Capítulo 5

Regresión polinomial y factores

5.1. Regresión polinomial

Si una función media tiene un predictor X pueden usarse sus potencias enteras para aproximar $\mathrm{E}(Y|X)$. El caso más simple es la *regresión cuadrática*, cuya función media es:

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{5.1.1}$$

Dependiendo de los signos de los coeficientes, la función media cuadrática puede tomar cualquiera de las formas mostradas en la Figura 5.1. Se puede considerar usar este tipo de función cuando se tenga evidencia de que la media tome un valor máximo o mínimo en el rango del predictor. Este valor se obtiene cuando $\frac{d}{dx}\mathrm{E}(Y|X) = 0$ y es tal que:

$$x_M = -\frac{\beta_1}{2\beta_2} \tag{5.1.2}$$

Una función media cuadrática también puede usarse para modelar curvas sin necesidad de tener un valor máximo o mínimo en el rango del predictor. Por ejemplo en la Figura 5.1(a) si el rango se encuentra entre las líneas punteadas se puede observar una función media no lineal decreciente, mientras que en la Figura 5.1(b) la función media es no lineal creciente.

La regresión cuadrática es un caso especial de la *regresión polinomial*. Si se tiene un solo predictor, la función media polinomial de grado d es:

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d \tag{5.1.3}$$

Si $d = 2$ el modelo es cuadrático, si $d = 3$ es cúbico y así sucesivamente. Cualquier función suave puede estimarse usando un polinomio de grado conveniente.

La función media 5.1.3 puede estimarse usando mínimos cuadrados ordinarios con $p^{'} = d + 1$ términos dados por un intercepto y $X, X^2, \cdots, X^d$. Si d es mayor que tres se podrían tener serios problemas numéricos con algunos paquetes de computación y en algunos casos no sería posible realizar la estimación directa de 5.1.3. Se puede obtener mayor exactitud en los cálculos centrando los términos, $Z_k = (X - \overline{x})^k$, $k = 1, \cdots, d$. Seber (1977) desarrolla un método eficiente usando polinomios ortogonales.

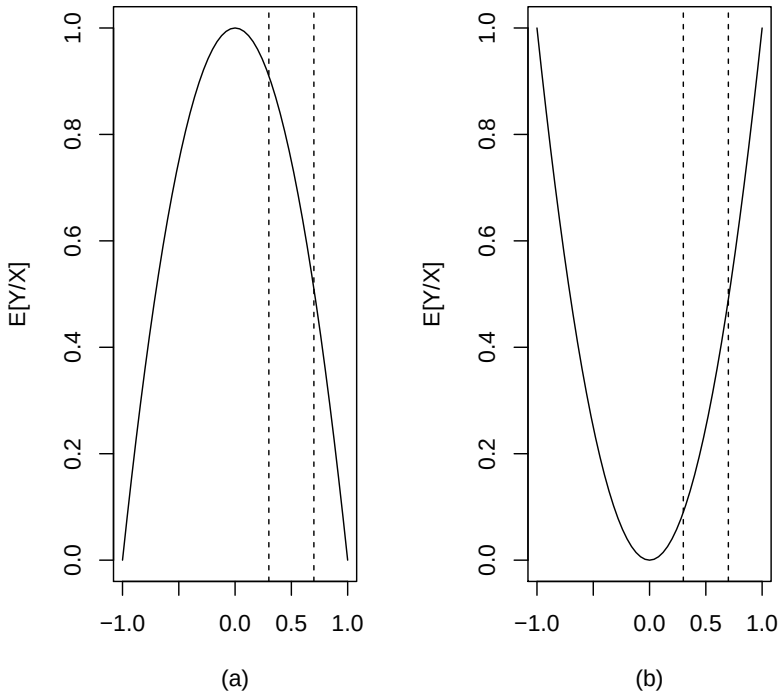

Figura 5.1: Curvas cuadráticas

Una prueba de falta de ajuste podría indicar si la función media lineal simple no resultada adecuada para la data. Cuando no se pueda realizarse esta prueba se podría comparar el modelo cuadrático:

$$\mathrm{E}\left(Y|X\right) = \beta_0 + \beta_1 X + \beta_2 X^2$$

con el modelo de regresión lineal simple:

$$\mathrm{E}\left(Y|X\right) = \beta_0 + \beta_1 X$$

usando una prueba t para $\beta_2 = 0$. Una estrategia para escoger d consiste en agregar términos a la función media hasta que la prueba t para el término de mayor grado resulte no significativa. También se puede utilizar una estrategia de eliminación en la que se fija un valor máximo para d y se eliminan los términos en la función media, uno a la vez empezando con el de mayor orden, hasta obtener un término con prueba t significativa. Kennedy y Bancroft (1971) sugieren utilizar un nivel de significación del 10 % para este procedimiento. En la mayoría de las aplicaciones para regresión polinomial solo se consideran $d = 1$ o $d = 2$. Para valores altos de d las curvas polinomiales estimadas tienen a *sobreestimar* la data modelando la variación aleatoria en lugar de la tendencia de la variable respuesta.

5.1.1. Polinomios con varios predictores

Con más de un predictor se puede considerar la posibilidad de agregar, además de las potencias, productos entre los predictores. Por ejemplo, en el caso de dos predictores la *función media de segundo orden* esta dada por:

$$\mathrm{E}\left(Y|X_1 = x_1, X_2 = x_2\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (5.1.4)$$

El nuevo término en 5.1.4, $x_1 x_2$, es llamado interacción. Con k predictores se tiene un intercepto, k términos lineales, k términos cuadráticos y $k\left(k+1\right)/2$ términos de interacción. Si $k = 5$ se tienen 26 términos y con $k = 10$ se tienen 76 términos. Una estrategia usual para el modelo de segundo orden es seleccionar sus términos usando la prueba t o algún proceso de selección para eliminar términos innecesarios.

Pasteles

Oehlert (2000) proporciona información para un pequeño experimento en horneado y mezclas de pasteles. Se consideraron dos variables, X_1 = tiempo de horneado (minutos) y X_2 = temperatura de horneado (°F). La variable respuesta Y = puntaje promedio para cuatro pasteles horneados usando una combinación particular de las variables predictoras.

La función media estimada basada en 5.1.4 para la data `Pastel` es:

$$\begin{aligned} \mathrm{E}(Y|X_1, X_2) &= -2204{,}4850 + 25{,}9176X_1 + 9{,}9183X_2 \\ &\quad -0{,}1569X_1^2 - 0{,}0120X_2^2 - 0{,}0416X_1X_2 \end{aligned} \tag{5.1.5}$$

Cada uno de los coeficientes estimados tiene nivel de significancia pequeño por lo que todos los términos son útiles en la función media. Desde que X_1 y X_2 aparecen en tres de los términos de 5.1.5, interpretar esta función es virtualmente imposible sin la ayuda de gráficos. La Figura 5.2 presenta una manera útil de resumir la información de la función media estimada. En la Figura 5.2(a), el eje horizontal es el tiempo de horneado X_1 y en el eje vertical la variable respuesta Y. Las tres curvas se obtienen fijando el valor de la temperatura X_2 en 340, 350 o 360. Por ejemplo, cuando $X_2 = 350$, 5.1.5 se simplifica a:

$$\mathrm{E}(Y|X_1, X_2 = 350) = -196{,}9664 + 11{,}3488X_1 - 0{,}1569X_1^2 \tag{5.1.6}$$

Cada una de las gráficas mostradas en la Figura 5.2 es una curva cuadrática ya que X_1^2 y X_2^2 se encuentran presentes en la función media. En la Figura 5.2(a), el valor máximo de la variable respuesta cuando $X_2 = 360$ es menor que el obtenido para $X_2 = 340$. En la Figura 5.2(b) con las curvas para tiempos de 35 y 37 minutos se obtiene una respuesta mayor que con 33 minutos. El puntaje es sensible a cambios en la temperatura de 10 o 15 grados y tiempos de horneado de pocos minutos.

Si se estima la función media 5.1.4, pero con $\beta_{12} = 0$ se obtienen las curvas de respuesta mostradas en la Figura 5.3. Sin interacción todas las curvas tienen la misma forma y son maximizadas con un mismo valor. Por ejemplo, para cualquier tiempo de horneado la respuesta se maximiza a una temperatura de aproximadamente 355 grados. A pesar que esta función media es más simple, la prueba F proporciona evidencia que esta función no estima adecuadamente la data.

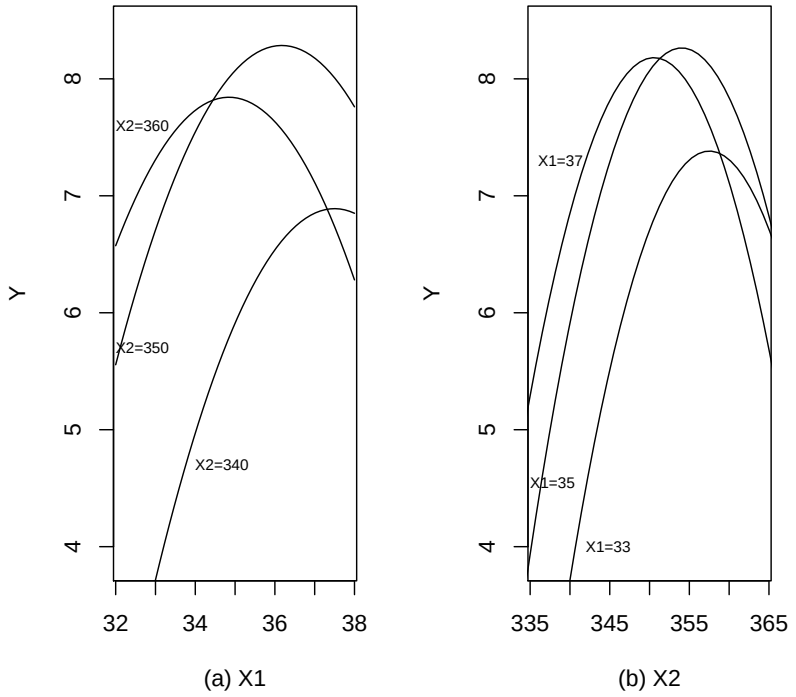

Figura 5.2: Curvas de respuesta estimadas para la data `Pastel` según 5.1.5

5.2. Factores

Los *factores* permiten la inclusión de predictores cualitativos o categóricos en la función media de un modelo de regresión lineal múltiple. Los factores pueden tener dos niveles como hombre y mujer, o más de dos niveles como color de ojos, distrito de residencia, etc.

Para incluir factores en la función media de un modelo de regresión múltiple se necesita una forma de indicar que nivel particular del factor esta presente para cada caso en la data. Para un factor con dos niveles puede usarse una *variable dummy*, es decir un término que toma el valor 1 para una de las categorías y 0 para la otra. La asignación de la categoría que toma el valor 1 es arbitraria y no afecta el resultado del análisis.

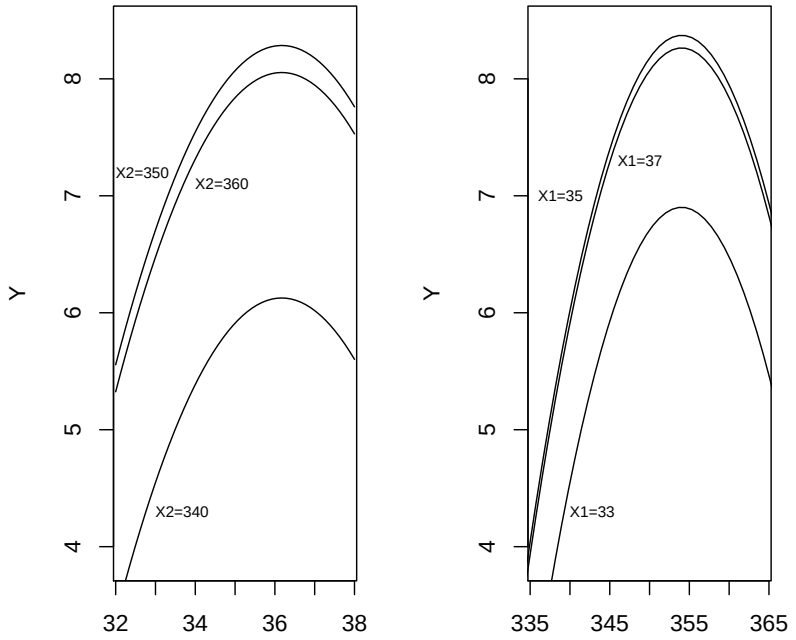

Figura 5.3: Curvas de respuesta estimadas para la data `Pastel` sin interacción

Como ejemplo considere la data `Mamiferos`. Se trata de un estudio de los patrones de sueño de 62 especies de mamíferos. Una de las variables respuesta de estudio es TS, el total de horas de sueño por día. Considere como predictor inicial la variable D que representa un índice para medir el peligro general que enfrenta cada especie. D tiene cinco categorías, con $D = 1$ se indica que la especie enfrenta menor peligro por parte de los otros animales, hasta $D = 5$ para especies que enfrentan mayor peligro. Las categorías se representan usando los números 1, 2, 3, 4 y 5 pero D no es una variable cuantitativa. Se pudo haber usado *muy poco*, *poco*, *medio*, *alto* y *muy alto* para esas cinco categorías. La data `Mamiferos` tiene tres valores perdidos para TS por lo que el análisis se lleva a cabo sobre las 59 especies restantes.

El objetivo es en determinar como cambia TS cuando D pasa de una ca-

tegoría a otra. Se debe estar en la posibilidad de escribir una función que permita que cada nivel de D tenga su propia media. Lo anterior puede realizarse haciendo uso de las variables dummy. Como D tiene cinco niveles, la $j-$ésima variable dummy para el factor, $J = 1, 2, \cdots, 5$ tiene el $i-$ésimo valor u_{ij}, para $i = 1, 2, \cdots, n$, dado por:

$$u_{ij} = \begin{cases} 1 & \text{si } \mathrm{D}_i = j - \text{ésima categoría de D} \\ 0 & \text{de otra manera} \end{cases} \tag{5.2.1}$$

Si el factor tuviera tres niveles en lugar de cinco, con $n = 7$ tal que los casos 1, 2 y 7 corresponden al primer nivel del factor, los casos 4 y 5 segundo nivel y los casos 3 y 6 al tercer nivel, entonces las tres variables dummy son:

U_1	U_2	U_3
1	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	1
1	0	0

Si estas variables dummy son agregadas al mismo tiempo se obtendría una columna de unos. Esta es una característica importante en un conjunto de variables dummy para un factor: la suma siempre da el mismo valor para cada caso.

Para la data `Mamiferos`, la función media puede escribirse como:

$$\mathrm{E}\left(\mathrm{TS}|\mathrm{D}\right) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5 \tag{5.2.2}$$

donde β_j puede ser interpretado como la media poblacional de todas las especies cuyo índice de peligro es j. La función media 5.2.2 aparentemente no incluye un intercepto. Sin embargo ya que la suma de los U_j es una columna de unos, el intercepto se encuentra implícito. Si se requiere incluir un intercepto de manera explícita, es posible dejar de lado una de las variables dummy según la siguiente regla: un factor con d niveles puede representarse usando como máximo d variables dummy. Si el intercepto se encuentra en la función media, pueden usarse como máximo $d - 1$ variables dummy.

Otra opción es eliminar la primera variable dummy y utilizar la siguiente función media:

$$\mathrm{E}\,(\mathrm{TS}|\mathrm{D}) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5 \tag{5.2.3}$$

donde los parámetros están representados de manera diferente ya que tienen significados diferentes. Las medias para los cinco grupos son ahora $\eta_0 + \eta_i$ para los niveles $j = 2, \cdots, 5$ de D, y η_0 para $\mathrm{D} = 1$. Aunque los parámetros tengan diferentes significados en 5.2.2 y 5.2.3, ambos estiman una media para cada nivel de D.

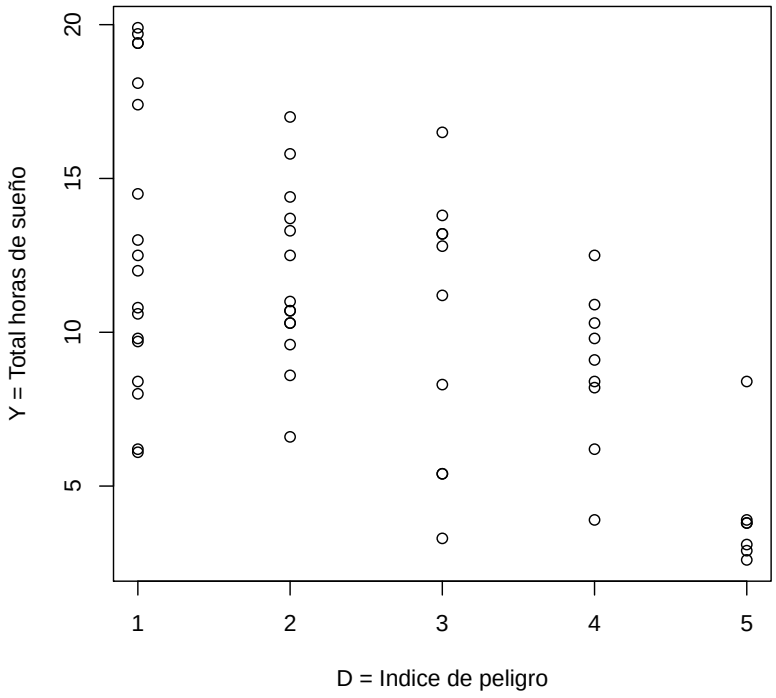

Figura 5.4: Diagrama de dispersión para la data `Mamíferos`

La mayoría de programas de computadora permiten usar factores en una función media sin necesidad de calcular las variables dummy.

En R el primer paso es declarar que D es un factor y luego estimar la función media 5.2.3 usando:

```
> D <- as.factor(D)
> m1 <- lm(TS ~ D, na.action=na.omit)
> coef(m1)
```

```
(Intercept)          D2          D3          D4          D5
  13.083333   -1.333333   -2.773333   -4.272222   -9.011905
```

y la funcion media 5.2.2 con:

```
> m2 <- lm(TS ~ -1 + D, na.action=na.omit)
> coef(m2)
```

```
       D1        D2        D3        D4        D5
13.083333 11.750000 10.310000  8.811111  4.071429
```

5.2.1. Agregando un predictor: comparación de lineas de regresión

Para la data Mamíferos, suponga que se agrega x, el logaritmo en base dos del peso promedio del cuerpo de las especies, como predictor. Se tienen dos predictores, el factor D con cinco niveles y x.

Se asume que para un valor dado de D:

$$\mathrm{E}\left(\mathrm{TS}|X=x,\ \mathrm{D}=j\right)=\beta_{0j}+\beta_{1j}x \tag{5.2.4}$$

Se pueden distinguir cuatro situaciones diferentes:

Modelo 1: *Más general* Cada nivel de D tiene diferente pendiente e intercepto, correspondiente a la Figura 5.5. Se puede especificar esta función media de muchas formas. Si se incluye una variable dummy para cada nivel de D, se puede escribir :

$$\mathrm{E}\left(\mathrm{TS}|X=x,\ \mathrm{D}=j\right)=\sum_{j=1}^{5}\beta_{0j}U_j+\sum_{j=1}^{5}\beta_{1j}U_jx \tag{5.2.5}$$

La función media 5.2.5 tiene $2d$ términos, las d variables dummy para d interceptos y d interacciones formadas multiplicando cada variable dummy por la variable continua para formar las d pendientes.

En R la función media 5.2.5 puede especificarse como:

```
> x <- logb(PesoCuerpo,2)
> coef(lm(TS ~ -1 + D + D:x, na.action=na.omit))

        D1         D2         D3         D4         D5       D1:x       D2:x
13.8681822 11.5089393 10.3673158  9.6408015  6.8370993 -0.4027435 -0.4266289
      D3:x       D4:x       D5:x
-0.6414392 -0.2851411 -0.4695489
```

Usando un símbolo diferente para los parámetros, esta función media puede escribirse como:

$$\mathrm{E}\left(\mathrm{TS}|X=x,\,\mathrm{D}=j\right)=\eta_0+\sum_{j=2}^{5}\eta_{0j}U_j+\sum_{j=1}^{5}\eta_{1j}U_jx \tag{5.2.6}$$

Comparando las dos parametrizaciones se tiene que $\eta_0=\beta_{01}$ y para $j>1$:

$$\eta_{0j}=\beta_{0j}-\beta_{01}\quad\text{y}\quad\eta_{1j}=\beta_{1j}$$

La parametrización 5.2.5 es más conveniente si se desea obtener parámetros interpretables, mientras que 5.2.6 es útil para comparar funciones media.

La función media 5.2.6 puede especificarse en R escribiendo:

```
> n1 <- lm(TS ~ D + D:x, na.action=na.omit)
> coef(n1)

(Intercept)          D2          D3          D4          D5        D1:x
 13.8681822  -2.3592429  -3.5008664  -4.2273808  -7.0310829  -0.4027435
       D2:x        D3:x        D4:x        D5:x
 -0.4266289  -0.6414392  -0.2851411  -0.4695489
```

En la Figura 5.5 algunas de las funciones medias estimadas para cada nivel de D parecen ser paralelas, luego es posible que una sola función media sea apropiada para toda la data.

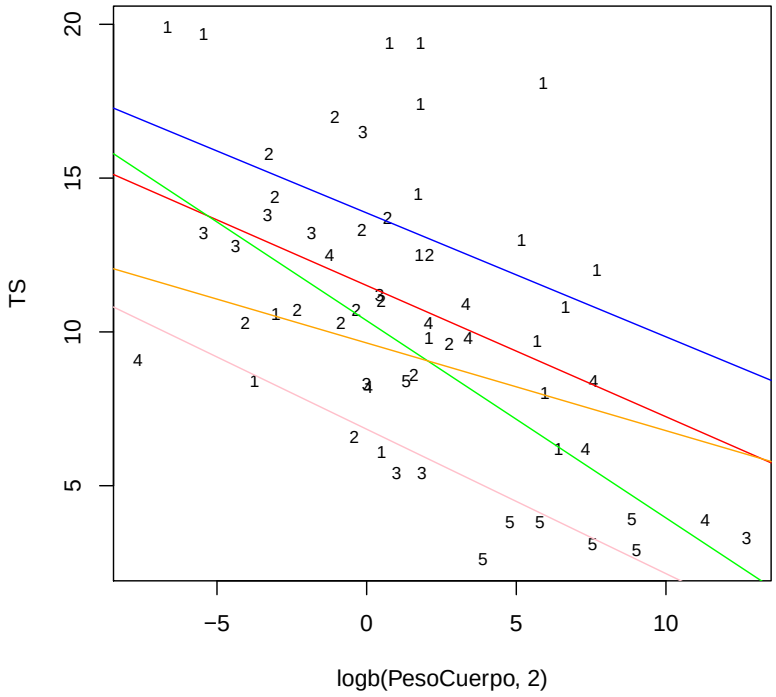

Figura 5.5: (a) Modelo de regresión general

Modelo 2: Regresiones paralelas Todas las funciones media dentro de cada grupo son paralelas como se muestra en la Figura 5.6, luego $\beta_{11} = \beta_{12} = \cdots = \beta_{15}$ en 5.2.5, o $\eta_{12} = \eta_{13} = \cdots = \eta_{15} = 0$ en 5.2.6. Cada nivel de D puede tener su propio intercepto. Esta función media puede especificarse como:

```
> n2 <- lm(TS ~ D + x, na.action=na.omit)
> coef(n2)
```

```
(Intercept)          D2          D3          D4          D5           x
  13.932502   -2.428716   -3.583566   -3.853468   -7.294486   -0.435749
```

La diferencia entre los niveles de D es la misma para todos los valores del predictor ya que no se incluye en la función media ninguna variable dummy

para la interacción con la variable predictora. Esta función media estima términos para el intercepto, x y D. El número de parámetros estimados es $d + 1$.

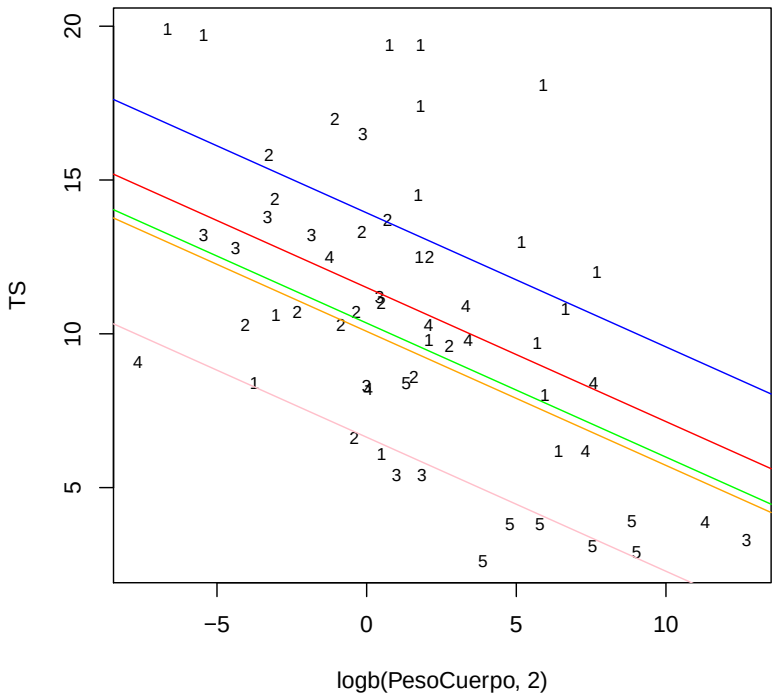

Figura 5.6: Regresión paralela

En la Figura 5.6 se puede observar que la función media estimada para D = 5 tiene el intercepto más pequeño, para D = 1 el intercepto es el más grande y para las tres categorías restantes las funciones media son aproximadamente las mismas. Esto podría sugerir que las tres categorías intermedias podrían combinarse.

Modelo 3: Intercepto común En esta función media todos los interceptos son iguales, $\beta_{01} = \beta_{02} = \cdots = \beta_{05}$ en 5.2.5, o $\eta_{02} = \eta_{03} = \cdots = \eta_{05} = 0$ en 5.2.6, pero las pendientes son arbitrarias, tal como se muestra en la Figura 5.7. Esta función media es particularmente inapropiada para la data `Mamíferos`,

ya que requiere que el número esperado de horas de sueño para una especie cuyo peso es de 1 kg, es $x = 0$, es el mismo para todos los niveles de peligro, y esto es totalmente arbitrario. La función media podría cambiar si se usan unidades diferentes como gramos o libras.

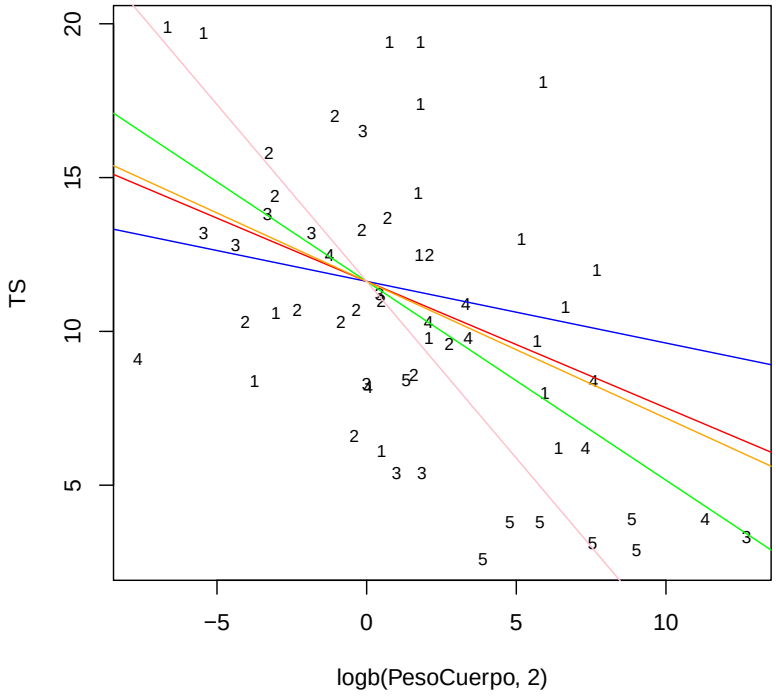

Figura 5.7: Regresión con intercepto común

Esta función media se estima en R usando:

```
> n3 <- lm(TS ~ x + x:D, na.action=na.omit)
> coef(n3)

(Intercept)           x        x:D2        x:D3        x:D4        x:D5
 11.6258529  -0.2004724  -0.2105761  -0.4458769  -0.2441391  -0.9491168
```

Modelo 4: Líneas de regresión coincidentes Aquí, todas las líneas son las mismas, $\beta_{01} = \beta_{02} = \cdots = \beta_{05}$ y $\beta_{11} = \beta_{12} = \cdots = \beta_{15}$ en 5.2.5, o $\eta_{02} = \eta_{03} = \cdots = \eta_{05} = \eta_{12} = \cdots = \eta_{15} = 0$ en 5.2.6. Este es el más riguroso de los modelos, ilustrado en la Figura 5.8. Esta función media requiere solo un término para el intercepto y otro para x, con un total de 2 parámetros y esta dado por:

```
> n4 <- lm(TS ~ x, na.action=na.omit)
> coef(n4)

(Intercept)           x
 11.4377412  -0.5497446
```

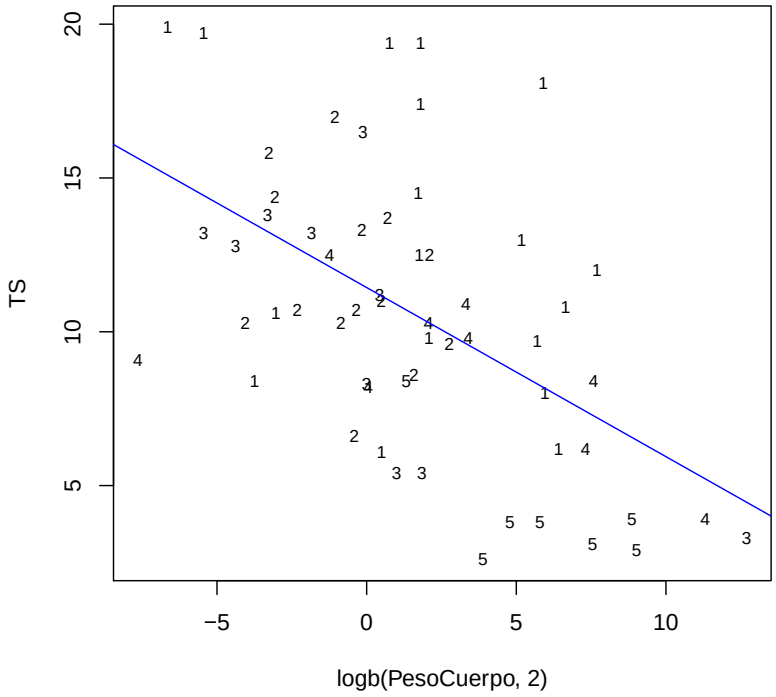

Figura 5.8: Regresión común