Tarea 7
Transformadas

Erick Cervantes Mendieta
Matrícula: 2032430

Modelos Probabilistas Aplicados

Octubre 2020

1. Regresión lineal simple

En esta sección se trata de describir la relación entre dos variables por medio del cálculo de la gráfica y la ecuación de la recta que representa dicha relación. Esta recta se conoce como **recta de regresión** y su ecuación como **ecuación de regresión**. La ecuación de regresión expresa una relación entre x (variable independiente)y y (variable dependiente).

La ecuación típica de una línea recta puede estar expresada en la forma $y = b_0 + b_1x$, donde b_0 es el *intercepto* y, y b_1 es la *pendiente*. Así, dado un conjunto de datos muestrales en pares, decimos que la ecuación de regresión $\hat{y} = b_0 + b_1x$, describe algebraicamente la relación entre dos variables. La gráfica de la ecuación de regresión se denomina recta de regresión (recta del mejor ajuste) [4].

Se define el coeficiente de correlación r como una medida de qué tan bien se ajusta la recta de regresión a los datos, ya que este nos indica qué tan fuerte o débil es una relación lineal. Sin embargo, cuando lo que interesa es analizar una relación de causalidad entre dos variables, primero debemos definir cuál de ellas es la variable dependiente, y cuál la independiente. La variable dependiente y es la que se busca explicar; en términos estadísticos, es la que se busca estimar o pronosticar. A su vez, la variable independiente x es la que brinda información para explicar y y recibe el nombre de variable de predicción [1].

Para saber si una variable x es *buena* para explicar la variable y se calcula el **coeficiente de determinación**, cuya representación se denota con r^2, dicho coeficiente tiene las características siguientes: es el cuadrado del coeficiente de correlación, su rango de valores varía de *cero* a *uno*, no da ninguna información sobre la dirección de la relación entre las variables, cuanto más cerca esté de uno, la variable independiente x será una buena variable para explicar y, por otra parte, conforme r^2 se acerca a cero, indica que x no es un factor significativo para explicar y.

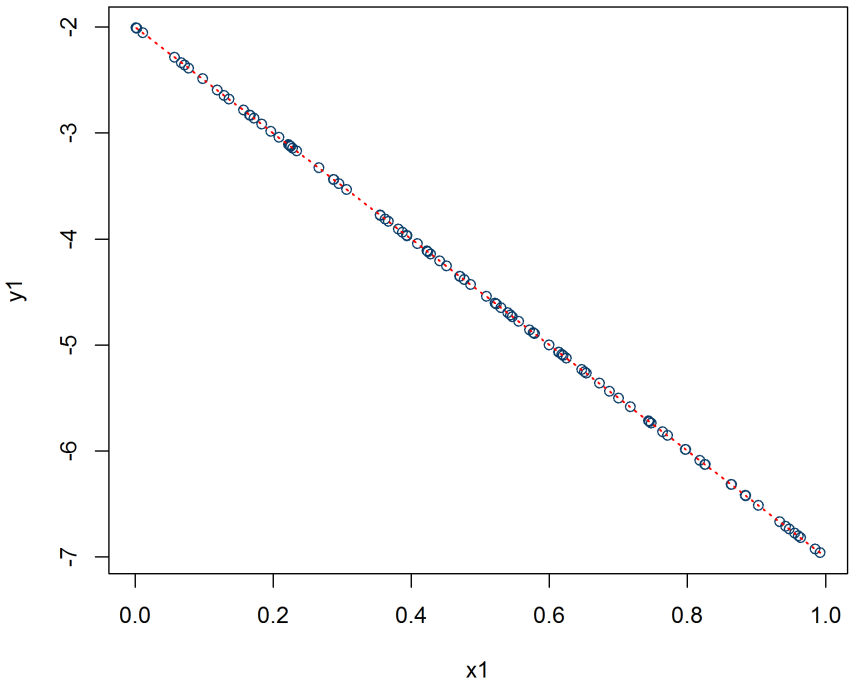


Figura 1: Gráfica de dispersión para datos1.

Para aterrizar bien estos conceptos, procedemos a analizar una estructura de datos llamada `datos1` generada en el programa R [2] (Versión 4.0.2), la gráfica de dispersión se muestra en la figura 1, en donde, es muy evidente apreciar que existe una relación lineal perfecta, por lo que se procede a obtener su ecuación de regresión con la ayuda del lenguaje R. El valor para los coeficientes son los siguientes: $b_0 = -2$ y $b_1 = -5$, con un $r^2 = 1$, es decir, la variable x explica muy bien el comportamiento de y, por lo que la ecuación de regresión para ese conjunto de datos es:

$$\hat{y} = -5x - 2.$$

Por otra parte, otra estructura de datos llamada `datos2` fue analizada, la gráfica de dispersión se muestra en la figura 2, en dónde es evidente que no existe una relación lineal entre las variables x_2 y y_2, por lo que procedemos a hacer alguna transformación que nos ayude a encontrar esta relación lineal. En R viene implementado la función `assumptions{trafo}`, la cual ofrece una primera descripción general de si una transformación es útil y qué transformación promete cumplir con los supuestos del modelo de normalidad, homocedasticidad y linealidad. Los resultados se muestran en el cuadro 1, en donde, se observa que varias transformaciones pueden ser utilizadas para poder encontrar una ecuación de regresión lineal para este caso, en particular, se analizó la transformación de Box–Cox (boxcox), una logarítmica (log) y una transformación utilizando el recíproco de y (reciprocal), ya que estas últimas

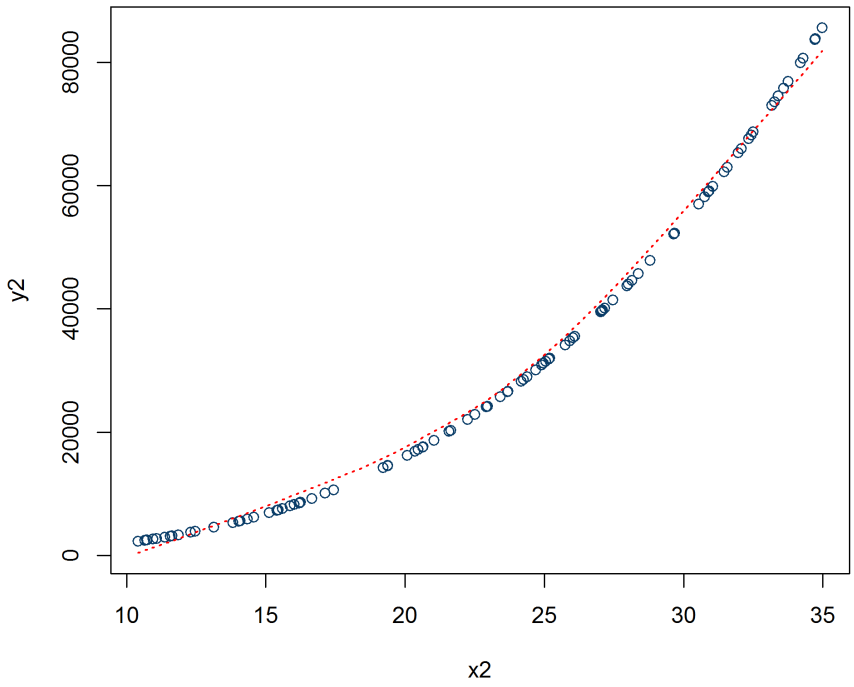


Figura 2: Gráfica de dispersión para datos2.

forman parte de la escalera de Tukey [3].

Transformación función recíproca

Para este caso se utilizó la transformación $\hat{y} = 1/y$, de está forma se pudo obtener la siguiente ecuación de regresión para ese conjunto de datos, con un $r^2 = 0.6994$:

$$\hat{y} = 1.283 \times 10^{-05} x + 9.996 \times 10^{-01}.$$

Transformación logarítmica

Para este caso se utilizó la transformación $\hat{y} = \ln(y)$, de está forma se pudo obtener la siguiente ecuación de regresión para ese conjunto de datos, con un $r^2 = 0.9752$:

$$\hat{y} = 6.61209x + 0.14389.$$

Transformación Box–Cox

Finalmente se utilizó la transformación Box–Cox y con un $r^2 = 1$ se obtuvo la siguiente ecuación de regresión:

Cuadro 1: Valores del coeficiente de correlación para las diferentes transformaciones de la estructura datos2.

Transformada	r
bickeldoksum	1.00
boxcox	1.00
dual	1.00
glog	0.99
gpower	1.00
log	0.99
logshiftopt	1.00
manly	–
modulus	1.00
neglog	0.99
sqrtshift	1.00
geojohnson	1.00
reciprocal	0.84

$$\hat{y} = 3.778x - 2.989.$$

El diagrama de dispersión para el conjunto de datos, luego de haber aplicado las tres transformaciones se muestra en la figura 3, en dicha figura se puede observar que el mejor ajuste lo ofrece la transformación Box–Cox, debido a que presenta un índice de correlación más alto que las otras dos transformaciones, que en este caso es igual a uno, es decir, x explica muy bien el comportamiento de y en dicha transformación. Nótese que la función que origina los datos es $y = 2x^3 + rnorm$.

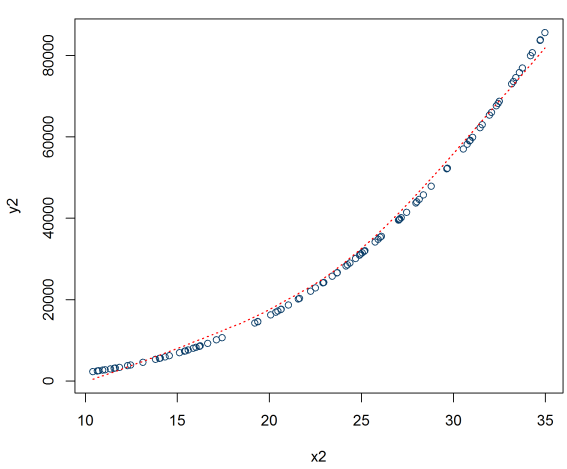
2. Regresión lineal múltiple

En la sección anterior se analizaron métodos de regresión lineal para investigar relaciones entre exactamente dos variables, pero en algunas ocasiones se requieren algunas más. Así, una **ecuación de regresión múltiple** expresa una relación lineal entre una variable dependiente y y dos o más variables independientes $(x_1, x_2, \ldots, x_k)$. La forma de esta ecuación es:

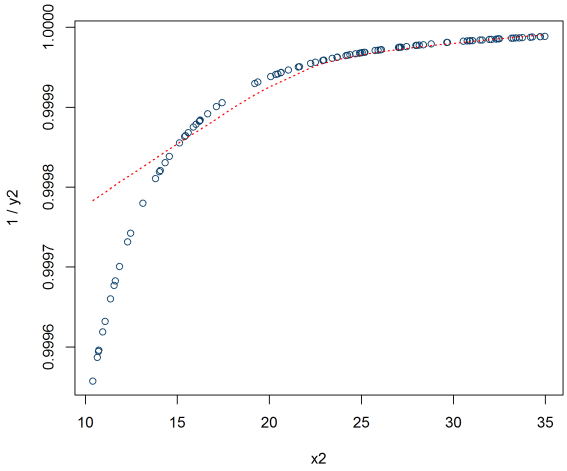
$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k.$$

En este caso se define un nuevo valor R^2 o **coeficiente múltiple de determinación**, el cual nos permite medir qué tan bien se ajusta la ecuación de regresión múltiple a los datos que se tienen. Un ajuste perfecto se tendría cuando $R^2 = 1$, y un ajuste muy bueno da como resultado un valor cercano a uno, por otra parte, un ajuste muy pobre se relaciona con un valor de R^2 cercano a cero.

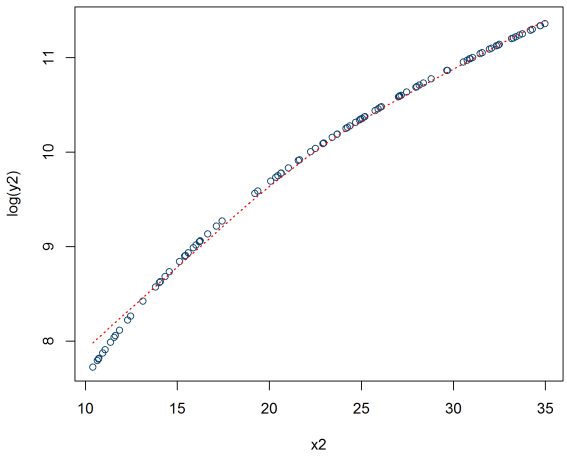
Se tiene una estructura de datos llamada `datos3`, la cual hace referencia a valores de y, x_2 y una x_3, la matriz de dispersión de correlaciones se muestra en la figura 4, en donde, se



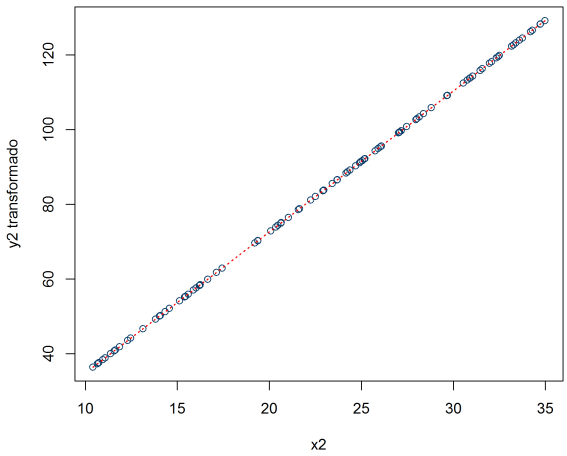
(a) Datos *no* transformados



(b) Transformación función recíproca



(c) Transformación logarítmica



(d) Transformación Box–Cox

Figura 3: Gráficas de dispersión para datos2.

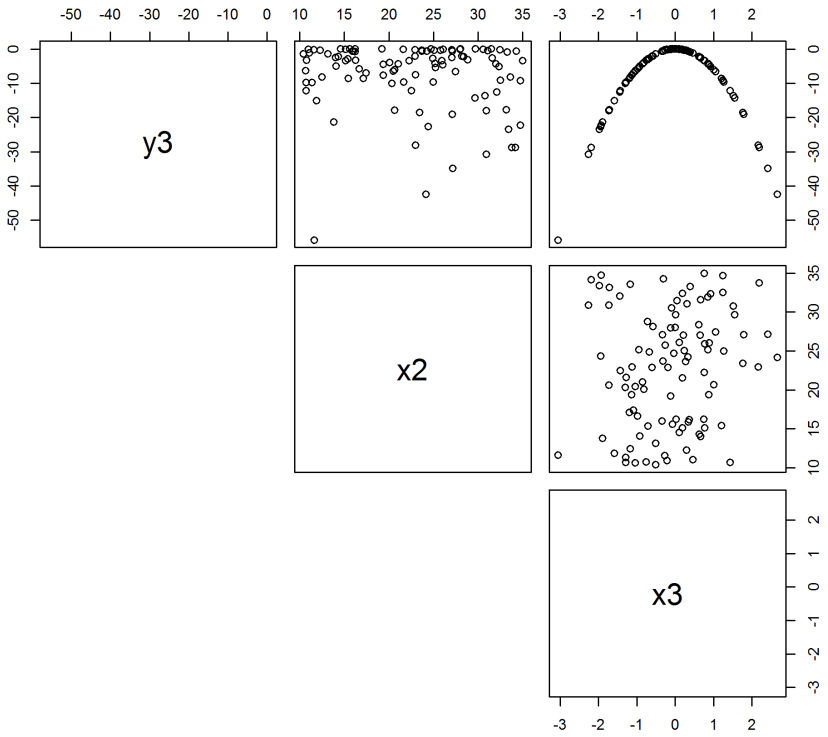


Figura 4: Matriz de dispersión de correlaciones para datos3.

puede apreciar que la variable y al parecer tiene una relación cuadrática con la variable x_3 y con respecto a la variable x_2 no se logra apreciar bien cuál sería dicha relación. Se procede a generar un modelo de regresión lineal para observar que tanto se ajustan los datos, el valor para los coeficientes son los siguientes: $b_0 = -2.9077$, $b_1 = -0.1973$ y $b_2 = 1.3579$, con un $R^2 = 0.03764$, es decir, el modelo lineal que se forma con esos valores explica el $3.764\,\%$ de la variabilidad observada en y , por lo que la ecuación de regresión presentada a continuación para ese conjunto de datos no es muy bueno:

$$\hat{y} = -2.9077 - 0.1973x_2 + 1.3579x_3.$$

Una idea de poder generar un buen modelo lineal, es aplicando transformaciones a las variables independientes, como en el caso de regresión lineal simple. La matriz de dipersión con correlaciones presentada en la figura 5 nos da otra información importante, y esta se ve reflejada en el índice de correlación que se presenta, ya que para y y x_2 el signo es negativo, mientras que con x_3, el signo es positivo.

A manera de ejemplo, se utilizó la transformación de las dos variables independientes como $\ln(x_i)$, de esta forma la ecuación de regresión quedaría como sigue:

$$\hat{y} = b_0 + b_1 \ln(x_2) + b_2 \ln(x_3),$$

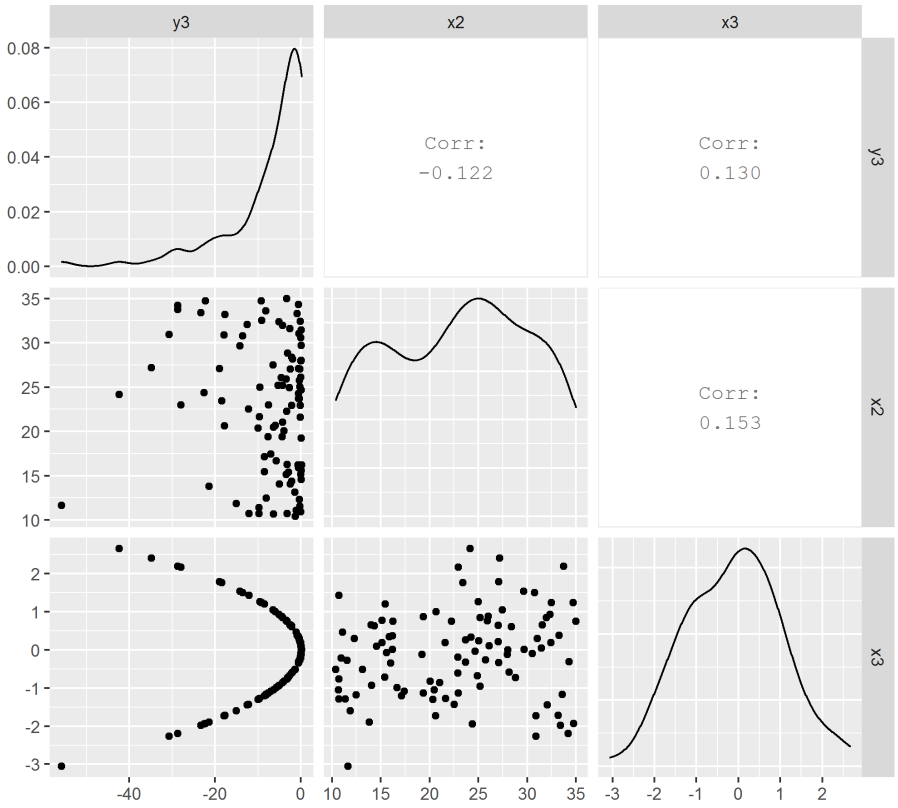


Figura 5: Matriz de dispersión *con* correlaciones para datos3.

de donde, se puede encontrar que el valor para los coeficientes son los siguientes: $b_0 = 0.2045$, $b_1 = -3.2310$ y $b_2 = -5.3535$, con un $R^2 = 0.4396$, es decir, este modelo explica el $43.96\,\%$ de la variabilidad observada en y , por lo que la ecuación de regresión para esta transformación propuesta es mejor que la que se tienen inicialmente.

Nótese que no es la única forma de poder transformar las variables independientes, un análisis más detallado requeriría analizar la mayoría de las transformaciones implementadas en R sobre cada una de las variables, tomando en cuenta la información que se puede leer visualmente en la matriz de dispersión con correlaciones y con esto poder tener un mejor modelo lineal que ajuste mejor a los datos con respecto a la variabilidad observada en la variable dependiente.

Referencias

[1] A. Gutiérrez. *Probabilidad y Estadística, Enfoque por competencias*. McGraw Hill, México, 2012.

[2] R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/, 2020.

[3] S. E. Schaeffer. Modelos probabilistas aplicados. https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html, 2020.

[4] M. F. Triola. *Estadística*. Pearson Education, México, 2004.