Introducción a la regresión lineal

Ramon Ceballos

6/2/2021

1. Rectas de regresión y transformaciones logarítmicas

No siempre encontraremos dependencias lineales. A veces nos encontraremos otro tipo de dependencias, como por ejemplo pontencias o exponenciales.

Estas se pueden transformar a lineales mediante un **cambio de escala**.

1.1 Escalas logarítmicas

Por lo general, es habitual encontrarnos gráficos con sus ejes en **escala lineal**. Es decir, las marcas en los ejes están igualmente espaciadas.

A veces, es conveniente dibujar alguno de los ejes en **escala logarítmica**, de modo que la misma distancia entre las marcas significa el mismo cociente entre sus valores. En otras palabras, un eje en escala logarítmica representa el logaritmo de sus valores en escala lineal.

Diremos que un gráfico está en **escala semilogarítmica** cuando su eje de abcisas (eje "x") está en escala lineal y, el de ordenadas (eje "y"), en escala logarítmica.

Diremos que un gráfico está en **escala doble logarítmica** cuando ambos ejes están en escala logarítmica.

1.2 Interpretación gráfica

1.2.1 Escala semilogarítmica

Si al representar unos puntos $(x_i, y_i)_{i=1,\ldots,n}$ en escala semilogarítmica observamos que siguen aproximadamente una recta, esto querrá decir que los valores $\log(y)$ siguen una ley aproximadamente lineal en los valores x, y, por lo tanto, que y sigue una **ley aproximadamente exponencial** en x.

En efecto, si $\log(y) = ax + b$, entonces, despejando el logaritmo de la ida de la igualda se obtiene:

$$y = 10^{\log(y)} = 10^{ax+b} = 10^{ax} \cdot 10^b = \alpha^x \beta$$

con $\alpha = 10^a$ y $\beta = 10^b$.

Por tanto, y sigue una tendencia exponencial con respecto a x para este tipo de escalas con regesión lineal.

1.2.2 Escala doble logarítmica

Si al representar unos puntos $(x_i, y_i)_{i=1,\ldots,n}$ en escala doble logarítmica observamos que siguen aproximadamente una recta, esto querrá decir que los valores $\log(y)$ siguen una ley aproximadamente lineal en los valores $\log(x)$, y, por lo tanto, que y sigue una **ley aproximadamente potencial** en x.

En efecto, si $\log(y) = a\log(x) + b$, entonces, por propiedades de logaritmos

$$y = 10^{\log(y)} = 10^{a\log(x)+b} = (10^{\log(x)})^a \cdot 10^b = x^a\beta$$

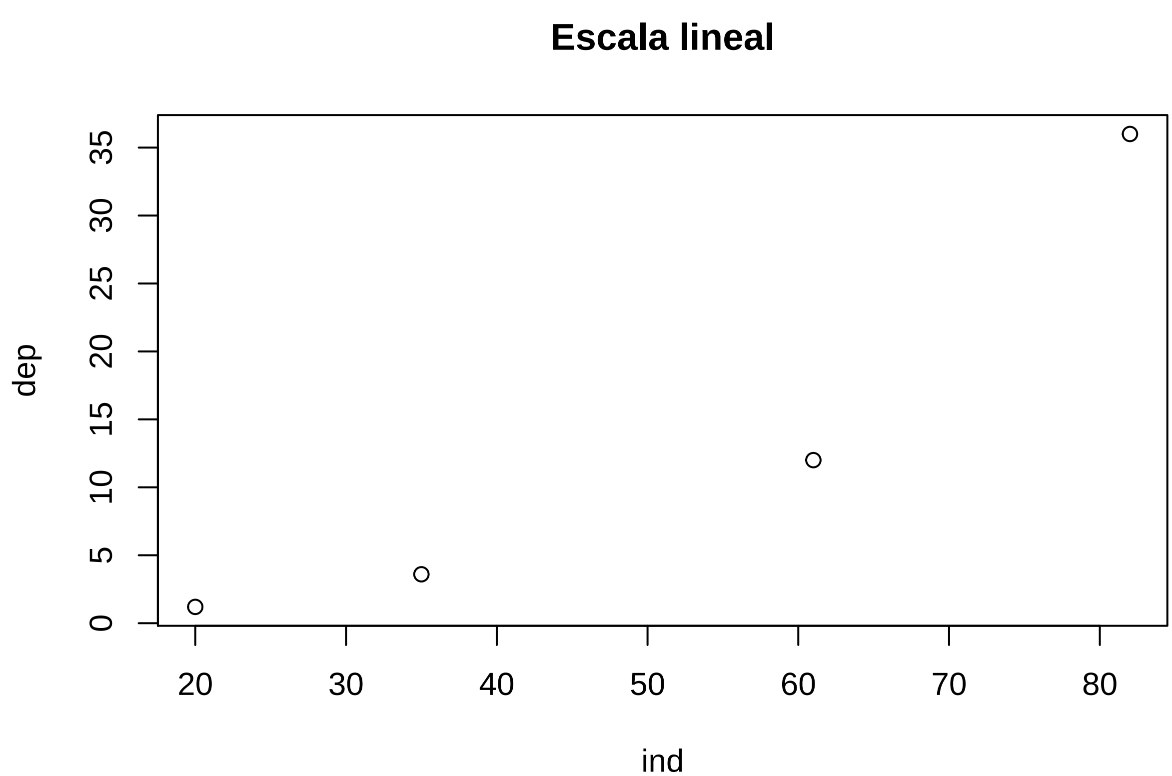
con $\beta = 10^b$

1.3 Ejemplo para relaciones exponenciales (escala semilogarítmica)

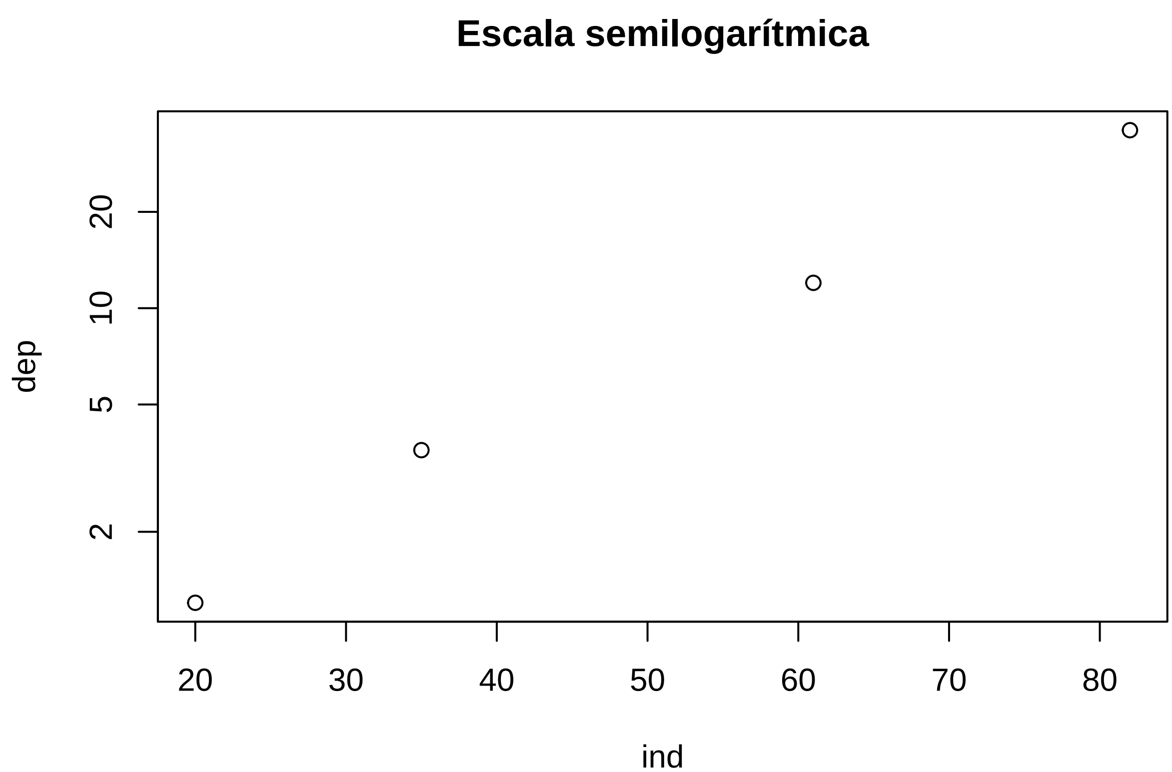
En este caso trabajaremos no con un data frame, sino directamente con los dos vectores siguientes:

```
#Datos de trabajo
dep = c(1.2,3.6,12,36) #variable dependiente
ind = c(20,35,61,82) #variable independiente
```

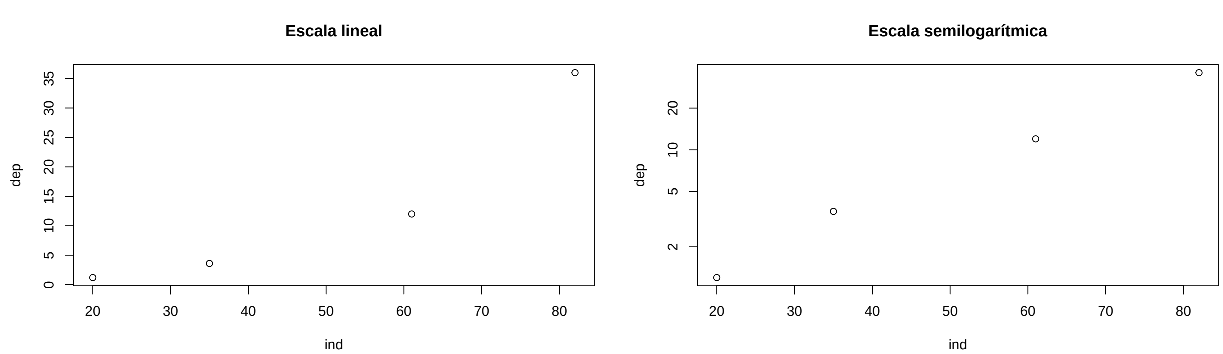
```
#plot sin cambio de escala
plot(ind,dep, main = "Escala lineal")
```



```r
#plot de escala logaritmica en eje "y"
plot(ind,dep, log = "y", main = "Escala semilogarítmica")
```




```r
#Gráficos representados en la misma figura
par(mfrow = c(1,2))
plot(ind,dep, main = "Escala lineal")
plot(ind,dep, log = "y", main = "Escala semilogarítmica")
```




```r
par(mfrow = c(1,1))
```

Hacemos el modelo lineal de transformar en logaritmo en base 10 a la variable dependiente, en función de la independiente (sin variar).

```r
#regresión lineal
lm(log10(dep)~ind)
```

```
##
## Call:
## lm(formula = log10(dep) ~ ind)
##
## Coefficients:
## (Intercept)          ind
##    -0.32951      0.02318
```

```r
#coeficiente de determinación
summary(lm(log10(dep)~ind))$r.squared
```

```
## [1] 0.9928168
```

Lo que acabamos de obtener es que $\log(dep) = 0.023 \cdot ind - 0.33$ es una buena aproximación de nuestros datos.

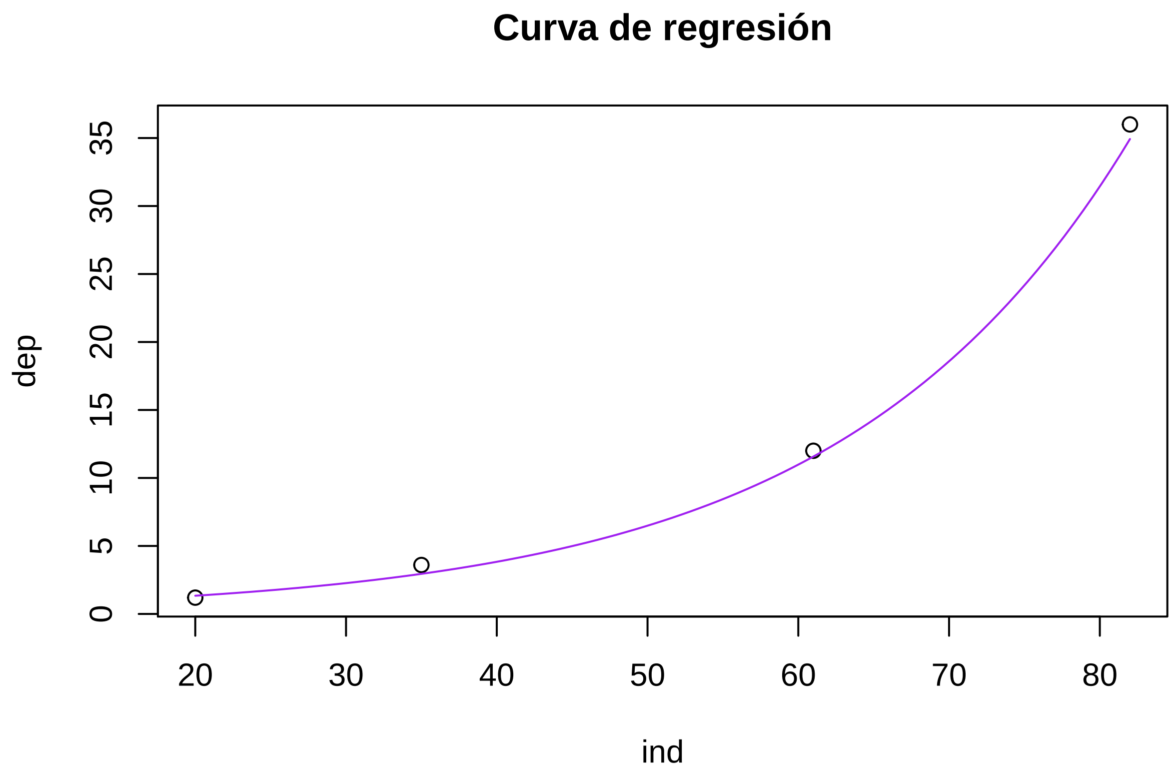
Con lo cual, la variable dependiente se obtiene con la siguiente fórmula:

$$dep = 10^{0.023 \cdot ind} \cdot 10^{-0.33} = 1.054^{ind} \cdot 0.468$$

Se puede discernir que la relación existente entre la varibale dependiente e independiente es una relación de tipo exponencial.

Para observar esto, vamos a representar la nube de ptos con plot() y luego añadimos la curva de la función dictada anteriormente.

```r
plot(ind,dep, main = "Curva de regresión")
curve(1.054^x*0.468, add = TRUE, col = "purple")
```



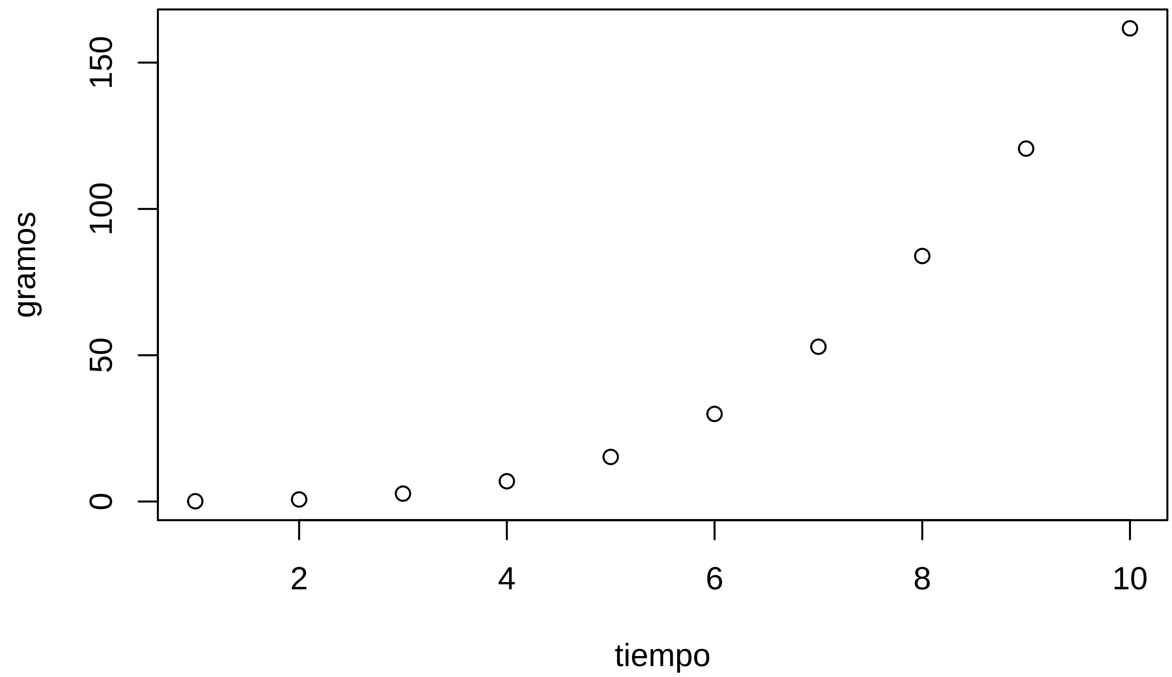
1.4 Ejemplo para relaciones potenciales (escala doble logarítmica)

En este caso trabajaremos con el siguiente data frame:

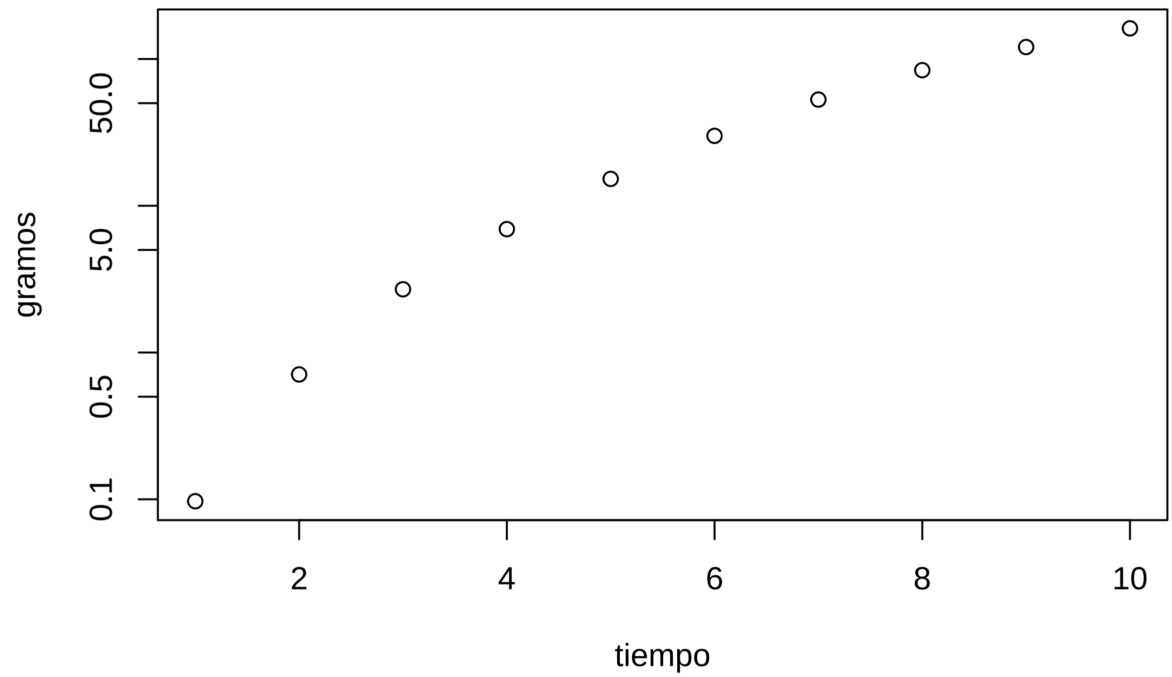
```
#Datos de estudios en forma DF
tiempo = 1:10 #variable independiente

#Variable dependiente (masa de virus o bacteria)
gramos = c(0.097,0.709,2.698,6.928,15.242,29.944,52.902,83.903,120.612,161.711)
d.f = data.frame(tiempo,gramos)
```

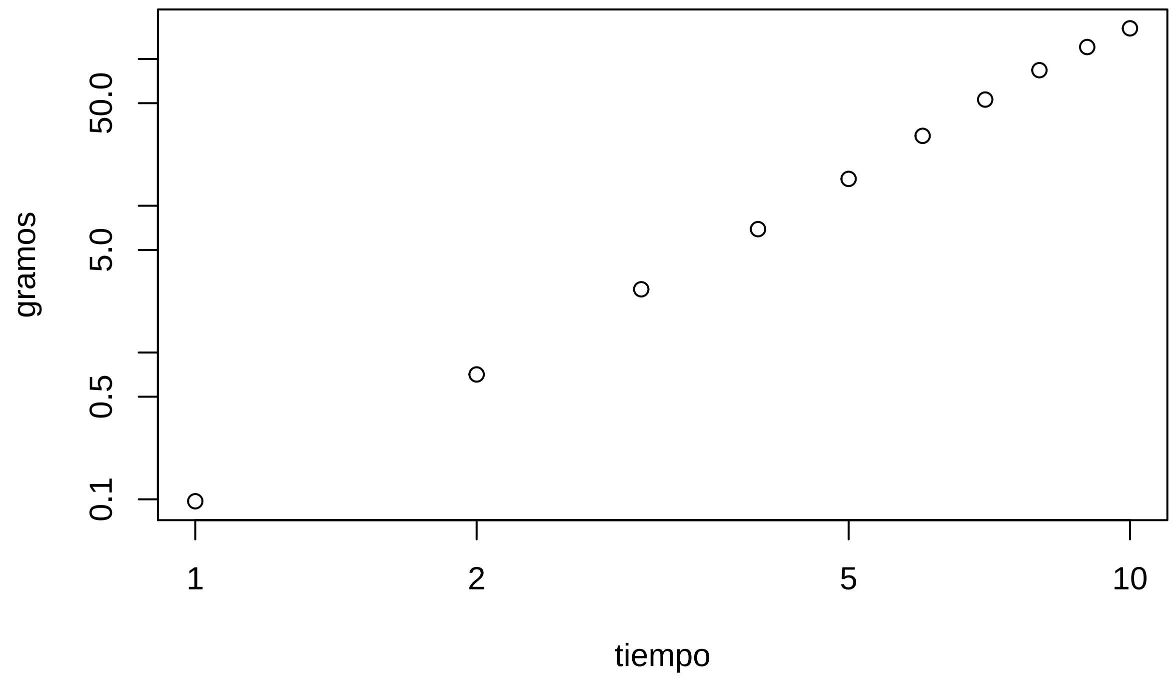
```
#Plot normal
plot(d.f)
```




```r
#Plot semilog
plot(d.f, log = "y")
```

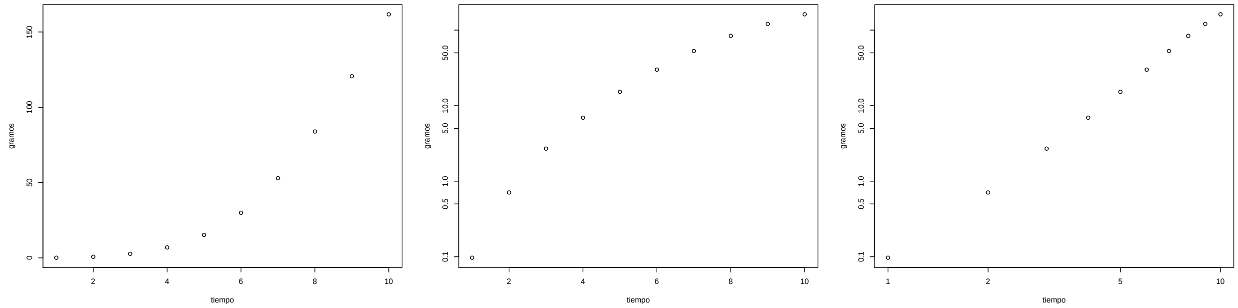



```
#Plot doble log
plot(d.f, log = "xy")
```



Represento todos los plots en la misma figura.

```
par(mfrow= c(1,3))
plot(d.f)
plot(d.f, log = "y")
plot(d.f, log = "xy")
```



```
par(mfrow= c(1,1))
```

Hacemos el modelo linea de transformar en logaritmo en base 10 a la variable dependiente, en función de la independiente (transformada a log en base 10).

```r
#regresión lineal
lm(log10(gramos)~log10(tiempo), data = d.f)
```

```
##
## Call:
## lm(formula = log10(gramos) ~ log10(tiempo), data = d.f)
##
## Coefficients:
##   (Intercept)  log10(tiempo)
##        -1.093          3.298
```

```r
#coeficiente de determinación
summary(lm(log10(gramos)~log10(tiempo), data = d.f))$r.squared
```

```
## [1] 0.9982009
```

Lo que acabamos de obtener es que:

$$\log(gramos) = 3.298 \cdot \log(tiempo) - 1.093$$

Es una buena aproximación de nuestros datos (r.squared = 0.998).

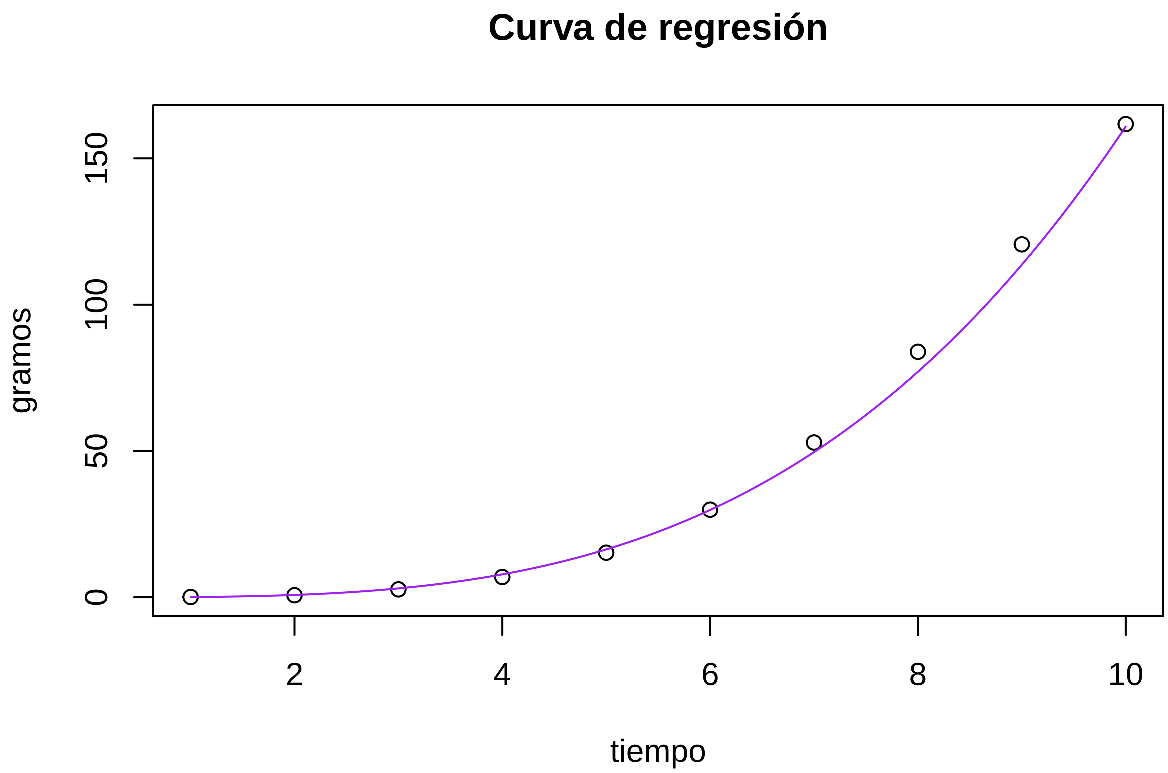
Con lo cual:

$$gramos = 10^{3.298 \cdot \log(tiempo)} \cdot 10^{-1.093} = tiempo^{3.298} \cdot 0.081$$

Se puede discernir que la relación existente entre la varibale dependiente e independiente es una relación de tipo potencial.

Para observar esto, vamos a representar la nube de ptos con plot() y luego añadimos la curva de la función dictada anteriormente con curve().

```r
plot(d.f, main = "Curva de regresión")
curve(x^(3.298)*0.081, add=TRUE, col = "purple")
```


Curva de regresión
gramos
0
50
100
150
2
4
6
8
10
tiempo