

Temas de Estadística Práctica

Antonio Roldán Martínez

Proyecto <http://www.hojamat.es/>

Tema 5

Distribuciones bidimensionales. Regresión.

[Recta de regresión](#)

[Predicciones](#)

[Varianzas en la regresión](#)

[Regresión no lineal](#)

Recta de regresión

El problema de la Regresión Lineal es el más importante, junto con la Correlación, que podemos considerar en las distribuciones bidimensionales. Nos ceñiremos a los casos de tablas simples que contengan pares de valores de X e Y , sin consideración de frecuencias. Supondremos también que los datos son de tipo cuantitativo.

Quizás debas repasar los conceptos del Tema 4, en el que se explican la variables bidimensionales.

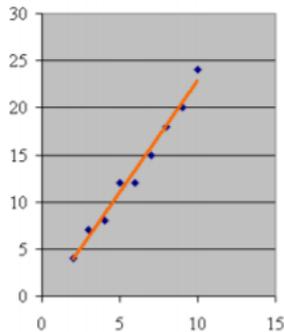
Llamaremos **Recta de Regresión de Y sobre X** a aquella que *mejor se adapta* al diagrama de dispersión XY , también llamado *Nube de puntos*. Este acercamiento se define de forma rigurosa como

La recta de regresión de Y sobre X es aquella que minimiza la suma de cuadrados de las diferencias entre los valores de Y y los correspondientes Y' medidos en dicha recta.

Así, la tabla

X	2	3	4	5	6	7	8	9	10
Y	4	7	8	12	12	15	18	20	24

da lugar a una nube de puntos en cuya gráfica hemos añadido la recta de regresión:



Efectivamente, esta recta sigue *lo mejor posible* la tendencia de los puntos. Matemáticamente, las diferencias al cuadrado de los valores verdaderos de Y y los incluidos en la recta, suman lo mínimo posible.

A la variable X se le suele llamar *predictora*, y a la Y, *criterio*.

La recta de regresión es un instrumento para efectuar predicciones, ya sea en el rango de datos como fuera de él. Llamaremos ***pronóstico o predicción*** para un valor de X a su imagen Y' en la recta de regresión.

La recta de regresión tiene una validez limitada. No debemos efectuar predicciones en valores de X muy alejados del rango considerado. Además, no todas las relaciones son de tipo ***lineal***.

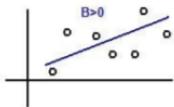
El origen de la palabra *regresión* es histórico. Cuando Galton estudió estas cuestiones, descubrió que los hijos de padres muy altos o muy bajos no lo eran tanto como sus padres, *regresaban* a valores medios. Después se vio que este fenómeno no siempre se daba.

Recordemos que la ecuación de una línea recta en dos dimensiones tiene la forma:

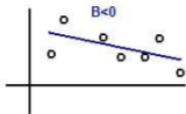
$$Y' = A + BX$$

donde el coeficiente B representa la tasa de cambio o **pendiente** y el coeficiente A es el valor correspondiente a $X=0$, y la llamaremos **ordenada en el origen**.

Según el signo de la pendiente, hablaremos de relación **positiva o creciente**



y de relación **negativa o decreciente**.



Mediante las técnicas de búsqueda de mínimos podemos demostrar que la recta que minimiza los cuadrados de los errores es la que viene dada por estas fórmulas:

$$B = \frac{S_{xy}}{S_x^2}$$

es decir, la **covarianza** dividida entre la **varianza de X**

$$A = \bar{Y} - B\bar{X}$$

que podemos expresar como la diferencia entre la media de Y y la de X multiplicada por B

Existen desarrollos simplificados de estas fórmulas para facilitar su cálculo, que no consideraremos aquí.

También se puede considerar la recta de X sobre Y, pero no lo haremos aquí.

Sí puede ser interesante estudiar la recta de regresión para puntuaciones típicas, porque en ese caso su fórmula es muy sencilla: $Z_y = R_{xy} \cdot Z_x$, donde R es el coeficiente de correlación.

Predicciones

Llamaremos **predicción o pronóstico** para un valor de X al dado por la expresión $Y' = A + BX$.

En los gráficos de dispersión XY las predicciones pertenecerán a la línea recta, mientras los valores reales Y figurarán más arriba o abajo de ella.

Llamaremos **error de predicción** a la diferencia $Y - Y'$

El promedio de las predicciones Y' coincide con el de los valores reales Y.

El promedio de los errores de predicción cometidos es cero.

Varianzas a considerar en la regresión

En el problema de la regresión es conveniente considerar distintas sumas de cuadrados para calcular varianzas también distintas:

Varianza total de Y

Si no consideramos la recta de regresión y deseamos medir la variabilidad del conjunto de datos que estamos usando, acudiremos a la varianza de Y, o **varianza total**.

$$S_y^2 = \frac{\sum(Y - \bar{Y})^2}{N}$$

que es la varianza en el sentido general.

Si sólo consideramos la variabilidad que presentan las predicciones (los valores situados en la recta), deberemos usar en la fórmula anterior los datos Y' en lugar de Y (la media no cambia, según se indicó más arriba). Al resultado le llamaremos **varianza explicada**.

$$S_{exp}^2 = \frac{\sum(Y' - \bar{Y})^2}{N}$$

Un resultado fundamental es el siguiente:

$$S_{exp}^2 = S_y^2 \cdot r^2$$

siendo r el **coeficiente de correlación de Pearson**. Esto significa que la relación entre la varianza explicada y la total es el cuadrado del coeficiente r . A este cuadrado lo conocemos como **Coeficiente de determinación** y expresa el porcentaje de varianza que explica la línea recta.

Por último, llamaremos **varianza de error o residual** a la que presentan los valores de Y comparados con sus pronósticos:

$$S_e^2 = \frac{\sum(Y - Y')^2}{N}$$

Se puede demostrar la relación:

$$S_e^2 = S_y^2 - S_{exp}^2 = S_y^2(1 - r^2)$$

A la raíz cuadrada de la varianza residual la llamaremos **error típico de estimación**, que es importante en la teoría de la Regresión.

Regresión no lineal

Cuando no está clara o bien fundamentada una tendencia lineal, tendremos que buscar otras formas de gráficas que se ajusten mejor a la distribución bidimensional que estemos estudiando. Por su facilidad de manejo y cálculo, se suelen estudiar las siguientes:

Función exponencial: Se usa para crecimientos y decrecimientos en los que la tasa es proporcional al valor actual (de forma aproximada). Cuanto mayor es el valor actual, mayor es el incremento que sufre. Según ese incremento sea positivo o negativo, la gráfica puede presentar una de estas variantes:

Exponencial creciente



Exponencial decreciente



La gráfica de color rojo corresponde a los datos reales y la de color verde al ajuste exponencial

Su expresión es $y = ae^{bx}$, en la que a y b son dos parámetros a determinar, y e es el número trascendente 2,71828...

Su ajuste se logra transformando la anterior expresión mediante logaritmos neperianos, con lo que queda de la forma $\text{Ln}y = bx + \text{Ln}a$, que al ser lineal, admite los cálculos de regresión explicados hasta ahora. No desarrollaremos esta técnica. Tan sólo hay que recordar que se sustituye Y por su logaritmo.

Función logarítmica: Si se da la proporcionalidad anterior entre el valor actual y la tasa, pero de forma inversa, es decir, que la tasa de variación sea proporcional al valor inverso del actual ($1/X$), el mejor ajuste es el logarítmico.

Su gráfica suele presentar este aspecto

Se observa un crecimiento de los valores, pero un decrecimiento progresivo de la tasa de variación:



La gráfica de color rojo corresponde a los datos reales y la de color verde al ajuste exponencial

Su expresión es $y = a + b \cdot \log(x)$

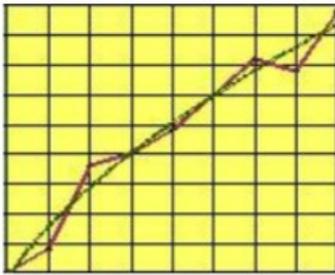
Su ajuste se realiza sustituyendo los valores de x por sus logaritmos.

Función potencial: Es la más potente, pues permite encontrar un exponente fraccionario, lo que abarca las potencias y raíces de todo tipo de exponentes. Su expresión es $y = a \cdot x^b$

Presenta múltiples formas de gráfica, pues depende del valor de b .

Para ajustar con este procedimiento deberemos tomar logaritmos tanto en la X como en la Y .

La gráfica siguiente corresponde a un ajuste a una raíz cuadrada.



Función polinómica: Suelen ajustarse bien a los datos, pero sus fórmulas pueden complicarse, ya que presentan forma de polinomios, que, a partir del tercer grado son muy complicados. Se usa a menudo el ajuste a un polinomio de segundo grado, o ajuste cuadrático, especialmente en ámbitos científicos.

Para elegir el mejor ajuste a los datos disponemos del coeficiente **R^2 de Determinación**, que es el cuadrado del coeficiente de correlación de Pearson, y nos informa del porcentaje de varianza que está explicado por la función que ajusta los datos. Así, basta elegir la modalidad que tenga el coeficiente mayor, o bien, que su fórmula sea sencilla y el coeficiente apreciable.

El significado más sencillo del valor de **R^2** es el de que representa el cociente entre la varianza explicada por el modelo y la varianza total (ver atrás para el caso lineal). Así, cuanto mayor sea su valor, más varianza explica el modelo y menor es la varianza de error.