Teórica 01

Peredo Leonel

En muchas disciplinas científicas interesa saber cómo se relacionan distintas variables entre sí. Una de las herramientas principales que tiene la estadística para hacer eso es la regresión.
El modelo de regresión lineal es un método conceptualmente simple para investigar la relación entre dos o más variables. Esta relación se expresa en la forma de una ecuación o un modelo que conecta una variable respuesta o variable dependiente (continua) y una o muchas variables explicativas o covariables. Es una técnica clásica y muy utilizada.
La teoría de modelos lineales es un caso especial de la teoría más general que cubre modelos más flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitar el aprendizaje, especialmente sin matemáticas avanzadas.
Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estadística se ha realizado sobre modelos lineales. También significa que muchos de los consumidores de estadística esperan modelos lineales o compararán los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto cómo funcionan, como cuáles son sus limitaciones.
La regresión lineal se ocupa de investigar la relación entre dos o más variables continuas.
Comenzaremos tratando de describir el vínculo entre dos variables aleatorias continuas. Medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc.
Ejemplo: Se miden en el año 2015, para 187 países:
Y: Expectativa de vida: Número promedio de años que un niño recién nacido espera vivir, si los patrones de mortalidad no cambiaran (`life`)
X: Mortalidad infantil: Número de niños de 0 a 5 años que mueren en un año, por cada 1000 niños vivos (`child`) La notación matemática para las observaciones será (X_i, Y_i), donde X_i = child del país i-ésimo Y_i = life del país i-ésimo, con $1 \leq i \leq n = 187$

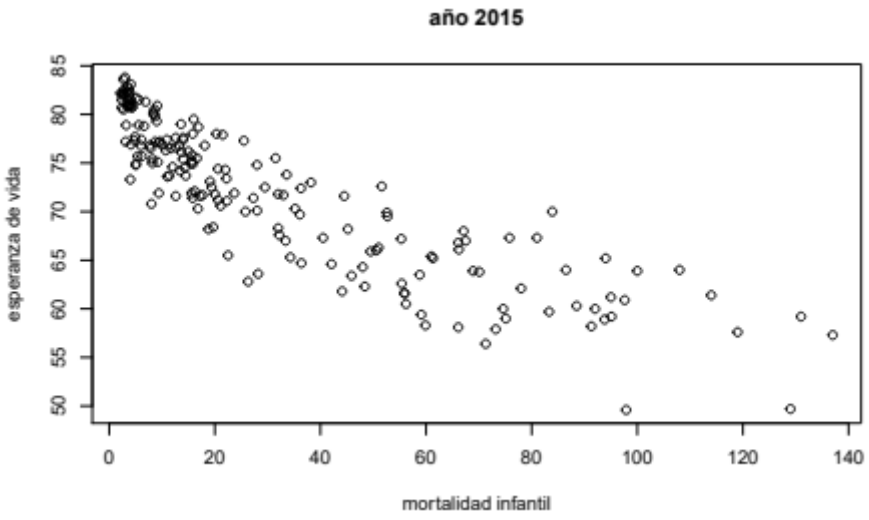


Figure 1: Expectativa de vida en función de Mortalidad en 187 países, año 2015

En un scatter plot se ubican los resultados de una variable (X) en el eje horizontal y los de la otra variable (Y) en el eje vertical. Cada punto en el gráfico representa una observación (X_i, Yi).
Se pierde la información del individuo (país). Con este gráfico podemos determinar si existe algún tipo de relación entre X e Y.
En este caso vemos que a medida que aumenta la mortalidad infantil, decrece la esperanza de vida. Queremos modelar la relación entre ambas variables. El objetivo es tratar de explicar la esperanza de vida a partir de la mortalidad infantil.
Otro ejemplo: Pearson-Lee data.
Karl Pearson organizó la recolección de datos de 1100 familias en Inglaterra en el período 1893 a 1898. El conjunto de datos `Heights`, en el paquete `alr4` de `R` da la altura de madres e hijas (en pulgadas), con hasta dos hijas por madre. Todas las hijas tienen 18 años o más, y todas las madres son menores de 65 años. En la fuente los datos aparecen redondeados a la pulgada más cercana. En la librería se les agrega un error de redondeo para que el gráfico no sea discreto. Mostramos datos en cm.

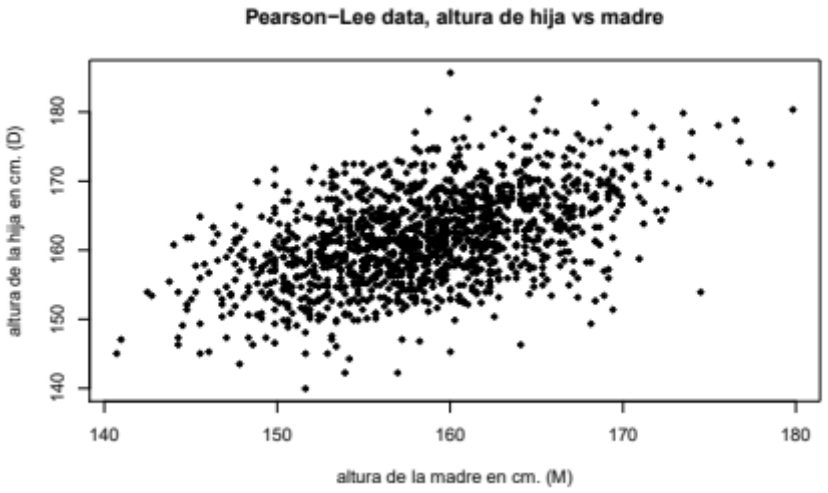


Figure 2: Altura de la hija (cm) en función de la altura de la madre (cm)

Ejemplo con más variables: Para $n = 193$ pa´ıses, medimos en 2015 las siguientes variables: $Y =$ `life` es la esperanza de vida al nacer (en años) $X1 =$ `income`: Producto Bruto Interno, per cápita (en USD) $X2 =$ `child` Tasa de Mortalidad de 0 a 5 años, por cada mil niños nacidos vivos en el año. $X3 =$ `dtp3`: porcentaje de niños de un año inmunizados con tres dosis de vacuna contra la difteria, tétanos y pertussi (DTP3) $X4 =$ `school`: n´umero de años de escolaridad promedio en hombres de 25˘34 años. $X5 =$ `status`: grado de desarrollo del país (”developed” o ”not.developed”) El objetivo es explicar a Y ¿Cómo lo visualizamos?

Correlación de Pearson

Correlación poblacional de un vector aleatorio (X, Y):

$$\rho(X,Y) = \frac{\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

Correlación muestral:

$$r = \hat{\rho}(X,Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

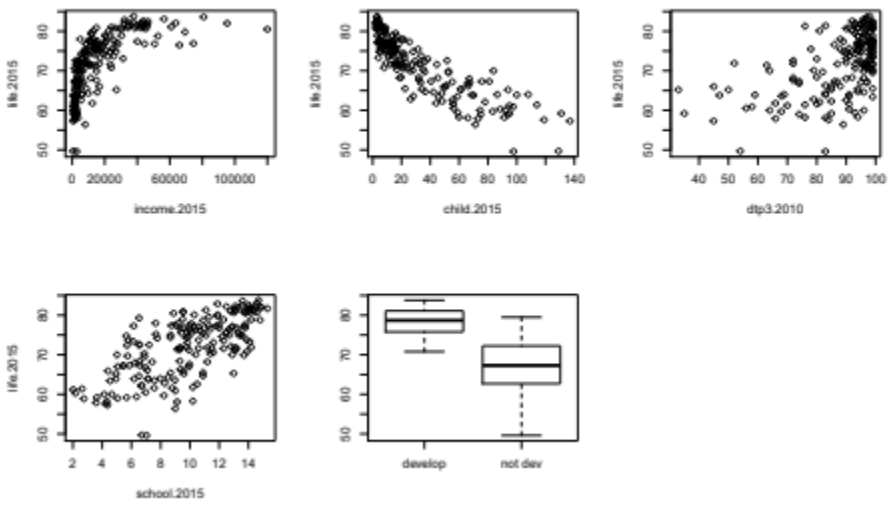
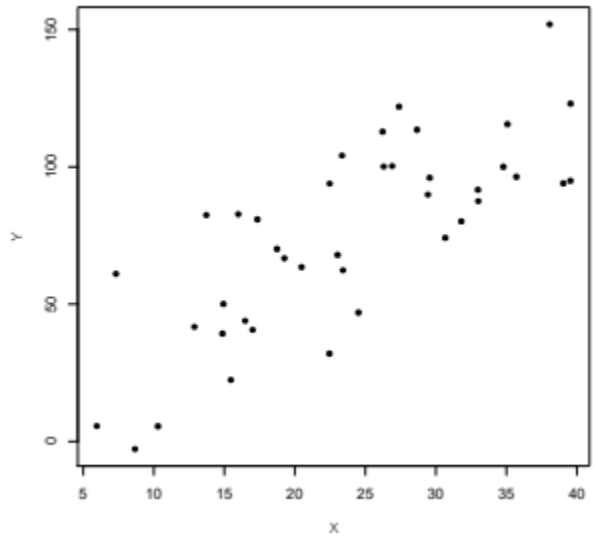


Figure 3: Gráficos de life en función de distintas variables.

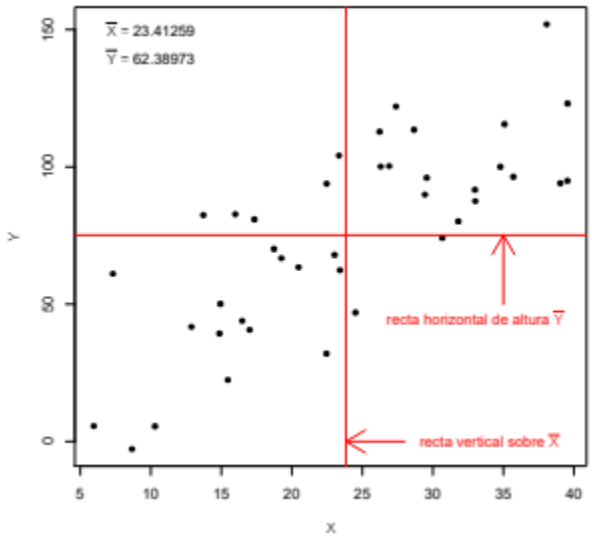
A su vez, el numerador se denomina covarianza muestral de X y de Y. Entonces $\widehat{cov}(X_i, Y_i) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$, y el denominador es el producto de los desvíos muestrales de cada muestra por separado S_X y S_Y. Si bien el numerador puede ser positivo o negativo, el denominador es siempre positivo. Entonces el signo de r está determinado por la covarianza muestral.

Interpretación de la correlación

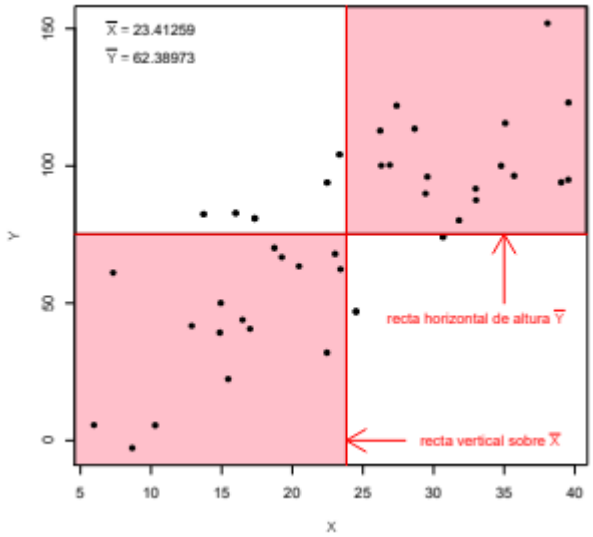
Supongamos que tenemos este gráfico y queremos analizarlo:



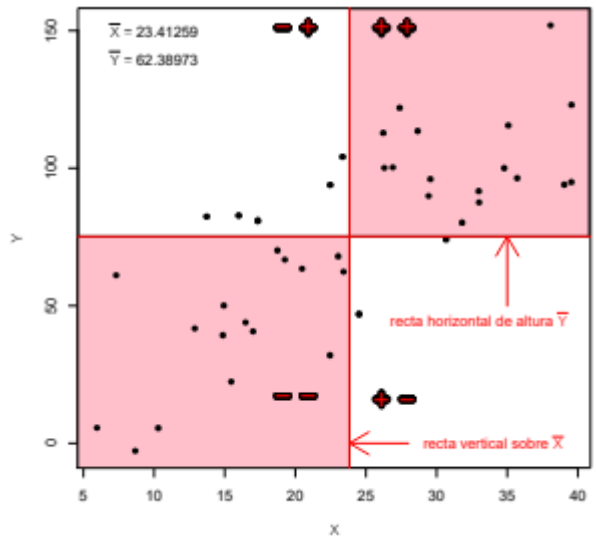
Le agregamos una línea horizontal a nivel de $\bar{Y}$, y una vertical a nivel de $\bar{X}$.



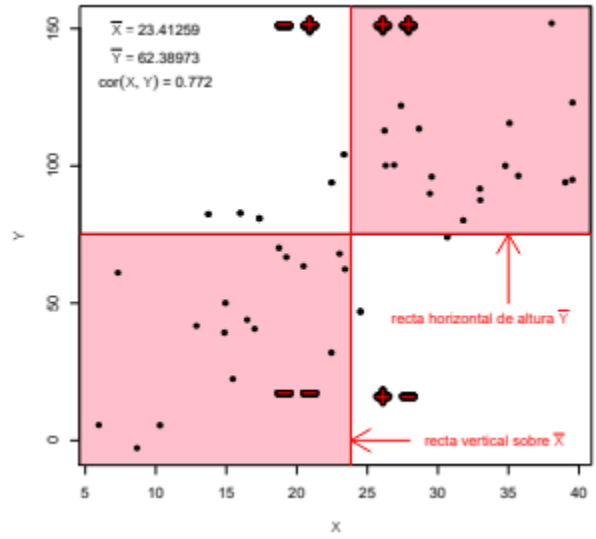
Notemos que la mayoría de los datos están en los cuadrantes rosas



Que a su vez, si lo pensamos como signos respecto a las medias como puntos de corte, se verían así.



La covarianza positiva implica que la mayoría de los puntos están en cuadrantes cuyo producto es positivo.



Propiedades del coeficiente de correlación

1. $-1 \leq r \leq 1$.
2. El valor absoluto de r, $|r|$, mide la fuerza de la asociación lineal entre X e Y, a mayor valor absoluto, hay una asociación lineal más fuerte.
3. El caso particular $r = 0$ indica que no hay asociación lineal entre X e Y.
4. El caso $r = 1$ indica asociación lineal perfecta. O sea que los puntos están ubicados sobre una recta de pendiente (o inclinación) positiva.
5. En el caso $r = -1$ tenemos a los puntos ubicados sobre una recta de pendiente negativa (o sea, decreciente).
6. El signo de r indica que hay asociación positiva entre las variables (si $r > 0$); o asociación negativa entre ellas (si $r < 0$).
7. $r = 0,90$ indica que los puntos están ubicados muy cerca de una recta creciente, $r = 0,80$ indica que los puntos están cerca, pero no tanto, de una recta creciente.
8. r no depende de las unidades en que son medidas las variables (milímetros, centímetros, metros o

kilómetros, por ejemplo) .

9. Los roles de X e Y son simétricos para el cálculo de r.
10. **Cuidado**: el coeficiente de correlación de Pearson es muy sensible a observaciones atípicas. Hay que hacer siempre un scatter plot de los datos antes de resumirlos con r.

Ejemplo

La temperatura corporal de mamíferos y pájaros tiende a fluctuar durante el díkka según un ritmo circadiano regular. En un estudio 1 se registra la temperatura corporal de 10 ardillas antílopes cada 6 minutos a lo largo de 10 días consecutivos en condiciones de laboratorio. Elegimos una ardilla y promediamos las temperaturas de los 10 días para obtener un conjunto de datos de 24×10 observaciones. Los autores trataban de contestar a la pregunta: ¿Hay una asociación entre la hora del día y la temperatura corporal?
Para contestarla, tenemos dos estrategias:
- Calcular la correlación entre la hora del día y la temperatura corporal de la ardilla
- Graficar ambas variables: horario y temperatura en un scatter plot

Xi = horario de la i´esima medici´on Yi = temperatura promedio a lo largo de 10 d´ıas de las 10 mediciones realizadas en el horario X

Calculamos la correlaci´on muestral: `cor(horario, temperatura6) [1] -0.05863851` Parece no haber relaci´on entre ambas. ¿Eso mide la correlaci´on? Casi no hay relaci´on lineal entre ambas variables.

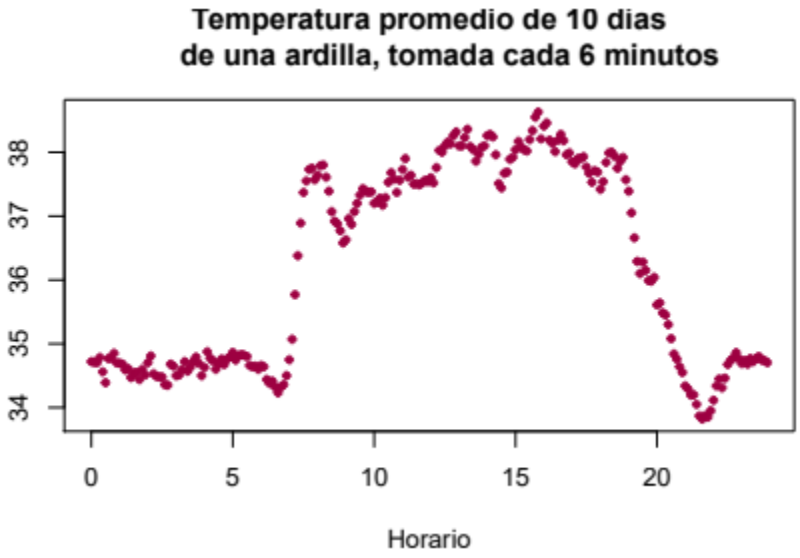
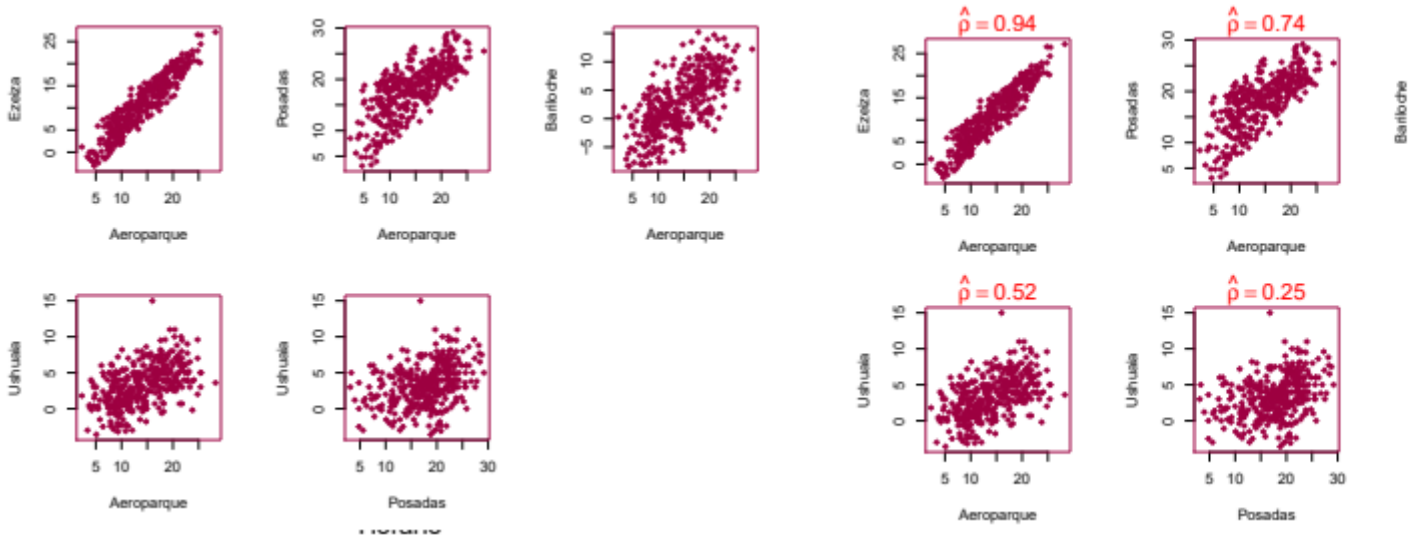
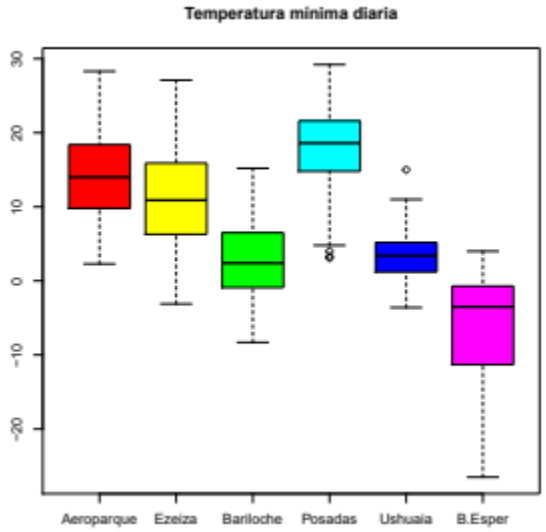


Figure 4: Scatterplot

Una correlaci´on cercana a cero no significa (necesariamente) que las dos variables no est´an asociadas: la correlaci´on mide s´olo la fuerza de una relaci´on lineal

En el sitio del Servicio Meteorol´ogico Nacional pueden bajarse los datos de temperaturas m´aximas y m´ınimas diarias de los distintos observatorios ubicados en el pa´ıs. 2 Elegimos 5 localidades, queremos ver c´omo se relacionan entre s´ı las temperaturas m´ınimas del mismo d´ıa. As´ı tenemos un vector aleatorio (Ai , Ei , Bi , Pi ,Ui), con 1 LEQ i LEQ n = 365 Ai = temperatura m´ınima del d´ıa i en Aeroparque Ei = temperatura m´ınima del d´ıa i en Ezeiza Bi = temperatura m´ınima del d´ıa i en Bariloche Pi = temperatura m´ınima del d´ıa i en Posadas Ui = temperatura m´ınima del d´ıa i en Ushuaia





La teor´ıa de modelos lineales es un caso especial de la teor´ıa m´as general que cubre modelos m´as flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitan el aprendizaje, especialmente sin matem´aticas avanzadas. 2 Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estad´ıstica se ha realizado sobre modelos lineales. Tambi´en significa que muchos de los consumidores de estad´ıstica esperan modelos lineales o comparar´an los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto c´omo funcionan como cu´ales son sus limitaciones.

El modelo de regresi´on lineal es un modelo para el v´ınculo de dos variables aleatorias que denominaremos X = variable predictora o covariable e Y = variable dependiente o de respuesta. El modelo lineal (simple pues s´olo vincula una variable predictora con Y) asume que 1 La distribuci´on de X no est´a especificada, incluso puede ser determin´ıstica. 2 Proponemos el siguiente modelo para las variables: FORMULA donde E es el t´ermino del error. 3 Asumimos que la variable aleatoria error E tiene esperanza 0, varianza constante desconocida que llamaremos SIGMA 2 , no est´a correlacionado con X y no est´a correlacionado con los errores de otras observaciones. En el modelo (2) los n´umeros BETA0 y BETA1 son constantes desconocidas que se denominan par´ametros del modelo, o coeficientes de la ecuaci´on. Los par´ametros se denominan BETA0 = ordenada al origen BETA1 = pendiente. El supuesto de la relaci´on funcional entre X e Y sea lineal es no trivial, ya dijimos que muchas variables no lo cumplen. El requisito de que el error tenga

varianza constante, se lo suele llamar homoscedasticidad, y tampoco es no trivial. Lo mismo pasa con las no correlaciones. Pero el supuesto de que los errores tengan esperanza cero s´ı es trivial. Verificar que los supuestos se cumplan para un conjunto de datos ser´a uno de los objetivos que atacaremos m´as adelante en la materia.

FORMULAS DIAPO 32, 33 y 34