

Laboratorio: Análisis exploratorio de varias variables. Modelos de regresión lineal.

Objetivos: El alumno al finalizar la práctica ha de ser capaz de:

- Construir gráficos de dispersión de variables cuantitativas e interpretar el suavizado.
- Interpretar el coeficiente de correlación lineal y comprender el significado de la covarianza.
- Utilizar un modelo de regresión lineal simple para describir una relación lineal entre dos variables numéricas.
- Reconocer la importancia del condicionamiento en la predicción.
- Obtener intervalos de confianza para los valores esperados de la variable respuesta.
- Realizar un análisis de residuos para comprobar las hipótesis del modelo.
- Introducir una variable cualitativa en un gráfico de dispersión y en la construcción del modelo de regresión.

En el análisis exploratorio de datos han de ser capaces de desarrollar las tres etapas del análisis. En primer lugar, dados unos datos, conocer qué análisis y gráficos son interesantes para describir su comportamiento; en segundo lugar, realizar dicho análisis y, finalmente, saber interpretar los resultados.

Muchos problemas que aparecen en la práctica y que se abordan mediante técnicas estadísticas tratan de estudiar relaciones entre varias variables. En esta práctica, se considera el análisis de la relación entre dos variables (X, Y) numéricas. ¿Por qué interesa estudiar la relación entre dos variables? Si se descubre una fuerte asociación entre ellas y se puede determinar, entonces es posible utilizar una de ellas para establecer previsiones sobre el comportamiento de la otra. Es decir, el conocimiento del valor de una de las variables permite mejorar la estimación sobre el valor de la otra. Así será interesante estudiar la distribución de Y condicionada por el valor de la componente X , por lo que se presentarán procedimientos adecuados para describir ese comportamiento condicionado.

Para el caso en que X e Y son variables continuas, se dispone de una técnica más avanzada. Los modelos de regresión como una herramienta para investigar y modelizar la relación entre una variable respuesta y una o más variables explicativas numéricas y su utilidad en la estimación de valores de la variable respuesta a partir del valor de las variables explicativas.

L.1. Análisis exploratorio para dos variables numéricas

En el análisis conjunto de dos o más variables es conveniente realizar gráficos que permitan visualizar la posibilidad de una relación de sus comportamientos. Se considera de nuevo el archivo `Altura.Peso.Teleco.mtw`,

que contiene datos del peso, en kilogramos, y la altura, en metros, de los alumnos de Ingeniería de Telecomunicaciones, indicando también su sexo y el año de inicio de los estudios.

Cuando las dos variables son numéricas, por ejemplo, el peso y la altura, interesa conocer el comportamiento conjunto de las dos variables, en particular, si existe alguna relación entre dichas variables (por ejemplo, lineal). El procedimiento **Graph/Scatterplot** representa un diagrama de dispersión del tipo YX , además permite incluir un suavizado **lowess** para explorar la forma de la relación entre ambas variables. La opción **Graph/Marginal Plot** incluye además gráficos sobre el comportamiento individual o marginal de cada una de las variables. La opción **Labels/Data Labels** permite añadir sobre cada dato el valor de la variable elegida. Si hay más de dos variables numéricas, **Graph/Matrix Plot** representa diagramas de dispersión para cada par de variables (“todas por todas”).

Los diagramas de dispersión permiten determinar visualmente si existe una relación entre variables, pero siempre es necesario disponer una medida numérica. En el caso del grado de asociación lineal dos medidas son la covarianza y el coeficiente de correlación, que en Minitab® se calculan con los procedimientos de **Stat/Basic Statistics/Covariance** y **Stat/Basic Statistics/Correlation**. El coeficiente de correlación es de más sencilla interpretación gracias a que se incluye en el rango $[-1,1]$.

El coeficiente de correlación, r , es una medida de la relación *lineal* entre X e Y , obtenida a partir de los datos $(x_1, y_1), \dots, (x_n, y_n)$ correspondientes a las variables X e Y .

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} = \frac{S_{xy}}{S_x S_y}$$

En la interpretación del coeficiente de correlación lineal se distinguen los siguientes casos

- $r \approx 0$: ausencia de relación lineal. No se excluyen otras relaciones.
- $r \approx 1$: fuerte relación lineal positiva, $X \uparrow, Y \uparrow$.
- $r \approx -1$: fuerte relación lineal negativa, $X \uparrow, Y \downarrow$.

L.2. Modelos de regresión lineal

Los modelos de regresión se utilizan para establecer la distribución de Y condicionada por el valor de X , por lo que un primer paso imprescindible es la identificación de cuál es la variable respuesta Y y cuál es la variable explicativa o covariable X . La importancia de la utilización de diagramas de dispersión en la regresión queda perfectamente ilustrada con la colección de datos creados, con este propósito, por F. Anscombe y que se recogen en el fichero ANSCOMBE.mtw. Estos datos ponen de manifiesto la necesidad de examinar los diagramas de dispersión antes de decidir si la estimación del modelo lineal. Una herramienta útil para valorar la forma de la relación entre la variable respuesta y la variable explicativa es el suavizado **lowess**, que se incluye en el diagrama de dispersión, proporciona el aspecto de la esperanza de Y condicionada por el valor de X , lo que permite disponer de una opción sencilla para explorar la existencia o no de una relación de tipo lineal o de otra forma.

La recta de regresión es aquella que “mejor” se aproxima a la nube de puntos (x_i, y_i)

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

siendo los valores \hat{y}_i aquellos que minimizan la distancia cuadrática, es decir,

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Las siguientes estimaciones de β_0 y β_1 proporcionan la mínima distancia:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Por consiguiente la expresión de la recta de regresión resulta ser:

$$\hat{y}_i = b_0 + b_1 x_i = \bar{y} + b_1 (x_i - \bar{x})$$

Cuando el gráfico y el valor del coeficiente de correlación lineal sugiere una relación lineal entre las dos variables numéricas, se continúa con la estimación del modelo lineal, **Stat/Regression/Regression**. **Minitab®** proporciona la estimación de cada coeficiente y el p-valor correspondiente. El valor de S expresa la desviación típica residual que indica la variabilidad de la variable respuesta cuando se extrae la parte sistemática expresada por la recta, el valor de R cuadrado o proporción de variabilidad que explica la recta. El modelo de regresión estimado explica parte de la variabilidad de los datos.

Por ejemplo, la variabilidad del peso de los estudiantes de telecomunicaciones que iniciaron sus estudios en 1993 (en TELECOS.mtw) medida por su desviación típica es 8.89 Kg. El modelo de regresión lineal estimado que relaciona el peso y la altura de un estudiante es:

$$Peso = -67.8 + 76.5Altura$$

El coeficiente 76.5 expresa que 1 cm de estatura tiene un incremento de 0.765 Kg en el peso, en término medio. La variabilidad explicada por el modelo es 60.5 %, es decir, se mejora la precisión de las predicciones cuando se considera además la información adicional dada por la variable Altura. La variabilidad restante se expresa con una desviación típica residual de 5.638.

Stat/Regression/Regression/Options permite obtener estimaciones puntuales y por intervalo del valor respuesta para un valor de la variable explicativa en el rango de valores utilizado en la regresión. Hay que tener especial cuidado y no extrapolar el modelo lineal fuera del rango de estimación. **Stat/Fitted Line Plot** permite representar el diagrama de dispersión junto con la recta de regresión y obtener una idea visual de la bondad del ajuste. Las opciones del gráfico permiten realizar transformaciones sobre los datos, ajustar modelos cuadráticos y cúbicos, representar las bandas de confianza y de predicción y detectar valores atípicos.

En la construcción de modelos de regresión, la etapa de estimación del modelo debe ir seguida de la etapa de verificación de las hipótesis asumidas. En esta etapa, la mayoría de las hipótesis se pueden comprobar mediante el análisis de los residuos (**Stat/Regression/Regression/Graphs**), en particular, se comprueba la normalidad de los residuos y la ausencia de patrones de comportamiento en el gráfico de los residuos frente a los valores ajustados. Los residuos presentan los comportamientos sistemáticos que no son recogidos por la recta de regresión y ayudan a detectar puntos con residuos elevados que expresan individuos mal "ajustados". Un aspecto importante de esta fase de crítica del modelo es la identificación de posibles datos influyentes, es decir, que tienen un efecto excesivo en el valor estimado de los parámetros del modelo. Los residuos son una de las herramientas para identificar este tipo de puntos, otra opción es utilizar el marginal plot. El procedimiento gráfico **Graph/Marginal Plot** permite observar además de la relación entre las variables Peso y Altura, el

comportamiento individual de las variables para detectar, por ejemplo, la existencia de valores atípicos, bien en las variables individuales como en la relación X-Y.

La relación entre dos variables numéricas también puede cambiar según el valor de la variable cualitativa. Por ejemplo, se analiza la relación entre la altura y el peso para cada uno de los sexos. En el procedimiento **Graph/Scatterplot** se elige el tipo de gráfico **With Groups** y en el espacio para **Categorical variables for grouping** se incluye Sexo. El coeficiente de correlación también se puede calcular por separado para cada uno de los grupos definidos por Sexo, para disponer una medida numérica del grado de asociación lineal entre las dos variables numéricas en cada grupo.

L.3. Ejercicios propuestos

Los ficheros **Desperdicios.mtw** y **Osos.mtw** se han obtenido del libro: Mario F. Triola. Estadística. Décima edición, Pearson Educación, México, 2009. ISBN: 978-970-26-1287-2. El fichero **Coches.mtw**, **Iberoamerica.mtw**, **Cerebro.mtw** y **Termica.mtw** se ha obtenido del libro: P. Grima, L.I. Marco, J. Tort-Matirell. Estadística Práctica con Minitab®. Pearson Educación, 2004. ISBN: 9788420543550. Los ficheros **Congelado.mtw** y **Pulso.mtw** son ficheros de muestra del software Minitab®.

1. El archivo **Epract2.mtw** contiene información dada por un taller dedicado a instalar autorradios se desea tener una estimación del tiempo de trabajo. Para ello, se ha recogido el tiempo que costó instalarlas. En las columnas UNIDADES y MINUTOS se recoge el número de autorradios y el tiempo de montaje en minutos, respectivamente. Ajustar un modelo de regresión lineal simple.
2. El archivo **Epract2.mtw** contiene información sobre un proceso químico. Se desea encontrar la relación, caso de existir, entre la temperatura de operación del proceso (dada en la columna $TC(X)$), medida en grados centígrados, y el porcentaje en la concentración de un determinado compuesto (columna C). Comentar los resultados que se obtienen al ajustar un modelo de regresión simple.
3. El archivo **Epract2.mtw** contiene información sobre la evolución a lo largo del tiempo de dos magnitudes de interés. Una de ellas es el porcentaje del presupuesto nacional dedicado a Educación y la otra es el número de asnos en el país, en miles. Se presentan los datos en las columnas M ASNOS y % EDUCACION. Se estudia la relación entre ambas variables y en apariencia una es altamente explicativa para la otra. Criticar el modelo.
4. El archivo **Ordenador.mtw** contiene la información dada por una revista en el año 99 sobre 200 ordenadores divididos en cuatro categorías. Para cada ordenador se indica la relación calidad/precio en su categoría, el precio en pesetas, la velocidad en Mh, el tipo de procesador, Mb de memoria Ram, Mb de disco duro, su categoría y si es ordenador portátil, su peso.
 - a) A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal, elegir dos variables numéricas, razonando cuál es la variable respuesta, y ajustar un modelo de regresión lineal.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿se aprecia un patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - d) ¿Se aprecian diferencias de precio entre los tipos de ordenador?
 - e) Analizar la relación entre el precio y la capacidad del disco según el tipo de ordenador.
 - f) Analizar el precio de un ordenador en relación con el procesador y del tipo de ordenador.
5. El archivo **Pulso.mtw** del Software Minitab® contiene los datos sobre 92 estudiantes de una clase. Para cada estudiante se recoge su altura, peso, sexo, si fuma o no, el nivel de actividad física habitual y su pulso en reposo. Se eligió al azar un conjunto de estudiantes y corrieron 1 minuto, a continuación, todos se volvieron a tomar el pulso.

- a) A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal elige dos variables numéricas, razonando cuál es la variable respuesta, y ajusta un modelo de regresión lineal.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - d) ¿Se aprecian diferencias en el pulso después de correr entre las personas que fuman y las que no fuman?
 - e) Analizar la relación entre el peso y la altura según el sexo del estudiante.
 - f) Analizar el incremento en el pulso en las personas que corrieron en relación con su sexo y si fuman o no.
6. El archivo **Congelado.mtw** del Software Minitab® contiene datos pertenecientes a una empresa de productos congelados. Se desea determinar cuál es la temperatura del horno y el tiempo de cocción de un plato congelado, para que el sabor que se obtenga sea el mejor. La calidad del plato es la puntuación media entre 0 (peor sabor) y 10 (mejor sabor) otorgada por unos jueces catadores. En los experimentos se registró el operario que manejaba el horno.
- a) ¿Existe relación entre la calidad del plato y el tiempo de cocción del mismo?
 - b) Calcular la matriz de correlación de las variables continuas e interpreta los resultados.
 - c) ¿Influye la temperatura del horno en la relación entre la calidad del plato y el tiempo de cocción?
 - d) Estudiar la variable calidad según el tiempo de cocción.
 - e) Analizar la calidad del plato según la temperatura del horno y el operario que lo manejó.
7. El archivo **Posta.mtw** contiene información del servidor de correo electrónico de la Universidad de Zaragoza sobre el número de mensajes recibidos desde el exterior, enviados al exterior y de tráfico interno.
- a) A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal, elegir dos variables numéricas, razonando cuál es la variable respuesta, y ajustar un modelo de regresión lineal.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - d) ¿Se aprecian diferencias en el número de mensajes internos según año?
 - e) Analizar la relación entre los mensajes recibidos y los enviados según el año.
 - f) Realizar para cada año un gráfico de barras con el tráfico total por meses.
8. El archivo **Coches.mtw** contiene información extraída de la revista *Coche Actual* de noviembre de 1994 sobre 247 coches.
- a) A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal, elegir dos variables numéricas, razonando cuál es la variable respuesta, y ajustar un modelo de regresión lineal.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - d) ¿Se aprecian diferencias en el precio según el número de cilindros?
 - e) Analizar la relación entre el peso y el consumo según el número de cilindros.
 - f) Elegir al azar 10 marcas y analiza la velocidad máxima en relación con la marca y el número de cilindros.

9. El archivo **Termica.mtw** contiene información sobre 48 días de funcionamiento de una central térmica. El objetivo del estudio es construir un modelo para explicar el rendimiento de la central a partir de las variables disponibles.
- A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal ajustar un modelo de regresión lineal que ayude a explicar el rendimiento de la central.
 - ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - Proporcionar el valor previsto por el modelo estimado del rendimiento para el valor de la variable explicativa elegida.
 - ¿Se aprecian diferencias en el rendimiento de la central según si el día de la semana es laborable o fin de semana?
 - Analizar la relación entre el rendimiento de la central y la temperatura del agua según el combustible utilizado.
 - Estudiar el rendimiento de la central según el combustible utilizado y el día de la semana (laborable o fin de semana).
10. El archivo **Osos.mtw** contiene los datos de 54 osos silvestres. Para cada oso se recoge su edad, el mes de medición, su sexo y medidas físicas de la cabeza y el cuerpo. El objetivo del estudio es construir un modelo para determinar el peso del oso a partir de otra medida de longitud más fácil de obtener en el bosque.
- A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal, ajustar un modelo de regresión lineal que ayude a determinar el peso del oso sin necesidad de pesarlo.
 - ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - ¿Se puede determinar el peso de un oso a partir de otras mediciones más fáciles de realizar en su ambiente natural? Proporciona el valor previsto por el modelo estimado del peso del oso para el valor de la variable explicativa elegida.
 - ¿Se aprecian diferencias en el peso según el sexo del animal?
 - Analizar la relación entre el peso y la longitud del oso según el sexo del animal.
 - Analizar el peso del oso en relación con su sexo y su edad (entre 0 y 2 años, entre 2 y 4 años y mayor de 4 años).
11. El archivo **Países.mtw** contiene información sobre 30 países extraída de El País Semanal en 1994 sobre diferentes aspectos.
- A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal elige dos variables numéricas, razonando cuál es la variable respuesta, y ajusta un modelo de regresión lineal.
 - ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - ¿Se aprecian diferencias en la esperanza de vida de las mujeres (*EsperMujer*) según las principales causas de muerte (*Causas*)?
 - Analizar la relación entre el consumo diario de calorías (% de las necesidades diarias)(*Calorias*) y la renta per cápita (*Renta*) según las principales causas de mortalidad.
 - Definir la variable *Hijos* con valor 1 si el número de hijos por mujer es menor o igual a 2, 2 si está entre 2 y 3 hijos y 3, más de tres hijos. Analizar la mortalidad infantil según la variable *Hijos* y la religión mayoritaria en el país.

12. El archivo **Desperdicios.mtw** recoge los pesos en libras de diferentes categorías de desperdicios desechados por una muestra de 62 hogares (metales, papel, plástico, vidrio, productos alimenticios, desperdicios del jardín, textiles y otros no incluidos en las categorías anteriores).
 - a) ¿Existe alguna relación entre el peso en desperdicios orgánicos y en papel?
 - b) Calcular la matriz de correlación de las variables continuas e interpretar los resultados.
 - c) ¿Se aprecian diferencias en los desperdicios de papel según el tamaño de la familia?
 - d) Analizar la relación entre el peso en desperdicios orgánicos y en vidrio según el tamaño de la familia.
 - e) Definir una variable con tres grupos según el peso de desperdicios orgánicos que producen y otra variable que diferencie entre familias de tamaño bajo, medio y alto. Analizar el consumo de papel según los valores de las variables anteriores.

13. El archivo **Iberoamerica.mtw** contiene datos de la página web del Instituto Nacional de Estadística sobre Indicadores sociales de países iberoamericanos en 1998.
 - a) A partir de los resultados de los diagramas de dispersión y el cálculo de los coeficientes de correlación lineal, elegir dos variables numéricas, razonando cuál es la variable respuesta, y ajustar un modelo de regresión lineal.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.
 - d) Definir una variable con tres grupos según si el porcentaje de menores de 15 años es menor del 25 %, está entre el 25 % y el 35 % y es mayor del 35 %, ¿se aprecian diferencias en la población en estos grupos?
 - e) Analizar la relación entre la esperanza de vida al nacer y la tasa de mortalidad infantil según los grupos definidos por la variable porcentaje de menores de 15 años.
 - f) Definir tres grupos en los países según el % de PIB aportado por la agricultura y en otros tres grupos según el % de PIB aportado por la industria. Analiza la variable población según los valores de los grupos definidos.

14. El archivo **Cerebro.mtw** contiene los datos sobre el nombre, el peso del cerebro en gramos y el peso del cuerpo en Kg. de 62 especies de mamíferos.
 - a) Realizar un gráfico de dispersión e interpreta el resultado. Considerar alguna transformación sobre los datos y ajustar el modelo lineal que explique el peso del cerebro de los mamíferos en función del peso de su cuerpo.
 - b) ¿Tienen los residuos un comportamiento normal?, ¿existe algún patrón de comportamiento?
 - c) Proporcionar el valor previsto por el modelo estimado para el valor de la variable explicativa elegida.