones lineales de la dependiente como independientes.

Pero, por construcción, cada vez que añado una variable al modelo, aunque no ajuste nada, me reduce los grados de libertad, con lo que aumenta el ajuste (en última instancia un modelo con tantas variables dependientes como datos ajustaría perfectamente aunque las variables independientes no tengan nada que ver), por ello es mejor corregir dicho coeficiente con los grados de libertad. El coeficiente de determinación ajustado es:

$$\bar{R}^2 = 1 - \frac{SCR\ /(n-k)}{SCT/(n-1)}$$

Es decir hacemos el mismo cálculo pero ponderando por los grados de libertad con los que hemos medido cada uno de los componentes de variabilidad. Si *n* es el total de observaciones y *k* el de variables en el modelo *(n-k)* son los grados de libertad de los residuos y *(n-1)* los grados de libertad del modelo.

$\bar{R}^2$ también es un parámetro que oscila entre 0 (nulo ajuste) y 1 (ajuste sospechosamente perfecto).

Algunos autores consideran que el test de significación global F también es una medida del ajuste. Este test consiste en comprobar si se cumple la Hipótesis nula de que todas las estimaciones son cero, es decir Ho: $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_k = 0$. Si el parámetro F es significativamente distinto de 0 se rechaza la Ho y se puede afirmar que alguna $\hat{\beta}_j$ es distinta de 0 por lo que el modelo ajusta algo. Sin embargo pensar que sólo porque el p-

valor de F sea menor de 0.000 ya tenemos un modelo ajustado puede ser pecar de optimismo.

Otra medida del ajuste es la que se denomina raíz cuadrada de los residuos (error cuadrático medio). Esta medida es muy interesante que aprendamos a calcularla porque algunos modelos más complejos no ofrecen una estimación del R2 y esta medida de ajuste nos la podemos fabricar muy fácilmente. El ECM (Root MSE en Stata) se calcula como

$$ECM = \sqrt{\frac{\sum(y - \hat{y})^2}{(n - k)}}$$

En el caso de la Regresión lineal suele venir calculada pero es sencillo obtenerla como $ECM = \sqrt{MSR}$. Pero en el caso de la regresión lineal ni siquiera es necesaria porque ya disponemos de los coeficientes de determinación. La importancia de conocer esta medida de ajuste será cuando nos enfrentemos a modelos en los que falte una medida del ajuste.

ECM es una medida mayor que 0, no está acotada superiormente y no conocemos su distribución pero sí que nos sirve para comparar modelos siempre que compartan la misma variable dependiente, incluso aunque no estén anidados. Incluso, en algunos modelos relativamente lineales, podremos "fabricar" una pseudo R2, aunque sólo sea para consumo interno, como:

$$Pseudo\ R^2 = 1 - \frac{ECM_1}{ECM_0}$$

Donde ECM_1 es el *ECM* del modelo final y ECM_o es el *ECM* del modelo en que sólo está la constante. Decimos, para consumo interno, porque no conocemos la distribución de ECM por lo que, por ejemplo, no estamos seguros de que una reducción en el 50% del ECM se corresponda con un incremento del 50% en el ajuste, pero sí que sabemos que si alguna variable ajusta algo, normalmente $ECM_1 < ECM_o$ y que cuanto

más se parezcan ECM_1 y ECM_o, *más cercano será el índice a 0 y cuanto mejor ajuste más cercano será a 1.*

Interpretación de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$

Si *x* es continua, el valor de $\hat{\beta}_j$ es la derivada de y respecto a x_j ($\hat{\beta}_j = \frac{\partial y}{\partial x_j}$) Es decir cuánto cambia *y* cuando *x* cambia en una unidad permaneciendo el resto de variables constante.

El término *ceteris paribus* es un latinajo que utilizan los economistas para permitirse especular sobre lo que ocurriría si aislamos el efecto de una única variable permaneciendo el resto constante. Por ejemplo la producción es función de muchas variables como puede ser el capital invertido, el trabajo y el conocimiento, entonces como cambiaría la producción per cápita si ceteris paribus, sólo aumenta el número de trabajadores. Los economistas dicen que la producción total aumentará, pero menos que proporcionalmente por lo que, en términos de producción por trabajador, la producción disminuirá. Como los profesores de los economistas abusaban tanto de este término y era tan confuso para algunos de sus estudiantes, probablemente los estadísticos pensaran en importarlo para sus clases de regresiones de forma que los valores de los parámetros estimados serían ceteris paribus.

Con la base "satisfacción" supongamos que queremos medir el efecto de la edad en el estado de salud. Una primera regresión indica que cada año de edad incrementa la variable estasalud (0; 5) en 0.017 de promedio. Sin embargo es posible que consideremos que este resultado se queda algo corto porque podemos pensar que el estado de salud también puede depender del género o del nivel educativo. Una segunda regresión nos indica que así es y que, controlado por género o nivel educativo la edad sólo tiene un efecto sobre estasalud de 0.013 por año de incremento. Es decir que, ceteris paribus, un año más de vida incrementa, en promedio, la variable estasalud en 0.013.

```
. generate edad_recode =(edad- 47)
(3 missing values generated)

. regress estasalud edad_recode

      Source |       SS           df       MS      Number of obs   =     7,725
-------------+----------------------------------   F(1, 7723)      =   1309.49
       Model |  694.014765         1  694.014765   Prob > F        =    0.0000
    Residual |  4093.09669     7,723  .529987918   R-squared       =    0.1450
-------------+----------------------------------   Adj R-squared   =    0.1449
       Total |  4787.11146     7,724  .619771033   Root MSE        =      .728

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 edad_recode |    .016782   .0004638    36.19   0.000     .0158729    .0176911
       _cons |   2.120836   .0082883   255.88   0.000     2.104589    2.137083
------------------------------------------------------------------------------

. regress estasalud edad_recode genero est_recode

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(3, 7357)      =    443.01
       Model |  669.316596         3  223.105532   Prob > F        =    0.0000
    Residual |  3705.06324     7,357  .503610608   R-squared       =    0.1530
-------------+----------------------------------   Adj R-squared   =    0.1527
       Total |  4374.37984     7,360  .594345087   Root MSE        =    .70966

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 edad_recode |    .013337   .0005167    25.81   0.000     .0123241    .0143498
      genero |  -.0972383   .0165592    -5.87   0.000     -.129699   -.0647775
  est_recode |  -.1029902   .0081015   -12.71   0.000    -.1188715    -.087109
       _cons |    2.39657   .0221388   108.25   0.000     2.353172    2.439969
------------------------------------------------------------------------------
```

Además dada la configuración dada a edad, como edad_recode (= edad-47) hemos configurado un individuo de referencia mujer, analfabeta y con edad de 47 años. Y sabemos que, en promedio su estasalud será de 2.40. De no haber recodificado la edad su coeficiente y su error estándar sería el mismo, sólo cambiaría la estimación de la constante.

Los coeficientes estandarizados

Existe la posibilidad de pedir a Stata que estime los coeficientes estandarizados. Estos son los resultantes de una regresión en la que se han estandarizado todas las variables y, en este sentido su interpretación no es intuitiva pero, dado que las variables estandarizadas no tienen unidades de medida, sí que nos sirve para indicarnos qué variable son más influyentes que otras.

En el caso anterior, con la base "satisfacción" obtenemos el mismo modelo pero, a la derecha, en lugar de los intervalos de confianza aparecen los betas

estandarizados que nos indican que, de las tres variables la más influyente es la edad, después los estudios y después el género. Si observamos los coeficientes no estandarizados el mayor corresponde a los estudios, el segundo al género y el último a la edad pero estos coeficientes nos confunden porque dependen de la unidad de medida de la variable. Los coeficientes estandarizados salvan dicho problema aunque su interpretación es menos intuitiva. Una interpretación no exacta puede ser la siguiente: como se han estandarizado, el 99.7% de la distribución se ha comprimido entre [-3; +3] luego el coeficiente $\widehat{\beta^*}$ *sería el cambio en estasalud (que sabemos que está distribuida entre 1 muy buena y 5 muy mala) cuando cada variable recorre 1/6 de su rango. O lo que es lo mismo la edad (desde casi el mínimo de 18 a casi el máximo de 97 años) influye 0.3*6= 1.8 puntos en estasalud.*

```
. generate edad_recode =(edad- 47)
(3 missing values generated)

. regress estasalud edad_recode  genero est_recode, beta

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(3, 7357)      =    443.01
       Model |  669.316596         3  223.105532   Prob > F        =    0.0000
    Residual |  3705.06324     7,357  .503610608   R-squared       =    0.1530
-------------+----------------------------------   Adj R-squared   =    0.1527
       Total |  4374.37984     7,360  .594345087   Root MSE        =    .70966

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
 edad_recode |    .013337   .0005167    25.81   0.000                 .3007861
      genero |  -.0972383   .0165592    -5.87   0.000                -.0630661
  est_recode |  -.1029902   .0081015   -12.71   0.000                -.1480374
       _cons |    2.39657   .0221388   108.25   0.000                         .
------------------------------------------------------------------------------
```

Sistemas de selección de variables (stepwise)

La mayoría de los programas informáticos y Stata no es una excepción, permite a los investigadores introducir un número elevado de variables en el modelo y dejar al software que haga las iteraciones necesarias para seleccionar aquellas que tienen un p-valor máximo (stepwise forward) o para desechar aquellas que no tengan un p-valor mínimo) (stepwise backward). El programa realiza cientos de iteraciones y determina qué variables superan el requisito de p-valor exigido para quedarse en el modelo.

Esta herramienta es meramente exploratoria, no hay que decir, a estas alturas, que la introducción o la exclusión de una variable en un modelo de regresión sólo puede justificarse por la razón y la literatura (que no siempre coinciden) y nunca porque exista una correlación en la práctica porque esta puede resultar espuria.

El resultado es el mismo en ambos casos pero se pueden hacer dos formulaciones, "hacia delante" o "forward" que consiste en ir introduciendo variables conforme quedan como significativas en el modelo (en nuestro caso decimos que añada la variable mientras que su p-valor sea menor de 0.05).

```
. stepwise, pe(0.05) : regress satisf_1_10 frec_prim_publ frec_urg_publ frec_esp_pub
frec_hosp_pub izq_der voto genero edad rentafam est_recode estasalud cronico ocup_1d
nac_esp

(output omitido)
```

y "hacia detrás" o "backward" que consiste en meter a todas las varibles al principio y extrae aquellas que pierden la significación. (en nuestro caso decimos que retire la variable cuando su p-valor sea mayor de 0.05).

```
. stepwise, pr(0.05) : regress satisf_1_10 frec_prim_publ frec_urg_publ frec_esp_pub
frec_hosp_pub izq_der voto genero edad rentafam est_recode estasalud cronico ocup_1d
nac_esp

                     begin with full model
p = 0.7296 >= 0.0500  removing ocup_1d
p = 0.7143 >= 0.0500  removing frec_prim_publ
p = 0.5640 >= 0.0500  removing est_recode
p = 0.2215 >= 0.0500  removing voto
p = 0.0891 >= 0.0500  removing genero
p = 0.0544 >= 0.0500  removing cronico

      Source |       SS           df       MS      Number of obs   =     3,947
-------------+----------------------------------   F(8, 3938)      =     35.54
       Model |  1038.48285         8  129.810357   Prob > F        =    0.0000
    Residual |  14385.1472     3,938  3.65290687   R-squared       =    0.0673
-------------+----------------------------------   Adj R-squared   =    0.0654
       Total |  15423.6301     3,946  3.90867463   Root MSE        =    1.9113

------------------------------------------------------------------------------
  satisf_1_10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      nac_esp |  -.7635506   .1166087    -6.55   0.000    -.9921697   -.5349314
frec_urg_publ |  -.0731514   .0254253    -2.88   0.004    -.1229994   -.0233035
 frec_esp_pub |    .058812   .0199054     2.95   0.003     .0197863    .0978378
frec_hosp_pub |   .2083507   .0792349     2.63   0.009     .0530054    .3636959
      izq_der |   .0949765   .0165397     5.74   0.000     .0625494    .1274036
    estasalud |  -.3536168   .0452662    -7.81   0.000    -.4423641   -.2648694
     rentafam |   .0543324    .018843     2.88   0.004     .0173895    .0912754
        edad |   .0246365   .0020073    12.27   0.000     .0207011    .0285719
       _cons |    5.98724   .1920789    31.17   0.000     5.610657    6.363823
------------------------------------------------------------------------------
```

Factor de ponderación

Uno de los requisitos fundamentales para estimar parámetros poblacionales a partir de muestras es que la muestra se haya obtenido mediante un m.a.s. (muestreo aleatorio simple). En ocasiones (y por motivos que van desde la imposibilidad física al presupuesto) no se realiza un m.a.s. de toda la población sino que se realiza un muestreo estratificado. Es decir se divide la población en grupos y se realiza un muestreo simple en cada grupo. Si no se entrevista a un número proporcionalmente igual de individuos en cada grupo (es decir si existen grupos infra-encuestados), entonces habrá individuos en la encuesta que serán representativos de más individuos que otros. Por ello el encuestador tiene que calcular e introducir, para cada observación, un factor de ponderación que nos indica el número de individuos de los que dicha observación es representativa, en función del tipo de muestreo que haya realizado.

Nosotros, para no sesgar nuestras estimaciones, tenemos que utilizar el factor de ponderación para ponderar con un mayor peso aquellas observaciones que representen a más individuos. Pero debe hacerse con cuidado, porque la mayoría de los programas informáticos permiten introducir muchas formas de ponderación. Lo más peligroso es confundirla con la ponderación que supone que una observación se corresponde con distintos valores repetidos. Es lo más peligroso porque si utilizamos esta última estaremos incrementando artificialmente el número de observaciones y, por consiguiente reduciendo los p-valores. Por ello es muy importante observar siempre que cuando utilicemos pesos no se incrementen el número de observaciones de la muestra.

En la base satisfacción y una vez que, mediante stepwise, tenemos una relación de variables que están relacionadas con la satisfacción repetiremos la operación pero ponderando por el factor de elevación. Primero comprobamos que no se ha incrementado el número de observaciones y después observamos que algunas variables cambian sensiblemente su estimación. Por ejemplo, cuando se pondera, la renta familiar deja de ser significativa y la opinión política se sitúa como tercera variable en importancia.

```
. regress satisf_1_10 edad   estasalud   nac_esp   rentafam   izq_der   frec_esp_pub
frec_urg_publ frec_hosp_pub, beta

      Source |       SS           df       MS      Number of obs   =     4,104
-------------+----------------------------------   F(8, 4095)      =     35.13
       Model |  1048.00632         8   131.00079   Prob > F        =    0.0000
    Residual |  15269.2481     4,095  3.72875411   R-squared       =    0.0642
-------------+----------------------------------   Adj R-squared   =    0.0624
       Total |  16317.2544     4,103  3.97690821   Root MSE        =     1.931

------------------------------------------------------------------------------
  satisf_1_10 |      Coef.   Std. Err.      t    P>|t|                     Beta
--------------+---------------------------------------------------------------
         edad |   .0242404   .0019556    12.40   0.000                .2077126
    estasalud |  -.3652539   .0445546    -8.20   0.000               -.1408294
      nac_esp |  -.7444213   .1168009    -6.37   0.000               -.0982157
     rentafam |   .0574368   .0187394     3.07   0.002                .0483961
      izq_der |   .0930722   .0164056     5.67   0.000                .0864495
 frec_esp_pub |   .0537909   .0197697     2.72   0.007                  .04548
frec_urg_publ |   -.062989   .0248274    -2.54   0.011               -.0418183
frec_hosp_pub |   .2122674    .077882     2.73   0.006                .0453574
        _cons |   5.999079   .1916498    31.30   0.000                       .
------------------------------------------------------------------------------

. regress satisf_1_10 edad   estasalud   nac_esp   rentafam   izq_der   frec_esp_pub
frec_urg_publ frec_hosp_pub [pweight = factor], beta
(sum of wgt is   4.2451e+03)

Linear regression                               Number of obs     =      4,104
                                                F(8, 4095)        =      27.40
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0641
                                                Root MSE          =     1.9261

------------------------------------------------------------------------------
             |               Robust
  satisf_1_10 |      Coef.   Std. Err.      t    P>|t|                     Beta
--------------+---------------------------------------------------------------
         edad |   .0220374   .0021826    10.10   0.000                .1896101
    estasalud |  -.3504589   .0573838    -6.11   0.000               -.1336643
      nac_esp |  -.6812052   .1261829    -5.40   0.000                -.091423
     rentafam |   .0338264   .0215688     1.57   0.117                .0289861
      izq_der |    .110823   .0209158     5.30   0.000                .1031019
 frec_esp_pub |   .0608049    .027268     2.23   0.026                .0514552
frec_urg_publ |  -.0838816   .0361632    -2.32   0.020               -.0533808
frec_hosp_pub |   .2923589   .0848188     3.45   0.001                .0635414
        _cons |   6.037104   .2252347    26.80   0.000                       .
------------------------------------------------------------------------------
```

El efecto tamaño

Otro caso relevante es el denominado “efecto tamaño”. En estudios agregados de países y regiones, en los que se mezclan países grandes como Argentina o Brasil, con otros pequeños como Costa Rica o Guatemala. En muchos casos se plantea la necesidad o no de ponderar en función de pesos como puede ser la población o la superficie (entre otros). La respuesta es clara y evidente. Si lo que se está estudiando son características estructurales de los países o regiones (tasa de crecimiento, endeudamiento y variables

macroeconómicas en general) no se debe ponderar porque cada país o región es un individuo. Pero si lo que se está estudiando son características de los individuos de cada país (renta per cápita, tasa de analfabetismo y variables características de los individuos) sí se debe ponderar si se quieren tener una interpretación individual.

Por ejemplo si queremos estimar si los países con menos impuestos atraen más capital extranjero, entonces no se puede ponderar, pero si se quiere saber si la inversión en educación mejora la esperanza de vida pero, en lugar de tener datos individuales tenemos datos agregados por regiones entonces tenemos que pesar cada región en función de sus habitantes.

A continuación estimamos dos modelos, en el primero se trata de testar la relación de la curva de Phillips entre inflación y desempleo en España, controlado por el PIB per cápita. En este caso cada región es un individuo y no se puede pesar. Por cierto la relación, aunque negativa no es significativa lo cual es relativamente esperable porque no tenemos datos temporales, con los que trabajó Phillips, sino un corte transversal de provincias.

En el segundo modelo nos preguntamos si la existencia o la posibilidad de acceder a servicios sociales como educación, sanidad etc, incrementa o diminuye la posibilidad de estar parado (controlando también por PIB per cápita). En este caso sí que procede pesar por alguna variable de "efecto tamaño" porque nos estamos preguntando por un problema individual ante un servicio individual y las conclusiones que saquemos tienen que ser individuales. En este caso el resultado es positivo y significativo es decir, a igualdad de renta, la existencia de más servicios sociales implica más desempleo.

*Ojo: es muy importante tener mucho cuidado con la interpretación individual de fenómenos de los que disponemos datos agregados porque podemos caer en la denominada "**falacia de la composición**". El caso anterior puede ser un ejemplo. No es lo mismo observar que en las provincias donde hay más servicios sociales hay más desempleo que argumentar que los individuos que hacen más uso de los servicios*

sociales tienen más probabilidad de estar parados. Sobre todo porque, a nivel agregado la causalidad puede ser al revés y no podemos decir nada a nivel individual.

```
. regress paro incr_ipc pib_pc

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     33.29
       Model |  1155.64273         2  577.821364   Prob > F        =    0.0000
    Residual |  850.541326        49  17.3579863   R-squared       =    0.5760
-------------+----------------------------------   Adj R-squared   =    0.5587
       Total |  2006.18406        51  39.3369423   Root MSE        =    4.1663

------------------------------------------------------------------------------
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    incr_ipc |  -1.590368   2.785332    -0.57   0.571    -7.187703    4.006967
      pib_pc |  -.0010399   .0001627    -6.39   0.000    -.0013668    -.000713
       _cons |   41.88826   4.108709    10.19   0.000      33.6315    50.14502
------------------------------------------------------------------------------

. regress paro n_s_educs_ss pib_pc [pweight = pob_total]
(sum of wgt is   4.6624e+07)

Linear regression                               Number of obs     =         52
                                                F(2, 49)          =      36.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6707
                                                Root MSE          =     3.4454

------------------------------------------------------------------------------
             |               Robust
        paro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
n_s_educs_ss |   .0001214   .0000449     2.70   0.009     .0000311    .0002117
      pib_pc |  -.0011793   .0001442    -8.18   0.000    -.0014691   -.0008895
       _cons |   45.33242    3.07657    14.73   0.000     39.14982    51.51502
------------------------------------------------------------------------------
```

Intermediación, interacción y confusión

La introducción de variables en un modelo es un asunto complejo y siempre ha de tener una justificación teórica. Los procesos más reconocidos son:

Intermediación: Se dice que una variable intermedia a otra cuando al introducirla como variable independiente desaparece la correlación previa de otra independiente. Suelen ser variables muy colineadas que, a su vez se producen una a la otra. Por ejemplo la ampliación del presupuesto de la sanidad debe mejorar la salud de los contribuyentes. Pero el presupuesto provocará mejoras en el capital físico y el capital humano sanitario y serán estos y no el presupuesto el que provoque la mejoría del servicio. Luego si introducimos éstas últimas en el análisis deberá desaparecer la correlación entre presupuesto y mejora en salud.

Interacción: Muchas veces las variables muestran una correlación de forma individual pero, a veces algunas variables cambian su grado de influencia en función del valor de una tercera variable. Esto se llama intermediación. Por ejemplo los litros de cerveza deben influir en la conducción, pero es posible que cuanto mayor sea el peso del individuo menor sea la influencia de la cerveza. Una vez comprobado que existe interacción, para medirla correctamente, se deben introducir todas las variables en estado (incluso aunque no sean significativas) además de en forma multiplicativa.

Confusión: La confusión de variables en un problema de especificación. Las correlaciones son falsas o están mal medidas porque falta introducir en el modelo una variable "confusora" que cambiaría las estimaciones. El caso más conocido es el de un estudio de una compañía de seguros que afirmaba que, tras una campaña de concienciación de conductores en Gran Bretaña, había crecido la probabilidad de tener un accidente. El estudio se basaba en una encuesta previa y otra posterior a la campaña. El problema era que no habían controlado por género y en el estudio previo habían encuestado a muchas más mujeres que hombres y en el estudio posterior habían encuestado a más varones que mujeres. Como la media de accidentes de hombre es superior a la de las mujeres el resultado era que había más accidentes. Sin embargo si hubiesen controlado por género habrían obtenido que tanto hombres como mujeres habían reducido su siniestralidad.

Se presentan tres prácticas. En los tres casos utilizaremos la base "satisfacción". Se supone que buscamos los determinantes del voto. En el primer caso parece observarse una intermediación de rentafam sobre nivel de estudios. De forma que cuando introducimos ambas el nivel de estudios deja de ser significativo en favor de la renta familiar. Una justificación puede ser el que el nivel de estudios determina un nivel de renta que es el que realmente incita a la participación política.

Un segundo caso parece indicar un proceso de interacción. El voto parece estar determinado por la edad, la ocupación y el estado de salud, pero también por sus interacciones que han sido introducidas mediante variables multiplicativas.

El tercer caso sería de confusión. En un primer modelo parece que el estado de salud influye en la probabilidad de voto con un coeficiente significativo. Sin embargo se ha omitido una importante variable de control, la edad. Una vez introducida esta variable confusora (confusora de la relación se entiende) observamos como la relación entre estasalud y voto es negativa y significativa.

Intermediación

```
. regress voto edad estasalud ocup_1d est_recode , beta

      Source |       SS           df       MS      Number of obs   =     7,361
-------------+----------------------------------   F(4, 7356)      =    144.12
       Model |  107.779478         4  26.9448695   Prob > F        =    0.0000
    Residual |  1375.28233     7,356  .186960622   R-squared       =    0.0727
-------------+----------------------------------   Adj R-squared   =    0.0722
       Total |  1483.06181     7,360  .201502964   Root MSE        =    .43239

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        edad |   .0072301   .0003331    21.70   0.000                 .2800425
   estasalud |   -.018707   .0070907    -2.64   0.008                -.0321279
     ocup_1d |  -.0131749   .0023377    -5.64   0.000                -.0727368
  est_recode |   .0351379   .0056993     6.17   0.000                 .0867421
       _cons |   .4119321   .0331784    12.42   0.000                        .
------------------------------------------------------------------------------

. regress voto edad estasalud ocup_1d est_recode rentafam , beta

      Source |       SS           df       MS      Number of obs   =     5,351
-------------+----------------------------------   F(5, 5345)      =     97.30
       Model |  87.0258212         5  17.4051642   Prob > F        =    0.0000
    Residual |  956.147417     5,345  .178886327   R-squared       =    0.0834
-------------+----------------------------------   Adj R-squared   =    0.0826
       Total |  1043.17324     5,350  .194985652   Root MSE        =    .42295

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        edad |   .0067403    .000389    17.33   0.000                 .2597893
   estasalud |  -.0190992   .0080574    -2.37   0.018                -.0336298
     ocup_1d |  -.0070514   .0028009    -2.52   0.012                -.0391972
  est_recode |   .0106533   .0069456     1.53   0.125                 .0268038
    rentafam |    .040714   .0039988    10.18   0.000                 .1539239
       _cons |   .2905787   .0424308     6.85   0.000                        .
------------------------------------------------------------------------------
```

Interacción

```
. regress voto edad estasalud ocup_1d [pweight = factor], beta
(sum of wgt is   7.7366e+03)

Linear regression                               Number of obs     =      7,725
                                                F(3, 7721)        =     141.52
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0696
                                                Root MSE          =     .42995

------------------------------------------------------------------------------
             |               Robust
        voto |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        edad |   .0066184   .0003419    19.36   0.000                 .2652622
   estasalud |  -.0301577   .0079603    -3.79   0.000                -.0530602
     ocup_1d |  -.0182515   .0022807    -8.00   0.000                -.1001809
       _cons |   .5759777   .0221512    26.00   0.000                        .
------------------------------------------------------------------------------
```

```
. regress voto edad estasalud ocup_1d c.edad#c.estasalud c.edad#c.ocup_1d [pweight =
factor], beta
(sum of wgt is   7.7366e+03)

Linear regression                               Number of obs     =      7,725
                                                F(5, 7719)        =      93.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0749
                                                Root MSE          =     .42877

------------------------------------------------------------------------------
                   |               Robust
             voto |      Coef.   Std. Err.      t    P>|t|                 Beta
------------------+-----------------------------------------------------------
             edad |   .0074467   .0011498     6.48   0.000             .2984586
        estasalud |   .0614793   .0239932     2.56   0.010             .1081683
          ocup_1d |  -.0440699   .0071456    -6.17   0.000            -.2418965
                  |
c.edad#c.estasalud |  -.0018321   .0004273    -4.29   0.000            -.2737476
                  |
  c.edad#c.ocup_1d |   .0005562   .0001358     4.10   0.000             .2068263
                  |
            _cons |   .5308189   .0606077     8.76   0.000                    .
------------------------------------------------------------------------------
```

Confusión

```
. regress voto  estasalud rentafam , beta

      Source |       SS           df       MS      Number of obs   =     5,632
-------------+----------------------------------   F(2, 5629)      =     69.81
       Model |  26.4653734         2  13.2326867   Prob > F        =     0.0000
    Residual |  1066.98189     5,629  .189550878   R-squared       =     0.0242
-------------+----------------------------------   Adj R-squared   =     0.0239
       Total |  1093.44727     5,631  .194183496   Root MSE        =     .43537

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|                 Beta
-------------+----------------------------------------------------------------
   estasalud |   .0297427    .007503     3.96   0.000             .0536782
    rentafam |   .0418965   .0035658    11.75   0.000             .1591002
       _cons |   .4884871   .0256046    19.08   0.000                    .
------------------------------------------------------------------------------

. regress voto edad estasalud rentafam , beta

      Source |       SS           df       MS      Number of obs   =     5,632
-------------+----------------------------------   F(3, 5628)      =    163.63
       Model |  87.7197937         3  29.2399312   Prob > F        =    0.0000
    Residual |  1005.72747     5,628  .178700688   R-squared       =    0.0802
-------------+----------------------------------   Adj R-squared   =    0.0797
       Total |  1093.44727     5,631  .194183496   Root MSE        =    .42273

------------------------------------------------------------------------------
        voto |      Coef.   Std. Err.      t    P>|t|                 Beta
-------------+----------------------------------------------------------------
        edad |   .0064067    .000346    18.51   0.000             .2551636
   estasalud |   -.018072   .0077293    -2.34   0.019            -.0326155
    rentafam |   .0485067   .0034806    13.94   0.000             .1842024
       _cons |   .2538681   .0279044     9.10   0.000                    .
------------------------------------------------------------------------------
```

---

6. Extensiones del Modelo lineal

El modelo lineal es suficiente para modelar muchas de las correlaciones más evidentes y siempre que los residuos tengan una forma esperada normal. Sin embargo hay algunas veces que necesitamos de modelos lineales con una estimación de los parámetros algo atípica. Repasamos los modelos básicos más usuales

Variables Truncadas y Censuradas

Una variable truncada es una variable normal de la que nos falta la información de un grupo completo de individuos por encima o por debajo de un cierto valor. Por ejemplo en la encuesta sólo se encuesta a individuos de 16 a 80 años. El truncamiento es un problema grave cuando justo queremos analizar lo que pasa en esos extremos, como puede ser el nivel educativo alcanzado o la necesidad de cuidados de dependencia. En el resto de los casos pues puede ser un problema o no.

Una variable censurada es una variable de la que o no se dispone del valor por encima o por debajo de una frontera o a esta se le asigna un valor constante. Por ejemplo en la encuesta se asigna valor 10 a los individuos que tienen una renta familiar superior a 6000€ mes.

Cuando las variables dependientes están censuradas o truncadas el estimador estará sesgado (al alza o a la baja depende del tipo de truncamiento o censura) y tendremos que utilizar la regresión truncada o la regresión censurada (tobit).

Como ejemplo de regresión truncada vamos a truncar nosotros mismos una variable y estimaremos dos modelos (con y sin truncamiento). Estimaremos los condicionantes dela asistencia a atención primaria. Truncaremos la base de datos sólo para los mayores de 30 años. Los resultados muestra resultados ligeramente distintos, aunque, salvo la edad al cuadrado, ninguna cambia el signo o significación.

```
. regress frec_prim_publ c.edad##c.edad estasalud cronico est_recode if edad >30

      Source |       SS           df       MS      Number of obs   =     5,865
-------------+----------------------------------   F(5, 5859)      =    155.67
       Model |  13750.9074         5  2750.18147   Prob > F        =    0.0000
    Residual |  103507.233     5,859  17.6663651   R-squared       =    0.1173
-------------+----------------------------------   Adj R-squared   =    0.1165
```

```
       Total |   117258.14     5,864  19.9962723   Root MSE        =    4.2031

------------------------------------------------------------------------------
frec_prim_p~l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         edad |  -.0560542   .0278648    -2.01   0.044    -.1106795    -.001429
              |
c.edad#c.edad |   .0004878   .0002473     1.97   0.049     3.11e-06    .0009726
              |
    estasalud |   1.204643   .0824174    14.62   0.000     1.043075    1.366212
      cronico |   1.487637    .134339    11.07   0.000     1.224283    1.750991
   est_recode |  -.3272761   .0550121    -5.95   0.000    -.4351202   -.2194321
        _cons |   1.855866   .7713635     2.41   0.016     .3437092    3.368023
------------------------------------------------------------------------------

. truncreg frec_prim_publ c.edad##c.edad estasalud cronico est_recode, ll(edad >30)
(note: 3,409 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood = -8369.9257
Iteration 1:   log likelihood = -8342.9172
Iteration 2:   log likelihood = -8340.5224
Iteration 3:   log likelihood = -8340.5148
Iteration 4:   log likelihood = -8340.5148

Truncated regression
Limit:   lower =          0                     Number of obs     =      3,941
         upper =       +inf                     Wald chi2(5)      =     522.91
Log likelihood = -8340.5148                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
frec_prim_p~l |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         edad |   .0424582    .021178     2.00   0.045     .0009501    .0839662
              |
c.edad#c.edad |  -.0003933   .0002094    -1.88   0.060    -.0008037     .000017
              |
    estasalud |   1.259428   .1052229    11.97   0.000     1.053195    1.465661
      cronico |   1.305438   .1649816     7.91   0.000     .9820804    1.628796
   est_recode |  -.3742258   .0729267    -5.13   0.000    -.5171596   -.2312921
        _cons |   1.092324   .5308578     2.06   0.040      .051862    2.132786
--------------+---------------------------------------------------------------
       /sigma |   4.341074   .0471113    92.14   0.000     4.248738    4.433411
------------------------------------------------------------------------------
```

Regresión lineal en dos etapas

Uno de los requisitos del modelo es que las variables independientes no estén correlacionadas con los residuos. Pero, en ocasiones, esto no sucede. Causas pueden ser:

- Variables relevantes omitidas.
- Errores de especificación.
- Endogeneidad de la/s variable/s independientes.

Una vez solucionados los dos primeros casos (linealizando, buscando las variables necesarias, etc.) queda el tercero. En estos casos una solución puede ser la

estimación en dos etapas con variables instrumentales. El más grave problema sería encontrar los instrumentos porque es necesario que estos están correlacionados con las variables independientes (que sean relevantes) y no lo estén con los residuos (que sean independientes).

La regresión bietápica realiza las siguientes operaciones. Supongamos el modelo

$$y = \beta_o + \beta_1 x1 + \beta_2 x2 + u$$

Donde *y* es la variable dependiente, *x1* y *x2* son las independientes pero parece que *x1* está correlacionada con los residuos *u*. Necesitaremos, al menos una variable instrumental *Z* para que el sistema esté identificado – es decir haya una solución- necesitamos tantas variables instrumentales como variables potencialmente endógenas). Es necesario que Z no esté correlacionada con los residuos u y que esté correlacionada con x1. El **primer paso** del modelo es estimar la regresión

$$x1 = \beta'_o + \beta'_1 x2 + \beta'_2 Z + v$$

Y obtener los valores previstos para la variable endógena ($\widehat{x1}$). El segundo paso será estimar ahora el modelo inicial pero reemplazando *x1* por sus valores estimados antes ($\widehat{x1}$). Es decir el modelo sería:

$$y = \beta''_o + \beta''_1 \widehat{x1} + \beta''_2 x2 + u'$$

Hay dos problemas graves con esta metodología:

a) Encontrar buenos instrumentos. Hay que tener en cuenta que en función del instrumento elegido los valores previstos de *x1* cambian y los parámetros estimados también cambian. No siempre es fácil. Hay que buscar variables que afecten a *x1* y no afecten a *y*. Suelen ser variables "naturales" que provocan shocks.

b) Encontrar pruebas de la endogeneidad. Saber si *x1* es o no endógena no siempre se obtiene de un test. Lo normal es que la endogeneidad (así como la independencia de los

instrumentos) sea el resultado de un debate intelectual. Existe un test de endogeneidad (propuesto por Hausman) que consiste en estimar las dos siguientes ecuaciones

$$y = \beta_o + \beta_1 x1 + \beta_2 x2 + \beta_3 Z + u \quad [1]$$

$$y = \beta'_o + \beta'_1 v + \beta'_2 x2 + u \quad [2]$$

En la ecuación [1] es necesario que el coeficiente $\widehat{\beta_3}$ no sea significativo (para que el instrumento sea independiente) y en la ecuación [2] donde v son los residuos estimados en $x1 = \beta'_o + \beta'_1 Z + v$. Si $\widehat{\beta'_1}$ es significativa es que *x1* es efectivamente endógena y si no lo es es que *x1* puede que no sea endógena. Sin embargo, como ya advertimos, los test no son siempre suficientes.

1er. ejemplo

Supongamos que en la base "satisfacción" queremos estimar la relación entre la satisfacción con el servicio público sanitaria con la renta familiar con la edad, el nivel y el estado de salud. La regresión inicial muestra un bajo ajuste pero las variables son todas significativas y tienen el signo esperado. Sospechamos, no obstante que puede haber cierta endogeneidad entre estado de salud y satisfacción porque una mayor satisfacción puede provocar más frecuenciación y una percepción de peor salud. Decidimos instrumentalizar autovaloración del estado de salud con padecer una enfermedad crónica. comprobamos que crónico no entra en la regresión y lo estimamos por 2 etapas. Los resultados en la regresión bietápica muestran un ajuste similar y una estimación similar en el resto de regresores. Respecto a la variable endógena "estasalud" parece que incrementa un poco su influencia, de 0.34 a 0.44.

```
. regress satisf_1_10 c.edad##c.edad nac_esp rentafam estasalud

      Source |       SS           df       MS      Number of obs   =     5,600
-------------+----------------------------------   F(5, 5594)      =     79.83
       Model |  1556.01284         5  311.202568   Prob > F        =    0.0000
    Residual |   21808.007     5,594  3.89846389   R-squared       =    0.0666
-------------+----------------------------------   Adj R-squared   =    0.0658
       Total |  23364.0198     5,599  4.17289156   Root MSE        =    1.9745

------------------------------------------------------------------------------
  satisf_1_10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         edad |  -.0293866    .008748    -3.36   0.001    -.0465361   -.0122371
              |
c.edad#c.edad |   .0005605   .0000839     6.68   0.000      .000396    .0007249
              |
      nac_esp |    -.87417   .0890797    -9.81   0.000    -1.048801   -.6995391
     rentafam |   .0776756   .0167819     4.63   0.000     .0447766    .1105747
```

```
    estasalud |  -.3449031   .0362704    -9.51   0.000    -.4160072   -.2737991
        _cons |   7.593775   .2281942    33.28   0.000     7.146426    8.041125
------------------------------------------------------------------------------

. regress satisf_1_10 c.edad##c.edad nac_esp rentafam estasalud cronico

      Source |       SS           df       MS      Number of obs   =     5,591
-------------+----------------------------------   F(6, 5584)      =     66.47
       Model |  1553.80397         6  258.967328   Prob > F        =    0.0000
    Residual |  21756.1188     5,584  3.89615307   R-squared       =    0.0667
-------------+----------------------------------   Adj R-squared   =    0.0657
       Total |  23309.9227     5,590  4.16993251   Root MSE        =    1.9739

------------------------------------------------------------------------------
  satisf_1_10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         edad |  -.0293466    .008751    -3.35   0.001    -.0465019   -.0121913
              |
c.edad#c.edad |   .0005636   .0000839     6.72   0.000     .0003991    .0007281
              |
      nac_esp |  -.8697246   .0891352    -9.76   0.000    -1.044464   -.6949851
     rentafam |   .0760211   .0167922     4.53   0.000     .0431018    .1089404
    estasalud |  -.3260467    .039735    -8.21   0.000    -.4039427   -.2481507
      cronico |  -.0810155   .0670275    -1.21   0.227    -.2124155    .0503845
        _cons |   7.571778   .2301599    32.90   0.000     7.120575    8.022981
------------------------------------------------------------------------------

. ivregress 2sls satisf_1_10 c.edad##c.edad nac_esp rentafam (estasalud = cronico)

Instrumental variables (2SLS) regression          Number of obs   =      5,591
                                                  Wald chi2(5)    =     331.37
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0652
                                                  Root MSE        =     1.9742

------------------------------------------------------------------------------
  satisf_1_10 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    estasalud |  -.4435617   .0887486    -5.00   0.000    -.6175057   -.2696177
         edad |     -.0271   .0089866    -3.02   0.003    -.0447134   -.0094866
              |
c.edad#c.edad |    .000552   .0000843     6.55   0.000     .0003868    .0007173
              |
      nac_esp |  -.8711651   .0891402    -9.77   0.000    -1.045877   -.6964536
     rentafam |   .0672438   .0181946     3.70   0.000     .0315831    .1029045
        _cons |   7.761376   .2614505    29.69   0.000     7.248943     8.27381
------------------------------------------------------------------------------
Instrumented:  estasalud
Instruments:   edad c.edad#c.edad nac_esp rentafam cronico
```

2º Ejemplo

Igual que antes se realiza una primera regresión por la que pretendemos explicar el valor del estado de salud en función de la edad, el género la renta familiar y el nivel de estudios. Todas las variables aparecen como significativas. Sin embargo pensamos que rentafam puede ser potencialmente endógena del estado de salud por lo que los resultados pueden estar sesgados. Entonces pretendemos instrumentalizarla con ocup_1d. Observamos que Ocup_1d no es significativa en el modelo general y sí que está correlacionada con rentafam. Al utilizarla como instumento el resultado muestra que la rentafam aumenta ligeramente su influencia y que el nivel de estudios (est_recode) deja de ser significativa (probablemente por problemas de colinealidad con renta familiar y ocup_1d)

```
. regress estasalud  edad genero rentafam est_recode

      Source |       SS           df       MS      Number of obs   =     5,351
-------------+----------------------------------   F(4, 5346)      =    238.15
       Model |  489.145082         4  122.286271   Prob > F        =    0.0000
    Residual |  2745.12851     5,346  .513492052   R-squared       =    0.1512
-------------+----------------------------------   Adj R-squared   =    0.1506
       Total |  3234.27359     5,350   .60453712   Root MSE        =    .71658

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |   .0130606   .0006268    20.84   0.000     .0118318    .0142894
     genero |  -.0904561   .0196136    -4.61   0.000    -.1289068   -.0520055
   rentafam |   -.061293   .0066012    -9.29   0.000    -.0742341    -.048352
  est_recode |  -.0602441   .0106784    -5.64   0.000    -.0811782     -.03931
       _cons |   1.958409   .0485802    40.31   0.000     1.863172    2.053646
------------------------------------------------------------------------------

. regress rentafam ocup_1d

      Source |       SS           df       MS      Number of obs   =     5,648
-------------+----------------------------------   F(1, 5646)      =    994.94
       Model |  2370.46694         1  2370.46694   Prob > F        =    0.0000
    Residual |  13451.6915     5,646   2.3825171   R-squared       =    0.1498
-------------+----------------------------------   Adj R-squared   =    0.1497
       Total |  15822.1585     5,647  2.80186975   Root MSE        =    1.5435

------------------------------------------------------------------------------
    rentafam |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ocup_1d |  -.2657006   .0084235   -31.54   0.000    -.2822139   -.2491872
       _cons |   5.868146   .0510094   115.04   0.000     5.768148    5.968144
------------------------------------------------------------------------------


. regress estasalud  edad genero rentafam est_recode ocup_1d

      Source |       SS           df       MS      Number of obs   =     5,351
-------------+----------------------------------   F(5, 5345)      =    190.91
       Model |  490.083656         5  98.0167311   Prob > F        =    0.0000
    Residual |  2744.18994     5,345  .513412523   R-squared       =    0.1515
-------------+----------------------------------   Adj R-squared   =    0.1507
       Total |  3234.27359     5,350   .60453712   Root MSE        =    .71653

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |   .0131886   .0006339    20.81   0.000     .0119459    .0144313
     genero |  -.0919558   .0196434    -4.68   0.000    -.1304649   -.0534467
   rentafam |  -.0595219   .0067294    -8.84   0.000    -.0727143   -.0463294
  est_recode |  -.0536325   .0117441    -4.57   0.000    -.0766556   -.0306093
     ocup_1d |   .0064253   .0047521     1.35   0.176    -.0028909    .0157414
       _cons |   1.894696   .0676768    28.00   0.000     1.762022     2.02737
------------------------------------------------------------------------------

. ivregress 2sls estasalud edad genero est_recode (rentafam = ocup_1d)

Instrumental variables (2SLS) regression          Number of obs   =      5,351
                                                  Wald chi2(4)    =     869.46
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.1439
                                                  Root MSE        =     .71935

------------------------------------------------------------------------------
   estasalud |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rentafam |  -.1062628    .034042    -3.12   0.002    -.1729839   -.0395416
       edad |   .0132934   .0006526    20.37   0.000     .0120144    .0145724
     genero |  -.0861584   .0199463    -4.32   0.000    -.1252524   -.0470645
  est_recode |  -.0281602   .0261236    -1.08   0.281    -.0793614    .0230411
       _cons |   2.071344   .0970064    21.35   0.000     1.881215    2.261473
```

```
------------------------------------------------------------------------------
Instrumented:  rentafam
Instruments:   edad genero est_recode ocup_1d
```

Modelos de probabilidad lineal y modelos de respuesta fraccional

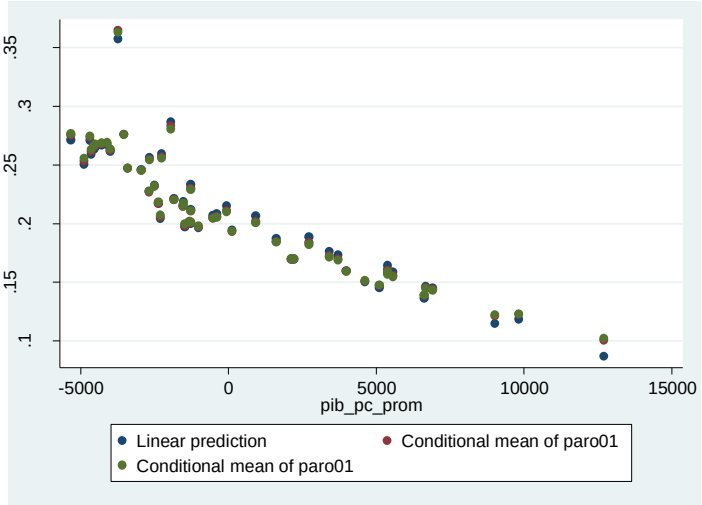
Cuando la variable dependiente es un porcentaje, es decir está acotado superior e inferiormente entre 0 y 1 se puede estimar un modelo de probabilidad lineal, sobre todo si el modelo sólo tiene valores intermedios. Sin embargo, cuando los porcentajes están muy próximos a 0 o a 1 ya no se comportan como cuando están en mitad de la tabla porque se "frenan", se "tuercen" acotados en 0 y 1 respectivamente. Entonces hay que estimar un modelo de respuesta fraccional, que es similar a un logit o probit. Hay dos versiones similare francreg y betareg que se pueden estimar como logit, probit log-log, etc. La única diferencia entre ambas es que la última no puede ajustar cuando se tienen 0 y 1 exactos en la variable dependiente (lo cual no deja de ser sorprendente).

En el ejemplo, utilizando la base "provincias" estimamos la regresión de pib per cápita y nacimientos por 1000 habitantes sobre paro. Como paro es una variable porcentual pero se mueve en valores intermedios podemos estimar los modelos de probabilidad lineal (regress) y de respuesta fraccional (fracreg y betareg). En los primeras instrucciones generamos y adaptamos las variables necesarias. Despues se realizan las estimaciónes y se predicen los valores de la dependiente (ŷ). En la última se genera el gráfico de las tres predicciones. El gráfico muestra que las tres regresiones son similares en los resultados previstos e incluso en los parámetros estimados. Sorprende el bajísimo pseudo R2 reportado por fracreg lo que nos hace dudar de estas medidas de ajuste en los modelos estimado por máxima verosimilitud. (Si lo calculamos mediante la proporción del ECM entre el modelo vacío y el completo el ajuste – ver en la sección de ajuste - es resultado es del 0.4442. Bastante más parecido al del modelo lineal)

```
. generate pib_pc_prom = pib_pc - 20500

. generate nacim1000_prom = nacim1000 - 9

. generate paro01 = paro/100
```

```
. regress paro01 pib_pc_prom nacim1000_prom

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     54.52
       Model |  .138419449         2  .069209724   Prob > F        =    0.0000
    Residual |  .062198962        49  .001269367   R-squared       =    0.6900
-------------+----------------------------------   Adj R-squared   =    0.6773
       Total |  .200618411        51  .003933694   Root MSE        =    .03563

------------------------------------------------------------------------------
       paro01 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
  pib_pc_prom |  -.0000109   1.15e-06    -9.51   0.000    -.0000132   -8.61e-06
nacim1000_prom |    .010018   .0023322     4.30   0.000     .0053312    .0147048
        _cons |   .2132477   .0049604    42.99   0.000     .2032795     .223216
------------------------------------------------------------------------------

. predict est1, xb

. fracreg probit paro01 pib_pc_prom nacim1000_prom

Iteration 0:   log pseudolikelihood = -41.646255
Iteration 1:   log pseudolikelihood = -26.261506
Iteration 2:   log pseudolikelihood = -26.233496
Iteration 3:   log pseudolikelihood = -26.233487
Iteration 4:   log pseudolikelihood = -26.233487

Fractional probit regression                    Number of obs     =         52
                                                Wald chi2(2)      =     113.75
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -26.233487               Pseudo R2         =     0.0159

------------------------------------------------------------------------------
             |               Robust
       paro01 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
  pib_pc_prom |  -.0000398   4.10e-06    -9.71   0.000    -.0000479   -.0000318
nacim1000_prom |   .0303472   .0073074     4.15   0.000      .016025    .0446695
        _cons |  -.8087627   .0166711   -48.51   0.000    -.8414375    -.776088
------------------------------------------------------------------------------

. predict est2, cm

. betareg paro01 pib_pc_prom nacim1000_prom

initial:       log likelihood =   88.43002
rescale:       log likelihood =   88.43002
rescale eq:    log likelihood =   88.43002
(setting technique to bhhh)
Iteration 0:   log likelihood =   88.43002
Iteration 1:   log likelihood =  102.94676
Iteration 2:   log likelihood =   104.0842
Iteration 3:   log likelihood =   104.5199
Iteration 4:   log likelihood =  104.59874
Iteration 5:   log likelihood =  104.60789
Iteration 6:   log likelihood =  104.60817
Iteration 7:   log likelihood =  104.60945
Iteration 8:   log likelihood =  104.60945

Beta regression                                 Number of obs     =         52
                                                LR chi2(2)        =      63.94
                                                Prob > chi2       =     0.0000

Link function  :  g(u) = log(u/(1-u))           [Logit]
Slink function :  g(u) = log(u)                 [Log]

Log likelihood =  104.60945

------------------------------------------------------------------------------
       paro01 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
paro01        |
  pib_pc_prom |  -.0000706   7.29e-06    -9.68   0.000    -.0000849   -.0000563
nacim1000_prom |   .0492118   .0116811     4.21   0.000     .0263173    .0721063
        _cons |  -1.334806   .0281772   -47.37   0.000    -1.390033    -1.27958
--------------+---------------------------------------------------------------
```

```
scale          |
         _cons |   5.013226   .1957241    25.61   0.000     4.629614    5.396838
------------------------------------------------------------------------------

. predict est3, cm

. twoway (scatter est1 pib_pc_prom) (scatter est2 pib_pc_prom) (scatter est3
pib_pc_prom)
```



7. Bibliografía

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. Stata Journal 3: 1–31.

Becketti, S. 1995. sg26.2: Calculating and graphing fractional polynomials. Stata Technical Bulletin 24: 14–16. Reprinted in Stata Technical Bulletin Reprints, vol. 4, pp. 129–132. College Station, TX: Stata Press.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Breusch, T. S., and A. R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47:1287-1294.

Cameron, A. C., and P. K. Trivedi. 1990. The information matrix test and its applied alternative hypotheses. Working Paper 372, University of California-Davis, Institute of Governmental Affairs.

Cong, R. 1999. sg122: Truncated regression. Stata Technical Bulletin 52: 47–52. Reprinted in Stata Technical Bulletin Reprints, vol. 9, pp. 248–255. College Station, TX: Stata Press.

Desbordes, R., and V. Verardi. 2012. A robust instrumental-variables estimator. Stata Journal 12: 169–181.

Greene, W. H. (2006). Análisis Econométrico (3ª ed.). Madrid, España: Prentice-Hall.

Gujarati, D. N. y Portes, D.C. (2009). Econometría (5ª ed.). México: McGraw-Hill.

Lindsey, C., and S. J. Sheather. 2010a. Power transformation via multivariate Box–Cox. Stata Journal 10: 69–81. 2010b. Optimal power transformation via inverse response plots. Stata Journal 10: 200–214.

Ramsey, J. B. 1969. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society, Series B 31: 350-371.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. Econometrica 26: 24–36.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48: 817-838.

Wooldridge, J. M. (2010). Introducción a la econometría: un enfoque moderno. Thomsom Learning.

Granada, marzo de 2016