

Práctica 8

Estadística descriptiva bidimensional: recta de regresión

En esta práctica aprenderemos a usar R Commander para calcularla recta de recta de regresión, representarla gráficamente sobre el diagrama de dispersión y predecir o aproximar, a partir de dicha recta, el valor de una variable a partir del valor de la otra.

Contenido de la práctica

8.1. Recta de regresión	56
8.1.1. Recta de regresión de Y sobre X	56
8.1.2. Representación gráfica	57
8.1.3. Predicción	57
8.2. Ejercicios propuestos	58

Para ilustrar esta práctica vamos a utilizar los datos que aparecen en el fichero *Pulso.RData*. Este fichero contiene información de diversos alumnos y el efecto que tiene el ejercicio sobre ellos. Para cada alumno se registra información que se codifica en las siguientes variables:

- *Pulso.Antes*: pulso antes de realizar ejercicio (reposo)
- *Pulso.Despues*: pulso después de realizar ejercicio
- *Corre*: si ha corrido entre las dos tomas del pulso
- *Fuma*: si fuma
- *Sexo*: hombre o mujer
- *Altura*: altura, en centímetros
- *Peso*: peso, en kilogramos
- *Actividad*: frecuencia actividad física (baja, moderada, alta)

Ejercicio 51: Realiza un breve estudio de la relación entre las variables numéricas contenida en el archivo *Pulso.RData*. Para ello, haz uso de los diagramas de dispersión y del coeficiente de correlación.

1. ¿Qué variables están relacionadas? ¿Qué tipo de relación observas?
2. ¿Para qué variables no observas relación?

8.1. Recta de regresión

Antes de calcular la recta de regresión es necesario determinar qué variable es la explicativa (X) y qué variable es la explicada (Y). A modo de ejemplo, supondremos que queremos explicar la variable $Y =$ “Peso” en función de la variable $X =$ “Altura”.

8.1.1. Recta de regresión de Y sobre X

Para calcular un modelo de regresión lineal con *R Commander* se utiliza la opción:

Estadísticos > Ajustes de modelos > Regresión lineal...

En la ventana que aparece se puede indicar un nombre para el modelo que sea representativo y nos sirva para recordar la recta con la que estamos trabajando. Por ejemplo, *RegPesoSobreAltura*. Además, hay que seleccionar en el cuadro de la izquierda la variable Y , que es la explicada, y en el de la derecha la variable X , que es la explicativa.

De todos los resultados que devuelve el programa nos interesan los marcados en la siguiente imagen:

```
> summary(RegPesoSobreAlt)
Call:
lm(formula = peso ~ alt, data = Dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-14.2879  -5.1432  -0.5136   3.9078  24.1010

Coefficients:
(Intercept) -92.86878
alt          0.90929

Std. Error t value Pr(>|t|)
13.22662   -7.021 4.02e-10 ***
0.07567   12.016 < 2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.71 on 90 degrees of freedom
Multiple R-squared: 0.616    Adjusted R-squared: 0.6117
F-statistic: 144.4 on 1 and 90 DF,  p-value: < 2.2e-16
```

Tendremos, por tanto:

- La recta de regresión Y sobre X :

$$y = 0.90929x - 92.86878$$

- El coeficiente de determinación lineal, $r_{XY}^2 = 0.616$.

Recuerda que el coeficiente de determinación lineal representa el porcentaje de la variabilidad de la variable Y que puede explicarse, mediante la recta de regresión, si conocemos la variable X . En este caso, el 61.6% de la variabilidad de la variable *Peso* puede explicarse, mediante la recta de regresión, si conocemos la variable *Altura*.

8.1.2. Representación gráfica

Para dibujar la recta de regresión sobre el diagrama es necesario dibujar dicho diagrama utilizando el menú:

Gráficas > Diagrama de dispersión...

y marcar la opción *Línea de mínimos cuadrados* en la pestaña *Opciones*, dentro de *Opciones gráfica*.

También es posible representar diferentes rectas de regresión si existe alguna variable cualitativa que permita agrupar los casos según los valores de dicha variable. Para ello bastará utilizar la opción *Gráfica por grupos...* de la pestaña *Datos* y seleccionar ahí la variable en cuestión.

Observa que, aunque la representación gráfica es sencilla, para calcular numéricamente las ecuaciones de las rectas u otros coeficientes es necesario filtrar los datos.

Ejercicio 52: Representa gráficamente, utilizando un diagrama de dispersión, las variables *Altura* y *Peso*, distinguiendo los casos correspondientes a hombres y a mujeres, e incluye la recta de regresión correspondiente a cada sexo.

8.1.3. Predicción

Para predecir el valor esperado de la variable Y dado un valor concreto de la variable X se utiliza el menú:

Modelos > Predecir usando el modelo activo > Introducir datos y predecir...

En la tabla que se abre introduciremos, en la columnas correspondiente, tantos valores como queramos predecir. En nuestro caso, por ejemplo, si queremos predecir el valor esperado del peso de una persona que mide 171 cm, entonces escribiremos 171 en la columna *Altura*. Después, para que calcule la predicción correspondiente hay que cerrar la ventana.

Nota: algunas versiones de R Commander no disponen de la opción *Introducir datos y predecir...* que se describe en este apartado. En ese caso, para realizar una predicción basta con utilizar la ecuación de la recta de regresión. Por ejemplo, si disponemos de la recta de regresión $y = 0.90929x - 92.86878$ obtenida en el apartado 8.1.1, para calcular el valor esperado del peso de una persona que mide 171 cm tendremos que escribir $0.90922 * 171 - 92.86878$ en la ventana de instrucciones y, después, pulsar el botón *Ejecutar*:

```
> 0.90922 * 171 - 92.86878
[1] 62.60784
> |
```

Ejercicio 53: Predice, utilizando la recta de regresión del *Peso* sobre la *Altura*, cuál es el peso esperado de dos personas adultas que midan 180 cm y 160 cm y un niño que mide 100 cm. Comenta los resultados obtenidos.

Ejercicio 54: Filtra los datos para cada sexo y, después, predice utilizando la recta de regresión del *Peso* sobre la *Altura* cual es el peso esperado para una mujer que mide 171 cm y para un hombre de esa misma estatura. Compara el resultado con el que se obtiene sin desagregar los datos.

8.2. Ejercicios propuestos

Ejercicio 55: Utiliza el fichero *Pulso.RData* para responde a las siguientes preguntas:

1. ¿Cuál es la altura esperada de una persona que pesa 80 kg?
2. Intenta predecir el peso de una persona con 65 pulsaciones en reposo, ¿te parece fiable esta predicción?
3. Estudia la relación entre las variables *Pulso.Antes* y *Pulso.Después* desagregando los datos según si el alumno ha corrido o no. ¿Para qué grupo de alumnos hay mayor relación entre las dos variables, para los que ha corrido o para los que no?
4. Intenta predecir el pulso después de correr para una de las personas que no ha corrido y que, en reposo, tuviese 65 pulsaciones. ¿Te parece fiable esta predicción?

Ejercicio 56: El fichero *DatosZgzMdr.RData* contiene los datos de la temperatura media mensual y el número de bodas al mes en las ciudades de Madrid y Zaragoza, entre los años 2015 y 2019.

1. ¿Entre qué variables se observa mayor relación lineal?
2. Intenta predecir qué temperatura habrá hecho en Madrid un mes en el que la temperatura media de Zaragoza ha sido de 20°C grados. ¿Te parece fiable la predicción?
3. ¿Qué porcentaje de la variabilidad de la temperatura media mensual en Madrid puede explicarse con la temperatura media de Zaragoza, utilizando la recta de regresión?
4. Intenta predecir el número de bodas que habrá habido en Madrid un mes en el que en Zaragoza hubo 115 bodas.
5. Intenta predecir el número de bodas que habrá habido en Zaragoza un mes en el que en Madrid hubo 697 bodas.
6. ¿Qué porcentaje de la variabilidad del número de bodas mensual en Zaragoza puede explicarse con el número de bodas mensual en Madrid, utilizando la recta de regresión?
7. En abril de 2020 en Madrid hubo tan solo 15 bodas, intenta predecir cuántas hubo en Zaragoza y comenta el resultado.
8. ¿Existe algún tipo de relación entre la temperatura media mensual en Zaragoza y el número de bodas se llevan a cabo en Madrid? ¿Crees que es una relación de causalidad?

Ejercicio 57: Introduce los siguientes datos y responde:

x_i	1	0	6	13	12	4
y_i	10	12	9	4	0	7

1. ¿Existe relación entre las variables? ¿de qué tipo?
2. ¿Cuál es el coeficiente de correlación?
3. ¿Cuál es la recta de regresión de Y sobre X ?
4. ¿Cuál es el valor esperado de la variable Y si X vale 10?
5. ¿Cuál es la varianza de Y ?
6. ¿Qué porcentaje de la variabilidad de Y se explica con la recta de regresión? ¿Cuál será entonces la varianza residual?

Calcula todos los residuos: puedes hacerlo de manera automática utilizando el menú *Modelos > Añadir las estadísticas de las observaciones a los datos* y marcando la opción *Residuos*. Después calcula la varianza de los residuos y observa que coincide con el valor que habías calculado previamente.