Curso de Análisis de Datos para Epidemiología Ambiental

Centro Panamericano de Ecología Humana y Salud
(ECO-OPS)
Centro de Investigaciones en Salud Poblacional
Instituto Nacional de Salud Pública

Instructor :
Dr. Mauricio Hernández-Avila

Belo Horizonte, Brasil
del 9 al 13 de junio 1997

Indice

INTRODUCCION

El presente texto esta dirigido al trabajador en salud pública interesado en conocer la metodología del análisis estadístico moderno y el uso de sistemas de cómputo para implementarlo en el campo de la epidemiología y, particularmente, en epidemiología ambiental. La lectura es fácil, el autor enfatiza en aspectos conceptuales y de interpretación pero sin descuidar el marco teórico, el cual introduce en forma sencilla.

El texto que se presenta en esta obra contiene una clara explicación de los métodos modernos de análisis estadístico. Incluye ejemplos con bases de datos reales provenientes de estudios en epidemiología ambiental. La obra se divide en dos grandes capítulos, cada uno de ellos conteniendo la presentación teórico-práctica de los elementos y herramientas de las dos grandes ramas del análisis estadístico: Análisis Exploratorio y Análisis Confirmatorio.

En la sección de Análisis Exploratorio se discute sobre los estadísticos puntuales, tales como la media y varianza, mediana, rango intercuaticular y otros estadísticos de orden con especial énfasis en su significado, su contraparte gráfica y su aplicación para determinar la forma distribucional de una variable aleatoria. Además, el autor ilustra con ejemplos desarrollados en el paquete estadístico Stata, los procedimientos para generar gráficos exploratorios y tablas con medidas de tendencia central, de dispersión, de sesgo y del comportamiento de las colas para explorar la distribución de la concentración de plomo en el aire. Cada ejemplo presenta la serie de comandos de Stata necesarios para completar un determinado objetivo así como una descripción detallada de las salidas que el paquete emite después de ejecutar cada comando. Continuando en este mismo esquema teórico-práctico el autor introduce el tema de transformación de variables dentro del marco de normalidad, así mismo introduce algunos métodos de diagnóstico tales como los gráficos quantil-normal y los de desviación de la mediana. De nuevo, los ejemplos presentados explican de manera clara el marco conceptual y los procedimientos y series de comandos necesarios para generar estos gráficos diagnósticos. Se presentan también estadísticos que permiten realizar pruebas de hipótesis sobre la normalidad de una muestra aleatoria.

Al igual que en el primer capítulo, en la sección de Análisis Confirmatorio el autor conduce al lector a través de una serie de ejemplos en los que el objetivo es introducirlo en la metodología del análisis comparativo bi-variado y multivariado. Después de una breve exposición sobre su uso en la vida cotidiana, el autor introduce algunos ejemplos de modelos en los que las variables dependientes no son necesariamente normales tal como lo es el número de consultas y su relación con la concentración de ozono en el aire y en las que el autor modela la razón de tasas de incidencia. Regresando al tema de la normalidad, se presenta otro ejemplo en el marco del cual es introducido el modelo lineal.

En este capítulo se discute sobre los tipos de variables y su implicación en la metodología de análisis aplicada. Después de establecer las medidas de comparación y establecer su relación con los diferentes modelos estadísticos, el autor lanza un nuevo ejemplo, esta vez sobre el peso al nacer y su relación con la concentración de plomo en hueso. Se analiza la información en diferentes formas al reparametrizar tanto la variable independiente como la independiente, reexpresando el modelo para así ilustrar los procedimientos y la secuencia de comandos de Stata necesarios para ejecutar la estimación de parámetros en los diferentes tipos de modelo presentados. Nuevamente, el énfasis del autor esta dirigido a la interpretación de los parámetros estimados por cada tipo de modelo de acuerdo a la parametrización de las variables introducidas y de la variable dependiente. Se presentan también conceptos básicos sobre la estimación del valor p y las prueba de hipótesis. El resto de esta sección esta dedicado a la presentación formal del método de estimación de mínimos cuadrados y su aplicación en la regresión lineal, continuando con un ejemplo, esta vez con la concentración de plomo en sangre y su relación con la concentración de plomo en rótula. Aquí se ilustra el proceso para ejecutar la estimación de los coeficientes del modelos de regresión así como la serie de herramientas de diagnóstico que permiten verificar el ajuste del modelo y la observancia de los supuestos del mismo. Más ejemplos, con el respectivo código de Stata y los comentarios a las salidas ilustran situaciones especiales del modelos de regresión lineal, tal es el caso del ANOVA, en la que se compara la media entre dos poblaciones y el análisis de covarianza, en el que el interés reside en comparar el efecto de una variable continua en dos poblaciones distintas y la posible interacción.

La obra presenta de manera concreta la metodología de análisis exploratorio y el modelo de regresión lineal ordinaria. Introduce y el modelo lineal presentando una regresión de poison, en la que la variable dependiente es el número de consultas y la estimación es de razón de tasas de incidencia. También introduce el modelo logístico en el que la variable dependiente es binaria y la estimación es de razón de momios. Aunque el modelo lineal generalizado no es tratado formalmente, se proporciona amplia información sobre la interpretación de este tipo de modelos y las situaciones en las que es valido aplicarlos. Esta obra es una valiosa fuente de referencia para el alumno del curso de bioestadística ambiental.

Algunos comentarios sobre el paquete Stata.

STATA es una marca registrada de Stata Corporation, es producido por STATA Corporation y se puede comprar directamente en Stata Corporation.

STATA es un paquete de programas muy completo, incluye comandos para realizar análisis muy sofisticados inclusive las nuevas metodologías para el análisis de datos longitudinales. Es un paquete al cual se le da mantenimiento y actualizaciones periódicas, esto con la consecuente introducción de nuevas técnicas y algoritmos según se hacen disponibles. Tiene una gran capacidad gráfica y una buena interfase de

comunicación de datos con procesadores de texto e impresoras. STATA ocupa una cantidad de espacio en disco relativamente pequeña y muy importante, de fácil aprendizaje. Aun cuando no es un programa enteramente orientado a menús, los comandos son consistentes en su sintaxis y las palabras clave de cada uno de ellos es nemotécnica. Aún cuando STATA presenta dificultades para la introducción de datos crudos al formato binario natal se cuenta con el apoyo de traductores como DBMSCOPY M. R.

Es posible realizar todos los procedimientos estadísticos presentados en este texto utilizando cualquier otro paquete estadístico con excelentes características gráficas, tal como el caso de S-plus M. R. la utilización de STATA para el desarrollo y ejecución de los ejercicios presentados en esta obra es meramente preferencia del autor.

ANALISIS EXPLORATORIO DE DATOS

El análisis exploratorio de datos agrupa una serie de técnicas estadísticas y gráficas que permiten "explorar" grandes cantidades de información. En general, las técnicas de análisis exploratorio de datos se utilizan en las primeras fases del análisis estadístico y sirven para:

a) Evaluar la calidad y consistencia de la información
b) Investigar la distribución de las variables de interés
c) Investigar adherencia a las suposiciones estadísticas necesarias, en etapas posteriores del análisis
d) Resumir información mediante diferentes estadísticos y gráficos
e) Evaluar la necesidad de realizar transformación de las variables de interés
f) Detectar valores "Fuera de serie "(OL)" no plausibles" (outlier)
g) Explorar formas de categorizar variables (puntos de corte)

La evaluación de la calidad y consistencia de la información es un paso importante que debe realizar antes de iniciar el análisis estadístico. En el campo de la investigación epidemiológica, se recolecta información sobre un gran número de variables, ya sea mediante cuestionarios o mediante sistemas de adquisición de información que en muchas ocasiones no dependen del investigador, estos datos de fuentes secundarias generalmente no están sujetos a controles de calidad estrictos, por lo que es conveniente realizar evaluaciones que permitan detectar patrones que se desvíen de los valores posibles. La evaluación que se realiza con mayor frecuencia es la búsqueda de valores no plausibles, de valores faltantes o errores de codificación y captura.

Varias técnicas estadísticas sirven para identificar valores que potencialmente podrían ser considerados como errores o valores aberrantes-outliers-. En general los valores aberrantes se identifican como valores que se encuentran lejos del total de observaciones, y estas se diferencian notablemente de la nube de puntos o como valores que no son posibles dentro del rango de variación biológica o de codificación. Existen diferentes criterios y técnicas estadísticas para el tratamiento de los valores aberrantes. En general se recomienda identificar la fuente de error. En este sentido, también es importante diferenciar si se trata de una observación con plausibilidad biológica, o de una que queda fuera del rango biológico de observación. Cuando se trata de un valor extremo, pero que si se encuentra dentro del rango posible de observaciones,

se recomienda mantener la observación y explorar su efecto en fases subsecuentes del análisis estadístico. En el segundo caso, cuando se trata de un valor no plausible, se recomienda excluir la observación de análisis subsecuentes. En ambos casos es recomendable consultar las fuentes primarias de información para descartar la posibilidad de error .

Mediante las técnicas de análisis exploratorio de datos, es posible estudiar la distribución de la información, detectar asimetría, rangos observados, así como los valores máximos y mínimos. La información sobre la distribución de las variables es importante, ya que muchas de las técnicas estadísticas utilizadas a menudo, asumen una serie de suposiciones sobre el comportamiento y distribución de la variables en estudio. Así por ejemplo, la regresión lineal simple considera que la variable dependiente debe estar normalmente distribuida.

Por otra parte los procedimientos utilizados proporcionan al investigador métodos gráficos, de fácil interpretación, que son muy útiles para resumir grandes volúmenes de información, así como para representar comparaciones entre grupos .

I. ESTADISTICAS UNIVARIADAS

Las estadísticas univariadas incluyen la media, la mediana, valores que definen los límites de los percentiles, moda, los valores máximos y mínimos, así como las medidas de dispersión (rango, desviación estandar), comúnmente utilizadas en estadística para resumir información.

Para ilustrar los diferentes estadísticos y gráficos utilizados se emplearán algunas bases de datos obtenidas de diferentes investigaciones epidemiológicas, además utilizaremos bases de datos derivadas de las estaciones de monitoreo ambiental de la Ciudad de México.

Inicialmente se utilizará la información relativa a las concentraciones de plomo partículas medidas en la Ciudad de México. La información se encuentra expresada en microgramos por metro cúbico. La norma actual para plomo en aire es de 1.5 $\mu g/m^3$, promedio anual.

Además incluiremos dentro de los ejemplos, los resultados de una investigación en la que se midieron los niveles de plomo en sangre, hueso y leche materna.

Dado que la base de datos se encuentra directamente en el formato de Stata, se puede acceder a la misma mediante el programa con el comando **use**. Esto quiere decir que la información sobre las variables, valores, campos etc. se encuentra ya definida.

La base de datos se llama plomoddf.**dta**. La terminación **dta** indica que el archivo tiene un formato compatible con Stata. Dado que Stata busca la terminación **dta** para reconocer archivos, no es necesario especificar esta terminación, como se observa en la instrucción utilizada.

```
. use a:plomoddf

. desc

Contains data from :untitled::plomoddf.dta
     Obs: 3951   (max=  16700)
    Vars:    3   (max=    236)            23 Jul 1996 22:00
   Width:   16   (max=    290)
 1. fecha        str8     %8s
 2. estacion     str4     %4s
 3. plomo        float    %6.3f
Sorted by:
Note:  Data has changed since last save
```

La instrucción **describe** proporciona alguna información general sobre la base de datos, el número de observaciones y de variables, así como de los límites establecidos para la partición de memoria. En este caso, se podrían tener 16,700 observaciones y 236 variables. La base que analizaremos tiene 3951 observaciones y tres variables.

También nos da información sobre el formato de cada variables, fecha y estación aparecen como variables con caracteres, mientras que plomo es una variable numérica con tres decimales.

con la instrucción **list** se peuden visualizar las observaciones:

```
     fecha          estación     plomo
1.   06/22/93        CHA          0.009
2.   05/05/93        CHA          0.012
3.   06/16/93        CHA          0.013
4.   08/09/93        CHA          0.014
5.   05/23/93        CHA          0.015
6.   05/17/93        CHA          0.016
7.   08/03/93        CHA          0.016
8.   06/29/94        CFE          0.017
9.   08/28/94        CHA          0.018
10.  07/04/93        CHA          0.018
11.  04/18/94        CHA          0.020
12.  08/15/93        CHA          0.022
13.  03/18/93        PED          0.023
```

En este paso sería conveniente documentar las variables, esto se puede hacer con la instrucción **label var** . La documentación de las variables es muy importante, ya que durante el análisis se crean diferentes expresiones de las mismas, lo cual puede ser una fuente de error si no existe una documentación apropiada.

```
. label var  fecha "fecha de muestreo"
. label var estacion "estacion de monitoreo ambiental"
. label var plomo "plomo en mcg/m3"
. desc
```

```
Contains data from :untitled::plomoddf.dta
     Obs: 3951   (max=   16700)
    Vars:    3   (max=     236)            23 Jul 1996 22:00
   Width:   16   (max=     290)
 1. fecha       str8      %8s     fecha de muestreo
 2. estacion    str4      %4s     estacion de monitoreo ambiental
 3. plomo       float     %6.3f   plomo en mcg/m3
Sorted by:
Note:  Data has changed since last save
```

Las estadísticas descriptivas se pueden estimar mediante la instrucción **sum.** Estas estadísticas son de mucha utilidad, ya que permiten evaluar los valores máximos y mínimos, así como los puntos de corte para los percentiles más utilizados. La "Skewness" y la "Kurtosis" proporcionan información sobre la simetría de la distribución.

La instrucción tiene dos versiones una de formato abreviado y otra de formato extenso. Esta última se obtiene mediante la opción **detail**

```
. sum plomo

Variable   Obs       Mean     Std. Dev.  Min        Max

plomo      3951      0.802    0.814      0.009      8.870
```

```
. sum plomo,detail

                plomo en mcg/m3
      Percentiles  Smallest
1%    0.046        0.009
5%    0.109        0.012
10%   0.150        0.013     Obs           3951
25%   0.263        0.014     Sum of Wgt.   3951
50%   0.600                  Mean          0.802
                   Largest   Std. Dev.     0.814
75%   1.042        7.600
90%   1.631        8.389     Variance      0.663
95%   2.260        8.784     Skewness      3.198
99%   4.050        8.870     Kurtosis      20.297
```

La información que se obtiene con esta instrucción permite estimar las estadísticas descriptivas tradicionales, los valores de los percentiles, los valores mínimos y máximos, la desviación estandar y la varianza, y los estadísticos de simetría (Skewness, valor esperado es 0 cuando la distribución es perfectamente simétrica y Kurtosis cuyo valor esperado es de 3, cuando la distribución es normal).

Los percentiles son estadísticas que indican la posición de diferentes valores en relación al resto de las observaciones. Se obtienen al ordenar las observaciones de menor a mayor de esta manera la mediana percentil 50 - 0.600 -, indica el punto medio de las observaciones, el percentil 25 - 0.263 - indica el punto donde terminan el 25% de las observaciones. Estas estadísticas son útiles ya que en ciertos tipos de análisis estadísticos es conveniente re-expresar las variables en percentiles, por ejemplo: en terciles, cuartiles o quintiles, dependiendo del número de observaciones.

Otras estadísticas que describen el centro de la distribución, se refieren a las medias armónica y geométrica. Se pueden obtener mediante la instrucción**means**

```
. means plomo

        Arithmetic          Geometric          Harmonic
      Mean        Obs     Mean        Obs     Mean        Obs

plomo .8019501    3951    .5285255    3951    .3157208    3951
```

La definición de las diferentes medias es la siguiente:

$$\text{media arithmétrica} \quad \frac{\sum x_i}{n}$$

$$\text{media geométrica} \quad e^{\frac{\sum \ln(x_i)}{n}}$$

$$\text{media harmónica} \quad \frac{n}{\sum \frac{1}{x_i}}$$

En general, se espera que la media aritmética sea mayor la geométrica y la geométrica sea mayor que la armónica. Esto cuando se refieren al mismo número de observaciones (ej valores positivos), en el caso de la geométrica.

Existen otros estimadores del centro de la distribución que se basan en la exclusión de cierta proporción de los valores extremos.Estos estimadores se conocen como *"trimmed means"* en ingles o medias recortadas en español.

La manera de estimar las *medias recortadas* se puede entender fácilmente comparando la manera de estimar la media y la mediana. Para estimar la media se utilizan todas las observaciones, se asume que todos los valores observados tienen un peso específico igual a 1. Mientras que para la mediana se utiliza solo un valor por lo que se asume un peso 0 para el resto de las observaciones. Cuando no existe un punto medio definido, se utilizan dos observaciones con peso de 0.5 cada una.

Para las medias recortadas se define peso 0 para cierta proporción de los datos. Es posible definir medias recortadas (MR), una MR (0.0) es equivalente a la media aritmética. La mediana se obtiene al eliminar **1/2-1/(2n)** observaciones. La media recortada MR (0.05) elimina el 5% de las observaciones (el percentil inferior a 2.5 y superior 97.5). Para eliminar las observaciones es necesario ordenar la variable de mayor a menor y eliminarlos valores extremos que corresponden al porcentaje de observaciones que se requiere eliminar Al comparar las medias con diferentes proporciones de exclusión de datos, se puede evaluar el efecto de los valores extremos sobre la media.

Para obtener los valores dè las *medias recortadas* es necesario excluir los valores de acuerdo al punto de corte que se requiere. Como ejemplo hemos utilizados los puntos MR (0.02), excluyendo los valores menores al punto de corte correspondiente al percentil 1 y superiores correspondientes al percentil 99. Estos valores 0.046 y 4.050 se pueden obtener del **sum, detail.** Los valores de la media y media recortada (0.02, 0.10, 0.20, y 0.50) para los datos de plomo en aire observados en la Ciudad de México 1988-93 son:

```
. sum plomo

Variable Obs      Mean    Std. Dev.   Min      Max

plomo    3951     0.802   0.814       0.009    8.870

. sum plomo if plomo>0.046 & plomo<4.050

Variable Obs      Mean    Std. Dev.   Min      Max

plomo    3872     0.761   0.652       0.046    3.928

. sum plomo if plomo>0.109 & plomo<2.260

Variable Obs      Mean    Std. Dev.   Min      Max

plomo    3555     0.699   0.482       0.109    2.260

. sum plomo if plomo>0.150 & plomo<1.631

Variable Obs      Mean    Std. Dev.   Min      Max

plomo    3161     0.659   0.386       0.150    1.628

. sum plomo if plomo>0.263 & plomo<1.042

Variable Obs      Mean    Std. Dev.   Min      Max

plomo    1977     0.612   0.232       0.263    1.041
```

Es interesante comparar estos valores con los de la mediana (0.600) y las medias armónica y geométrica (.528 y .315). Se puede observar como estos estimadores de la media son mas resistentes al efecto de valores extremos, y como tienden a disminuir conforme eliminamos algunas observaciones. Sin embargo, pese a excluir casi al 50% de los datos la media recortada es de 0.612.

El comando **sum**, se puede combinar con la instrucción**tab**, esto puede ser de utilidad para describir los valores que toma las variables de acuerdo a las categorías de otras variables.

En nuestro ejemplo:

```
. tab est,sum(plomo)

        Summary of plomo
est     Mean      Std. Dev. Freq.
CES     0.627     0.533     362
CFE     0.622     0.512     436
CHA     0.218     0.242     195
FAN     0.651     0.640     454
LPR     2.357     0.000     1
MCM     0.834     0.573     397
MER     0.849     0.627     413
NET     0.599     0.370     35
PED     0.514     0.427     415
POT     0.336     0.314     7
SHA     0.543     0.000     1
TAX     1.260     0.000     1
TEC     0.973     0.716     384
TLA     0.738     0.641     366
UIZ     0.745     0.268     40
XAL     1.768     1.510     411
XCH     0.286     0.103     33
Total   0.802     0.814     3951
```

```
. tab  aa,sum(plomo)

        Summary of plomo
aa      Mean      Std. Dev. Freq.
1986    0.697     0.501     65
1987    1.242     0.645     95
1988    1.541     1.209     388
1989    1.336     0.893     470
1990    1.184     0.752     467
1991    0.900     0.698     676
1992    0.547     0.451     508
1993    0.292     0.274     538
1994    0.250     0.268     596
1995    0.226     0.142     148
Total   0.802     0.814     3951
```

Este tipo de tabulaciones es muy útil ya que proporciona información sobre la frecuencia relativa de la distribución (freq) y el promedio y desviación estandar asociada a esta frecuencia.

Con el comando aNOVA, se podría obtener la prueba estadística para evaluar diferencia de medias.

II. GRAFICO DE LETRAS (INSTRUCCION 1v)

El diagrama de letras se basa principalmente en el ordenamiento de los datos, **de menor** a mayor, y en el cálculo de diferentes estadísticos que evalúan el impacto de los extremos de la distribución, "de las colas", de los datos, asumiendo diferentes puntos de corte. El nombre de diagrama de letras se origina en el hecho de que a cada punto de corte se le ha asignado una letra.

El procedimiento para obtener los estadísticos de diagrama de letras, consiste en ordenar los datos -de menor a mayor- y en extraer información sobre los valores que definen el punto medio (la mediana), los que definen los cuartos, es decir los percentiles 25 y 75; los octavos con los percentiles 12.5 y 87.5, los y dieciseisavos, los treintadosavos, y así sucesivamente, en el siguiente cuadro se muestran percentiles de corte superior e inferior

				Punto de corte en %	
Fracción de corte	Símbolo	%	fracción	Inferior	superior
Mediana	M	0.5	1/2	50.0	50.0
Cuartiles	F	0.25	1/4	25.0	75.0
Octiles	E	0.125	1/8	12.5	87.5
Dieciseiciles	D	0.0625	1/16	6.25	93.75
Treintaidosciles	C	0.03125	1/32	3.125	96.87
Sesentaicuatrociles	B	0.01562	1/64	1.56	98.44
Cientoveintiochoavos	A	0.00781	1/128	0.78	99.22

Como ya se mencionó, a cada punto de corte se le ha asignado una letra, esta asignación es arbitraria, es decir no sigue un orden particular pero es la que se usa convencionalmente en la representación gráfica.

A continuación se examinará el diagrama de letras para una de las variables de estudio:

```
. lv plomo

#   3951    plomo en mcg/m3

M   1976            0.600                spread   pseudosigma
F   988.5   0.263    0.652    1.041       0.778    0.577
E   494.5   0.170    0.815    1.461       1.291    0.561
D   247.5   0.122    1.073    2.023       1.901    0.620
C   124     0.084    1.405    2.727       2.643    0.710
B   62.5    0.059    1.778    3.497       3.439    0.799
A   31.5    0.042    2.141    4.240       4.198    0.869
Z   16      0.030    2.760    5.490       5.460    1.028
Y   8.5     0.017    3.368    6.719       6.702    1.168
X   4.5     0.015    3.775    7.536       7.521    1.223
    1       0.009    4.439    8.870       8.861    1.240

                                          # below  # above
inner fence -0.904           2.209        0        210
outer fence -2.072           3.376        0        67
```

La primera línea **# 3951 plomo en mcg/m³** muestra el número de observaciones y la etiqueta de la variable.

La segunda línea, **M 1976 | 0.600** contiene información sobre la mediana y el número de observaciones que se encuentran por debajo de la mediana. En este caso la mediana es de 0.600 y por debajo de este valor existen 1976 observaciones. En la segunda línea aparecen las estadísticas asociadas con los cuartos, lo que corresponde a la letra **F 988.5 0.263 0.652 1.041**. Existen 988.5 observaciones del punto de corte al valor mas extremo, el punto de corte del percentil 25 es 0.263 y el del percentil 75 es 1.041. La columna del centro presenta el promedio de los valores de corte, en este caso (0.263+ 1.041)/2= 0.652. Si la distribución fuese perfectamente simétrica, se esperaría que los punto medios fueran iguales a la mediana. En este ejemplo se puede observar que el punto medio varía de 0.600 a 4.43, lo que sugiere que la distribución de la variable no se ajusta a una distribución normal. El "spread" o dispersión, se obtiene al calcular la diferencia entre el valor del límite superior y el inferior en este caso 1.041-0.263=0.778. En este caso se trata de una estadística que tiene mucha utilidad, la diferencia intercuartil. Posteriormente ejemplificaremos su uso. El estadístico pseudosigma es una estimación de la desviación estándar -para el cálculo se asume que la variable se distribuye normalmente- utilizando los valores que quedaron en los extremos de cada punto de corte.

Cuando la variable tiene una distribución normal, los valores para los diferentes puntos de corte deben ser similares. En la interpretación de los valores de la pseudosigma se puede inferir lo siguiente: a) si se observan valores decrecientes, se puede concluir que tiene menor dispersión que la distribución normal; b) si se incrementa ello indicaría mayor dispersión; ambos comportamientos indican asimetría en la distribución.

En la parte inferior del diagrama se presenta información sobre los valores que se encuentran separados de la nube de puntos. Es importante detectar y estudiar estos valores, ya que dentro del análisis estadístico merecen atención especial puesto que pueden tener un impacto importante sobre los resultados y conclusiones, es decir pueden ser observaciones "influyentes". Como ya se mencionó, estos valores pueden deberse a errores reales, en cuyo caso deben corregirse o excluirse del análisis, o a valores reales, con cierta plausibilidad, en cuyo caso deben incluirse en el análisis y evaluarse en términos del impacto que tienen sobre los resultados y conclusiones. Una alternativa es excluirlos del análisis final y evaluar la diferencia en los resultados; más adelante veremos la diferentes acciones que se pueden realizar para evaluar y ajustar el impacto de las observaciones influyentes.

Como convención, utilizando medidas robustas (es decir poco sensibles a los efectos de los valores extremos) , se definen dos puntos de corte para clasificar las observaciones; las observaciones que quedan separadas por estos puntos de corte merecen atención especial.

Los puntos de corte presentan dos categorías que marcan lejanía hacia la nube de puntos. Se manejan dos puntos de corte basados en el rango intercuartil, el rango que existe entre el valor del percentil 25 y el del percentil 75. Para el caso de la variable plomo en aire, del diagrama de caja podemos obtener estos valores, son:.0.263 y 1.041. Los puntos de corte se definen como **límite interno,** que identifica los puntos que podrían ser considerados como valores aberrantes o "outliers" y el **limite externo,** que identifica los valores con una alta probabilidad de ser aberrantes. Si las observaciones se originaran de una distribución normal, los valores para el límite interno equivaldrían a -2.698σ y a +2.68 σ y para los límites externos a -4.721 σ y a +4.721 σ.

Para definir los puntos de corte de identificación de observaciones que potencialmente pueden ser valores aberrantes, se utiliza el valor del rango intercuartil dado que es una medida robusta que no se afecta por la presencia de valores extremos, a diferencia de la desviación estándar o la dispersión (rango). Los límites interno (inner fence) y externo (outer fence) se

definen de la siguiente manera:

Diferencia intercuartil	*DI= C75 - C25*
Limite interno inferior	*LIi = C25 - 1.5 x DI*
Limite interno superior	*LIs = C75 + 1.5 x DI*
Limite externo inferior	*LEi = C25 - 3.0 x DI*
Limite externo superior	*LEs = C75 + 3.0 x DI*

Diferencia intercuartil DI = 1.041-0.263=0.778

Limite interno inferior (inner fence inferior) = 0.263 - 1.5 x 0.778 = -0.904
Limite interno superior (inner fence superior) = 1.041 + 1.5 x 0.778 = 2.209
Limite externo inferior (outer fence inferior) = 0.263 - 3.0 x 0.778 = -2.072
Limite externo superior (outer fence superior)= 1.041 + 3.0 x 0.778 = 3.376

Por ejemplo, para los percentiles 25, se obtiene con +-.6745, lo que daría 0.778/1.349 = 0.577. La pseudo-sigma se obtiene al dividir la dispersión (el spread) entre la desviación derivada de una distribución normal.

Para identificar las observaciones se puede realizar un "list", estableciendo los puntos de corte calculados para los valores de los puntos de corte. En el ejemplo anterior:

```
.      . list if plomo>3.376

        fecha          estacion     plomo
135.    01/11/88       XAL          6.777
163.    01/12/90       XAL          3.778
202.    01/14/88       XAL          4.422
203.    01/14/89       CES          3.530
211.    01/14/89       XAL          4.180
228.    01/16/91       XAL          3.621
235.    01/17/88       XAL          5.036
245.    01/17/92       XAL          3.778
267.    01/18/90       XAL          6.661
277.    01/18/94       XAL          3.435
306.    01/20/88       XAL          4.664
315.    01/20/89       XAL          5.490
333.    01/22/91       XAL          3.534
387.    01/25/91       XAL          5.298
414.    01/26/89       XAL          4.160
429.    01/28/91       TEC          3.552
462.    01/30/90       XAL          4.124
481.    01/31/91       XAL          5.578
514.    02/02/90       XAL          4.679
626.    02/10/92       XAL          3.830
698.    02/14/90       XAL          3.631
```

740.	02/17/90	XAL	5.741
945.	03/02/91	XAL	4.134
964.	03/03/89	XAL	3.640
1015.	03/07/90	XAL	4.775
1032.	03/08/88	XAL	3.559
1051.	03/09/89	XAL	4.250
1067.	03/11/88	XAL	6.261
1075.	03/11/91	XAL	4.650
1117.	03/14/88	XAL	5.501
1139.	03/15/89	XAL	4.500
1154.	03/17/88	XAL	4.131
1231.	03/21/89	XAL	7.600
1311.	03/28/90	XAL	5.216
1377.	04/03/90	XAL	4.230
1650.	05/02/89	XAL	3.880
1659.	05/03/90	XAL	3.790
1713.	05/08/89	XAL	3.430
1831.	05/19/88	XAL	3.883
1948.	05/31/88	XAL	3.928
2414.	07/20/90	XAL	3.583
2496.	07/27/92	XAL	3.833
2713.	08/18/89	FAN	8.870
2945.	09/10/88	XAL	3.692
3117.	09/28/88	XAL	4.951
3224.	10/13/91	XAL	7.395
3317.	10/25/91	XAL	5.371
3367.	10/31/91	XAL	3.576
3406.	11/03/88	XAL	6.846
3461.	11/09/88	CFE	3.857
3464.	11/09/88	MER	5.593
3466.	11/09/88	TEC	4.050
3467.	11/09/88	TLA	7.471
3468.	11/09/88	XAL	8.389
3477.	11/10/89	XAL	3.910
3667.	12/03/88	MCM	3.465
3668.	12/03/88	MER	3.447
3671.	12/03/88	XAL	8.784
3690.	12/06/91	XAL	4.160
3787.	12/16/89	TEC	3.660
3789.	12/16/89	XAL	5.300
3791.	12/18/87	TEC	3.380
3835.	12/21/88	MER	4.230
3837.	12/21/88	TEC	3.606
3839.	12/21/88	XAL	5.115
3847.	12/22/89	XAL	5.690
3903.	12/28/88	XAL	4.540

Se puede observar que los valores ocurrieron durante 80's en las estación del norte de la ciudad, lo que es posible y concuerda con lo esperado, por lo que los valores tienen plausibilidad. Al verificar en las bases originales, los valores son los mismos, de tal manera que se optó por dejar los valores para análisis subsecuentes. Sin embargo es importante tomar nota y evaluar el impacto de estas observaciones en las fases subsecuentes del análisis. En resumen el diagrama de caja presenta un método útil y sencillo para describir la información.

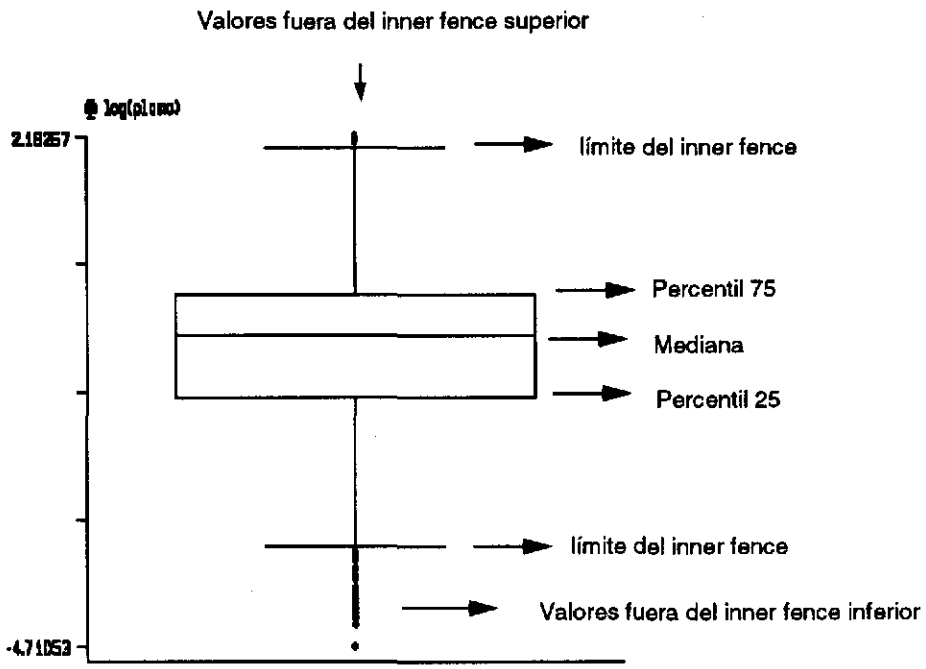
III. GRAFICO DE CAJA (BOXPLOTS) (GRAPH VARIABLE, BOX)

Este tipo de gráfico es una representación simple de la información, de la cual se puede obtener fácilmente la siguiente información:

- la localización del centro de los datos (la mediana)
- la dispersión
- la simetría
- la extensión de los extremos (colas de la distribución)
- la existencia de valores aberrantes (outliers)

La sencillez de este gráfico lo convierte en un buen instrumento para realizar comparaciones entre diferentes categorías, por ejemplo, entre los niveles de contaminación observados en diferentes años.

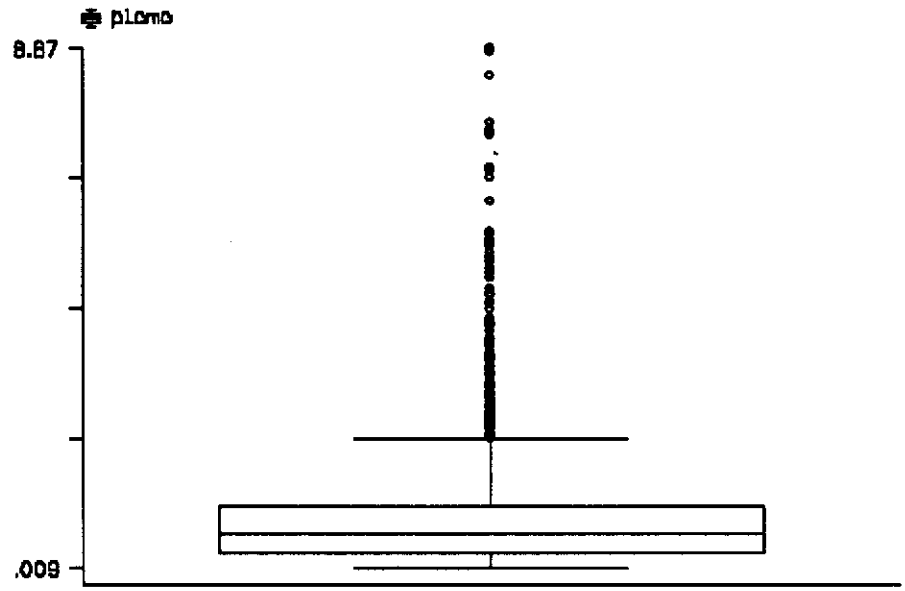
Estructura del diagrama de caja:



En algunos programas de análisis estadístico se pueden definir diferentes tipos de diagramas de caja. Por ejemplo, se pueden substituir los límites intercuartiles por el error estandar de la media y los bigotes por la desviación estandar de la muestra. Sin embago, es pertinente mencionar que este tipo de gráficos es susceptible de sesgo debido a la influencia de valores extremos. La ventaja del diagrama de caja, basado en los rangos intercuartiles, es que es resistente al impacto de valores extremos. De hecho, podrían presentarse valores extremos en el 25% de las observaciones y no tener un impacto importante sobre los límites de la caja. En relación con los límites para detectar valores aberrantes, éstos se definen de manera arbitraria. De una distribución normal, se esperaría que únicamente el 0.7% de las observaciones tomarán valores superiores a los puntos de corte definidos con el rango intercuartil. Los valores de corte, concuerdan con los obtenidos en el diagrama de caja. Los bigotes del diagrama de caja, corresponden a los límites establecidos para el inner fence superior e inferior

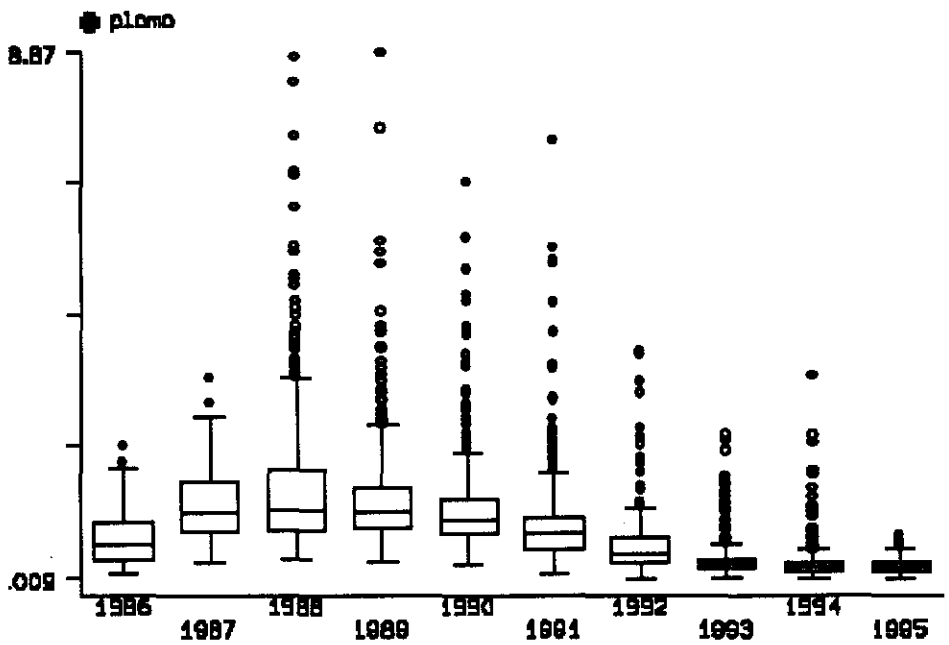
Como ejemplo hemos graficado utilizando el diagrama de caja, la información sobre plomo atmosférico. Se observa una gran asimetría, con un número considerable de valor que potencialmente pueden ser valores aberrantes. La distribución observado concuerda con la información derivada del diagrama de letras.

Ejemplo de los datos de plomo



Ejemplo de la utilidad que puede tener el gráfico de caja para comparar la distribución de los valores observados, se graficaron los valores observados de acuerdo con el año de medición. Para poder graficar por el año de medición primero se deben ordenar los datos de acuerdo a la variable de clasificación, en este caso el año (en la base se llama aa), posteriormente se usa el comando graph como se anota a continuación:

```
. sort aa
. graph plomo, box by(aa)
```



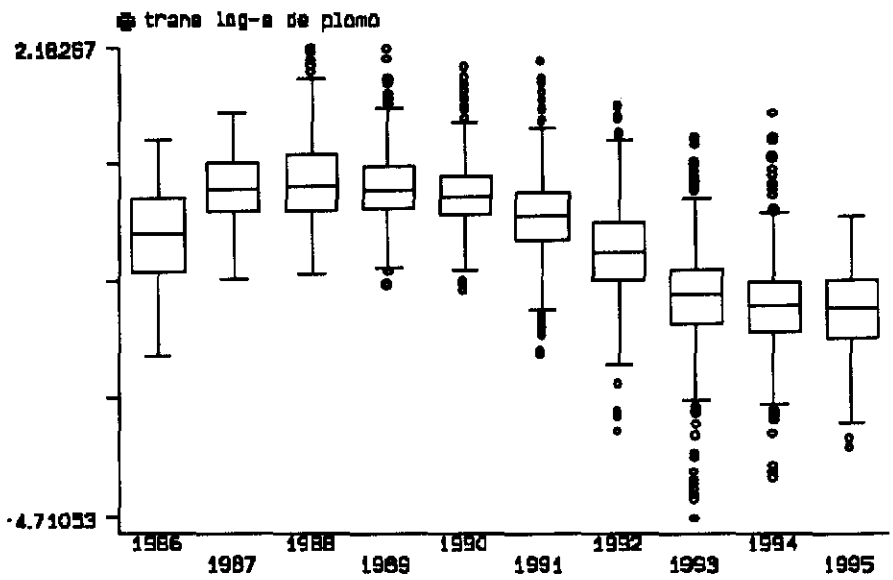
En este gráfico se pueden observar claramente las tendencias de la concentración de plomo en aire, así como las distribuciones anuales de los valores. También se puede apreciar que existe una relación entre la dispersión y el tiempo. La dispersión tiende a disminuir conforme avanza el tiempo.

En años recientes se observa una dispersión menor de los valores. Este patrón podría sugerir la necesidad de una transformación, es decir, de re-expresar los valores observados para lograr una dispersión similar, logrando una mejor representación gráfica y datos mas apropiados para los análisis estadísticos tradicionales, como el de varianza y la regresión lineal.

En el análisis de varianza se hace la suposición sobre igualdad de varianzas dentro de los diferentes grupos de comparación. Con el fin de ejemplificar el efecto de la transformación , utilizaremos la transformación logarítmica (log-e). Esto se puede hacer generando una nueva variable.

Las instrucciones en stata son:

```
gen lnplomo=ln(plomo)
label var lnplomo "trans log-e de plomo"
graph lnplomo, box by(aa)
```



Se observa que mejora considerablemente la imagen comparativa, también se observa que disminuye las diferencias en la dispersión de los valores, de acuerdo con el año calendario.

IV. DIAGRAMA TALLO-HOJA INSTRUCCION STEM

En su estructura más simple, se trata de una serie de números. La presentación del tipo de tallo-hoja permite explorar la estructura de los datos, mediante este gráfico se puede evaluar:

- Si la estructura es simétrica
- La dispersión
- Situación especial de algún valor
- Concentración de datos
- Valores faltantes dentro de la serie
- Patrones de dispersión y errores de dígitos

El procedimiento para construir este tipo de gráfico es muy simple y consiste en una presentación de los datos ordenados de mayor a menor Así por ejemplo, en el caso de los datos de plomo, que son muy numerosos este gráfico no es de mucha ayuda, ya que gráfica cada observación. Ejemplificaremos este tipo de gráfico con las mediciones de plomo en sangre. La variable mpb[3] contiene información sobre las concentraciones de plomo en sangre de 173 mujeres.

```
. stem mpb3

Stem and leaf plot for mpb3 in units of 0.1

 2 | 9
 3 | 222
 4 | 000022578889
 5 | 111222344566799
 6 | 0133344555666899
 7 | 00112334444678888899
 8 | 0112222334455777889
 9 | 1111122223344677889
 1 ||
00,00,00,00,00,03,03,04,05,06,07,07,07,10,12,13,14,14,15,15,16,16,17,17
 1 ||
18,19,19,19,20,23,24,24,24,25,27,29,29,31,31,33,34,40,42,43,44,44,44,45
 1 || 46,49,50,58,59,61,62,63,66,68,69,70,70,71,89,95
 2 || 03,10,24,30,31,33,39,45,91
```

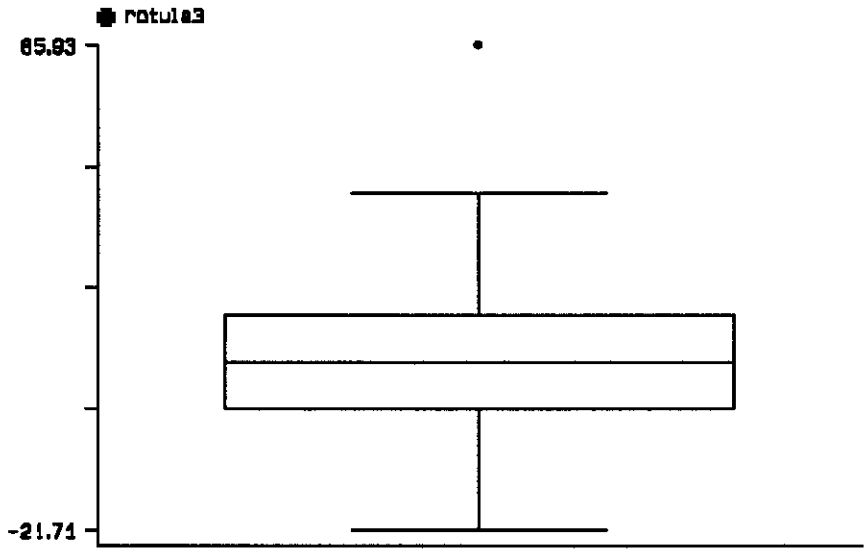
2 | 9 este primer valor corresponde a el valor mínimo de 2.9
3 | 222 estos valores corresponden a tres observaciones de 3.20

2 || 03,10,24,30,31,33,39,45,91. Estos valores corresponde a 20.3, 21.0, 22.4, 23.0, 23.1, 23.3, 23.9, 24.5 y 29.1.

La instrucción que se utiliza es: . **Stem VARIABLE,** para otra variable, por ejemplo la distribución de plomo en hueso. Actualmente es posible determinar la concentración de plomo en hueso mediante refracción de rayos fluorescente. La medición tiene cierto grado de error, de hecho cuando se aproxima al nivel de detección puede estimar valores negativos.A continuación se presenta el diagrama de tallo-hoja para los valores que se obtuvieron al medir plomo en hueso en una muestra de 176 mujeres en edad reproductiva.

```
.Stem and leaf plot for rotula3

-2 | 10
-1 | 330
-0 | 8776553311111000
 0 | 00000011112222233344444555556667777788889999
 1 | 000111111112222233444455555566778899
 2 | 0000111111122222333333333344555556667777888899 9
 3 | 01111122233444567788
 4 | 001126
 5 | 3
 6 |
 7 |
 8 | 5
```

Se puede observar que la distribución es simétrica y que va desde -21.0 hasta 85. Que existe una discontinuidad de los datos de 53 a 85. Cuando se realizan los diagramas de tallo-hoja a mano, la manera de calcular el número de intervalos y la amplitud de los intervalos es la siguiente: para el número de intervalo es $\mathbf{L=[10x\log_{10}n]}$ y para la amplitud del intervalo se divide L entre la amplitud de valores observados en los datos. Para el caso de los datos de plomo en hueso $\mathbf{L= L=[10x\log_{10}(173)]=22.4,}$ se estiman 22 intervalos como máximo; como la amplitud de los datos va de -21.0 a 85.93, se estima una amplitud de 4.7. Otro método para estimar el número de intervalos es raíz de n, en este ejemplo 13. Para fines comparativos, vamos analizar la misma información con el diagrama de caja y con el gráfico de caja.




```
. lv rotula3

#    176                rotula3

M    88.5               15.22               spread    pseudosigma
F    44.5     4.98      15.32     25.67     20.69     15.38
E    22.5     0.31      16.15     31.99     31.68     13.81
D    11.5    -4.80      16.43     37.67     42.48     13.90
C    6       -8.73      16.41     41.55     50.28     13.59
B    3.5    -13.67      15.55     44.77     58.44     13.93
A    2      -20.56      16.24     53.04     73.60     15.68
Z    1.5    -21.13      24.18     69.49     90.62     18.29
     1      -21.71      32.11     85.93     107.64    20.25

                                             # below   # above
inner fence -26.07               56.71       0         1
outer fence -57.11               87.75       0         0
```

Se puede observar la concordancia entre las diferentes medidas sumarias o gráficos. Se detecta un posible valor aberrante, la distribución parece ser simétrica y parece que los datos se pueden aproximar a una distribución normal.

V. TRANSFORMACIONES

La transformación de los datos -reexpresión- se utiliza principalmente con el fin de mejorar la representación gráfica o estadística o para facilitar el análisis estadístico.

Una de las aplicaciones del análisis exploratorio de datos, es la evaluación de la necesidad de realizar transformaciones. Las principales razones para realizar transformaciones son:

a) Normalizar las distribuciones
b) Ganar interpretabilidad
c) Corregir asimetrías fuertes
d) Categorías con dispersiones diferentes
e) Residuales influyentes (detectados en regresión lineal)

Las transformaciones más frecuentemente usadas son:

$$T_p(x) = \begin{cases} ax^p + b & \text{cuando } p \neq 0 \\ c\log x + d & \text{cuando } p = 0 \end{cases}$$

Se trata de transformaciones fuertes y, en general, cambian la forma de los datos; forman parte de un grupo conocido como transformaciones de potencia, que tienen la siguiente forma:

Se requiere que a, b, c, d y p sean números reales; y que $a>0$ para $p>0$ y $a<0$ para $p<0$. Con estas condiciones se asegura lo siguiente:

a) Se conserva la secuencia original de orden en los datos
b) Se conservan los valores asociados a las letras, en el diagrama de letras.
c) Son funciones continuas
d) Son funciones sin variaciones bruscas
e) Se utilizan transformaciones simples, que pueden re-expresarse sin dificultad

En general trabajaremos con transformaciones mas simples, que se definen con la siguiente función

$$T_P(x) = \begin{cases} x^p & \text{cuando } p > 0 \\ \log x & \text{cuando } p = 0 \\ -x^p & \text{cuando } p < 0 \end{cases}$$

Las transformaciones llevan la información a escalas que no resultan familiares por lo que, en general, se pierde interpretación. Los problemas sugen principalmente en el área de la interpretación y no tanto en la de análisis. Por las razones anteriores, solo se deben transformar los datos cuando:

a) Existe una dispersión muy amplia en los datos. Si la relación entre el valor menor y el mayor es superior a 20, es probable que la transformación tenga un buen efecto.
b) Se encuentran residuales con valores grandes (caso de regresión lineal)
c) Existen asimetrías importantes

Entre los usos que se pueden hacer de las transformaciones, está el de lograr "normalidad", es decir, que los datos se distribuyan de acuerdo con la distribución normal. Para evaluar en forma inicial si las observaciones se apegan a esta distribución, se mencionaron anteriormente los resultados que se obtienen del diagrama de letras. En este gráfico, si la distribución se apega a la normalidad, se esperaría que los valores de la pseudosigma fuesen constantes en las estimaciones asociadas a las diferentes letras.Veamos el ejemplo para plomo en leche

```
. lv lpb3

 #  179      conc pb en leche materna

M   90              0.48              spread    pseudosigma
F   45.5    0.31     0.58     0.85     0.54      0.40
E   23      0.21     0.66     1.12     0.91      0.40
D   12      0.13     0.77     1.42     1.29      0.43
C   6.5     0.05     1.06     2.08     2.03      0.56
B   3.5     -0.04    1.56     3.15     3.19      0.76
A   2       -0.17    1.67     3.51     3.68      0.78
Z   1.5     -0.19    1.70     3.60     3.79      0.76
    1       -0.21    1.74     3.68     3.89      0.73

                                       # below   # above
inner fence -0.50             1.66     0         10
outer fence -1.31             2.47     0         4

.
```

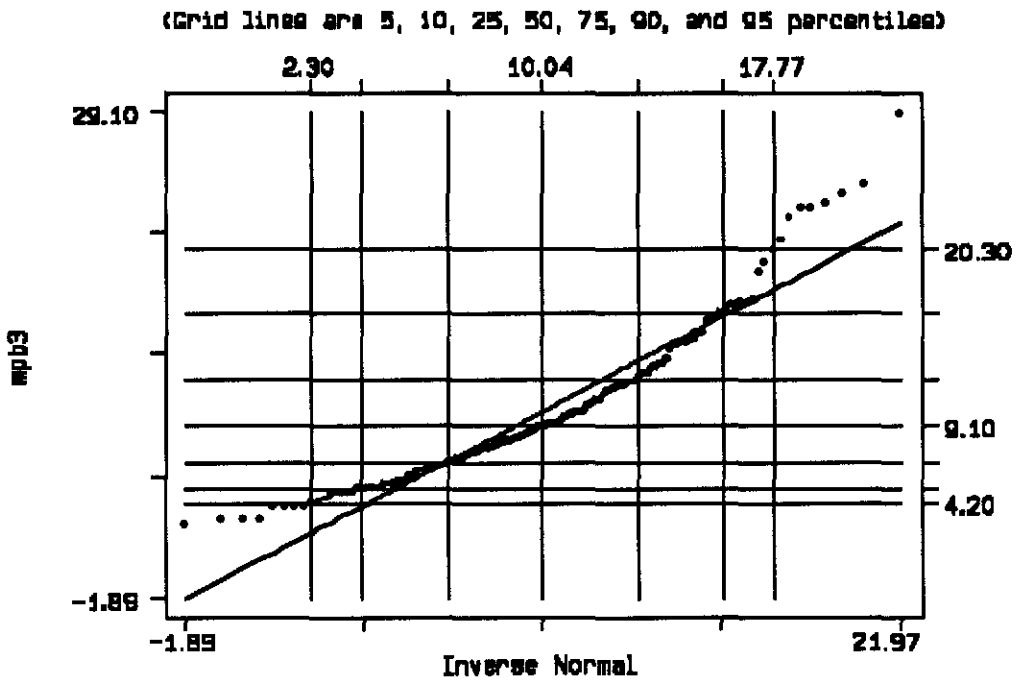
En este caso se aprecia un cambio importante en los valores de la pseudo-sigma lo que sugiere que la distribución no es normal.

Diagnósticos gráficos de normalidad

Los diagnósticos gráficos de normalidad mas utilizados son:

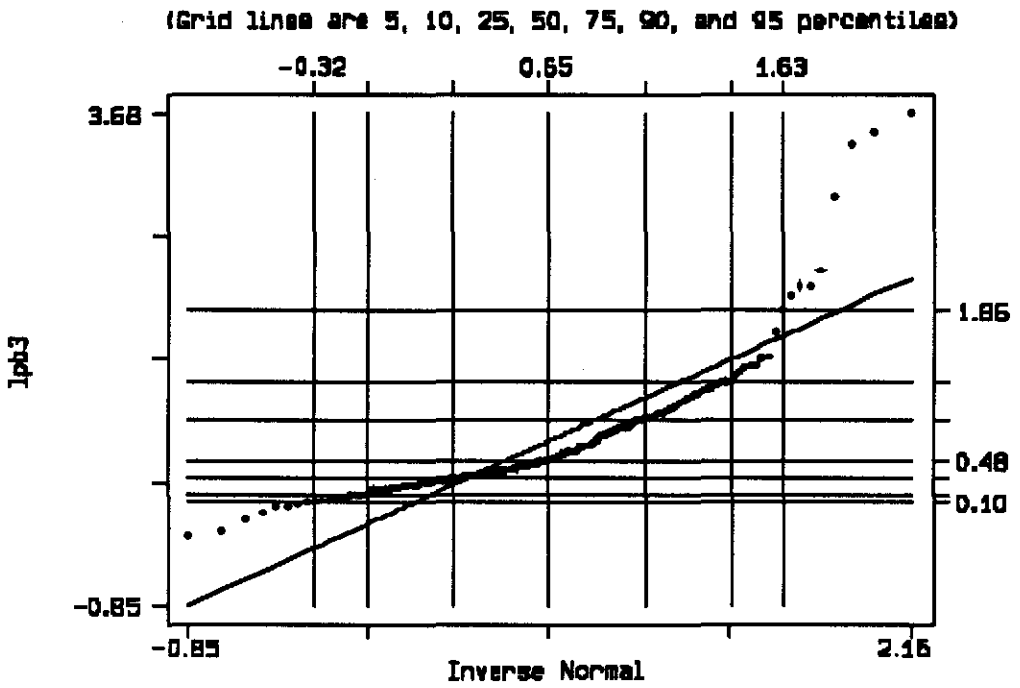
1) Gráfico de la distribución observada vs una distribución normal, este gráfico se puede obtener con la instrucción **qnorm**. Mediante esta técnica se grafican los valores observados con los que se obtienen de la siguiente transformación $(F[x_i-\mu]/\sigma)$.

Por ejemplo para los datos de plomo en sangre:



Se observa que la distribución se ajusta medianamente a la distribución normal, sin embargo existen algunas discrepancias importantes en los extremos de la distribución de los valores observados. En este gráfico el valor esperado de la distribución se marca con la línea sólida que va en diagonal. Conforme los datos se apeguen a esta línea, la distribución es normal.

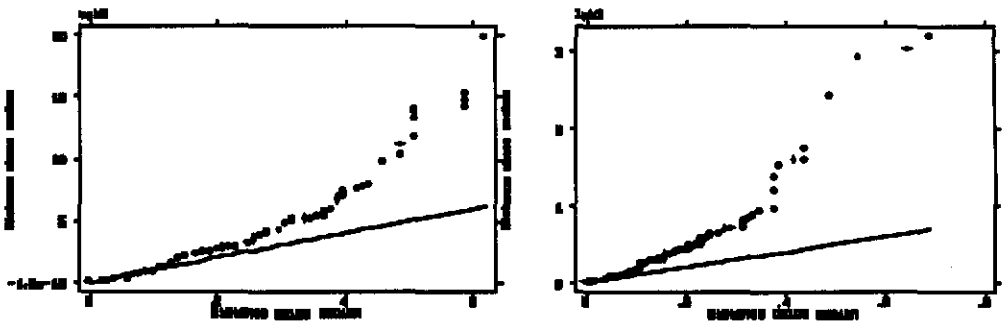
Veamos ahora como comparación la representación gráfica de los valores de plomo en leche materna.



Se puede observar que la distribución no es normal, que existe sobre dispersión especialmente en lo que corresponde a los valores extremos.

Otro diagnóstico gráfico comúnmente utilizado se refiere a los gráficos de simetría:

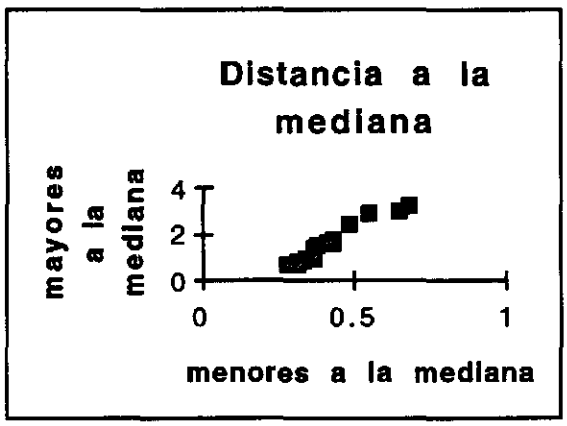
```
symplot mpb3, ylabel xlabel border saving(g1);
symplot lpb3, ylabel xlabel border saving(g2)
.graph using g1 g2
```



En estos gráficos se estima el rango de variación de los diferentes puntos al centro de la distribución, la mediana. Si la distribución es simétrica la distancia entre los puntos que se encuentran por debajo de la mediana y la distancia de los puntos que se encuentran sobre la mediana debe ser igual, la linea solida refleja el valor esperado. Los puntos que se grafican son: mediana-y_i vs $y_{i(\_N+1-i)}$-mediana.

Por ejemplo para los datos de plomo en leche se pueden construir la gráfica de la siguiente manera, para algunos de los puntos observados.

num de obs	valor observado	mediana	num de obs	valor observado	med-obs	obs-med
	-0.21	0.48	179	3.68	0.69	3.2
2	-0.17	0.48	178	3.51	0.65	3.03
3	-0.07	0.48	177	3.40	0.55	2.92
4	-0.01	0.48	176	2.90	0.49	2.42
5	0.04	0.48	175	2.23	0.44	1.75
6	0.04	0.48	174	2.08	0.44	1.6
7	0.06	0.48	173	2.08	0.42	1.6
8	0.09	0.48	172	1.99	0.39	1.51
9	0.10	0.48	171	1.86	0.38	1.38
10	0.10	0.48	170	1.67	0.38	1.19
11	0.10	0.48	169	1.43	0.38	0.95
12	0.13	0.48	168	1.42	0.35	0.94
13	0.14	0.48	167	1.36	0.34	0.88
14	0.14	0.48	166	1.35	0.34	0.87
15	0.15	0.48	165	1.33	0.33	0.85
16	0.16	0.48	164	1.28	0.32	0.8
17	0.16	0.48	163	1.25	0.32	0.77
18	0.16	0.48	162	1.20	0.32	0.72
19	0.19	0.48	161	1.19	0.29	0.71
20	0.19	0.48	160	1.19	0.29	0.71



A veces es difícil decidir en base a un gráfico, por lo que también se puede realizar una prueba estadística de ajuste. En este caso se asume que la distribución es normal y se estima la probabilidad de que los valores observados se deriven de una distribución normal. Este procedimiento tiene la desventaja de que el resultado dependerá del tamaño muestral. Para muestras grandes, pequeñas diferencias serán altamente significativas, mientras que para muestras pequeñas diferencias importantes con una distribución normal pueden pasar desapercibidas.

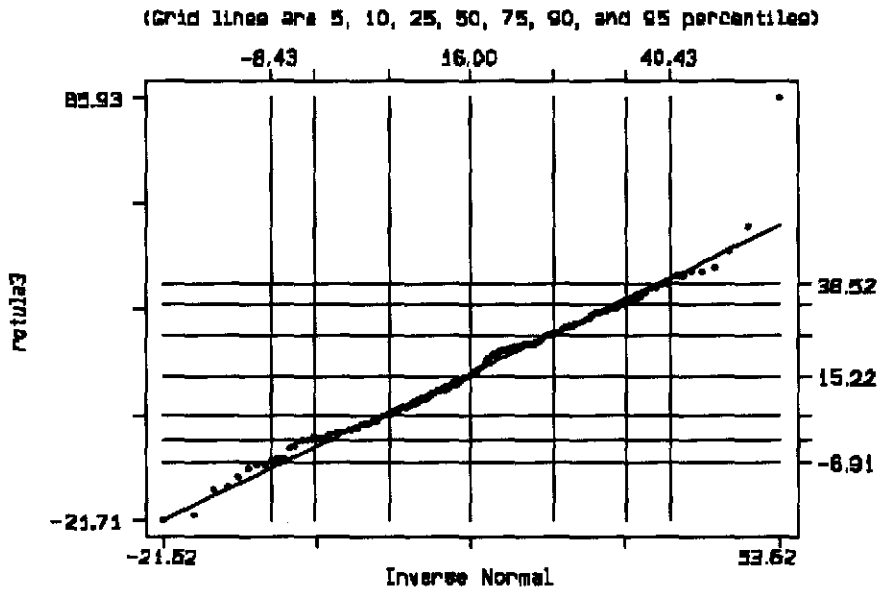
El comando para realizar la prueba de normalidad es**sktest,** esta prueba se basa en la kurtosis (curvatura) y la skewness (simetría). Para las variables de plomo en sangre y plomo en leche se obtienen los siguientes valores:

			—— joint ——	
Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi-sq(2)	Pr(chi-sq)
lpb3	0.000	0.000	.	0.0000
mpb3	0.000	0.001	33.79	0.0000
rotula3	0.007	0.001	14.99	0.0006

Otro estadístico para determinar si la variable se distribuye normalmente es la prueba de Shapiro-Wilk. Esta también se puede estimar fácilmente en Stata. La instrucción para obtener la prueba se Shapior-Wilk es **spwilk**. Para las variables que hemos trabajado los resultados son:

```
. swilk lpb3 mpb3 rotula3

                    Shapiro-Wilk W test for normal data

Variable Obs       W          V          z       Pr > z
lpb3     179       0.76513    31.822     7.916   0.00000
mpb3     178       0.90939    12.218     5.725   0.00000
rotula3  176       0.97373    3.507      2.868   0.00206
```

En este caso, se puede observar que para todas las variables se rechaza la hipótesis de que se ajustan a una distribución normal. Tomando encuesta que el valor esperado para el estadístico V es de 1.0, se puede observar que la variable lpb3 (plomo en leche) presenta los valores mas extremos y que la variable (rotula3) plomo en hueso se acerca mas a una distribución normal, como se observa en el gráfico siguiente.



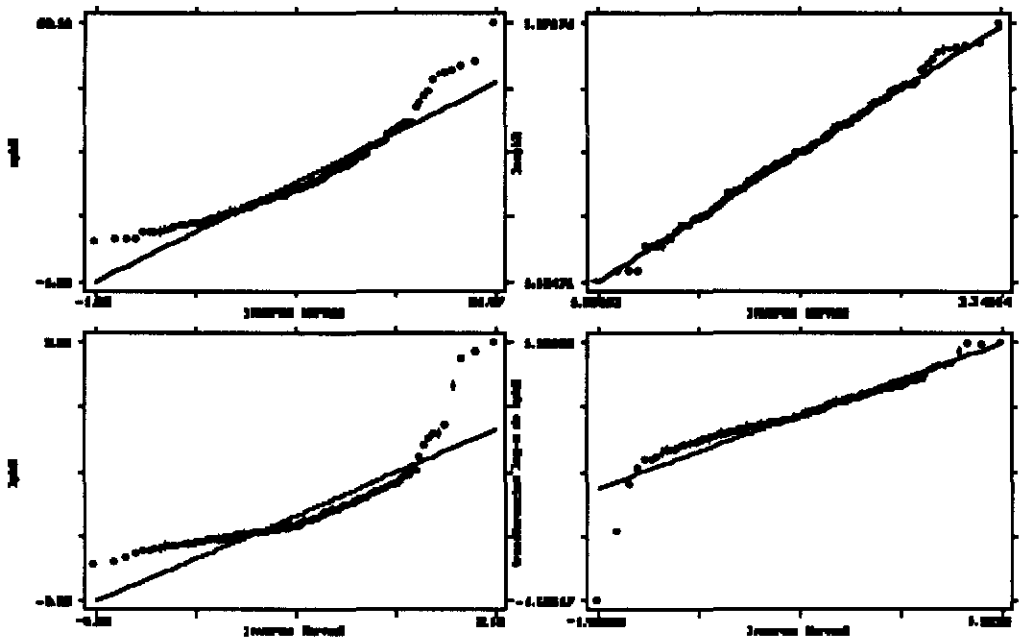
Otra manera de encontrar la mejor re-expresión de la variable para normalizarla (corregir simetría) es ensayar diferentes transformaciones y evaluar cual se ajusta mejor a la distribución normal. Para esto se recomienda utilizar transformaciones a diferentes potencias. Stata mediante el comando **ladder** permite evaluar las transformaciones mas comunmente utilizadas. Aplicando la instrucción **ladder** se obtienen los siguientes resultados:

```
. ladder lpb3

Transformation     formula        Chi-sq(2)    P(Chi-sq)

cube               lpb3^3         .            0.000
square             lpb3^2         .            0.000
raw                lpb3           .            0.000
square-root        sqrt(lpb3)     .            .
log                log(lpb3)      .            .
reciprocal root    1/sqrt(lpb3)   .            .
reciprocal         1/lpb3         .            0.000
reciprocal square  1/(lpb3^2)     .            0.000
reciprocal cube    1/(lpb3^3)     .            0.000


. ladder mpb3

Transformation     formula        Chi-sq(2)    P(Chi-sq)

cube               mpb3^3         .            0.000
square             mpb3^2         .            0.000
raw                mpb3           33.79        0.000
square-root        sqrt(mpb3)     10.82        0.004
log                log(mpb3)      0.03         0.987
reciprocal root    1/sqrt(mpb3)   10.71        0.005
reciprocal         1/mpb3         34.68        0.000
reciprocal square  1/(mpb3^2)     .            0.000
reciprocal cube    1/(mpb3^3)     .            0.000


. sum lpb3

Variable   Obs     Mean     Std. Dev.  Min     Max

lpb3       179     0.65     0.59       -0.21   3.68

. gen lpb3_p=lpb3+0.22
```

```
. ladder lpb3_p

Transformation      formula          Chi-sq(2)    P(Chi-sq)
cube                lpb3_p^3         .            0.000
square              lpb3_p^2         .            0.000
raw                 lpb3_p           .            0.000
square-root         sqrt(lpb3_p)     35.19        0.000
log                 log(lpb3_p)      65.86        0.000
reciprocal root     1/sqrt(lpb3_p)   .            0.000
reciprocal          1/lpb3_p         .            0.000
reciprocal square   1/(lpb3_p^2)     .            0.000
reciprocal cube     1/(lpb3_p^3)     .            0.000
```

Dado que existen algunos valores negativos para plomo en leche (cuando la concentración de plomo en leche es muy baja), se requiere agregar una constante para evitar problemas de indefinición y perdida de los valores que toman las observaciones con valores negativos o iguales a 0. Si seleccionamos en ambas variables la transformación logarítmica veamos que efecto pueden tener sobre la forma de la distribución.

```
gen lnlpb3=ln(lpb3_p)
label var  lnlpb3 "transformacion log-e de lpb3"
gen lnmpb3=ln(mpb3)
label var  lnlpb3 "transformacion log-e de mpb3"
qnorm mpb3, saving(n1)
qnorm lnmpb3, saving(n2)
qnorm lpb3, saving(n3)
qnorm lnlpb3, saving(n4)

graph using n1 n2 n3 n4
```



En ambos casos se puede observar como la transformación logarítmica mejora sustancialmente la distribución de la variables, aun en el caso de la variable plomo en leche (lpb3) en la que aun con la transformación rechazamos estadísticamente la hipótesis sobre su ajuste a una distribución normal.

Cuando el objetivo es encontrar la mejor representación gráfica de una serie de datos que presentan diferencias en los rangos de dispersión, se puede buscar la transformación que mejor ajuste a los datos. Para este fin se puede graficar los rangos intercuartiles y la mediana de los diferentes grupos, utilizando la siguiente técnica:

log(DI) = log c + b log(M)
donde DI es la diferencia intercuartil y
M es la mediana en las diferentes categorías

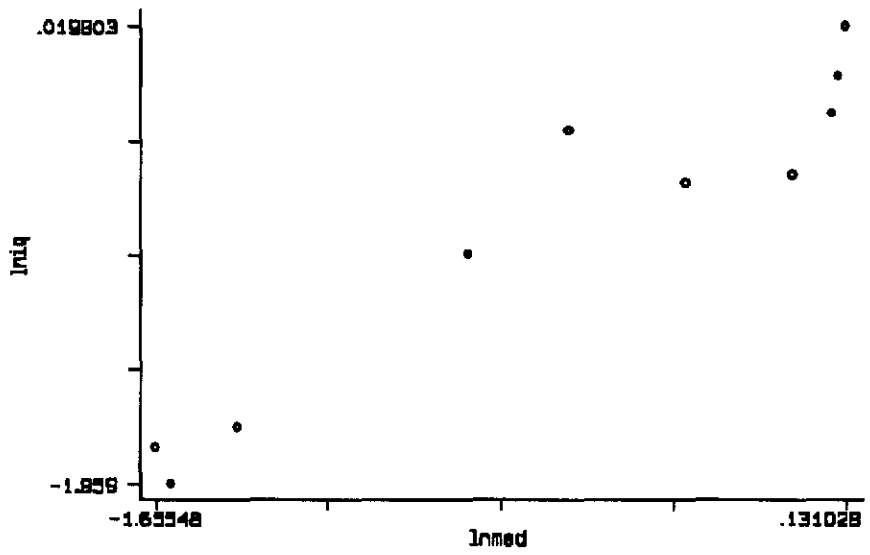
la transformación apropiada se define como: x^p
cuando p = 0 se utiliza el logaritmo (ln), p se define como p = 1 - β β *se* deriva al estimar la pendiente de la relación entre los rangos intercuartiles y las medianas.

Para el ejemplo de plomo en aire, se obtienen los siguientes datos:

```
. lv plomo if aa==1986

#    65        plomo

 M   33                  0.563                spread    pseudosigma
 F   17        0.325     0.649     0.974      0.649     0.493
 E   9         0.190     0.710     1.230      1.040     0.467
 D   5         0.177     0.925     1.672      1.495     0.510
 C   3         0.148     0.999     1.850      1.702     0.489
 B   2         0.120     1.044     1.968      1.848     0.474
 A   1.5       0.107     1.107     2.108      2.001     0.476
     1         0.094     1.170     2.247      2.153     0.467

                                               # below   # above
inner fence  -0.648               1.947        0         2
outer fence  -1.622               2.921        0         0
```

	año	rango inter-cuartil	mediana
1.	86	0.649	0.563
2.	87	0.82	1.12
3.	88	1.02	1.14
4.	89	0.700	1.1
5.	90	0.536	0.995
6.	91	0.518	0.755
7.	92	0.379	0.43
8.	93	0.18	0.236
9.	94	0.141	0.199
10.	95	0.165	0.191

Si graficamos los valores obtenemos:



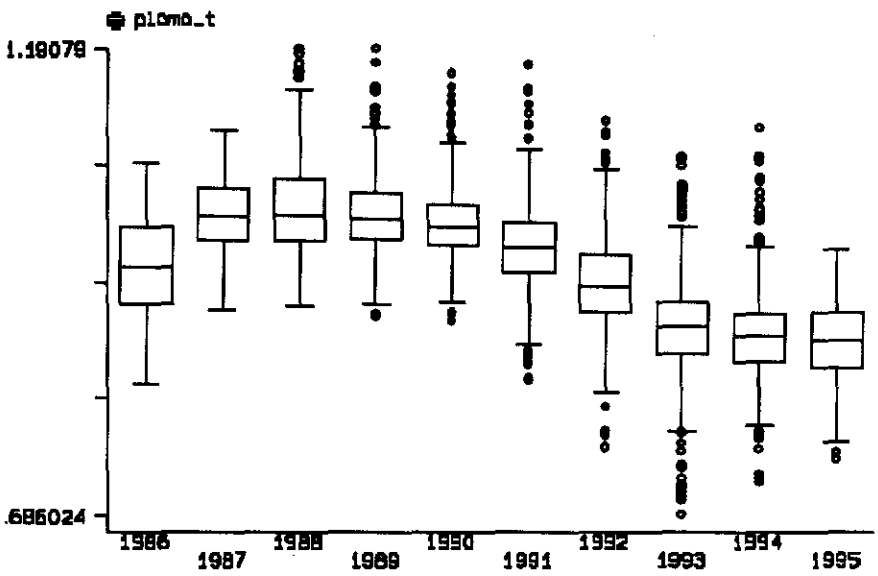
El ajustar una b de 1 podría resultar adecuado. Una manera de encontrar el valor es utilizando la regresión lineal y estimando la b.

```
. reg lniq lnmed

Source   SS          df  MS          Number of obs =  10
Model    4.20014745  1   4.20014745  F(  1,     8) =  88.13
Residual .381289114  8   .047661139  Prob > F      =  0.0000
Total    4.58143656  9   .509048507  R-squared     =  0.9168
                                     Adj R-squared =  0.9064
                                     Root MSE      =  .21831

lniq    Coef.      Std. Err.  t       P>|t|    [95% Conf. Interval]

lnmed   .9261176   .0986543   9.388   0.000    .6986203    1.153615
cons    -.3060345  .0914806   -3.345  0.010    -.5169892   -.0950799
```

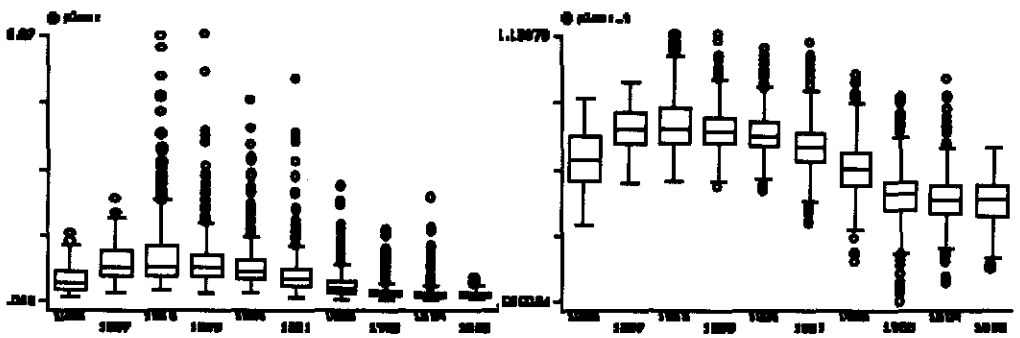
En nuestro caso, la ecuación estimada es **0.926**, lo cual es cercano a 0. (1-0.92=0.08). Podemos generar la nueva variable con la transformación encontrada:

```
gen plomo_t=plomo^0.08
sort aa
graph plomo_t, box by(aa)
```

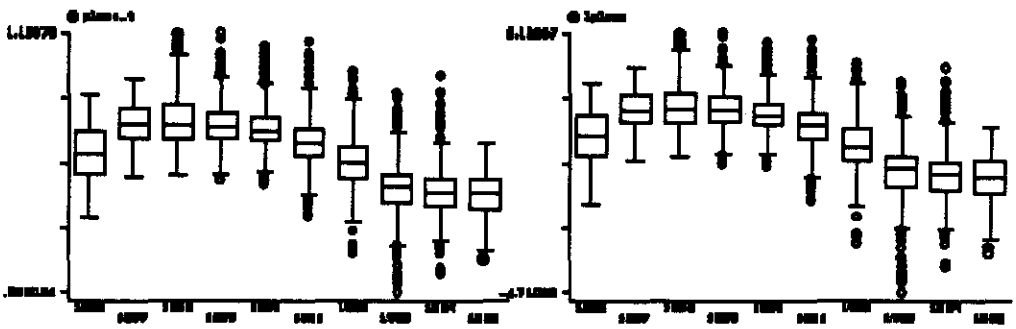
.



Comparando con la gráfica sin transforma se puede ver la ganancia importante en el despliegue de la información.



Simplificando, se podría elegir la transformación logarítmica. La relación con la variable transformada se aprecia en la siguiente forma:



Con las diferentes transformaciones se obtienen los siguientes beneficios

Transformación: Se gana simetría.
Se pierde "interpretabilidad"

Si la media > mediana desviación positiva
Si la media = mediana simétrica
Si la media < mediana desviación negativa

Cubo: y3	Reduce asimetría negativa muy fuerte
Cuadrado y2	Reduce asimetría negativa leve, si la varianza disminuye conforme aumenta y.
Raíz cuadrada	Reduce asimetría positiva leve moderada, útil cuando la variable se distribuye como una Poisson
Logaritmo	Reduce asimetría positiva (solo para valores positivos). Es útil cuando la varianza aumenta conforme aumenta el valor de la variable
Inverso (1/y)	Minimiza el efecto de valores muy altos en y

Transformación de variables.

Una de las aplicaciones del análisis exploratorio de datos, es la evaluación de la necesidad de realizar transformaciones. Las principales razones para realizar transformaciones son:

a) Normalizar las distribuciones
b) Ganar interpretabilidad
c) Corregir asimetrías fuertes
d) Categorías con dispersiones diferentes
e) Residuales influyentes (detectados en regresión lineal)

INTRODUCCION AL ANALISIS COMPARATIVO BI-VARIADO Y MULTIVARIADO

I- Usos de la Estadística:

a) En la vida diaria.

Cada día se utiliza con mayor frecuencia la estadística con el fin de formar opinión o de informar a la opinión pública. Los periódicos reportan con mucha frecuencia los resultados de encuestas. Recientemente se publicaba una sobre el programa "Hoy no Circula". Para este estudios se entrevistaron a un grupo de 200 automovilistas y se les preguntaron su opinión sobre le programa. El 80% estaba de acuerdo con la medida y pensaba que era proteger la salud de la población.

Otras encuestas de opinión no se reportan, ya que las realizan los partidos políticos para tomas decisiones en cuanto a sus candidatos o para preparar los discursos de los candidatos.

b) Para demostrar diferencias estadísticamente significativas

En épocas recientes, en especial en el campo de la investigación científica, se ha dado un culto por la significancia estadística. Esta devoción por la significancia se ha desarrollado a tal grado que los resultados se califican en base a si son o no son estadísticamente significativos. Los investigadores y editores de revistas toman decisiones sobre políticas de publicación en base a la significancia estadística, lo que ha ocasionado una preferencia de publicación basada en el valor p.

Hay quien afirma que la estadística ha servido para demostrar que durante la época de guerra hay significativamente ($p<0.000001$) mas muertes por arma de fuego que durante los tiempos de paz.

c) Para modelar fenómenos biológicos, aumenta el conocimiento y la capacidad de predicción.

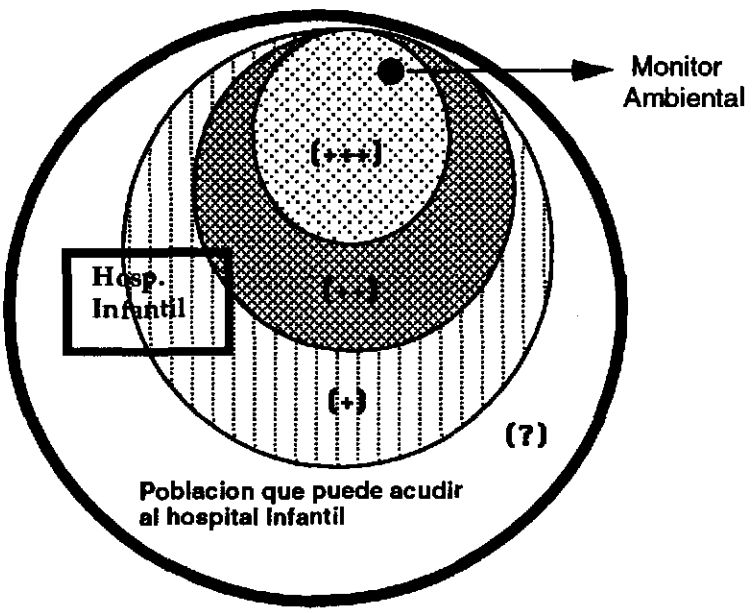
Ejemplos de este tipo de uso se encuentran frecuentemente en el área de salud pública y medicina. Por ejemplo, actualmente se puede predecir el riesgo que tiene una anciana de sufrir una fractura en base a las mediciones de densidad ósea del hueso.

La estadística representa una herramienta muy importante para comprender los fenómenos biológicos, nos permite:

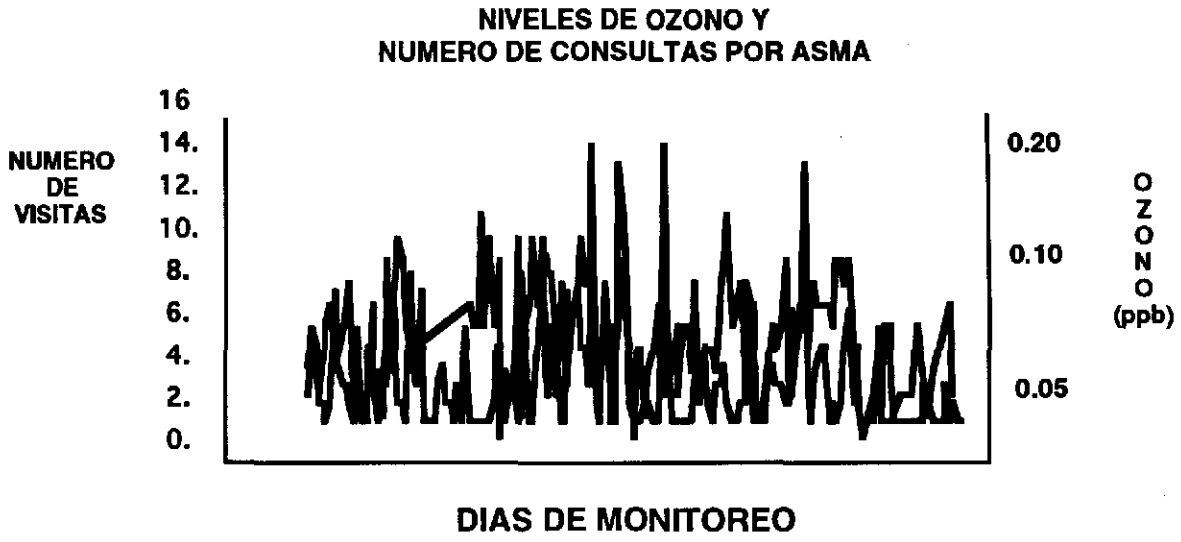
a) Comunicar y describir información en forma estandarizada
b) Contrastar hipótesis
c) Modelar y cuantificar diferentes relaciones entre parámetros.

Sin lugar a dudas la estadística es una herramienta muy útil, sin embargo es importante recordar que su aplicación se basa en una sobre simplificación de los fenómenos biológicos y una serie de suposiciones -que frecuentemente son poco reales- , sobre el comportamiento de las variables en las que se ha operacionalizado la medición de los fenómenos biológicos.

Por ejemplo, si analizamos el efecto de la contaminación ambiental sobre la demanda de servicios de urgencia por asma en un hospital pediatrico de la ciudad de México



Hipótesis: Los días con niveles de contaminación ambiental se asocian con un incremento en la demanda de serviicios de urgencias. Latencia 24 y 48 horas.
Eventos en estudio: Consultas por asma en urgencias n=395. Durante Enero-junio 1990.
Exposición: nivels máxiimos de ozono evaluados en una zona de monitoreo cercana al hospital.
Fuentes de error: Diagnóstico y exposición



En el caso de estudio, los investigadores estaban interesados modelar el efecto del ozono sobre la demanda de consultas por asma. En este caso la operacionalización del objeto de estudio -la influencia del ozono sobre el aparato respiratorio- se realizó estudiando la influencia del ozono ambiental, medido mediante la red de monitoreo ambiental cercana al hospital infantil, sobre la demanda diaria de consultas por asma.
La segunda operacionalización que se requiere es la de conceptualizar la hipótesis de estudio, con la operacionalización de las variables, en un modelo estadístico que nos permita resumir y entender la información recolectada. En este caso buscamos una representación estadística (matemática) que nos permita modelar el efecto que tienen los niveles diarios de ozono sobre la demanda diaria de consultas de asma. El modelo estadístico se puede representar mediante la siguiente ecuación:

n mero de consultas = α + contaminaci n ambiental * efecto

$y_i = \alpha + \beta x$

donde : y_i es el n mero de consultas

α es la media de las consultas

βx es el efecto de la contaminaci n ambiental

Como se puede suponer este modelo es muy sencillo ya que es poco probable que el único determinante de la demanda de servicio por asma sea la contaminación por ozono. Otros factores que pueden contribuir son: otros contaminantes como por ejemplo las partículas suspendidas, la temperatura, el día de la semana, el sexo y la edad entre otros. Al incluir estos factores el modelo estadístico se complica y podría quedar representado con la siguiente fórmula:

Si consideramos otros factores importantes el modelo es el siguiente:

$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_{ij}$

donde : y_i es el n mero de consultas

α es la media de las consultas

$\beta_1 x_1$ es el efecto del ozono

$\beta_2 x_2$ es el efecto del sexo

$\beta_3 x_3$ es el efecto de la temperatura

$\beta_4 x_4$ es el efecto de la edad

otro modelo mas complicado podr a ser :

$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2 x_5 + \varepsilon_{ij}$

donde : y_i es el n mero de consultas

α es la media de las consultas

$\beta_1 x_1$ es el efecto del ozono

$\beta_2 x_2$ es el efecto del sexo

$\beta_3 x_3$ es el efecto de la temperatura

$\beta_4 x_4$ es el efecto de la edad

$\beta_5 x_5$ es el efecto de las part culas suspendidas

$\beta_6 x_1 x_5$ es el efecto de la interacci n entre ozono y part culas

Con este ejemplo se ilustrar la importancia de los dos niveles de acción necesarios para utilizar los métodos estadísticos. El primero consiste en la operacionalización de la hipótesis de estudio en observaciones, es decir lo que corresponde al ejercicio de medición y el segundo consiste en la conceptualización de la hipótesis en un modelo estadístico que sea adecuado para el tipo de datos desde el punto de vista estadístico y desde el punto de vista conceptual.
En el estudio del efecto de la contaminación ambiental sobre las consultas de asma, los investigadores modelaron los datos utilizando un modelo de regresión lineal, asumiendo que la variable dependiente se distribuye como una variable aleatoria piosson y que la relación entre los niveles de contaminación ambiental y asma sigue una relación lineal, es decir a mayor contaminación mayor demanda de servicios. En este ejemplo, la b se interpreta como la razón de tasas de incidencia.

Resultados reportados en el estudio de asma y contaminación ambiental

Correlación entre # de consultas y los niveles de ozono:
mismo día **r=0.16**
24 hrs despues **r=0.25**

Razon de Tasas de Incidencia(24 hrs):
1.43 per 50 ppb
IC 95% (1.24-1.66)

Fuente: Effects of urban air pollutants on emergency visits for childhood asma in México City . Am J Epidemiol 1995;141:546-53.

Analicemos otro ejemplo.

Investigadores en el área de salud ambiental están interesados en determinar los efectos nocivos del plomo en el recién nacido. La hipótesis en cuestión es que los niveles de plomo plasmático determinan la toxicidad de este metal sobre el producto en gestación. Dado que actualmente es difícil realizar las mediciones de plomo plasmático, ya que requiere de equipo muy sofisticado y de condiciones *ultra limpias* para evitar contaminación externa y es sumamente costos, los investigadores midieron las concentraciones de plomo en hueso. Esto bajo el supuesto que durante el embarazo, dadas las necesidades de calcio, el plomo acumulado en el hueso seria liberado a la circulación plasmática y tendría un impacto importante sobre las concentraciones de plomo plasmático. Como indicador del daño al producto, los investigadores registraron las variaciones en el peso al nacer. Con la hipótesis de que el plomo en hueso podría explicar las variaciones observadas en el peso al nacer en niños sanos que nacen en la Ciudad de México.

En este ejemplo el modelo que se plantea es el siguiente:

peso al nacer = α - efecto del plomo

$y = \alpha - \beta x$

donde y es la media del peso al nacer

α es la media esperada del peso al nacer cuando $x = 0$

x es la concentracion de plomo en hueso

β es la medida de efecto, la unidad de cambio

por unidad de plomo en hueso que se produce en el peso al nacer

En este caso, el modelo planteado es una sobre simplificación de la realidad, primero por que existen otros factores que se relacionan con el plomo y que también se relaciona con el peso al nacer. Ejemplos de estos pueden ser el fumar cigarrillos. Dado que para el cultivo del tabaco se utilizan plaguicidas con alto contenido de plomo, los niveles de plomo en las madres fumadoras pueden ser mas elevados, ademas sabemos que el fumar cigarrillos es uno de los factores mas importantes que disminuye el peso al nacer. Esta variable es un factor de confusión. Ademas existen otros factores que determinan el paso al nacer y que deben ser controlados con el fin de reducir la variabilidad en el estudio y así aumentar la precisión de los estimadores. Ejemplos de estos factores son las antropometría materna, la clase social, el sexo del producto entre otros.
En este caso seria conveniente plantar un modelo mas complicado pero mas cercano a la realidad:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + e$$

donde α es la media esperada cuando todas las $x = 0$

x_1 es el valor de plomo en hueso

x_1 es el valor de la circunferencia de pantorrilla de la mam

x_1 es el valor de paridad

x_1 es el valor de a os aprovados de escuela

x_1 es el valor de la edad gestacional del reci n nacido

x_1 es el valor de tabaquismo durante el embarazo

e representa el error residual del modelo

las β's representan la unidad de cambio esperada en el peso al nacer por unidad de cambio en los predictores

Como interpretamos este modelo, en el caso de nuestro parámetro de interés, el efecto del plomo en hueso $b_1 x_1$. Nos interesa estimar el efecto, manteniendo las otras variables constantes. Es decir cual es el efecto esperado de las concentraciones de plomo en hueso para mujeres con el mismo número de hijos, con el mismo nivel educativo, con el mismo nivel nutricional (estimado por la circunferencia de pantorrilla), con el mismo nivel de tabaquismo y con productos de la misma edad gestacional.

Para este ejemplo en particular los valores asociados a los coeficientes fueron los siguientes:

Resultados del análisis del efecto del plomo en hueso sobre el peso al nacer

variable	beta (b)	valor p
Plomo en hueso (mg/g hueso)	-7.29	0.003
Circunferencia de pantorrilla	40.42	<0.001
Paridad(1=1 2=2 o mas)	205.87	<0.001
Años de escuela	17.00	0.016
Edad Gestacional	75.49	<0.001
Tabaquismo durante el embarazo si=1 0=no	-239.0	0.023
Constante (a)	-1576.02	0.033

¿Como interpretamos estos resultados?. Es importante reparar en las unidades de medición que tiene cada variable y en el tipo de variable. Si analizamos la variable de interés, el efecto del plomo en hueso, se trata de una variable continua, por lo que se puede interpretar de la siguiente manera, por cada mg de plomo/g de hueso, la media del peso al nacer disminuye 7.29 g. Por cada 10 mg de plomo/g de hueso, la media del peso al nacer disminuye 70.29 g. En este caso la unida de cambio es de -7.29 g por cada mg de plomo/g de hueso. En el caso del tabaquismo, se trata de una variable indicadora, que toma valores 0,1. Cuando las mujeres fumaron durante el embarazo, esta variable toma valor 1(x=1), por lo que el efecto es (1)*(-239.0), lo que quiere decir que para las mujeres que fumaron durante el embarazo se espera una reducción en el peso al nacer de 239.0 g. Cuando las mujeres no fumaron durante el embarazo la variable toma valor 0 (x=0), por lo que el efecto es (0)*(-239.0)=0, es decir el valor esperado es el de la media.

En este contexto es importante mencionar que en ambos casos se estima un efecto que compara dos grupos. En el primer caso, la b asociada a la variable plomo en hueso, estima el efecto de comparar un grupo de mujeres con una diferencia de 1 mg de plomo/g de hueso, el efecto estimado es una disminución de 7.29 g. En el segundo caso la b estima el efecto asociado al fumar durante el embarazo. Estima la diferencia en las medias de los pesos al nacer entre los productos de madres que fumaron durante el embarazo y madres que no fumaron, en este estudio en particular, el efecto asociado al fumar durante el embarazo es una reducción de 239.0 g.

El valor p asociado a los coeficientes, indica que la asociación observada es diferente a la magnitud de asociación que se podría observar simplemente por el azar, esto bajo el supuesto de que se cumplen las suposiciones estadísticas del modelo empleado.

En este contexto es importante revisar los conceptos sobre:

a) Tipos de variables
b) Medidas de comparación o de efecto
y su relación con los diferentes modelos estadísticos
c) Conceptos sobre la estimación de efectos
y las pruebas de hipótesis

II- Tipos de variables:

Los textos de estadística presentan diferentes clasificaciones de las variables en estudio. Es importante recordar que el ejercicio de medición y de operacionalización tiene como resultado la generación de las variables en estudio. De esta manera, un mismo objeto conceptual de estudio puede ser operacionalizado de diferente manera por los investigadores en el área. Por ejemplo, el estudio de la osteoporosis, la desmineralización ósea asociada con la edad, puede operacionalizarce mediante el estudio de la densidad ósea. También podría operacionalizarce el estudio de esta enfermedad por una de sus consecuencias mas graves, que es la fractura asociada a trauma mínimo. La operacionalización de los objetos de estudio determina de manera importante los modelos estadísticos o matemáticos que se pueden emplear para resumir la información, describir el fenómeno en estudio o para contrastar y poner a prueba diferentes hipótesis sobre el objeto de estudio. En el ejemplo anterior, los investigadores que opracionalizaron el estudio de la osteoporosis mediante la medición de la masa ósea, que se expresa como una variable continua normalmente distribuida, probablemente optaran por un modelo de regresión lineal de mínimos cuadrados, estimado como parámetro de comparación la diferencia de medias. Mientras que los investigadores que optaron por estudiar las fracturas, se verían en la necesidad de utilizar un modelo de regresión logística donde el parámetro de estudio es la razón de momios o riesgo relativo.

Las diferentes nomenclaturas empleadas para las variables se refieren a: 1) Su interpretación conceptual en el modelo o 2) De acuerdo a su escala de medición.

En relación a la interpretación conceptual en el modelo, se utiliza frecuentemente la clasificación de variable dependiente y de variable independiente. En el siguiente modelo:

$$y = \alpha + \beta x$$

y es la variable dependiente o respuesta

x es la variable independiente, explicativa o determinante

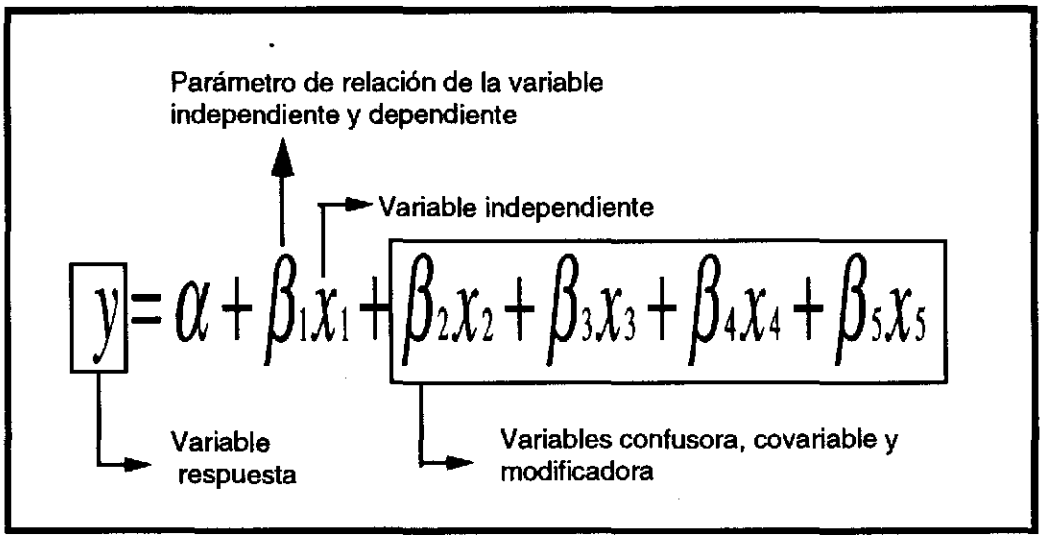
La variable independiente, explicativa o determinante es la que utilizamos para comprender mejor la variabilidad de la variable dependiente o respuesta. Conceptualmente es m uy importante establecer la relación entre las variables. Especialmente en el contexto de la investigación epidemiológica en la que se busca establecer relaciones de causalidad, de causa-efecto.

El función de las variables puede cambiar dependiendo del objeto de estudio. Por ejemplo si quisiéramos estudiar la variabilidad de las concentraciones de ozono en la ciudad de México. La variable respuesta sería las concentraciones de ozono, mientras que las variables explicativas sería la temperatura, la estación de año, el tiempo calendario etc. Sin embargo, para estudiar el efecto de la contaminación ambiental sobre la salud, el ozono sería utilizada en el modelo como una variable explicativa para predecir los cambios en salud.
Similarmente, en el estudio de los efectos del plomo en hueso sobre el peso al nacer, el plomo en hueso, en este caso, es una variable explicativa. Sin embargo, podría ser de interés estudiar los factores que determinan la concentración de plomo en el hueso, en cuyo caso la variable plomo en hueso sería estudiada como la variable respuesta o dependiente.
Como hemos mencionado anteriormente, los modelos bi-variados son una sobre simplificación de la realidad, en general trabajamos con modelos multivariados y las funciones de las variables en el modelo pueden variar de acuerdo a los objetivos de la investigación.

Tipología de las variables según su conceptualización en modelos epidemiológicos

Nombre	Definición
Dependiente o respuesta	Representa la operacionalización de del objeto principal de estudio.
Independiente	Representa la operacionalización de la información utilizada para comprender o predecir la variable dependiente. Cuantificar la asociación entre esta variable y la variable respuesta se considera como uno del los objetivos principales estudio.
Covariable	Información que se incluye en el modelo por que es importante para comprender mejor la relación entre la variable independiente y la dependiente. Reduce la variabilidad no-explicada de la variable dependiente. Mejora la precisión en la medición de la variable independiente
Confusora	Información que se debe incluir en el modelo para tener una estimación correcta -no sesgada- de la relación entre la variable dependiente y la independiente. Su inclusión es determinante de la validez de los resultados.
Modificadora	Información que se incluye para modelar a diferentes niveles la relación entre la variable dependiente y la independiente. Su inclusión es determinante de la validez de los resultados.

Variables en los modelos epidemiológicos



De acuerdo a la escala de medición las variables se pueden clasificar en continuas y categóricas:

Escala continua:

Escala de razón: (un término sobre otro). Por ejemplo número de hojas que tiene un árbol por la altura. El número de hospitalizaciones por tiempo persona en riesgo. Este tipo de escala tiene las siguientes características:

a) El intervalo de medición es constante. La diferencia entre 39 y 40 es la misma que entre 1 y 2.

b) El cero existe y tiene significado biológico. Esta propiedad permite decir que la una tasa de mortalidad de 5 por mil es del doble de una tasa de 2.5 por mil.

En la escala de medición representan la información mas amplia, se puede transformar o re-expresar en otro tipo de variable, lo que en general implica perder información.

Escala de Intervalo: Algunas mediciones poseen intervalos constantes pero no existe el cero. La temperatura de 40 grados centígrados no es el doble de una de 20 grados. Por ejemplo en el tiempo, horas del día el cero es un punto arbitrario.

Escala continua discreta: Son escalas numéricas cuantitativas, cuando es continua, es por que es posible encontrar un valor entre cada punto. Ejemplo talla, peso, nivel de colesterol, presión arterial. Cuando es discreta solo puede tomar cierto tipo de valores. Por ejemplo el número de muertes, número de camas, el número de personas.

Escala categórica:

Escala Ordinal: Indican cantidades, peor no se aprecia la magnitud de la cantidad.

Ejemplo +, ++, +++, ++++. Se tiene información sobre diferencias relativas pero no cuantitativas. La información permite ordenar de mayor a menor, pero no existen unidades claramente indetificables. Ejemplo nivel socio-económico.

Escala Nominal: La información indica una cualidad o atributo. Ejemplo color de los ojos, la raza, el género.

Escala Binaria: (variables dummy o indicadoras). Este tipo de variables ocurre frecuentemente ya sea de manera natural (por ejemplo: vivo o muerto, sano o enfermo, caso o control) o como una re-expresión de los tipos de escala antes mencionadas. En general, las variables nominales se re-expresar como variables binarias al modelarlas estadísticamente. Este transformación es un procedimiento ampliamente utilizado.

Como ejemplo de la re-expresion de variables continuas en este tipo de escala utilizaremos los datos presentados anteriormente sobre el efecto de plomo en hueso sobre el peso al nacer. Dado que no existe datos en la literatura médica sobre alguna propuesta de agrupación de los valores de plomo en hueso, estos se han agrupado de acuerdo a los cuartiles. Mediante esta categorización la población se divide en cuatro grupos que contiene al 25% de las observaciones. Esto se realiza ordenando de menor a mayor e identificando los puntos de corte que marcan los diferentes percentiles. Los valores también se pueden obtener del diagrama de letras (lv). La variable se re-expresa de la siguiente manera, de acuerdo al valor que toma el percentil(%):

si plomo en hueso <=25% se re-expreso como 1
si 25%< plomo en hueso <=50% se re-expreso como 2
si 50%< plomo en hueso <=75% se re-expreso como 3
si 75%< plomo en hueso se re-expreso como 4

En un segundo paso, creamos las variables indicadoras. Para este ejemplo se requiere de 4 variables indicadores (x1,x2,x3,x4) que **indican** la pertenencia a un grupo en particular con el siguiente esquema de codificación.

x1=1 si plomo en hueso=1, otra condición x1=0
x2=1 si plomo en hueso=2, otra condición x1=0
x3=1 si plomo en hueso=3, otra condición x1=0
x4=1 si plomo en hueso=4, otra condición x1=0

De acuerdo a las transformaciones hechas, re-expresamos la información contenida en una variable continua en cuatro variables indicadoras.

Veamos como se hace este proceso en STATA

```
. desc

Contains data from :Macintosh HD:datos:pesomha.
  Obs:   272 (max=  5231)                    seleccion de karen2b
 Vars:     1 (max=  2046)                    13 May 1996 12:57
Width:     4 (max=  3100)
  1. hueso_pb     float  %9.0g               conc plomo en hueso
Sorted by:
Note:  Data has changed since last save

. lv hueso_pb

 #    272          conc plomo en hueso
                ---------------------------------
 M    136.5 |                 9.14            |    spread  pseudosigma
 F     68.5 |      4.45      9.885      15.32 |     10.87    8.06946
 E     34.5 |       .53    9.4475     18.365 |    17.835   7.767028
 D     17.5 |    -3.665     10.25     24.165 |     27.83   9.096074
 C      9   |     -6.62     11.19         29 |     35.62   9.602936
 B      5   |     -9.29     10.97      31.23 |     40.52   9.570779
 A      3   |    -10.66     12.07       34.8 |     45.46   9.737699
 Z      2   |    -10.75     13.51      37.77 |     48.52   9.684046
 Y    1.5   |   -11.845   13.0325      37.91 |    49.755   9.463337
      1     |    -12.94    12.555      38.05 |     50.99    9.10386
            |                                |
            |                                |   # below     # above
inner fence |   -11.855                31.625 |         1           4
outer fence |    -28.16                47.93 |         0           0

. gen hueso_4= hueso_pb
. label var hueso_4 "conc de hueso en cuartiles"
. recode hueso_4 min/4.45=1 4.451/9.14=2 9.141/15.32=3
15.321/max=4
. tab hueso_4
```

```
   conc de|
    hueso en|
   cuartiles|      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         68       25.00       25.00
          2 |         68       25.00       50.00
          3 |         68       25.00       75.00
          4 |         68       25.00      100.00
------------+-----------------------------------
      Total |        272      100.00
```

```
. tab hueso_4,sum( hueso_pb)
```

```
    conc de|    Summary of conc plomo en hueso
    hueso en|
   cuartiles|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |  -.89485292   4.3929134          68
          2 |   6.6997059    1.387378          68
          3 |   12.530882   1.7474731          68
          4 |   21.016177   5.6926024          68
------------+------------------------------------
      Total |    9.837978    8.863548         272
```

```
. tab hueso_4,gen(h)
. desc

Contains data from :Macintosh HD:datos:pesomha.
  Obs:   272 (max=  5231)                    seleccion de karen2b
 Vars:     6 (max=  2046)                    13 May 1996 12:57
Width:    12 (max=  3100)
  1. hueso_pb     float  %9.0g               conc plomo en hueso
  2. hueso_4      float  %9.0g               conc  de hueso en cuartiles
  3. h1           byte   %8.0g               hueso_4==      1.0000
  4. h2           byte   %8.0g               hueso_4==      2.0000
  5. h3           byte   %8.0g               hueso_4==      3.0000
  6. h4           byte   %8.0g               hueso_4==      4.0000
Sorted by:
Note:  Data has changed since last save
```

```
. tab hueso_4 h1
    conc de| hueso_4==     1.0000
   hueso en|
  cuartiles|         0          1 |     Total
-----------+----------------------+----------
         1 |         0         68 |        68
         2 |        68          0 |        68
         3 |        68          0 |        68
         4 |        68          0 |        68
-----------+----------------------+----------
     Total|       204         68 |       272

. tab hueso_4 h2
    conc de| hueso_4==     2.0000
   hueso en|
  cuartiles|         0          1 |     Total
-----------+----------------------+----------
         1 |        68          0 |        68
         2 |         0         68 |        68
         3 |        68          0 |        68
         4 |        68          0 |        68
-----------+----------------------+----------
     Total|       204         68 |       272

. tab hueso_4 h3
    conc de| hueso_4==     3.0000
   hueso en|
  cuartiles|         0          1 |     Total
-----------+----------------------+----------
         1 |        68          0 |        68
         2 |        68          0 |        68
         3 |         0         68 |        68
         4 |        68          0 |        68
-----------+----------------------+----------
     Total|       204         68 |       272

. tab hueso_4 h4
    conc de| hueso_4==     4.0000
   hueso en|
  cuartiles|         0          1 |     Total
-----------+----------------------+----------
         1 |        68          0 |        68
         2 |        68          0 |        68
         3 |        68          0 |        68
         4 |         0         68 |        68
-----------+----------------------+----------
     Total|       204         68 |       272
```

El tipo de variable, según la escala de medición y/o la conceptualización en el modelo determina de manera importante el tipo de análisis estadístico, así como el parámetro de comparación que se utiliza. En el cuadro siguiente se ilustran algunos ejemplos.

Escala de medición y modelos estadísticos

Dependiente	Independiente	Tipo de análisis
razón, intervalo, continua o discreta	razón, intervalo, continua o discreta	regresión lineal regresión lineal de poisson
razón, intervalo, continua discreta	ordinal Nominal Binaria	análisis de varianza
ordinal	razón, intervalo, continua discreta	regresión ordinal
nominal	razón, intervalo, continua discreta	regresión logística politómica Modelos logarítmicos lineales
Binaria	ordinal Nominal Binaria	regresión logística modelos de riesgos proporcionales

III- Medidas de comparación o de efecto y su relación con los diferentes modelos estadísticos.

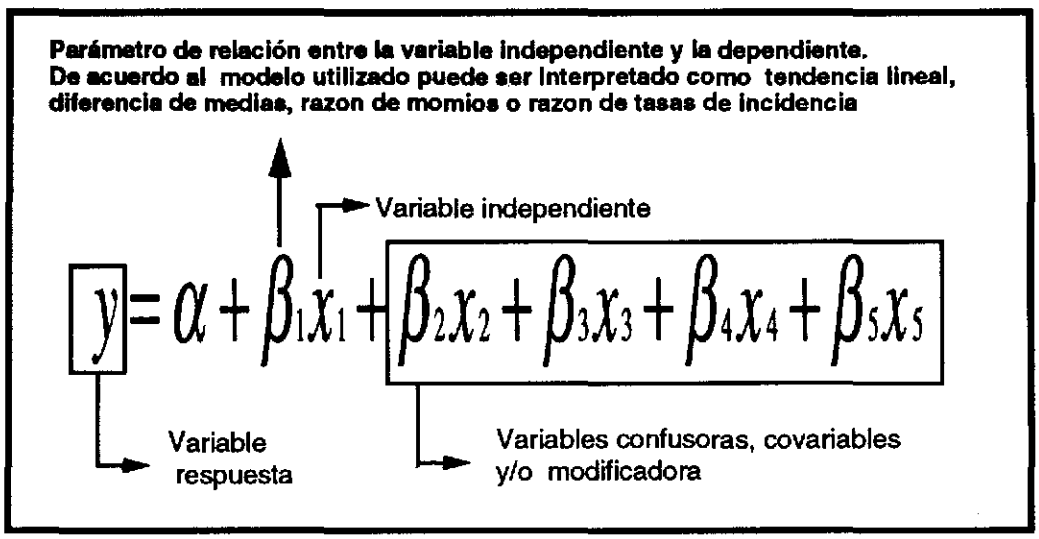
Una de los usos mas importantes de la estadística es la estimación de medidas de efecto o estadísticos que permiten comparar uno o mas grupos. La posibilidad de estimar esta medidas no solo permite contrastar hipótesis (realizar pruebas de significancia estadística) sino cuantificar (estimar) las diferencias diferencias para establecer su significado biológicos o social.

En los diferentes campos de investigación existen un buen número de medidas de comparación, por lo que sería imposible abarcar todas en este curso, por lo que discutiremos únicamente las más utilizadas en estudios epidemiológicos y biomédicos.

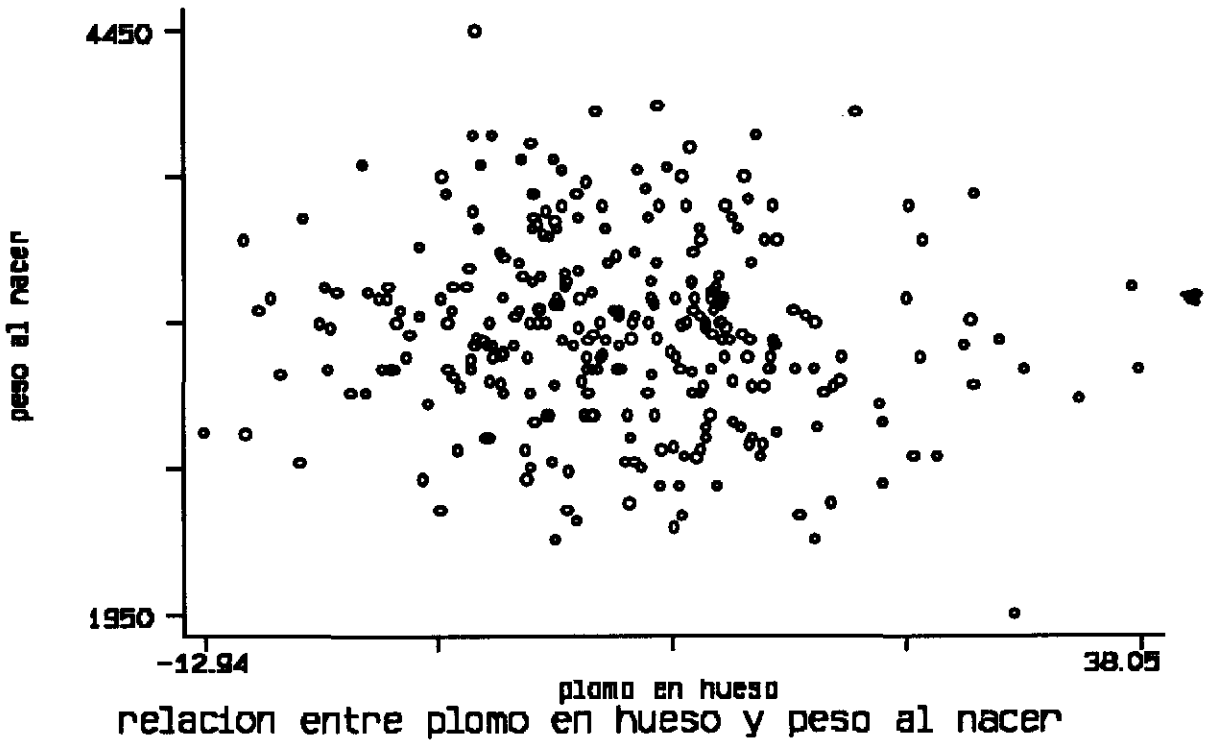
La conceptualización de una medida de comparación necesariamente implica la suposición de la existencia de dos grupos que difieren en base a una característica. La medida de efecto que se utilice dependerá en gran medida de la escala de medición de la variable.

Parámetro de comparación	valor de nulidad o no efecto	escala de medición de la variable dependiente (y)	escala de medición de la variable independiente (x)	Modelo estadístico
diferencia de medias	0	continua	categórica	regresión lineal anova
tendencia lineal	0	continua	continua	regresión lineal
tendencia lineal	0	continua	continua categórica	regresión lineal
tendencia lineal	0	Indicadora (0,1)	continua	regresión logística
razón de momios	1	indicadora (0,1)	categoría	regresión logística
razón de momios	1	categórica	categoría	regresión logística politómica
tendencia lineal	0	continua discreta	continua	regresión poisson
razón de tasas de incidencia	0	continua discreta	categórica	regresión poisson
tendencia lineal	0	indicadora (0,1)	continua	regresión de riesgos proporcionales
razón de tasas de incidencia	1	indicadora (0,1)	categórica	regresión de riesgos proporcionales

Un paso importante dentro del tipo de análisis que discutiremos durante este curso es el de operacionalizar la hipótesis de interés en un modelo estadístico lineal apropiado. La forma general de los modelos que utilizaremos en el curso es la siguiente:



Analicemos en este contexto el ejemplo antes mencionado del efecto del plomo en hueso sobre el peso al nacer. En este estudio a manera de recordatorio la hipótesis era que: El plomo acumulado en el hueso, liberado durante el embarazo tiene un efecto sobre el producto gestante. Los investigadores operaciojnalizaron este efecto en el peso al nacer. La hipótesis estadística que se pretende evaluar es que a mayor plomo en hueso menor será el peso al nacer. En este caso el parámetro de comparación es una tendencia lineal, que se observa en la siguiente gráfica:



Un modelo sencillo de para representar esta asociación es la siguiente:

$y = \alpha - \beta x$

donde y es la valor del peso al nacer

x es el valor de plomo en hueso y β

representa la constante de cambio.

Si ajustamos el modelo propuesto en stata obtenemos los siguientes resultados:

```
. reg pesona hueso_pb

      Source |       SS       df       MS              Number of obs =     272
-------------+------------------------------           F(  1,   270) =    3.24
       Model |  525009.029     1  525009.029            Prob > F      =  0.0732
    Residual |  43804515.1   270  162238.945            R-squared     =  0.0118
-------------+------------------------------           Adj R-squared =  0.0082
       Total |  44329524.2   271   163577.58            Root MSE      =  402.79

------------------------------------------------------------------------------
      pesona |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hueso_pb |   -4.965822   2.760484    -1.799   0.073          -10.40063
.4689895
       _cons |   3217.034   36.52398     88.080   0.000       3145.126    3288.942
------------------------------------------------------------------------------
```

En este caso, la ***b*** estimada corresponde a **-4.96,** lo que indica que por cada micro-gramo de plomo en hueso, el peso esperado al nacer disminuye 4.96 g. Esta constante de cambio se asume lineal, lo que quiere decir que se espera la misma magnitud de cambio de 1.0 a 2.0, que de 50.5 a 51.5.

Ahora, analicemos estos mismos datos, pero con una variable transformada. Utilizaremos la variable re-expresada en cuartiles. En este caso la hipótesis estadística que se postula es la siguiente:
La media estimada de peso al nacer será diferente para cada cuartil de plomo en hueso, ademas se espera que la media de peso al nacer sea menor conforme se pertenece a un cuartil superior de plomo en hueso.

Analicemos el ejemplo:

```
. gen hueso_4=hueso_pb
. recode hueso_4 min/4.45=1 4.451/9.14=2 9.141/15.32=3 15.321/max=4
. label var  hueso_4 "conc  de hueso en cuartiles"
. tab hueso_4,sum(pesona)
```

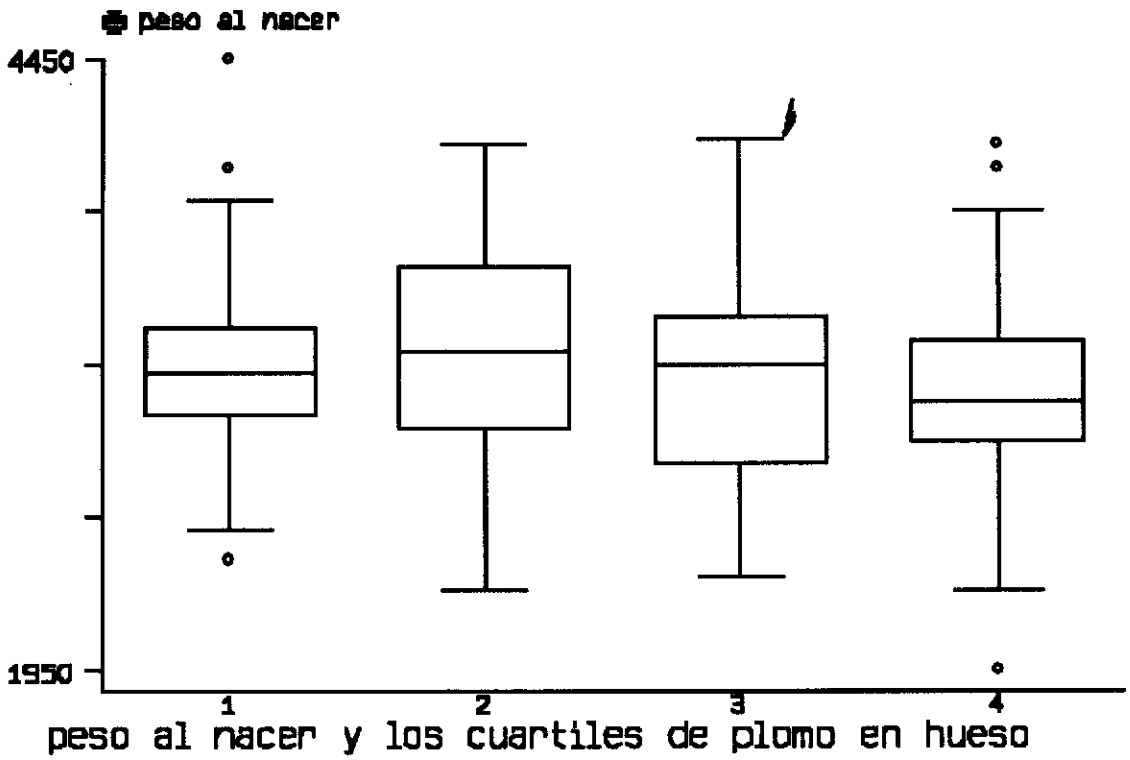
conc de hueso en cuartiles	Summary of peso al nacer		
	Mean	Std. Dev.	Freq.
1	3190.9	362.73449	68
2	3237.5	422.96793	68
3	3149.2	415.14902	68
4	3095.0	409.39708	68
Total	3168.1801	404.44725	272

Iniciamos creando la variable que contiene la información sobre los cuartiles. Posteriormente estimamos las medias de peso al nacer para cada cuartil. En este caso fueron 3190, 3237, 3149 y 3095, para el primer, segundo, tercer y cuarto cuartil respectivamente. Si tomamos como referencia el primer cuartil -lo que es lógico ya que representa al peso esperado de los recién nacidos, hijos de mujeres con menor contenido de plomo en hueso-. Las diferencias por cuartil serian:

Q_1-Q_1=0, Q_1-Q_2=-46-6 Q_1-Q_3=41.7 Q_1-Q_4=95.9

La representación gráfica de la diferencia de medias es la siguiente:

```
graph pesona,box by(hueso_4) title("peso al nacer y los cuartiles de plomo en
hueso")
```



Para modelar la comparación de los cuartiles se requiere de estimar tres diferencias de medias por lo que es necesario incluir tres parámetros en el modelo. El modelo se podría expresar de la siguiente manera:

$$y = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$$

donde x_{11}, x_{12} y x_{13} son las variables indicadoras x_{11} toma valor 1 si pertence al primer cuartil, x_{12} toma valor 1 si pertence al segundo cuartil, y x_{13} toma valor 1 si pertence al tercer cuartil. β_1, β_2 y β_3 estiman las diferencias de medias entre los cuartiles 2,3 y 4 con el cuartil 1.

Ajustamos el modelo usando stata, inicialmente generamos las variables indicadores de cada cuartil, seleccionamos una de referencia y la dejamos fuera del modelo de regresión, ya que cuando todas las variables indicadoras en el modelo toman valor 0, entonces la constante corresponde a la media estimada para el primer cuartil. Analicemos los resultados obtenidos al ajusta el modelo planteado:

```
tab hueso_4,gen(h)
reg pesona h2 h3 h4

  Source |       SS       df       MS                  Number of obs =     272
---------+------------------------------                F(  3,   268) =    1.54
   Model |   750523.07     3  250174.357                Prob > F      =  0.2049
Residual |  43579001.1   268  162608.213                R-squared     =  0.0169
---------+------------------------------                Adj R-squared =  0.0059
   Total |  44329524.2   271   163577.58                Root MSE      =  403.25

------------------------------------------------------------------------------
  pesona |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      h2 |   46.54412   69.1563       0.673   0.502     -89.61462    182.7029
      h3 |  -41.69118   69.1563      -0.603   0.547     -177.8499    94.46756
      h4 |  -95.95588   69.1563      -1.388   0.166     -232.1146    40.20286
   _cons |   3190.956   48.90089     65.254   0.000      3094.677    3287.235
------------------------------------------------------------------------------
```

Los valores que toma el coeficiente son:

contraste	valor estimado mediante el modelo de regresión	estimador empírico de la diferencia de medias
cuartil_1 - cuartil_2	46.54	-46.6
cuartil_1 - cuartil_3	-41.69	41.70
cuartil_1 - cuartil_4	-95.95	-95.95
media en el cuartil 1	3190.95	3190.95

Como se puede apreciar en este último ejemplo, hemos adecuado el modelo estadístico empleado para estimar el parámetro de interés. Ahora supongamos que nos interés re-expresar la variable peso al nacer y que nos interesa predecir el riesgo de tener un producto de bajo peso al nacer, definido como un peso menor de 2500 g. En este caso se plantea la hipótesis estadística de que el riesgo de tener un producto de bajo peso varía de acuerdo con el percentil de plomo en hueso. El parámetro de interés en este caso es la razón de momios. Re-expresamos la variable peso al nacer

```
recode pesona_d min/2500=1 2501/max=0
tab pesona_d
peso 1<=2500|
      otro=0|      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        258       94.85       94.85
          1 |         14        5.15      100.00
------------+-----------------------------------
      Total |        272      100.00
```

Quedan 14 nacimientos con peso al nacer menor o igual que 2500 y 258 nacimientos con peso superior a los 2500. Construimos ahora la tabla de contraste de riesgo:

obtenemos las razones de momios comparando los cuartiles 2,3,y 4 con el cuartil 1.

```
           |   Bajo peso     Normal |     Total
-----------------+-----------------------+----------
 Cuartil 2       |         3         65  |        68
 Cuartil 1       |         1         67  |        68
-----------------+-----------------------+----------
           Total |         7        129  |       136
```

```
                |   Bajo peso     Normal |     Total
-----------------+-----------------------+----------
 Cuartil 3       |         6         62  |        68
 Cuartil 1       |         1         67  |        68
-----------------+-----------------------+----------
           Total |         7        129  |       136
```

categoría	peso normal	bajo peso	razón de momios
cuartil 1	67	1	1.0
cuartil 2	65	3	3.09
cuartil 3	62	6	6.48
cuartil 4	64	4	4.18

Para modelar la comparación de los cuartiles y el riesgo de bajo peso al nacer se requiere de estimar tres razones de momios (riesgo relativo) por lo que es necesario incluir tres parámetros en el modelo. El modelo se podría expresar de la siguiente manera:

$y = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$

donde x_{11}, x_{12} y x_{13} son las variables indicadoras x_{11} toma valor 1 si pertence al primer cuartil, x_{12} toma valor 1 si pertence al segundo cuartil, y x_{13} toma valor 1 si pertence al tercer cuartil. β_1, β_2 y β_3 estiman las diferencias de medias entre los cuartiles 2,3 y 4 con el cuartil 1.

Ajustamos el modelo usando stata, pero con la indicación de asumir una distribución logística, podemos utilizar las variables indicadoras ya generadas. Igualmente en este caso dejamos fuera del modelo de regresión la categoría de referencia.
Analicemos los resultados obtenidos al ajustar el modelo planteado:

```
logistic pesona_d h2 h3 h4

Logit Estimates                                   Number of obs =    272
                                                  chi2(3)       =   4.31
                                                  Prob > chi2   = 0.2301
Log Likelihood = -53.014097                       Pseudo R2     = 0.0390

------------------------------------------------------------------------------
pesona_d | Odds Ratio   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
      h2 |   3.092308    3.61102      0.967   0.334      .3135437    30.49771
      h3 |   6.483871   7.095919      1.708   0.088      .7590726    55.38414
      h4 |     4.1875   4.738597      1.266   0.206      .4557541    38.47504
------------------------------------------------------------------------------
```

Los valores que toma el coeficiente son:

contraste	Estimador derivado del modelo de regresión logística	estimador empírico de la razon de momios
cuartil_1 - cuartil_2	3.09	3.09
cuartil_1 - cuartil_3	6.48	6.48
cuartil_1 - cuartil_4	4.18	4.18

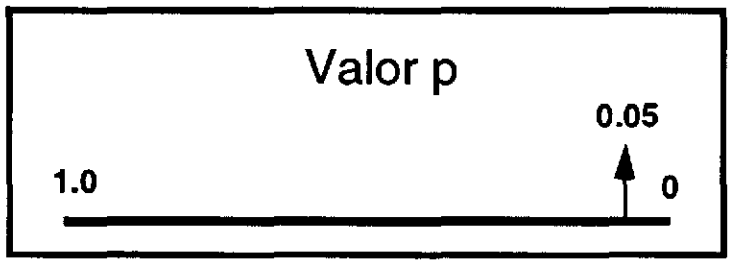
Como se puede apreciar en este último ejemplo, hemos adecuado el modelo estadístico empleado para estimar el parámetro de interés.

Hemos analizado la relación entre el peso al nacer y la concentración de plomo en hueso de diferentes maneras. La operacionalización del efecto adverso del plomo acumulado en el hueso sobre el producto gestante, se ha estimado mediante tres parámetros diferentes (**tendencia lineal inversa** entre el plomo en hueso y el peso al nacer, **diferencia de medias** de acuerdo a la cantidad de plomo en hueso y en base al **riesgo -razón de momios-** de dar a luz un producto de bajo peso al nacer, de acuerdo a la cantidad de plomo acumulado en hueso), asi mismo se utilizaron tres modelos estadísticos lineales, de la misma forma, en los que se variaron las suposiciones de distribución de las variables en estudio o se se re-expresaron de diferentes formas.

IV- Conceptos sobre la estimación del valor p y las pruebas de hipótesis.

Durante los últimos años, con la posibilidad de realizar cálculos estadísticos de gran complejidad en casi cualquier computadora personal, se ha revolucionado la práctica de la estadística. Esta disponibilidad de paquetes, ha propiciado también ciertos cambios en la manera en que se utilizan los métodos estadísticos. Antiguamente, dado el gran trabajo que representaba ajustar diferentes modelos estadísticos, la practica estadística se centraba sobre las pruebas de hipótesis y particularmente sobre la estimación del valor p. Esta práctica ha condicionado una sobre simplificación de la interpretación de datos. Muchas decisiones sobre la validez de las hipótesis planteadas se toman en base a la significancia estadística, lo cuál no es apropiado. La significancia estadística es un dato cualitativo que puede ser útil en la contrastación de hipótesis, pero no debe sustituir el razonamiento biológico o social que acompaña a las hipótesis planteadas.

Analicemos a que nos referimos con significancia epidemiológica. El valor P puede tomar valores entre 0 y 1.



Tradicionalmente el valor p se ha utilizado para definir la significancia estadística, el valor p estima la probabilidad de encontrar un resultado como el observado o mas extremo asumiendo que la hipótesis de nulidad es cierta. Como punto de corte para definir lo que es estadísticamente significativo se ha utilizado el 0.05. Los valores pro debajo de 0.05 se consideran como estadísticamente significativos (lo que se interpreta de la siguiente manera: la asociación o diferencia observada no se debe al azar) y los valores superiores al 0.05 se consideran como no-significativos (lo que se interpreta de la siguiente manera: las diferencias observadas o asociaciones observadas se deben al azar). El valor p se obtiene ajustando un modelo estadístico, a los datos, lo que implica hacer una serie de suposiciones sobre los mismos. Independientemente de que se cumplan las suposiciones, las cuales varían de acuerdo al modelo propuesto, la interpretación del valor p tiene serias limitaciones, por lo que se recomienda basar la interpretación estadística en la estimación de efectos y la variabilidad de los mismos y no en el valor p.

Los principales problemas asociados al valor son :

a) La ambigüedad de la interpretación. La diferencias entre 0.051, 0.060 ó 0.049

b) Para calcularlo se asume que la hipótesis de nulidad es cierta y que las observaciones se midieron sin error

c) Depende del tamaño de muestra:

En estudios pequeños el **valor p** es poco informativo, ya que en general va ha ser mayor de 0.05

En estudios con muestras muy grandes el **valor p** es poco informativo ya que cualquier diferencia detectada va ha estar asociada a valores p muy pequeños.

En estudios con tamaño de muestra adecuada el **valor p** se puede interpretar de la siguiente manera: a) Si el valor p es pequeño **(<=0.05)** se puede decir que los datos no son compatibles con la hipótesis de nulidad. b) Si el valor p es intermedio **(0.20 a 0.05)** se puede decir que los datos no permiten discriminar, apoyan ambas hipótesis. c) Si el valor p es grande **(>0.20)** se puede decir que los datos son compatibles con la hipótesis de nulidad.

La alternativa a los valores p, son los intervalos de confianza ya que esta sujeto a los mismos puntos de corte que el valor p, pero contiene información cuantitativa sobre la variabilidad observada y sobre la magnitud de la diferencia.

El intervalo de confianza nos indica la serie de valores que podemos observar. Por ejemplo un intervalo de confianza del 95% nos indica que al repetir n veces el experimento que dio origen a los datos, en el 95% de las veces los valores observados estarán contenidos entre x_1 y x_2. Si el valor de nulidad es contenido en el intervalo, entonces el valor de p será mayor de 0.05, si el valor de nulidad no esta contenido en el intervalo, entonces el valor de p será menor de 0.05.

Ejemplos de estudios hipotéticos con diferentes tamaños de muestra para el parámetro de comparación de razón de momios:

Tamaño de muestra pequeño

razón de momios	1.0	8.0	25.0
IC 95%	0.3-7.0	0.1-20.0	1.1-50.2
valor p	0.20	0.30	0.049

Tamaño de muestra intermedio

razón de momios	1.5	2.1	6.0
IC 95%	0.8-2.0	1.1-3.0	4.0-8.0
valor p	0.06	0.04	0.0001

Tamaño de muestra muy grande

razón de momios	1.2	6.1	7.0
IC 95%	1.1-1.3	5.9-6.2	6.9-7.2
valor p	0.00	0.00	0.000

REGRESION

Modelos de regresión

Estos modelos se utilizan para describir y/o modelar la relación entre dos variables. Es una de las técnicas de modelaje estadístico mas desarrolladas y mas usadas.

Modelar implica el desarrollo o aplicación de un modelo matemático que describe, de alguna manera, el comportamiento de una variable aleatoria.

Para describir las mediciones de los elementos de una población, se podría utilizar un modelo matemático muy simple:

$$y_i = \mu + \varepsilon_i$$

donde y_i es la medición en el mismo elemento de la población y μ es el promedio de todos los elementos de la población, dependiendo de los elementos que definan a la población. En general este valor se estima mediante muestras. ε_i es el error aleatorio.

Si se consideran simultáneamente varias poblaciones, que dieren entre si en alguna característica, llámese x, y cuya variable x contenga información sobre el comportamiento de y, se podría extender el modelo a cada población y suponer que los valores de $\mu(y)$ se pueden representar por una función lineal de x (es decir, se supone que los promedios de μ cambian al cambiar x en forma de línea recta).

Se puede utilizar el siguiente modelo:

$$y_i = \mu(y) + \varepsilon_i = \alpha + \beta x + \varepsilon_i$$

El modelo describe la ecuación de una recta con una ordenada al origen (alfa) y una pendiente (beta), la cual indica cómo cambian los promedios de la población por unidad de cambio de x.

Son medias las aplicaciones de la regresión, y se pueden utilizar en la mayoría de los campos de investigación , como por ejemplo:

Este proceso se conoce como "ajustar el modelo a los datos". Por esta razón es necesario evaluar si el modelo se ajusta a los datos y la calidad con que el modelo ajusta o describe los datos. La relación básica de las variables se expresa con la relación

$$y = \alpha + \beta x$$

$$y_i = \bar{y} + \beta x + e_{ij}$$

El valor de y para la misma observación puede estimarse por la media poblacional y el efecto del cambio en la variable x y la constante b.

El modelo puede extenderse y mejorar, pasando de un parámetro a dos.

Así por ejemplo, el peso no sólo va a depender de la talla, sino también de la edad, sexo, condición socioeconómica, etc.

Los modelos de regresión ayudan a representar la variabilidad natural mediante una expresión matemática.

En estadística se proponen modelos y se supone que se ajustan a los datos. Se estima la y (media) y la constante de cambio.

La interpretación del modelo, así como los métodos empleados para estimar los parámetros, dependerán del tipo de variables y de la distribución de las mismas.

En el modelo $y = \alpha + \beta x$ se estimaría, alfa y beta ya que son los parámetros desconocidos. Dependiendo del planteamiento del modelo β puede representar una tendencia lineal, una diferencia de medias, una razón de momios o una tendencia de logarítmica.

Los procedimientos estadístico para estimación varían. En el caso de las variables que son continuas y que se distribuyen de manera normal, se utiliza el método de:
Mínimos cuadrados (regresión lineal simple.)

En otros casos como en la distribución binomial, logística y poisson se utiliza el método de: máxima verosimilitud.

La variable x, en el caso del modelo de mínimos cuadrados, puede o no ser normal; sin embargo, la y sí debe ser normal en cuanto a distribución.

Si el modelo se formula correctamente, lo que se calcula es y. $y = \mu + b X$

El valor promedio de la variable dependiente aumentará o disminuirá linealmente por cada cambio de unidad en la variable X1, misma que es continua y puede ir de -• a +• ya que, de acuerdo con lo que se mida, estará el rango de posibilidad.

Beta: es una constante desconocida que indica el cambio esperado en la variable y por unidad de cambio en X.

Suposición: el cambio es constante a través de toda la escala de valores (esto no siempre es fácil de sostener).

ß: proporciona una idea de la fuerza de la relación entre y y x, Se asume que la fuerza es la misma para todos los posibles valores de x. b se estima mediante el modelo.

Sobre-simplifica la realidad; las relaciones biológicas no son tan fáciles ni tan simples.

La interpretación de α es una constante desconocida (se estima mediante el modelo) que mide el valor que toma la variable dependiente (y) cuando el valor de la independientes (x) es 0.
Por esta razón el parámetro α sólo tiene interpretación cuando el rango de variación posible de x incluye 0.

E: es un componente de error aleatorio, y se asume que este error tiene una media igual a cero (=0) y una varianza desconocida. Por otra parte, se supone que los errores no están correlacionados, es decir, que son independientes.

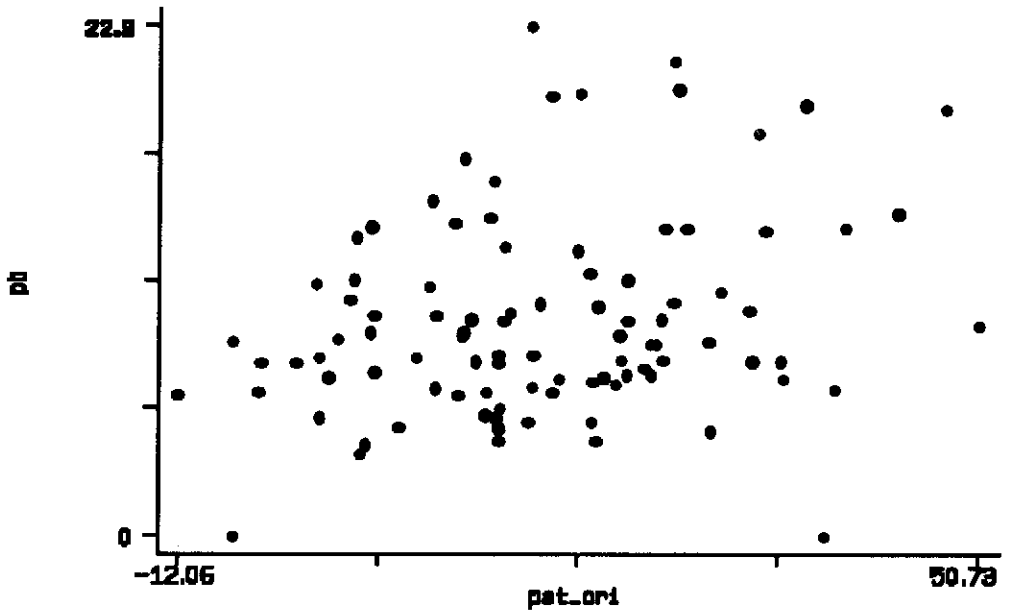
Relaciones posibles: positiva, inversa (negativa), ninguna.

Ejemplo: En un estudio epidemiológico sobre la determinación de los niveles de plomo en hueso se intenta describir la relación entre el plomo en hueso y el plomo en sangre. Se busca determinar el impacto del plomo en hueso sobre los niveles de plomo en sangre. El plomo en hueso se mide en dos sitios: la tibia y la rótula.
A continuación se presentan algunas de las estadísticas descriptivas de las variables

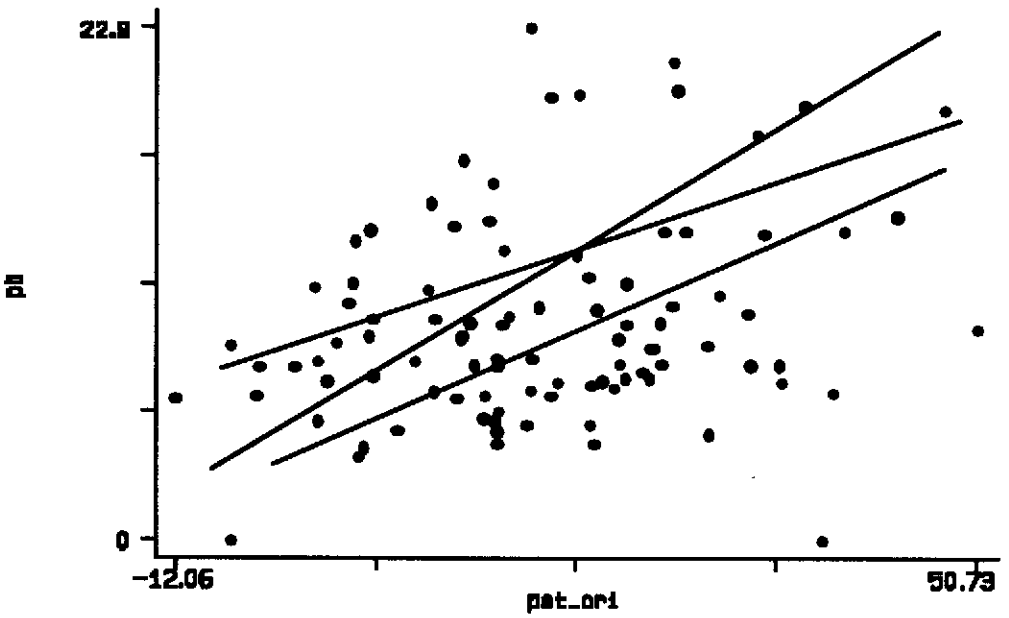
. sum sangre tibia rotula

Variable	Obs	Mean	Std. Dev.	Min	Max
sangre	95	9.657892	4.501572	0	22.89999
tibia	95	12.55147	11.62795	-20.25999	53.09
rotula |	95	16.67999	13.15133	-12.06	50.72998

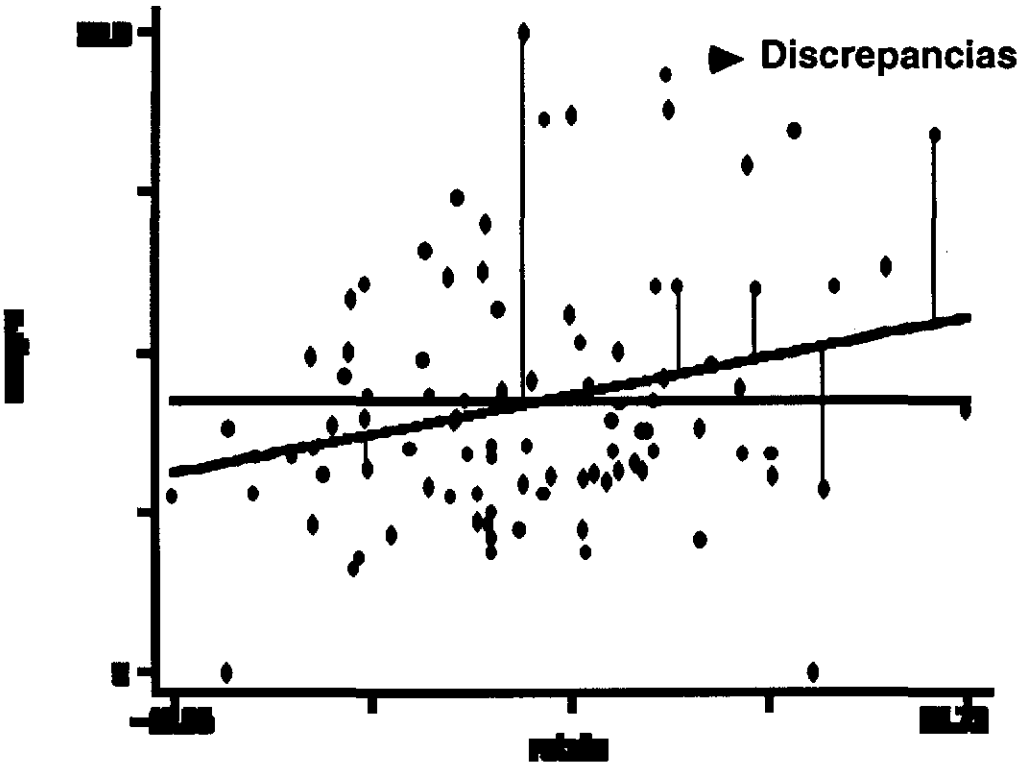
El gráfico de la relación entre la rotula y el plomo en sangre es la siguiente:



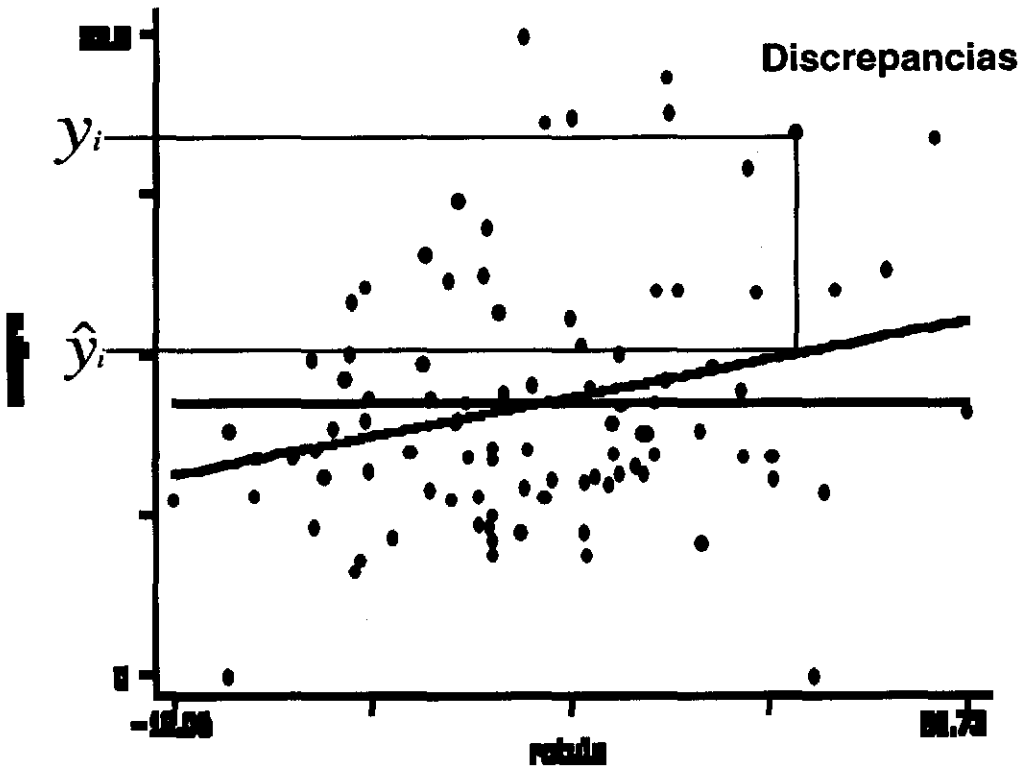
Se pueden ajustar diferentes rectas para resumir la relación entre la dos variables; sin embargo se requiere un método para encontrar la que mejor resume la relación entre el plomo en hueso y el plomo en sangre.



Discrepancias entre la recta y los puntos observados



Discrepancias entre la recta y los puntos observados



La discrepancia está dada por $y_i - \hat{y}_i$, puesto que las discrepancias pueden ser negativas, se utiliza el cuadrado de las discrepancias. Se puede substituir $\hat{y}_i$ por $\alpha + \beta x$, resultando la siguiente relación:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \alpha + \beta x)^2$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}$$

$$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

Se puede demostrar, por medio de cálculo, que los valores de α y β que minimizan la expresión anterior son:

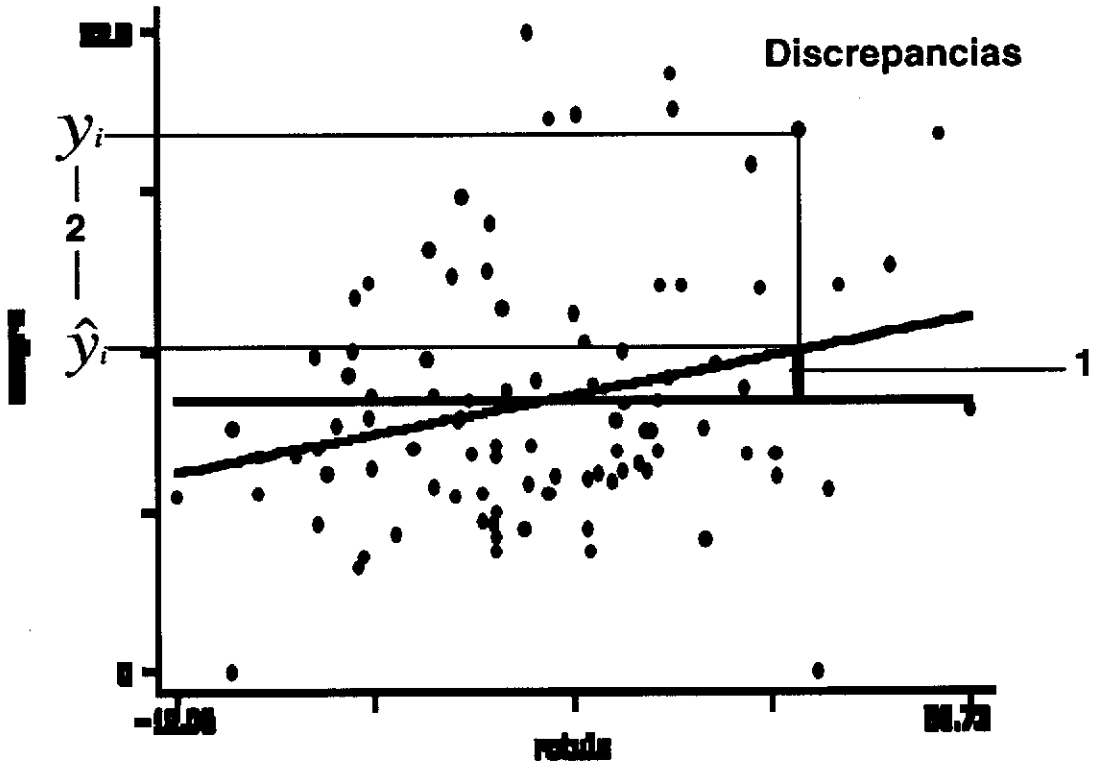
$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}$$

$$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

Actualmente existe un gran número de paquetes estadísticos que facilitan el cálculo de la regresión; en STATA la instrucción utilizada es **REG**

```
. reg sangre rotula

  Source |       SS       df       MS              Number of obs =      95
---------+------------------------------           F(  1,    93) =    6.66
   Model |   127.23982     1   127.23982           Prob > F      =  0.0114
Residual |  1777.59059    93  19.1138774           R-squared     =  0.0668
---------+------------------------------           Adj R-squared =  0.0568
   Total |  1904.83041    94  20.2641533           Root MSE      =  4.3719

------------------------------------------------------------------------------
  sangre |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  rotula |   .0884663   .0342879      2.580   0.011       .0203774    .1565553
   _cons |   8.182275   .7268379     11.257   0.000       6.738919    9.625631
------------------------------------------------------------------------------
```

Discrepancias entre la recta y los puntos observados



1 variación explicada por la regresion
2 error aleatorio experimental o de muestreo

Si los valores de discrepancia entre el valor estimado y la media de y son cercanos acero, la recta de regresión será horizontal, lo que indica que existe nulo o poco cambio en y por cada cambio de unidad en x. La comparación entre la variación explicada por el error y la explicada por la regresión se hace por medio de una razón de suma de cuadrados, lo que se conoce como la prueba de **F. Esta prueba proporciona información sobre la prueba de hipótesis en cuanto a la pendiente es cero. Mientras el valor de la razón de cuadrados se acerca a 1, se apoya la hipótesis de que b=0.**

De los modelos de regresión se puede obtener información sobre la relación entre las dos variables y si esta relación es estadísticamente significativa, es decir, que no se explica únicamente por el azar y que puede tener un significado biológico importante.

La regresión se usa también para predecir los valores que tomaría un individuo con x características.

I.SUPOSICION DE LOS MODELOS DE REGRESEION Y DIAGNOSTICOS UTILIZADOS

Antes de continuar con la aplicación de estos modelos, es conveniente revisar dos puntos:

1) ¿Cuáles son las suposiciones realizadas en el modelo?

2) ¿Cuáles son las ventajas de expresar las relaciones de interés en términos de modelos lineales?

Suposiciones del modelo de regresión lineal

La sobre simplificación de las relaciones biológicas que normalmente se estudian, a las que describe un modelo como el de regresión lineal, implica una serie de suposiciones sobre las variable incluidas en el modelo que deben ser consideradas, ya que si no se cumplen los resultados pueden ser erróneos, sesgados. Por esta razón es importante considerar y analizar estas suposiciones para poder prevenir la introducción de sesgo por el uso inapropiado de modelos.

Cuáles son las suposiciones del modelo de regresión simple:

1) Las observaciones de **y, la variable dependiente,** son una muestra aleatoria de una población de variables aleatorias con una media para cada población de $E(Y_i)$; cada observación es independiente y se puede derivar la observación Y_i de la población $E(Y_i)$ con cierto grado de error con el modelo:

$$Y_i = \alpha + \beta x_i + e_i$$

2) La variable **x, la independiente,** se mide sin error y es conocida

3) Las variables **x** y **y** se miden en el misma unidad de análisis

4) La relación entre y y x es linear o es razonablemente bien aproximada por una línea recta

5) El término del error tiene media 0

6) El término del error tiene varianza constante

7) Los errores no están correlacionados

8) El error se distribuye de manera normal (esto es particularmente importante para la estimación de intervalos de confianza y para las pruebas de hipótesis).

Cuando estas suposiciones no son correctas, es decir que no se cumplen, los resultados pueden ser erróneos. Podría suceder que al repetir los experimentos o estudios arrojen resultados incompatibles con los que se obtuvieron inicialmente. Las estimaciones de efectos o relación podrían estar muy sesgadas.

En el siguiente cuadro se describen algunos de los efectos cuando las suposiciones no se cumplen:

Problemas de la regresión y su impacto

problema	betas sesgadas	es sesgado	invalida pruebas de H	aumenta varianza
No linearidad	si	si	si	-
error en x	si	si	si	-
varianza no constante	no	si	si	si
Autocorrelación	no	si	si	-
correlación x y e	si	si	si	-
No normalidad de e	no	no	si	si
Multicolinearidad	si	si	no	si

En general la F, la R^2 y otras estadísticas no ayudan a resolver este problema, ya que son estadísticas globales. En este caso, tradicionalmente se recurre al estudio de los residuales.

¿Qué son los residuales? Son la discrepancias que existe entre el valor estimado con el modelo y el valor observado. En realidad los residuales se pueden ver como la variabilidad que no puede explicarse mediante el modelo de regresión.

Los residuales se pueden interpretar también como el valor del error. De esta manera, se pueden observar los residuales para saber si se cumplen o no las suposiciones básicas del modelo.

Gráficos de diagnóstico:

Curvas normales

Cuando la variable dependiente presenta desviaciones pequeñas de la distribución normal, no tienen efectos importantes sobre el modelo de regresión; se dice que en este sentido el modelo es robusto. Las desviaciones graves o las asimetrías importantes impactan la estimación de los intervalos de confianza y las pruebas de hipótesis.

Distribuciones que muestren asimetría hacia las "colas los valores extremos de la distribución" en general afectan la estimación de los mínimos cuadrados.
Un método muy sencillo para detectar la suposición de normalidad es el de gráficar los residuales y evaluar si se distribuyen de manera normal

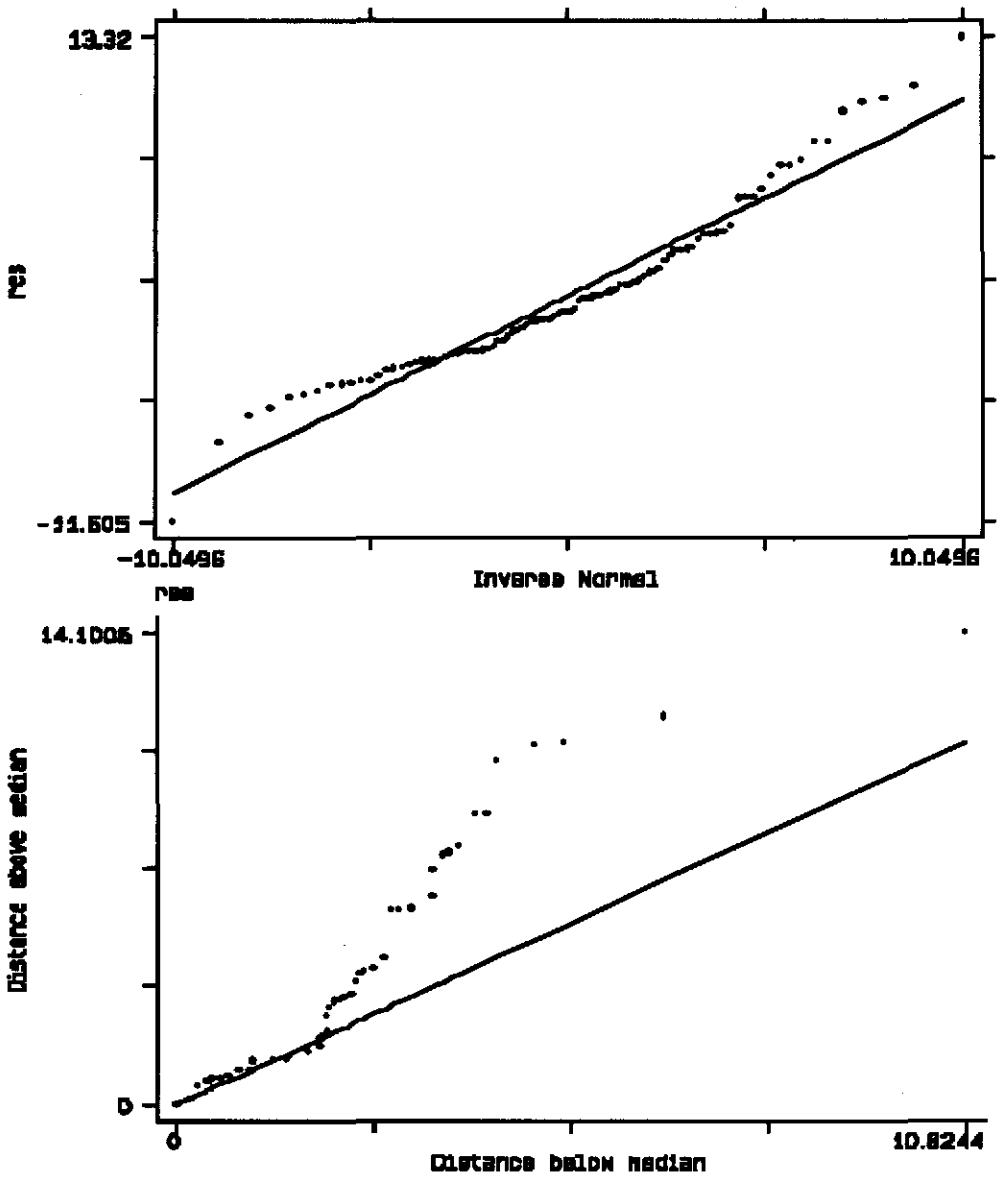
Los comandos que se utilizan son:
regres

predict nombre, resid

```
. . sum res

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
```

```
     res |        95    2.12e-08    4.348625   -11.60503    13.31995
. qnorm res
. syplot res
```

Gráficos de los residuales del modelo plomo en sangre vs plomo en rotula

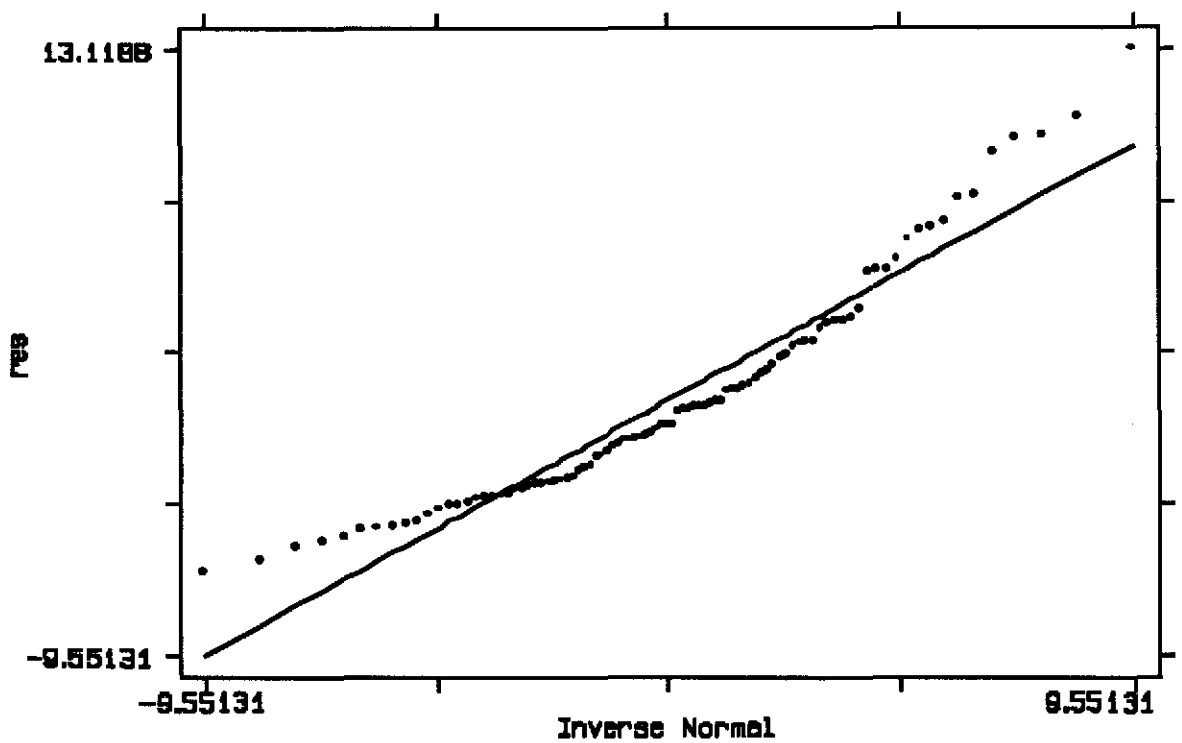


A continuación se presenta información sobre las variables incluidas en la regresión:

```
. sum sangre rotula,detail

                          sangre
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0             0
 5%     4.299999             0
10%     5.099998      3.699999       Obs                  95
25%     6.799999      4.099998       Sum of Wgt.          95

50%     8.699997                     Mean           9.657892
                       Largest       Std. Dev.      4.501572
75%         11.5      19.89999
90%           16      20.09999       Variance       20.26415
95%     19.79999      21.39999       Skewness        .864959
99%     22.89999      22.89999       Kurtosis       3.688746

                          rotula
-------------------------------------------------------------
      Percentiles      Smallest
 1%       -12.06        -12.06
 5%         -5.5      -7.699997
10%         -.14      -7.639999       Obs                  95
25%         7.98      -5.699997       Sum of Wgt.          95

50%         15.8                     Mean           16.67999
                       Largest       Std. Dev.      13.15133
75%        25.56      40.35999
90%        34.12      44.51999       Variance       172.9574
95%     39.57999      48.25998       Skewness        .228229
99%     50.72998      50.72998       Kurtosis       2.725546


. ladder sangre

Transformation         formula             Chi-sq(2)     P(Chi-sq)
-------------------------------------------------------------------
cube                   sangre^3                47.75         0.000
square                 sangre^2                32.34         0.000
raw                    sangre                  10.92         0.004
square-root            sqrt(sangre)            15.62         0.000
log                    log(sangre)                 .             .
reciprocal root        1/sqrt(sangre)              .             .
reciprocal             1/sangre                    .             .
reciprocal square      1/(sangre^2)                .             .
reciprocal cube        1/(sangre^3)                .             .

. ladder rotula

Transformation         formula             Chi-sq(2)     P(Chi-sq)
-------------------------------------------------------------------
cube                   rotula^3                56.41         0.000
square                 rotula^2                32.93         0.000
raw                    rotula                   1.06         0.589
square-root            sqrt(rotula)                .             .
log                    log(rotula)                 .             .
reciprocal root        1/sqrt(rotula)              .             .
reciprocal             1/rotula                    .         0.000
reciprocal square      1/(rotula^2)                .         0.000
reciprocal cube        1/(rotula^3)                .         0.000
. gen sangre2=sangre
```

```
. recode sangre2 0=.
(2 changes made)

. regress sangre2 rotula

  Source |       SS       df       MS                  Number of obs =      93
---------+------------------------------                F(  1,    91) =    7.58
   Model |  131.888632     1  131.888632                Prob > F      =  0.0071
Residual |  1582.38019    91  17.3887933                R-squared     =  0.0769
---------+------------------------------                Adj R-squared =  0.0668
   Total |  1714.26882    92  18.6333568                Root MSE      =    4.17

------------------------------------------------------------------------------
 sangre2 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  rotula |   .0932138   .0338463      2.754   0.007       .0259823    .1604454
   _cons |   8.308407    .711811     11.672   0.000       6.894482    9.722332
------------------------------------------------------------------------------

. predict res,resid
(2 missing values generated)

. qnorm res
```

Gráfico de normalidad incluyendo los ceros en la variable plomo en sangre



Se puede concluir que el impacto de estos valores no es importante. La interpretación de los gráficos de normalidad puede ayudar a descubrir problemas en los datos. Cualquier patrón de concentración de la nube de puntos que se separe de la linea recta, indica algún tipo de asimetría con respecto a la distribución normal. Una curva en forma de S sugiere problemas en los extremos de la distribución (colas de la distribución). Se pueden realizar pruebas de normalidad; sin embargo éstas deben interpretarse con cautela dado que los residuales no son independientes. No obstante, para tamaños superiores a 100 se consideran adecuadas cuando el número de parámetros estimados en el modelo no es muy grande.
En Stata se puede utilizar el comando **swilk,** que realiza la prueba de Shapiro-Wilk, esta prueba da información sobre el grado de concordancia entre la gráfica normal y la distribución esperada sobre la línea recta.

```
. swilk res

                    Shapiro-Wilk W test for normal data
 Variable |    Obs           W          V          z    Pr > z
 ---------+--------------------------------------------------
      res |     93      0.92618     5.737      3.860  0.00006
```

La W y representan los valores de las pruebas de Shapiro-Wilkins y la V el valor de la prueba. El valor esperado de V para distribuciones normales es de 1.

Como se puede observar, se rechaza la hipótesis sobre la normalidad de los residuales.

A continuación se llevará a cabo, el análisis utilizando la variable plomo en sangre, con una transformación logarítmica (loge).

```
. gen lsangre=ln(sangre2)
(2 missing values generated)

. regress lsangre rotula

  Source |       SS       df       MS                  Number of obs =      93
---------+------------------------------                F(  1,    91) =    7.23
   Model |  1.15420442     1  1.15420442                Prob > F      =  0.0085
Residual |  14.5236659    91  .159600724                R-squared     =  0.0736
---------+------------------------------                Adj R-squared =  0.0634
   Total |  15.6778703    92  .170411634                Root MSE      =   .3995

------------------------------------------------------------------------------
 lsangre |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  rotula |     .00872   .0032426      2.689   0.009        .002279    .0151611
   _cons |   2.057246   .0681942     30.167   0.000       1.921786    2.192705
------------------------------------------------------------------------------
```

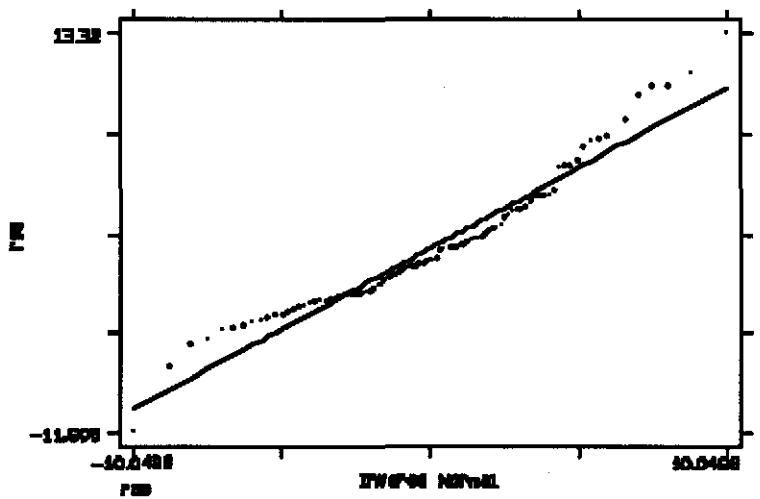
```
. predict res2,resid
(2 missing values generated)

. qnorm res2

. swilk sangre2 lsangre rotula res res2

                    Shapiro-Wilk W test for normal data
 Variable |    Obs           W          V          z    Pr > z
 ---------+--------------------------------------------------
  sangre2 |     93     0.90296      7.542      4.464  0.00000
  lsangre |     93     0.98393      1.249      0.492  0.31146
   rotula |     95     0.99087      0.723     -0.719  0.76388
      res |     93     0.92618      5.737      3.860  0.00006
     res2 |     93     0.98676      1.029      0.064  0.47454
```

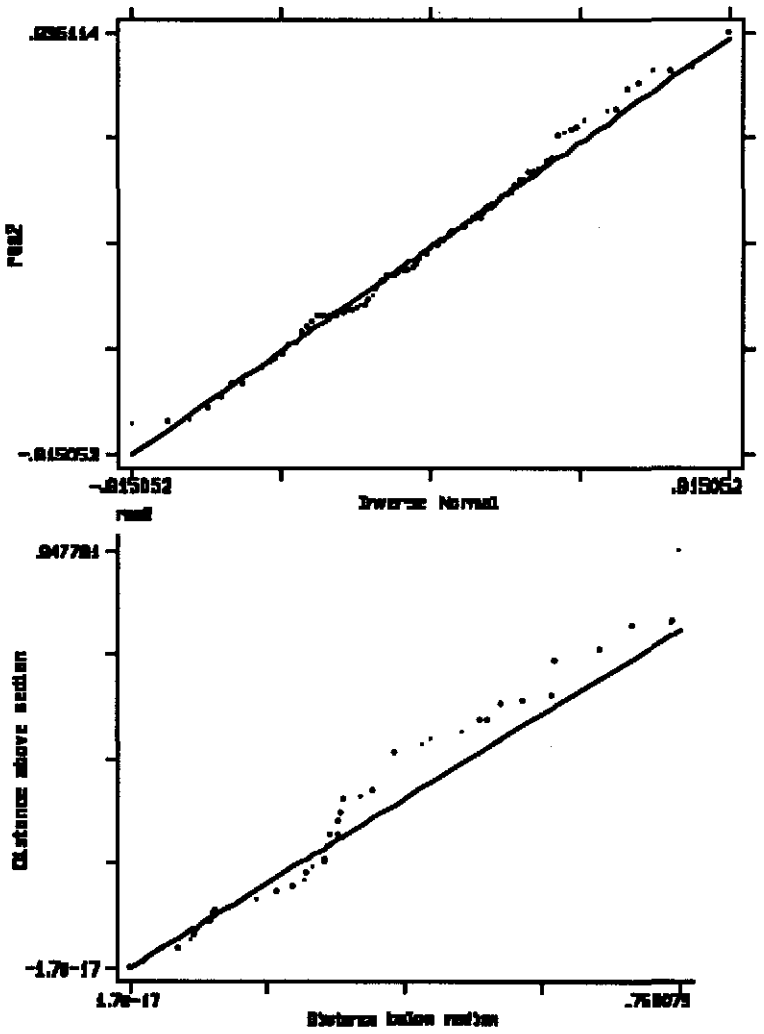
res se refiere a los residuales del modelo sin transformar, y res2 a los resduales del modelo utilizando el logaritmo de plomo en sangre.

La diferencia en la distribución de los residuales se observa en la siguiente figura

Gráficos de los residuales del modelo plomo en sangre vs plomo en rotula

transformación loge





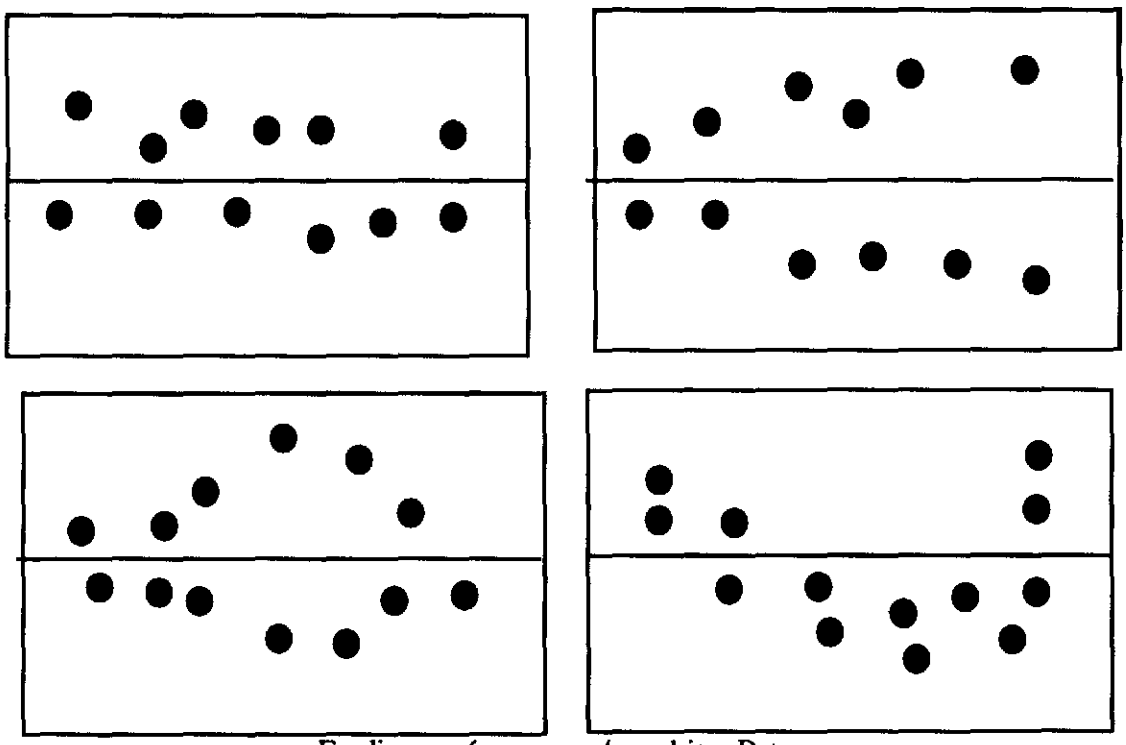
Gráficos de los residuales y el valor estimado mediante el modelo

$$e_i \text{ vs } \hat{y}_i$$

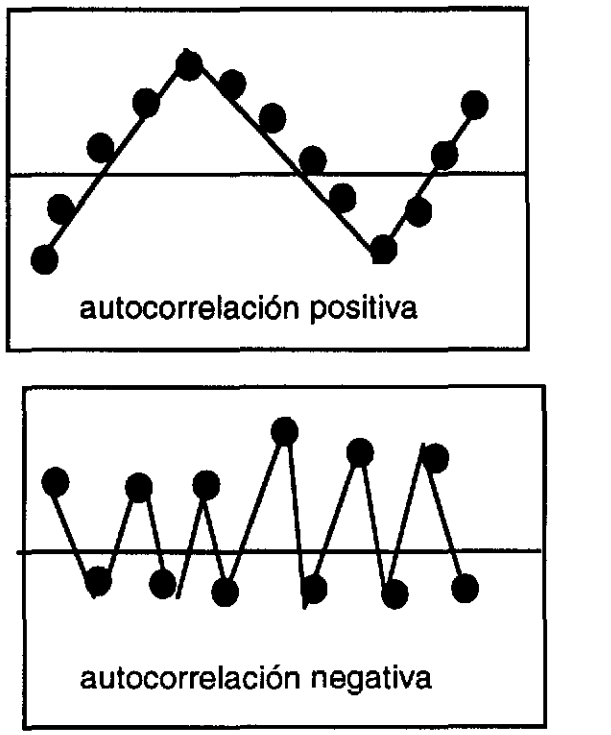
Dado que los valores observados en la variable independientemente y los residuales no son independientes, no se recomienda realizar gráficos diagnósticos utilizando estas variables.

Lo esperado en los gráficos de e_i vs $\hat{y}_i$ es que no exista relación entre los residuales y el valor esperado. Cualquier patrón de dependencia indica la existencia de algún problema.

La gráfica que se debe observar es de no dependencia

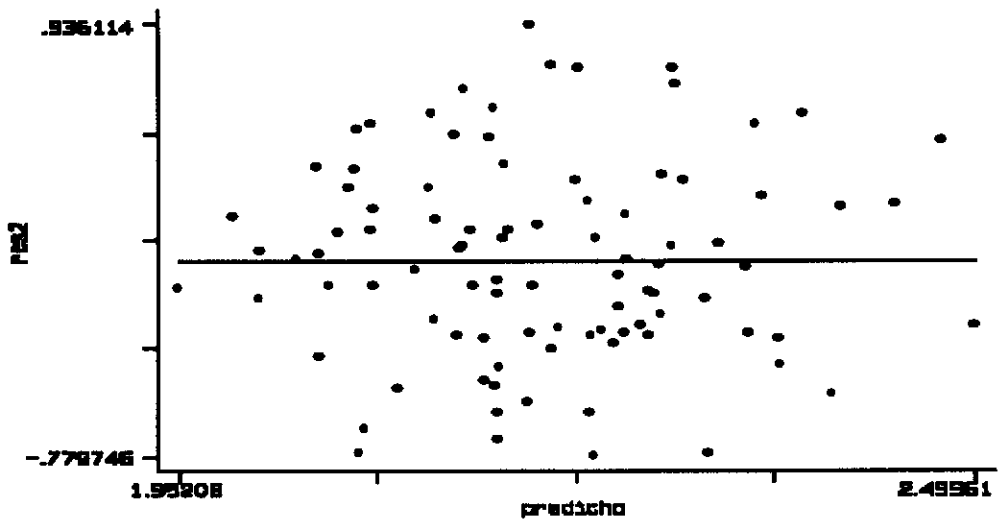


Explicar qué pasa en c/ cuadrito: Patrones



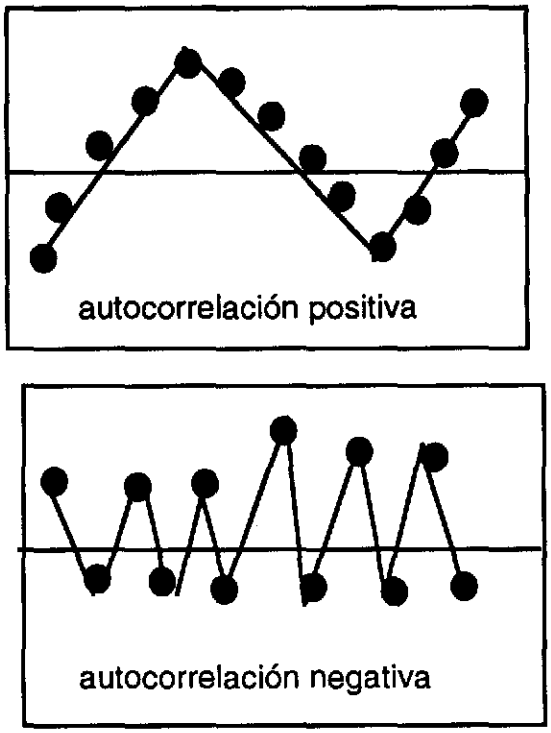
En nuestro ejemplo podemos observar lo siguiente:

Gráfico de residuales vs el valor predicho



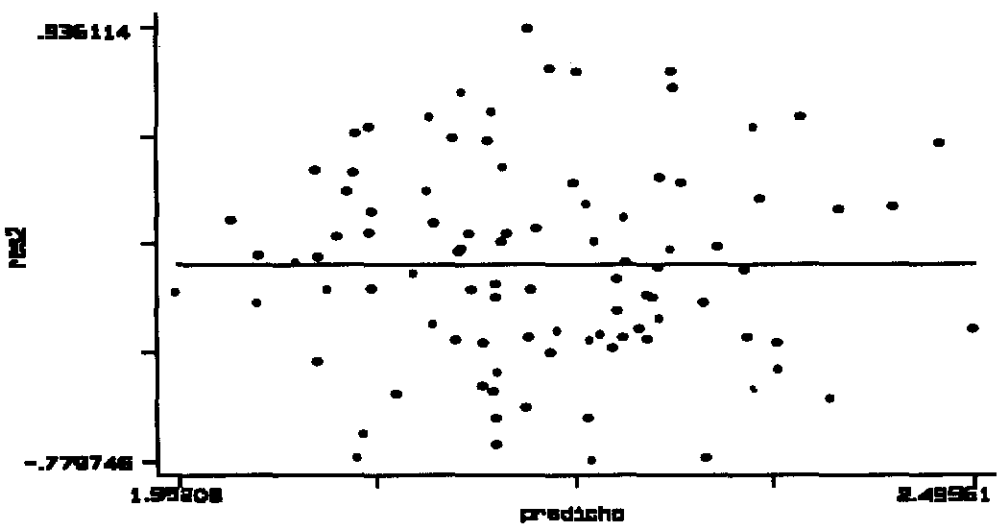
E

xplicar que pasa en c/ cuadrito Patrones



En nuestro ejemplo podemos observar lo siguiente:

Gráfico de residuales vs el valor predicho



En otra situación:

Gráfico de residuales vs el valor predicho

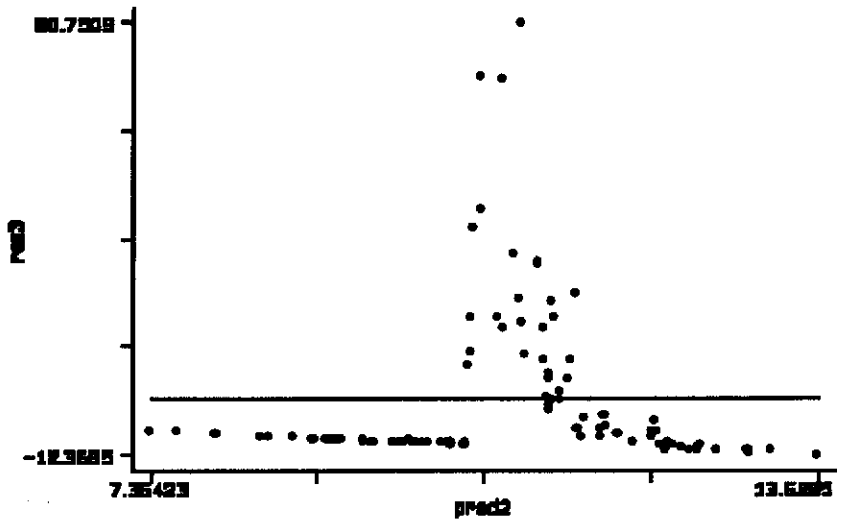


Gráfico de residuales vs el valor predicho
Plomo sangre y plomo en hueso (n=800)

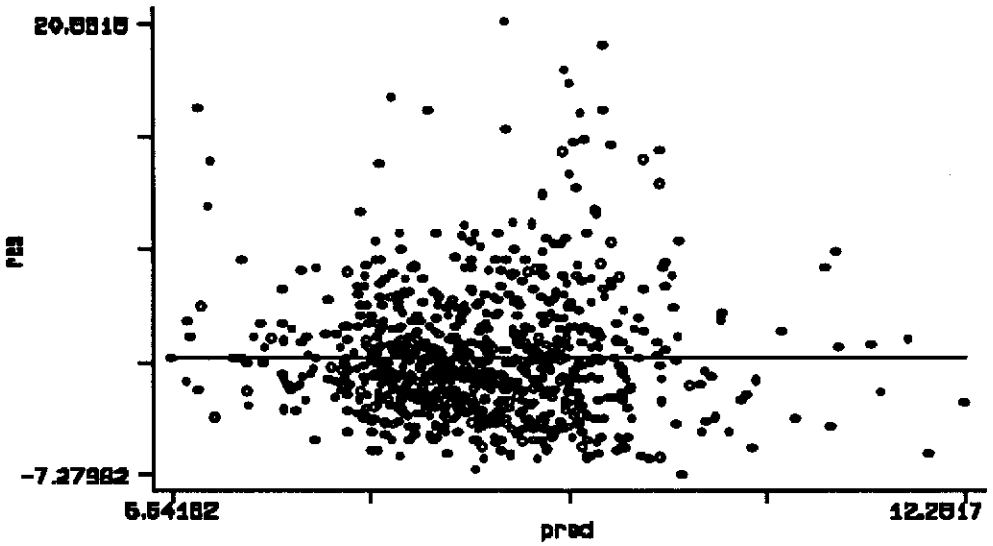
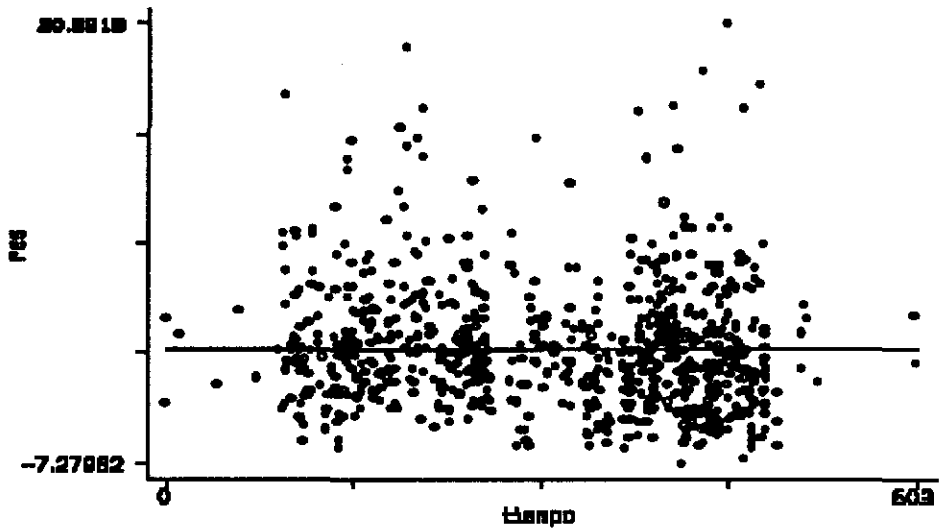
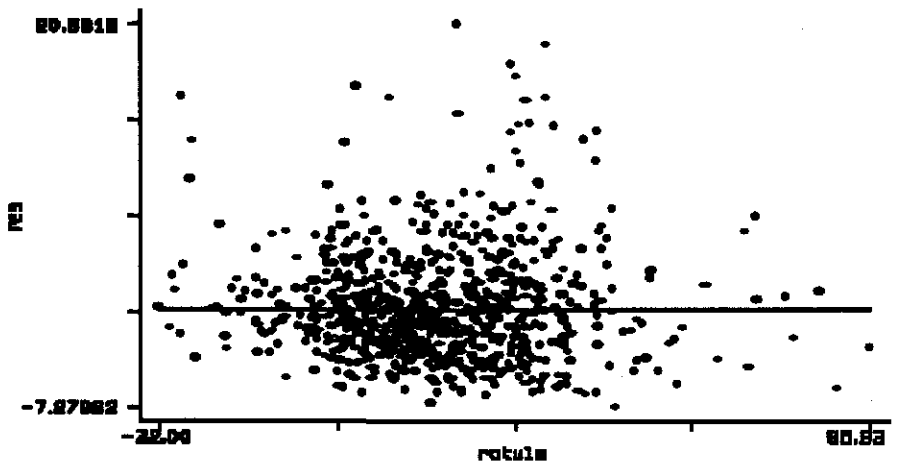


Gráfico de residuales vs el tiempo
Plomo sangre y plomo en hueso (n=800)



En este estudio de plomo en hueso se usa un equipo con una fuente radiactiva, la cual puede estar sujeta a error según para el tiempo. Debido a esto es interesante estudiar los residuales en relación al tiempo de estudio.

Otros gráficos que pueden ayudar en el diagnóstico son los de **error vs x**



Otros aspectos de los residuales son los residuales Studentizados que se definen de la siguiente manera:

predict nombre, rstudent

$$r_{ij} = \frac{e_{ij}}{S_{y|x}\sqrt{1-h_{ij}}}$$

interno

$$r_{ij} = \frac{e_{ij}}{S_{y|x(-i)}\sqrt{1-h_{ij}}}$$

Si el valor que se quita no tiene un efecto importante, entonces el estimador de la varianza permanece constante. Si por el contrario tiene efecto, es mejor eliminarlo de la estimación. Este estadístico es más sensible para detectar posibles valores aberrantes o outliers.

RESIDUALES INFLUYENTES (LEVERAGE):

Son estadísticas que ayudan a encontrar observaciones "influyentes" o de mucho peso. En general estos puntos aparecen claramente fuera de la nube de puntos. Una medida de la distancia de cada punto al centroide de puntos al "centroide" de puntos, se conoce como "the Hat Matrix", y los valores que puede tomar van desde:

$$\frac{1}{n} \leq h_{ij} \leq \frac{1}{c}$$

El valor mínimo se obtiene si todos los elementos de x_i son iguales a la media de la variable y si los datos caen en el centroide de la distribución. El valor máximo se presenta en observaciones alejadas del centroide. Si se obtiene el valor más alto, de 1, entonces el punto es tan influyente que forza la recta hasta su dirección, hasta pasar por el punto.

Para la regresión linear simple el LEVERAGE o peso se calcula de la siguiente manera:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

Idealmente todos los valores deben tener el mismo peso, por lo que el valor esperado debe ser pequeño y cercano a cero.

Si el valor se acerca a 1 esto significa que el residual se aproxima al valor esperado.

$h_{ij} = (k+1)/n$ donde k es
el numero de variables independientes

Se considera que las observaciones que toman valores dos veces por arriba del valor esperado potencialmente serán de gran peso para los parámetros estimados.

predict nombre, hat

$$h_{ij} > 2(k+1)/n$$

Para el ejemplo de plomo en hueso y sangre en el estudio completo, se tiene un parámetro en la regresión, por lo que el valor de corte para detectar observaciones influyente sería de **2*(1+1)/849=0.0047**

. lv hat

#	849	hat				
M	425	.0016754			spread	pseudosigma
F	213	.0013111	.001877	.002443	.0011319	.0008406
E	107	.0012189	.0023369	.0034548	.0022359	.0009742
D	54	.0011953	.0032367	.0052782	.0040829	.0013355
C	27.5	.0011919	.004343	.0074941	.0063022	.0017011
B	14	.0011908	.005628	.0100653	.0088745	.0020714
A	7.5	.0011905	.0074318	.0136731	.0124825	.0026121
Z	4	.0011905	.0092136	.0172367	.0160462	.003055
Y	2.5	.0011905	.0107916	.0203928	.0192023	.0034283
X	1.5	.0011905	.0120574	.0229244	.0217339	.0036287
	1	.0011905	.0128831	.0245758	.0233853	.0037127

			# below	# above
inner fence	-.0003868	.0041409	0	91
outer fence	-.0020847	.0058388	0	48

.
Existen cerca de 70 observaciones con una unfluencia alta

```
. list plomo rotula hat mes folio if hat>0.0047
```

	plomo	rotula	hat	mes	folio
1.	6.7	-32.00	.0110632	5	31849
2.	5.3	-29.89	.0101866	6	31890
3.	9.1	-29.59	.0100653	4	31830
4.	8.1	-29.17	.0098968	7	33128
5.	22.4	-28.16	.0094982	5	32349
6.	4.9	-28.12	.0094826	4	31002
7.	10.1	-27.60	.0092813	5	32325
8.	16.3	-26.59	.0088972	5	32345
9.	19.2	-26.28	.0087813	4	30806
10.	3.2	-25.74	.0085813	5	31876
11.	7	-23.03	.0076181	4	32225
12.	7.1	-22.30	.0073701	7	32565
13.	7.1	-21.71	.0071733	7	20032
14.	13.2	-21.63	.0071468	5	31813
15.	5.1	-20.86	.0068953	3	30353
16.	6.9	-20.79	.0068727	3	20282
17.	4.2	-20.56	.0067988	8	30242
18.	8.6	-19.76	.0065454	6	31700
19.	9.4	-18.81	.006252	5	30555
20.	7	-18.37	.006119	4	31725
21.	7.9	-18.11	.0060412	6	30503
22.	8.6	-17.28	.005797	7	32582
23.	6.5	-15.64	.0053331	7	33204
24.	11.8	-15.51	.0052974	3	10115
25.	9.6	-15.48	.0052892	6	31985
26.	6.2	-15.44	.0052782	5	30499
27.	8.5	-15.40	.0052673	6	31605
28.	4.2	-15.31	.0052427	4	32315
29.	5.8	-14.73	.0050863	6	33015
30.	5.5	-14.25	.0049592	1	30060
31.	9.3	-14.15	.0049329	5	30844
32.	.	-13.78	.0048367	11	30600
33.	5.7	-13.56	.0047801	9	30384
34.	4.2	-13.45	.004752	6	31986
815.	15.4	42.70	.0048437	2	10113
816.	13.4	42.75	.0048566	10	30510
817.	6	43.17	.0049662	7	31688
818.	10	43.30	.0050005	6	33044
819.	11.6	43.47	.0050455	6	33019
820.	17.6	43.59	.0050775	5	32361
821.	3	44.07	.0052066	5	31049
822.	8.6	45.33	.0055555	3	20214
823.	8.8	46.80	.0059809	9	20256
824.	.	46.84	.0059927	5	31947

825.	5.7	47.09	.0060672	6	30981
826.	9.6	47.53	.0061995	4	32274
827.	6.5	47.58	.0062147	6	33011
828.	9.3	48.39	.0064633	3	30296
829.	6.8	49.07	.0066767	9	30325
830.	12.8	49.90	.0069429	6	32534
831.	13.1	49.91	.0069461	3	30326
832.	13.4	50.08	.0070014	6	33030
833.	6	51.15	.0073557	8	32612
834.	8.1	53.04	.0080069	8	10052
835.	8.4	53.84	.0082925	6	31935
836.	5.2	54.41	.0084995	4	31033
837.	.	54.66	.0085912	7	33143
838.	9.4	55.13	.0087652	5	31899
839.	12.7	59.03	.0102871	3	20236
840.	7.3	61.11	.0111557	5	31888
841.	16.9	65.48	.0131095	12	30940
842.	7	66.16	.0134293	3	30628
843.	18.1	67.18	.0139168	7	33232
844.	12	67.45	.0140475	6	30940
845.	12.4	72.25	.0164816	6	32514
846.	9.5	73.66	.0172367	5	32316
847.	13.1	77.72	.0195126	5	31765
848.	6.1	80.69	.021273	6	32438
849.	9.4	85.93	.0245758	9	30628

El leverage o influencia identifica elementos lejanos al centroide de la variable independientemente (x); por lo tanto pueden no tener mucha influencia sobre la línea de regresión.

Una manera de identificar la influencia de una observación es excluyendo ésta de la regresión y comparando los resultados que se obtienen con y sin la observación.

Otra forma es la de excluir todos y comparar resultados:

. reg lpb rotula

Source	SS	df	MS		
Model	5.08123875	1	5.08123875	Number of obs	= 840
Residual	151.584807	838	.180888791	F(1, 838)	= 28.09
Total	156.666046	839	.186729494	Prob > F	= 0.0000
				R-squared	= 0.0324
				Adj R-squared	= 0.0313
				Root MSE	= .42531

lpb	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
rotula	**.0048223**	**.0009099**	**5.300**	**0.000**	**.0030364**	**.0066081**
_cons	2.018848	.0197006	102.476	0.000	1.980179	2.057516

. reg lpb rotula if hat <0.0047

Source	SS	df	MS		
Model	5.09744461	1	5.09744461	Number of obs	= 774
Residual	139.403865	772	.180574955	F(1, 772)	= 28.23
Total	144.50131	773	.186935717	Prob > F	= 0.0000
				R-squared	= 0.0353
				Adj R-squared	= 0.0340
				Root MSE	= .42494

lpb	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
rotula	**.0066295**	**.0012478**	**5.313**	**0.000**	**.0041801**	**.0090789**
_cons	1.990848	.0233987	85.084	0.000	1.944916	2.036781

Asimismo existen otros métodos que ayudan a diagnosticar el efecto de cada observación:

1) Distancia de Cook: mide el efecto sobre la beta
2) DFFITS mide el efecto en el valor predicho o estimado
3) DFBETAS mide el efecto sobre beta

DISTANCIA DE COOK

predict nombre, cooksd

La distancia de Cook permite detectar posibles valores aberrantes; la medida de Cook cuantifica el impacto de la observación o del punto sobre el modelo; cuantifica qué tanto cambia el modelo, es decir, los coeficientes de regresión, al excluir cada uno de los puntos.

Se espera que los resultados de la regresión no dependan de una sola observación o punto de regresión.

Distancia de Cook:

$$D_i = \frac{r_i^2}{p'}\left(\frac{h_{ij}}{1-h_{ij}}\right)$$

donde r_i^2 es el residual estandarizado, h_{ij} la diagonal de la matriz sombrero (hat) y p' el número de parámetros en el modelo.

La distancia de Cook combina una medida de influencia o peso y de falta de ajuste, y se distribuye como una F con p+1 y n-p-1 grados de libertad.

Los puntos que toman valor por arriba de uno ameritan investigación; los que pasan de dos indican serios problemas.

En el ejemplo:

```
. predict cook,cooksd
(38 missing values generated)

. sum cook

Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
    cook |       840    .0019108    .0052054   8.62e-09   .0817741
```

Gráfico de residuales vs la variable independiente. Area proporcional a la distancia de Cook. Plomo sangre y plomo en hueso (n=800)

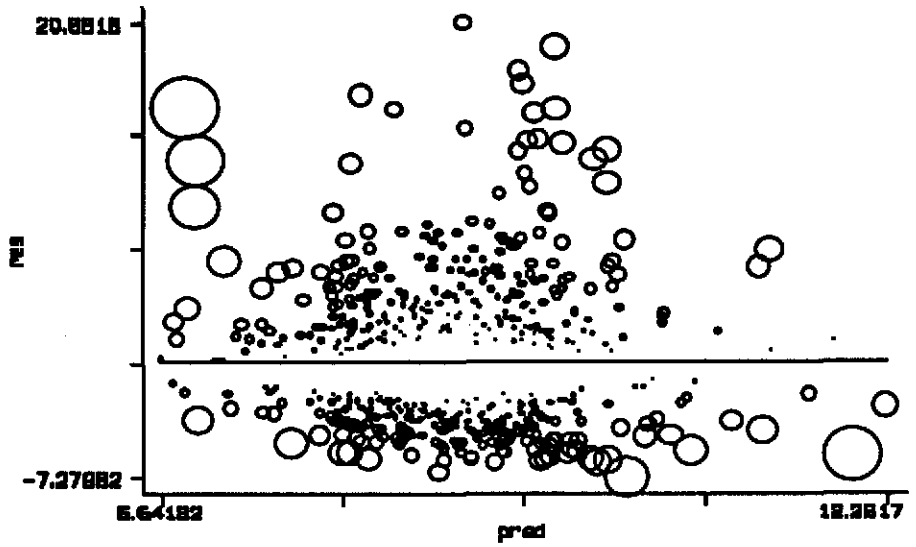
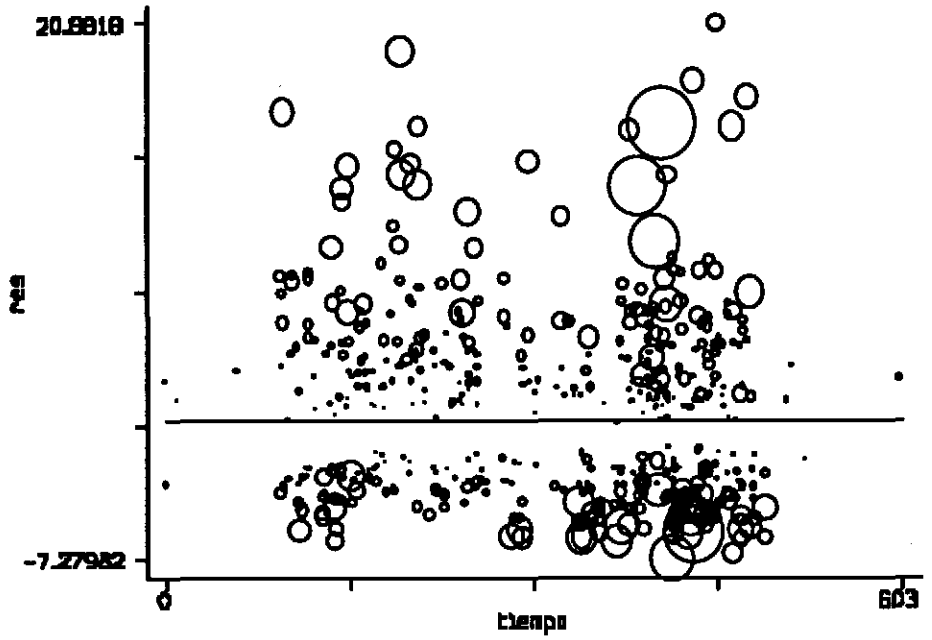


Gráfico de residuales vs la variable tiempo Area proporcional a la distancia de Cook Plomo sangre y plomo en hueso (n=800)



2) DFFITS

Los DFFITS informan acerca de cómo cambia el valor predicho al excluir la $\mathbf{x_i}$ observación. Su interpretación es muy similar a la distancia de Cook, por lo que no amerita utilizar mas de uno.

3) dfbetas

Este diagnóstico ayuda e evaluar el impacto sobre el vector de betas (β). Cabe recordar que no todos los outliers o valores aberrantes influyen en los todos estimadores.

Este diagnóstico indica el impacto que ejerce sobre las betas el eliminar la observación en cuestión y expresa la magnitud de cambio en unidades desviación estandar.

$$\text{DFBETAS} = \frac{b_k - b_{k(i)}}{S_{e(i)} / \sqrt{RSS_K}}$$

SI DFBETAS $>$ 0 sobre estima las b' s

SI DFBETAS $<$ 0 sub estima las b' s

Alternativas de interpretación:

1.- Como valores normales, |dfbetas| > 2 o para corregir por el tamaño de muestra:

$$\text{SI } |\text{DFBETAS}| > 2/\sqrt{n}$$

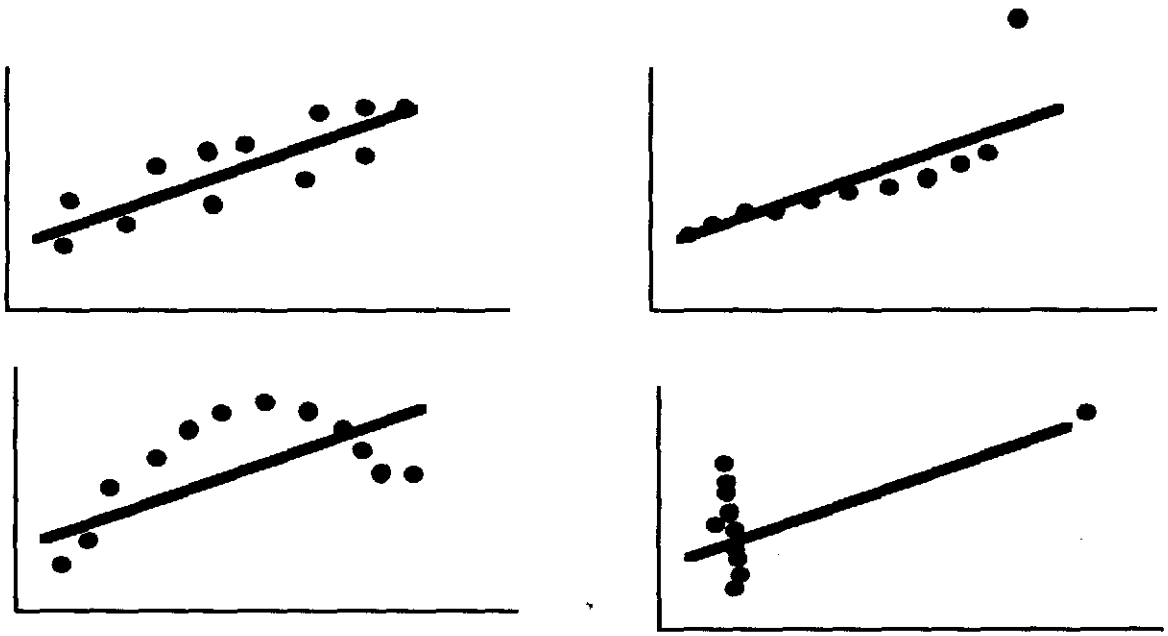
. fit lpb rotula

Source	SS	df	MS	Number of obs = 840
				F(1, 838) = 28.09
Model	5.08123875	1	5.08123875	Prob > F = 0.0000
Residual	151.584807	838	.180888791	R-squared = 0.0324
				Adj R-squared = 0.0313
Total	156.666046	839	.186729494	Root MSE = .42531

lpb	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]
rotula	.0048223	.0009099	5.300	0.000	.0030364	.0066081
_cons	2.018848	.0197006	102.476	0.000	1.980179	2.057516

Suposición de linearidad de la reacción entre y y x

Diferentes situaciones para las cuales se obtiene el mismo resultado de regresión, la misma pendiente



II. EXTENSIÓN DE LOS MODELOS BI-VARIADOS A LOS MODELOS MULTI-VARIADOS

Los métodos bi-variados son útiles para examinar la relación entre dos variables, sin embargo en el área de investigación observacional (no experimental) los modelos utilizados representan una sobre simplificación de las relaciones estudiadas y frecuentemente pueden arrojar resultados erróneos. Esto se debe a que en los estudios observacionales, el investigador no tiene control sobre todas las variables que pueden afectar los resultados observados.
Como ejemplo de los problemas asociados a los modelos bivariados analizaremos el siguiente ejemplo.

Se recolectaron datos sobre dos escuelas una rural y una urbana, con el objeto de evaluar el estado nutricional de los niños que acuden a cada escuela.
Los resultados obtenidos son los siguientes: en la escuela urbana se midieron 18 niños y en la rural 14.

obs	urbana	obs	rural
1.	132.7	1.	131
2.	133	2.	134.6
3.	133.2	3.	134.9
4.	135	4.	135.8
5.	136.8	5.	139.9
6.	137.6	6.	139.9
7.	140.7	7.	140.9
8.	145.4	8.	142.3
9.	147.5	9.	142.9
10.	147.8	10.	147.7
11.	148.3	11.	147.7
12.	148.3	12.	148.5
13.	148.7	13.	148.7
14.	148.8	14.	149.5
15.	149.9		
16.	150.6		
17.	152.2		
18.	165.3		

En stata el archivo tiene la siguiente estructura

. list escuela talla

	escuela	talla
1.	1	132.7
2.	1	133
3.	1	133.2
4.	1	135
5.	1	136.8
6.	1	137.6
7.	1	140.7
8.	1	145.4
9.	1	147.5
10.	1	147.8
11.	1	148.3
12.	1	148.3
13.	1	148.7
14.	1	148.8
15.	1	149.9
16.	1	150.6
17.	1	152.2
18.	1	165.3
19.	2	131
20.	2	134.6
21.	2	134.9
22.	2	135.8
23.	2	139.9
24.	2	139.9
25.	2	140.9
26.	2	142.3
27.	2	142.9
28.	2	147.7
29.	2	147.7
30.	2	148.5
31.	2	148.7
32.	2	149.5

Podemos analizar los datos en términos de una diferencia de medias utilizando la técnica de anova y obtenemos los siguientes resultados:

```
tab escuela,sum(talla) anova

    1=urbana|       Summary of  talla en cms
     2=rural|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   144.54444   8.5911878          18
          2 |   141.73571   6.0949666          14
------------+------------------------------------
      Total |   143.31562   7.6195892          32

                      Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      62.1256945      1   62.1256945      1.07     0.3086
 Within groups      1737.67665     30   57.9225551
------------------------------------------------------------------------
    Total           1799.80235     31   58.0581403

Bartlett's test for equal variances:  chi2(1) =   1.5995  Prob>chi2 = 0.206

.
```

Las medias de talla son 144 cms y 141 cm. respectivamente para la escuela urbana y la rural. La prueba de anova indica que la posibilidad de observar una diferencia de 2.81 cm. o mas extrema por azar es de 0.30, valor superior al punto de corte de 0.05. Por lo que podemos concluir que la diferencia observada puede únicamente deberse al azar.
Para fines de comparación vamos a re-expresar la relación entre las escuelas y la talla de los sujetos de estudio en base a un modelo de regresión.

$$y_i = \alpha + \beta x_i$$

donde y_i es la medici n de talla

x_i es la medici n escuela

y α y β son los par metros que estimaremos de modelo

```
reg talla escuela

  Source |       SS       df       MS                  Number of obs =      32
---------+------------------------------                F(  1,    30) =    1.07
   Model |  62.1256945     1  62.1256945                Prob > F      =  0.3086
Residual |  1737.67665    30  57.9225551                R-squared     =  0.0345
---------+------------------------------                Adj R-squared =  0.0023
   Total |  1799.80235    31  58.0581403                Root MSE      =  7.6107

------------------------------------------------------------------------------
   talla |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 escuela |  -2.808732   2.712056     -1.036   0.309      -8.347489    2.730024
   _cons |   147.3532   4.124197     35.729   0.000       138.9304    155.7759
------------------------------------------------------------------------------
```

Veamos como cambian los coeficientes cuando re-codificamos el valor de escuela como urbano=1 y rural=0.

```
. reg talla escuela

  Source |       SS       df       MS                  Number of obs =      32
---------+------------------------------                F(  1,    30) =    1.07
   Model |  62.1256945     1  62.1256945                Prob > F      =  0.3086
Residual |  1737.67665    30  57.9225551                R-squared     =  0.0345
---------+------------------------------                Adj R-squared =  0.0023
   Total |  1799.80235    31  58.0581403                Root MSE      =  7.6107

------------------------------------------------------------------------------
   talla |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 escuela |   2.808732   2.712056      1.036   0.309      -2.730024    8.347489
   _cons |   141.7357   2.034042     69.682   0.000       137.5816    145.8898
------------------------------------------------------------------------------
```

analicemos como obtener las medias para las escuelas apartir de los modelos ajustados:

$y_i = \alpha + \beta x_i$

para estimar la media $\bar{y}$ esperada de talla seg n escuela :

para el primer caso x toma valores 1 o 2, la media para

la escuela rural es $\bar{y} = \alpha + 2\beta$ y la media para

la escuela urbana es $\bar{y} = \alpha + \beta$.

Para el caso en que escuela toma valores 0 y 1, la media para

la escuela rural es $\bar{y} = \alpha + \beta$ y la media para

la escuela urbana es $\bar{y} = \alpha$.

Si realizamos las substituciones, en el primer modelo es para la escuela rural: 147.3532 - 2(2.808732)=141.73571 para la escuela urbana :147.3532 - 1(2.808732)=144.544.
Para el segundo caso, la media de la escuela rural esta dada por la contante del modelo que es 141.7357 y la media de la escuela urbana se puede obtener mediante el siguiente cálculo 141.7357 + 2.808732=144.544.
Nótese que en ambos casos la b que representa la diferencia de medias de talla entre las escuelas toma el mismo valor absoluto de 2.80, la diferencia en el signo se debe a la diferente codificación.
Que pasaría si la escuela fuera codificada con valores de -1 y +1.

$$\bar{y} = \alpha + \beta x$$

$$\bar{y}_{urbana} - \bar{y}_{rural} = \alpha + \beta x_{urbana} - (\alpha + \beta x_{rural})$$

$$= \alpha + \beta(1) - (\alpha + \beta(-1))$$

$$= \alpha - \alpha + \beta - (-\beta)$$

$$= 2\beta$$

En este caso, el estimador de la diferencia de medias estaría dado por $2b$ Ajustemos el modelo para ver los resultados:

```
. gen esc_1=escuela

. recode esc_1 0=-1
(14 changes made)

. tab escuela esc_1

   1=urbana|  esc_1
    2=rural|        -1          1 |     Total
-----------+----------------------+----------
         0 |        14          0 |        14
         1 |         0         18 |        18
-----------+----------------------+----------
     Total|        14         18 |        32


. regress talla esc_1

  Source |       SS       df       MS                  Number of obs =      32
---------+------------------------------                F(  1,    30) =    1.07
   Model |  62.1256945     1  62.1256945                Prob > F      =  0.3086
Residual |  1737.67665    30  57.9225551                R-squared     =  0.0345
---------+------------------------------                Adj R-squared =  0.0023
   Total |  1799.80235    31  58.0581403                Root MSE      =  7.6107

------------------------------------------------------------------------------
   talla |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   esc_1 |   1.404366   1.356028       1.036   0.309      -1.365012    4.173744
   _cons |   143.1401   1.356028     105.558   0.000       140.3707    145.9095
------------------------------------------------------------------------------
```

En este modelo el estimador de b es 1.404366 ademas sabemos que el estimador de la diferencia de medias esta dado por $2b$ por lo que la diferencia de medias es 2*1.404366=2.8087, valor que es igual al obtenido en los modelos anteriores.

La comparación entre la talla de las escuelas arroja una diferencia de 2.80 cm., la cual no es estadísticamente significativa. Sin embargo existen otros factores que podrían determinar la talla de los niños en estudio y que no han sido considerados en el modelo. Un ejemplo de una de estas variables es la edad. Diferencia en la edad podría estar distorsionando las comparaciones. Analicemos la edad por grupo de escuela.

```
. tab escuela,sum(edad) anova

    1=urbana|       Summary of  edad en meses
     2=rural|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   133.07143   6.5452776          14
          1 |   126.83333   10.205247          18
------------+------------------------------------
      Total |    129.5625   9.2175761          32

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      306.446429       1   306.446429      3.95     0.0561
 Within groups      2327.42857      30   77.5809524
------------------------------------------------------------------------
    Total             2633.875      31   84.9637097

Bartlett's test for equal variances:  chi2(1) =   2.6241  Prob>chi2 = 0.105
.
```

Observamos que la media de edad es diferente para cada escuela, los niños de la escuela rural son en promedio 6.2 meses mas grandes. Sabemos ademas que existe una relación entre la edad y la talla, a mayor edad mayor talla, al menos en este grupo de edad.

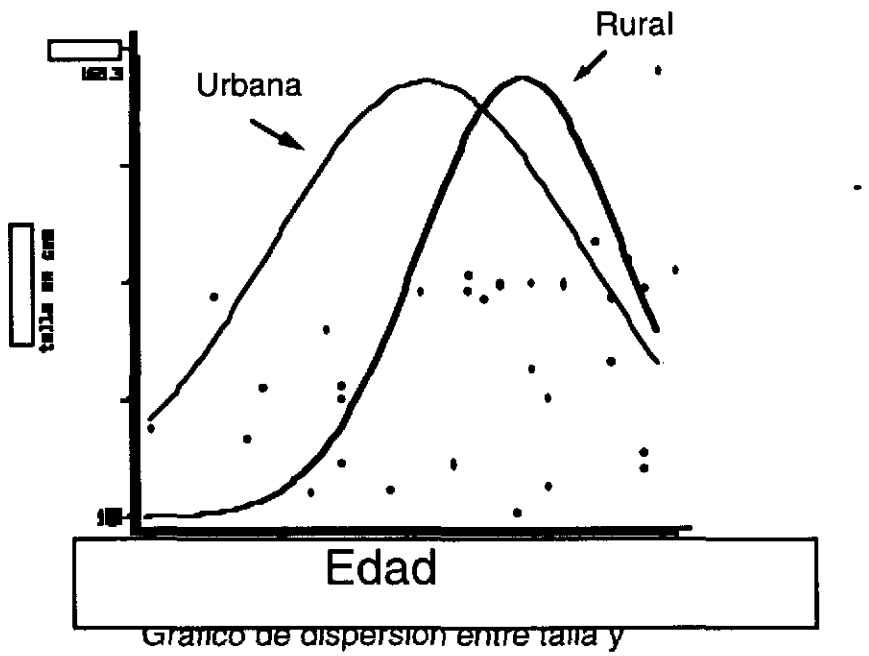


Grafico de dispersion entre talla y edad para las escuelas

Para que la comparación entre talla y tipo de escuela sea valida, -no este sesgada por diferencias en la edad- la edad de los niños que acuden a cada escuela debe ser la misma. Es poco probable que esto se pueda lograr mediante el diseño del estudio. Lo que podemos hacer es realizar la comparación de medias de talla tomando en cuanta el efecto de la edad. Esto se puede hacer incorporando la edad dentro del modelo de predicción.

$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2$

donde y_i es la medici n de talla

x_1 es la medici n de escuela

x_2 es la medici n de edad

y α β_1 y β_2 son los par metros que estimaremos del modelo solo que ahora β_1 es la diferencia de medias de talla entre escuelas, corregida por el efecto de edad

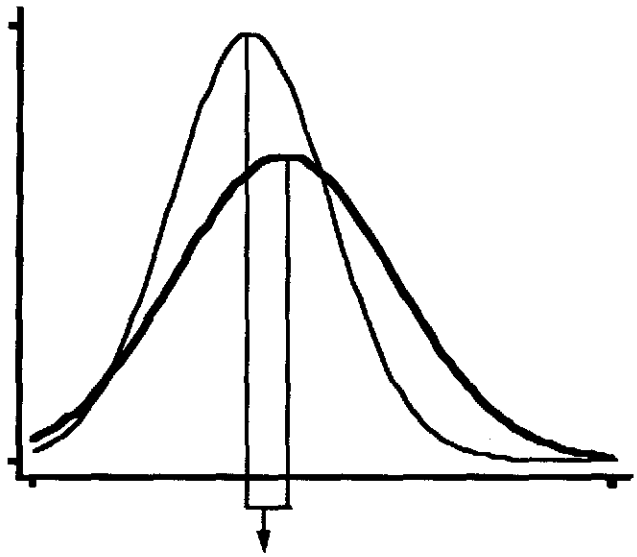
Ajustamos el modelo tomando en cuenta el efecto de edad:

```
regress talla escuela edad

  Source |       SS       df       MS                  Number of obs =      32
---------+------------------------------                F(  2,    29) =    4.94
   Model |  457.190349     2  228.595175                Prob > F      =  0.0143
Residual |    1342.612    29  46.2969655                R-squared     =  0.2540
---------+------------------------------                Adj R-squared =  0.2026
   Total |  1799.80235    31  58.0581403                Root MSE      =  6.8042

------------------------------------------------------------------------------
   talla |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 escuela |    5.37882   2.579351      2.085   0.046       .1034555    10.65418
    edad |   .4119988   .1410386      2.921   0.007       .1235424    .7004551
   _cons |   86.91045   18.85611      4.609   0.000       48.34538    125.4755
------------------------------------------------------------------------------
```

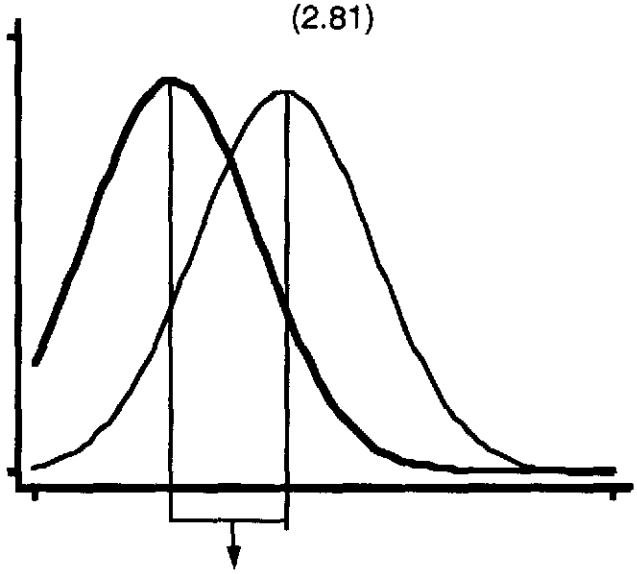
Los resultados indican que si tomamos en cuenta el efecto de edad, la diferencia entre la escuela urbana y rural es de 5.37 cm., ademas esta diferencia es estadísticamente significativa.

Las estadística asociadas al modelo cambian. Aumenta el coeficiente de determinación r^2 pasa de 2.3% a 20.2%, esto se debe a que edad es un determinante importante de la talla y a que las diferencias por escuela son también importantes al corregir por la diferencia en edad. Esto se puede observar gráficamente en el siguiente esquema:

.



Diferencia de medias cruda
(2.81)

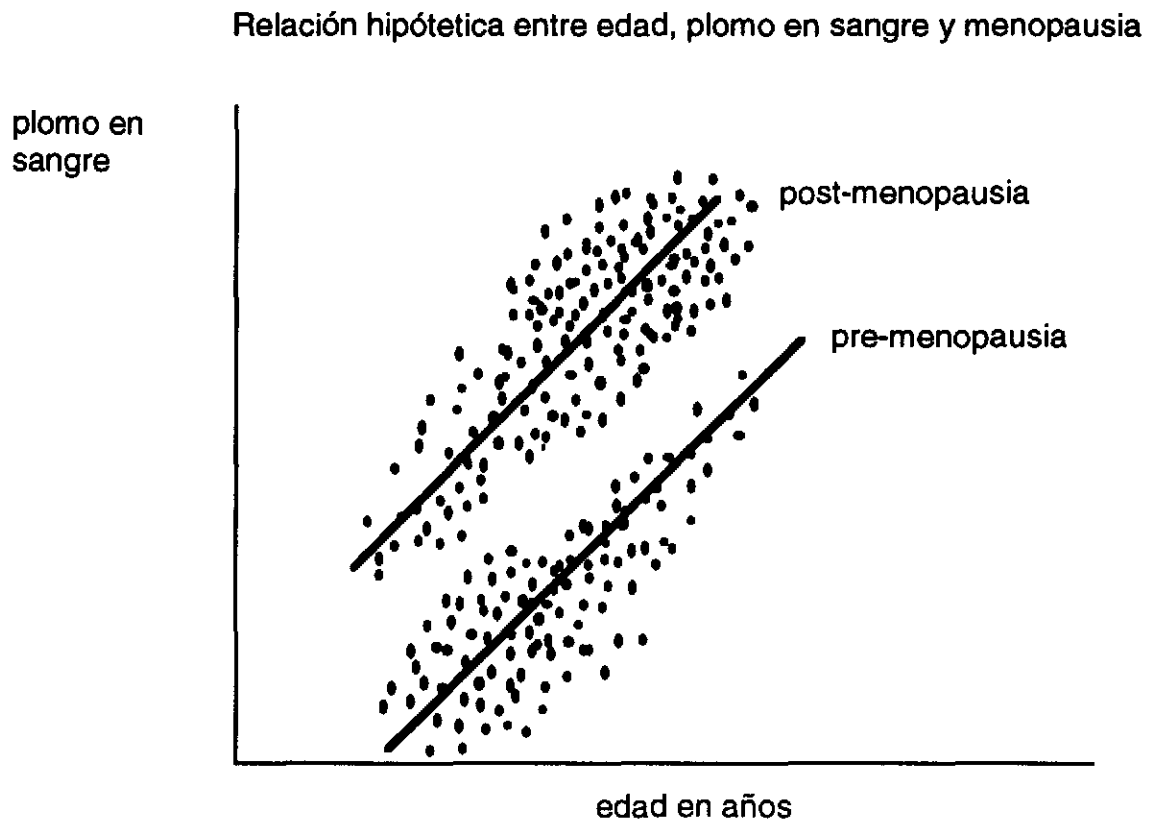


Diferencia de medias ajustada
por edad (5.37)

En este ejemplo hemos analizado el efecto que puede tener el ignorar ciertas variables que pueden ser importantes. En este caso hemos realizado la suposición sobre igualdad de efectos.

Analicemos ahora otras situaciones hipotéticas sobre la relación

entre tres variables:

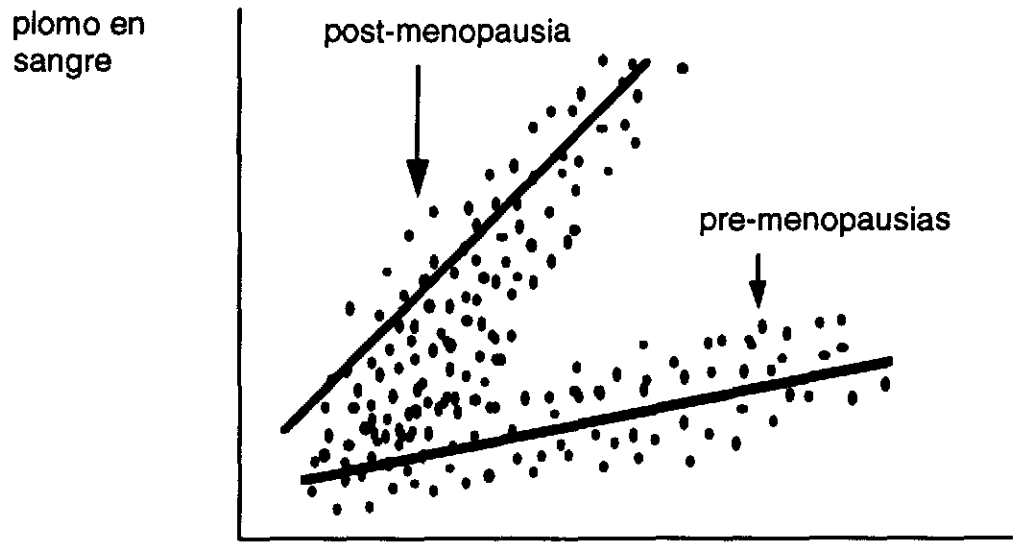


En este caso, se puede apreciar que las concentraciones de plomo en sangre aumentan conforme la edad, la edad es un determinante importante de las concentraciones de plomo en sangre. El aumento conforme la edad es parecido en mujeres pre y postmenopausicas, sin embargo las mujeres premenopausicas tienen un nivel menor de plomo en sangre. Si comparamos los niveles de plomo en sangre entre mujeres sin considerar la edad, la comparación no sería valida, estaría confundida por el efecto de la edad. El modelo planteado para modelar esta relación sería:

$y_1=a+\beta_1 x_1+\beta_2 x_2$

Donde y_1 indica los niveles de plomo en sangre
x_1 la edad y x_2 la ocurrencia de menopausia

Relación hipotética entre plomo en sangre, edad y menopausia



Para este segundo caso, hipotetico sobre los niveles de colesterol en sangre, se observa que en las mujeres post menopausicas el incremento relacionado con la edad es considerablemente mayor, por lo que la diferencias entre mujeres pre y post menopausicas va a depender de la edad. A mayor edad mayor será la diferencia entre mujeres pre y post menopausicas. En este caso, tendremos que tomar en cuenta esta relación ya que es biologicamente importante, lo que implica incluir un termino en el modelo que tome en cuenta esta relación.

En este caso el modelo que se poeude plantear es el siguiente:

$y_1=a+ß_1x_1+ß_2x_2+ß_3x_1x_2$

Donde y_1 indica los niveles de plomo en sangre
x_1 la edad, x_2 menopausia y x_1x_2 el efecto combinado de la menopausia

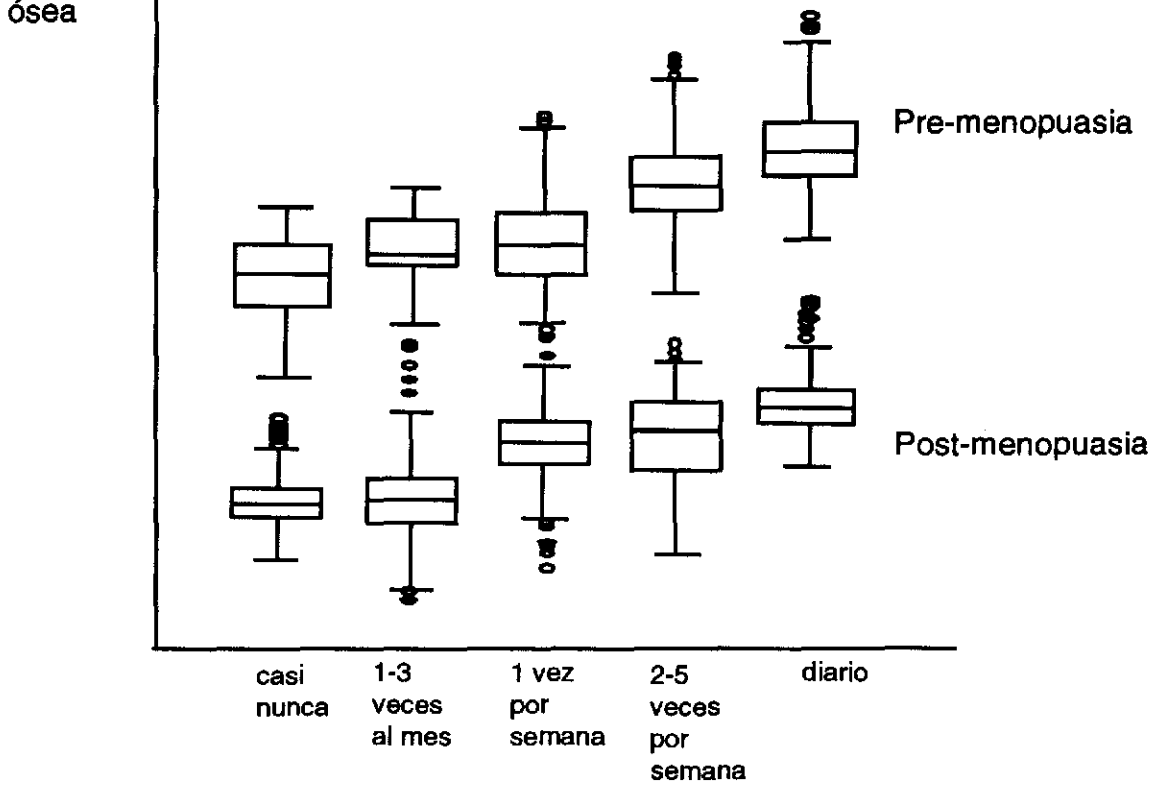
Es importante notar que en este modelo, la diferencia o el efecto atribuido a la menopausia esta estimada por dos coeficientes, uno que representa el efecto de la menopausia y otro que cuantifica como varía este efecto en relación a la edad. En este contexto, la interpretación del coeficiente ligado a la menopausia tendría interpretación solo cuando la edad es igual a 0. de otra manera siempre tendría un componente de cambio ligado a la edad. Por ejemplo:

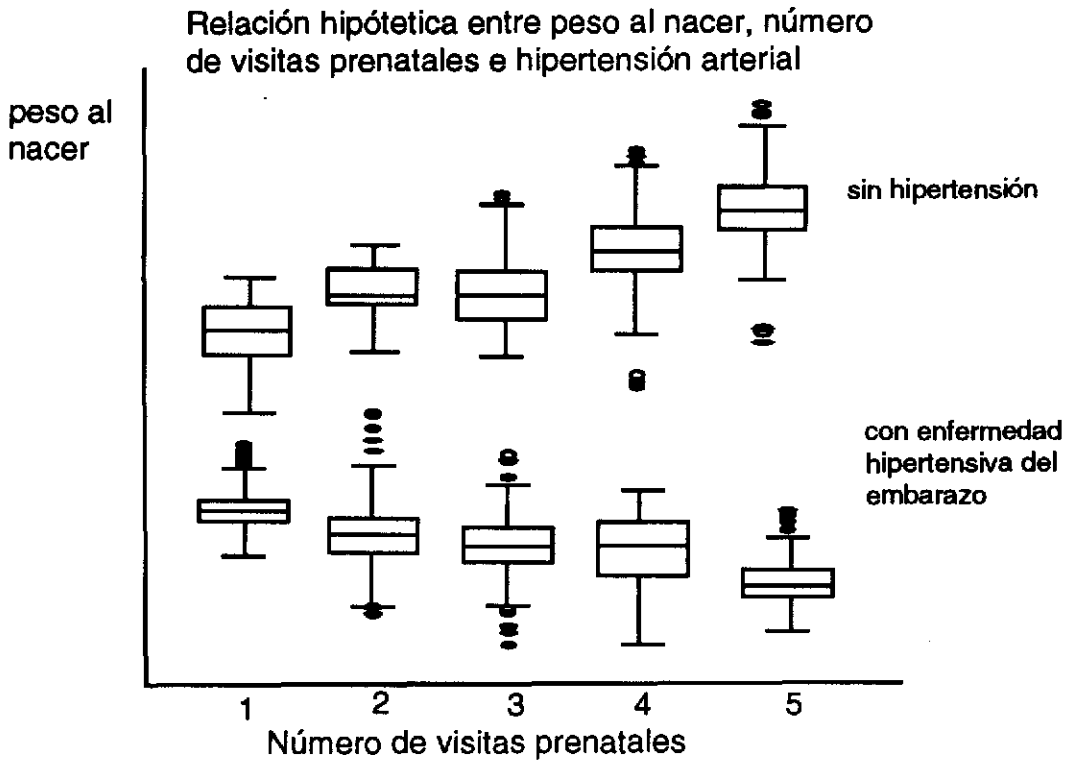
$y_1=a+ß_1x_1+ß_2x_2+ß_3x_1x_2$

Para un sijeto de 30 años el nivel de plomo en sangre
estar a estimado por:
$a+30ß_1$
Para una mujer menupausica de 30 años a os el nivel de plomo
estar a estimado por:
$a+30ß_1+ß_2+30ß_3$

Otros dos ejemplos hipotéticos sobre este tipo de relacionas:

Relación hipotética entre densidad ósea, niveles de ejercicio y menopausia




Relación hipótetica entre peso al nacer, número de visitas prenatales e hipertensión arterial
peso al nacer
sin hipertensión
con enfermedad hipertensiva del embarazo
1
2
3
4
5
Número de visitas prenatales

REFERENCIAS

1. STATA User's Guide. Stata Corp. Stata Statistical Software: Realease 5.0. College Station TX: Stata Corportation.

2. STATA Graphics Volume 1 al 3. Reference A-Z . Stata Corp. Stata Statistical Software: Realease 5.0. College Station TX: Stata Corportation.

Lecturas Complementarias

1. Anders Ahlbom. Descriptive Epidemiologic Measures. Bioestatistics for Epidemiologists pág. 55-60

2. Anders Ahlbom. Probability Theory. Bioestatistics for Epidemiologists pág. 3-32

3. Anders Ahlbom. The P Value, The P-Value, The P-Value Function and The confidence Interval Bioestatistics for Epidemiologists pág 35-53

4. David C. Hoaglin, Frederick Mosteller, John W. Tukey. Boxplots and Batch Comparison. Understanding Robust and Exploratory Data Analysis pág. 58-96

5. David C. Hoaglin, Frederick Mosteller, John W. Tukey. Letter Values: A Set of Sclected Order Statistics. Understanding Robust and Exploratory Data Analysis pag. 33-57

6. David C. Hoaglin, Frederick Mosteller, John W.Tukey. Stem-and-Leaf Displays. Understanding Robust and Exploratory Data Analysis pág. 7-32

7. David G. Kleinbaum, Lawrence L. Kupper, Keith E. Mueller. Classification of Variables and the Choice of Analysis. Applied Regression analysis and other Multivariable Methods pág. 7-15

8. Dawson-Saunders B, Trapp RG. Exploring & Presenting Data A Lang Medical Book, 1990: pp 20-42

9. Elwood Mark Causal relationship in medicine. A practice System for critical appraisal. USA 1988 pp 84-182.

10. Emerson JD, Stoto AM. Transforming Data Hoaglin DC. En: Understanding Robust and Exploratory Data Analysis. Mosteller F, Tukey JW, 1983: pp 97-128

11. Emerson JD, Strenio J. Boxplot and batch coparison. En: Undestanding Robust and Expoloraty Data Analysis. Hoaglinn DC, Mosteller F, Tukey JW. 1983 pp 58-96

12. Harold K, Cristopher T. Adjustment of Data Without Use of Multivariate Models Statistical Methods in Epidemiology, pp 85-135

13. Kenneth J. Rothmam. Analysis of Crude Data. Modern Epidemiology pág. 153-176

14. Kenneth J. Rothman. Analysis of crude Data. Modern Epidemiology 177-236

15. Lawrence C. Hamilton Variable Distributions. Regression with Graphics pág. 1-28

16. Lawrence C. Hamilton. Bivariate Regression Analysis. Regression with Graphics pág 29-64

17. Lawrence C. Hamilton. Basics of Multiple Regression. Regression with Graphics pág. 65-107

18. Lawrence C. Hamilton. Regression. Regression Criticism. Regression with Graphics pág.109-144

19. Lawrence C. Harnilton Robust Regression.. Regression with Graphics pág. 183-216

20. Miettinen Olli. Theoretical Epidemiologic. Regression analysis. USA. Wiley Medical. 1985. pp 216-244

21. Walker AM. Study types Observation and Inference, pp 27-44