# Práctica 3
## Modelo de regresión lineal simple[1]

## 1. Introducción

En esta práctica veremos cómo ajustar con SPSS un modelo de regresión lineal simple. Utilizaremos los datos del fichero `fracaso.sav`. Este fichero contiene, para 23 poblaciones de la Comunidad de Madrid, los datos correspondientes a renta per cápita y porcentaje de fracaso escolar. El objetivo del estudio es explicar el fracaso escolar en función de la renta.

## 2. Descripción de los datos

En primer lugar podemos representar un diagrama de dispersión de la variable respuesta y la variable regresora para tener una idea general de la relación entre ambas. En `SPSS` podemos ir al menú:

`Gráficos` ↪ `Cuadro de diálogo antiguos` ↪ `Dispersión/Puntos...`

Elegimos el tipo `Dispersión simple` y pulsamos `Definir`. En el cuadro de diálogo, como variable $Y$ elegimos la que contiene el porcentaje de fracaso escolar y como variable $X$ la que contiene las rentas. En `Etiquetar los casos mediante` situamos la variable que contiene los nombre de las poblaciones. Al pulsar `Aceptar` obtenemos el gráfico.

Haciendo doble click en el gráfico podemos cambiar su aspecto o añadir nuevos elementos. Las siguientes opciones son útiles:

- En `Elementos`, la opción `Modo de etiquetas de datos` nos permite situarnos sobre un punto y etiquetarlo con el nombre de la población al que corresponde.
- En `Elementos`, la opción `Línea de ajuste total` permite añadir al gráfico la recta de mínimos cuadrados. Podemos también añadir bandas de confianza para la media o para valores individuales (bandas de predicción) del nivel deseado.

### Cuestiones

- Añade una etiqueta a los dos puntos que muestren los porcentajes de fracaso escolar más altos y a los dos que muestren los porcentajes de fracaso escolar más bajos.
- Añade bandas de confianza de nivel 0.95 para el fracaso escolar medio. ¿Para qué valores de la renta se estima este porcentaje con mayor precisión?
- Añade bandas de predicción de nivel 0.95 para valores individuales. ¿Qué diferencia hay con respecto a las obtenidas en el apartado anterior?

[1] Para escribir estas notas se ha utilizado la versión SPSS 19

## 3. Ajuste del modelo de regresión lineal simple

Para ajustar el modelo de regresión lineal simple,

`Analizar` ↪ `Regresión` ↪ `Lineales...`

En el correspondiente cuadro de diálogo debemos elegir las opciones siguientes:

- En `Dependientes` elegimos la variable respuesta (*Fracaso*). En `Independientes` la variable regresora (*Renta*). En `Etiquetas de caso` la variable *Ciudad*.
- En `Gráficos` marcamos `Histograma` y `Gráfico de prob. normal`. Pasamos a `Y` la variable `*ZRESID` (residuos tipificados) y a `X` la variable `*ZPRED` valores pronosticados tipificados.
- En `Guardar` elegimos residuos tipificados.

Un vez elegidas las opciones anteriores pulsamos `Aceptar` para obtener el resultado.

Con las opciones anteriores se obtienen los gráficos más habituales para llevar a cabo el diagnóstico del modelo. De forma complementaria es posible aplicar también algún contraste de bondad de ajuste. Uno muy utilizado para saber si es razonable la hipótesis de normalidad es el de Kolmogorov-Smirnov. Para aplicarlo, vamos al menú:

`Analizar` ↪ `Pruebas no paramétricas` ↪ `K-S de una muestra...`

Se elige la variable que contiene los residuos y la distribución normal, y se pulsa `Aceptar`. El p-valor resultante sirve para constrastar la hipótesis nula de que los datos tienen distribución normal.

### Cuestiones

- A partir de los resultados anteriores, ¿qué se puede decir sobre las hipótesis habituales del modelo de regresión lineal simple?
- ¿Cuánto vale el p-valor del estadístico de Kolmogorov-Smirnov? Cambia la hipótesis de contraste (por ejemplo, selecciona la distribución uniforme en lugar de la normal) y comprueba cómo cambia el p-valor.

## 4. Transformación de variables

A veces, una relación no lineal se puede convertir en lineal transformando apropiadamente las variables. Para crear una nueva variable como resultado de aplicar una transformación a alguna de las ya existentes, se procede de la forma siguiente:

- `Transformar` ↪ `Calcular variable...`.
- En `variable de destino` se escribe el nombre de la nueva variable que vamos a crear. En los cuadros `Grupo de Funciones` y `Funciones y variables especiales`, se elige la transformación deseada. Por ejemplo, si queremos calcular el logaritmo neperiano de una variable, seleccionamos `Aritméticas` y `LN` respectivamente.
- Aparece en el recuadro de arriba un signo de interrogación que tenemos que sustituir por la variable que queremos transformar.

Cuestiones

- Calcula los logaritmos neperianos de las dos variables del fichero y ajusta los modelos siguientes:

$$\begin{aligned} \ln(\texttt{Fracaso}) &= \beta_0 + \beta_1 \ln(\texttt{Renta}) + u \\ \texttt{Fracaso} &= \beta_0 + \beta_1 \ln(\texttt{Renta}) + u \\ \ln(\texttt{Fracaso}) &= \beta_0 + \beta_1 \texttt{Renta} + u \end{aligned}$$

  ¿Cuál de los modelos parece ajustarse mejor a los datos? Interpreta el parámetro $\hat{\beta}_1$ obtenido en cada caso.

## 5. Modelo potencial, modelo logarítmico y modelo exponencial

Un resumen de los principales resultados y gráficos correspondientes a los modelos anteriores se puede obtener en:

`Analizar` ↪ `Regresión` ↪ `Estimación curvilínea...`

En el correspondiente cuadro de diálogo debemos elegir las opciones siguientes:

- En `Dependientes` elegimos la variable respuesta (*Fracaso*). En `Independientes` la variable regresora (*Renta*). En `Etiquetas de caso` la variable *Ciudad*.
- Marcamos los modelos lineal, logarítmico, potencia y exponencial.

Cuestiones

- ¿Para qué modelo es menor el coeficiente de determinación?
- ¿Cuánto vale el coeficiente de correlación entre el logaritmo del porcentaje de fracaso y la renta per cápita?
- ¿Corresponden los parámetros estimados con los que has calculado en la sección anterior?

## 6. Problema 8 del tema 3

El siguiente conjunto de datos corresponde a una muestra de las aguas de 20 lagos en Estados Unidos. Para cada lago se calculan el número de factorías por kilómetro que hay situadas en la orilla ($X$) y el porcentaje de impurezas en el agua ($Y$):

| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
|---|---|---|---|---|---|---|---|
| 5,00 | 0,005 | 8,20 | 0,910 | 1,14 | 0,003 | 1,47 | 0,009 |
| 8,90 | 0,430 | 1,01 | 0,008 | 1,18 | 0,010 | 3,00 | 0,150 |
| 7,15 | 0,009 | 0,61 | 0,030 | 0,73 | 0,001 | 2,40 | 0,250 |
| 6,10 | 0,240 | 0,72 | 0,010 | 0,40 | 0,950 | 4,10 | 0,200 |
| 7,70 | 0,007 | 0,68 | 0,005 | 0,87 | 0,200 | 4,00 | 0,010 |

(a) Plantear un modelo de regresión para explicar el porcentaje de impurezas en el agua en función del número de factorías. Hallar la recta de regresión, y decidir si el número de factorías influye sobre el porcentaje de impurezas (al nivel 0,05).
(b) Repetir el estudio eliminando el dato atípico $x = 0,40$, $y = 0,950$. Comentar las diferencias que se observan.

## 7. Problema 2 del tema 3

El muestreo de áreas contiguas se utiliza en Ecología para contar el número de especies distintas de plantas por área. El recuento se realiza de manera que cada siguiente área contigua tiene el doble de superficie, empezando por un área de 1 metro cuadrado. El modelo que relaciona $Y$, el número de especies distintas, con $X$, la superficie en metros cuadrados, es $Y = a \ln X + b$, donde $a$ es el *índice de diversidad* y $b$ es el *número de especies por unidad de área*. Ajustar dicho modelo a los datos:

| Superficie | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|---|
| Especies distintas | 2 | 4 | 7 | 11 | 16 | 19 | 21 |