Práctica 7

Modelos de regresión

7.1. Introducción

El término Análisis de Regresión describe una colección de técnicas estadísticas que sirven como base para realizar inferencias sobre la relación existente entre dos o más variables en estudio cuando dicha relación se modela mediante una función.

Por ejemplo: la lluvia caída por metro cuadrado, (x), y la cosecha obtenida, (y). Está claro que la relación entre estas variables no es perfecta (en el sentido de que no existe una función f tal que $y = f(x)$), aunque no se puede negar que existe cierta relación entre ambas con lo que se podría estudiar la posibilidad de encontrar una función matemática f tal que $y \approx f(x)$.

En esta práctica veremos cómo utilizar R para el estudio de modelos de regresión. Comenzaremos por los modelos de regresión lineal, que son los más sencillos.

7.2. Modelos de regresión lineal

7.2.1. Regresión lineal simple

El modelo de regresión lineal simple resulta cuando sólo hay una variable independiente x (también llamada variable regresora) y la función f es lineal. Con la notación anterior podemos escribir la relación entre las variables como $y \approx \beta_0 + \beta_1 x$. A la hora de plantear el modelo la condición de similaridad se traduce en la adición de un error en forma de variable aleatoria. De este modo el modelo de regresión lineal simple es $y = \beta_0 + \beta_1 x + \varepsilon$, donde

- y es la variable respuesta medida o variable dependiente.
- x es la variable regresora o variable independiente.
- β_0 y β_1 son los parámetros de la regresión.
- ε es el error del modelo.

En situaciones prácticas dispondremos de un conjunto de n pares de observaciones experimentales $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Nuestro objetivo será encontrar estimaciones de β_0 y β_1, que denotaremos por b_0 y b_1, que se obtendrán aplicando el método de mínimos cuadrados a los pares de valores observados. De esta forma, para cada par (x_i, y_i), $i = 1, \ldots, n$ tendremos la relación: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, donde ε_i es el error asociado a la observación i-ésima.

7.2.2. Regresión lineal múltiple

El **modelo de regresión lineal múltiple** es una generalización del modelo lineal simple que resulta cuando se dispone de más de una variable regresora. Denotaremos a dichas variables por $x_1, \ldots, x_k$. Entonces, el modelo podemos escribirlo en la forma

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

Al igual que en el caso del modelo de regresión lineal simple, la estimación de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ se realizará también mediante el método de los mínimos cuadrados.

Como ya hemos dicho, el modelo de regresión lineal múltiple es una generalización del modelo de regresión lineal simple. Por este motivo, en lo que sigue, desarrollaremos los distintos conceptos para el modelo de regresión lineal múltiple.

Hipótesis sobre el modelo

En el caso de la regresion lineal simple es conveniente, en primer lugar, representar gráficamente la nube de puntos que determinan los pares de valores observados para asegurarse de que un modelo lineal simple puede ser adecuado para describir la relación entre las variables. El estudio de la nube de puntos también puede ser útil a la hora de encontrar en la muestra pares de valores extraños que puedan alterar los resultados. Asimismo, nótese que aunque dicha representación es muy sencilla en el caso de la regresión lineal simple, resulta más complicado cuando tenemos dos variables regresoras e imposible cuando tenemos mas de dos.

Por otra parte, de cara a estimar los parámetros no es necesaria la hipótesis de normalidad sobre los errores, pero si se desea realizar contrastes de hipótesis y construir intervalos de confianza, entonces es necesario que se cumplan las siguientes condiciones:

- Errores distribuidos normalmente con media cero.
- Errores con varianza constante.
- Errores incorrelados entre sí[1].

[1] Al seguir los errores distribuciones normales independientes, decir que son incorrelados es equivalente a decir que son independientes.

Además de estimar los parámetros del modelo, en un análisis de regresión, puede ser necesario responder a las siguientes preguntas:

- ¿Realmente existe relación lineal entre las variables independientes y la variable dependiente?
- ¿Existe un ajuste adecuado de los datos y el modelo?

La primera cuestión se resuelve mediante el contraste de hipótesis

$$\begin{cases} H_0: & \beta_1 = \ldots = \beta_k = 0 \\ H_1: & \beta_i \neq 0 \text{ para algún } i \end{cases}$$

Al realizar dicho contraste, R nos mostrará el correspondiente *p-valo*r. Rechazar la hipótesis nula implica que, efectivamente, existe influencia lineal de las variables regresoras sobre la variable dependiente.

Para responder a la segunda cuestión usaremos el coeficiente de determinación R^2 que nos dará una medida de la bondad del ajuste. El coeficiente R^2 toma valores en el intervalo $[0, 1]$ e indica la proporción de la variabilidad de la variable respuesta, que es explicada por las variables regresoras, a través del modelo. Por tanto, cuanto mayor sea el valor de R^2 mejor será el ajuste. Para comparar la bondad de ajuste de dos modelos de regresión con distinto número de variables es preferible utilizar el denominado coeficiente de determinación corregido (Adjusted R-squared), en lugar de usar directamente el coeficiente de determinación, para hacer la comparación.

Para realizar un análisis de regresión con R-Commander debemos seleccionar el menú `Estadísticos`→`Ajuste de modelos`→`Regresión lineal`. A continuación indicamos el nombre que le daremos al modelo, la `variable explicada` (esto es, la variable dependiente) y las `variables explicativas` (esto es, las variables independientes). La ventana `expresión de selección` permite indicar las condiciones que deben cumplir los casos utilizados para realizar el análisis de regresión.

La salida proporcionada por R-commander nos muestra:

- Un resumen de los residuos (diferencias entre valores observados y valores estimados de la variable dependiente).
- Una tabla en la que se muestra, entre otra información, la estimación de cada coeficiente, valores de los t-estadísticos utilizados para contrastar si el correspondiente coeficiente es 0 o hay evidencia para rechazar esta hipótesis, así como el p-valor asociado a cada uno de estos estadísticos.
- La raíz cuadrada de los cuadrados medios del error (Residual standard error).
- Los valores de los coeficientes R^2 y R^2-corregido (Adjusted R-squared).

- El valor del F-estadístico y el correspondiente p-valor obtenidos al resolver el contraste:

$$\begin{cases} H_0: & \beta_1 = \ldots = \beta_k = 0 \\ H_1: & \beta_i \neq 0 \text{ para algún } i \end{cases}$$

Ejemplo 1: Abrir el archivo *Rcars*. Se piensa que la variable *acel* (aceleración) depende linealmente de las variables *cilindr, cv, motor y peso*. Tras realizar un análisis de regresión lineal:

a) Indicar e interpretar el coeficiente de determinación.

b) Se desea estudiar si existe relación lineal entre las variables regresoras y la variable dependiente. ¿Qué conclusión se obtiene para un nivel de significación $\alpha = 0.05$?

c) ¿Puede asumirse como nulo alguno de los coeficientes del modelo, para una significación de 0.05?

d) Expresar el modelo tras obtener una estimación de los valores de los coeficientes. Obtener intervalos de confianza al 95 % para los coeficientes del modelo.

El modelo de regresión que tenemos que estudiar es:

$$acel = \beta_0 + \beta_1 \cdot cilindr + \beta_2 \cdot cv + \beta_3 \cdot peso + \beta_4 \cdot motor + \varepsilon$$

Seleccionamos el modelo `Estadísticos`→`Ajuste de modelos`→`Regresión lineal`. Le damos como nombre al modelo *Modelo1*, por ejemplo. En la ventana `Variable explicada` seleccionamos la variable *acel*. En la ventana `Variables explicativas` seleccionamos las variables *cilindr, cv, motor y peso*. Finalmente, pulsamos en el botón `Aceptar`.

A partir de la salida podemos responder a las distintas cuestiones planteadas:

a) En la salida observamos que el valor del coeficiente de determinación (coeficiente R^2) es igual a 0.6382. Esto significa que el modelo explica el 63.82 % de la variabilidad de la variable *acel*. Esto es, el 63.82 % de la variabilidad de la aceleración se explica mediante un modelo lineal a partir del número de cilindros, la potencia en caballos del motor del vehículo, la cilindrada en centímetros cúbicos del motor y el peso del vehículo.

b) Para responder a la segunda pregunta tenemos que resolver el contraste

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: & \beta_i \neq 0 \text{ para algún } i \end{cases}$$

Los datos que, sobre este contraste, nos proporciona la salida de R se encuentran en la última línea de la misma. Así, observamos que el p-valor obtenido es inferior

a 2.2e-16, esto es, inferior a $2.2 \cdot 10^{-16}$. Por tanto, puesto que el p-valor es inferior a 0.05, que es la significación del contraste, tenemos evidencia significativa para rechazar la hipótesis nula y afirmar que, al menos, uno de los coeficientes de la regresión es distinto de cero. Esto significa que, efectivamente, las variables regresoras y la variable dependiente se relacionan linealmente.

c) Para responder a esta pregunta tenemos que resolver los contrastes

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

para cada $i = 0, 1, 2, 3, 4$.

En cada fila de la tabla `Coefficients` encontramos los datos necesarios para realizar inferencias sobre los distintos parámetros del modelo de regresión. La primera fila (`Intercept`) presenta los datos sobre β_0. El resto de filas corresponden a los parámetros que acompañan a las variables independientes. Así la segunda fila (`cilindr`) tiene la información sobre el parámetro β_1, la tercera fila (`cv`) sobre el parámetro β_2, etc. Es necesario ser cuidadoso en este punto porque el orden de los parámetros en el modelo puede no corresponder con el orden en la tabla, dado que en ésta las variables que denotan las distintas filas se ordenan por orden alfabético.

El p-valor necesario para resolver los contrastes se encuentra en la última columna de la tabla `Coefficients`. Puesto que el nivel de significación que debemos considerar es 0.05, se observa que debemos rechazar la hipótesis nula y por tanto tenemos evidencia significativa para afirmar que son distintos de cero los coeficientes β_0 y los que multiplican a las variables *cv* y *peso*, esto es, β_2 y β_3. En todos estos casos el p-valor obtenido es inferior a $2 \cdot 10^{-16}$ y, por tanto, a la significación del contraste. Por el contrario, no tenemos evidencia para afirmar que sean distintos de cero los coeficientes β_1 (p-valor 0.116) y β_4 (p-valor 0.146).

d) Las estimaciones de los coeficientes las podemos encontrar en la columna `Estimate` de la tabla `Coefficients`. Las estimaciones de β_0, β_1, β_2, β_3 y β_4 son, respectivamente (redondeamos al cuarto decimal): 17.9285, -0.2568, -0.0869, -0.0003 y 0.0092. De este modo, una vez estimados los coeficientes, el modelo queda:

$$y = 17.9285 - 0.2568 \cdot cilindr - 0.0869 \cdot cv + 0.0092 \cdot peso - 0.0003 \cdot motor$$

Para obtener los intervalos de confianza, en el menú `Modelos`→`Seleccionar el modelo activo`, seleccionaremos el modelo con el que vamos a trabajar (en nuestro caso *Modelo1*. Si sólo tenemos un modelo en memoria, aparecerá como activo por defecto y no se desplegará el menú de selección. A continuación, en el menú `Modelos`→`Intervalos de confianza`, indicaremos el nivel de confianza (dividido por 100) para el que deseamos calcular los intervalos, en nuestro caso, 0.95.

Obtenemos una salida formada por tres columnas. De izquierda a derecha, la primera columna contiene el valor estimado de los parámetros; la segunda contiene el extremo izquierdo del intervalo de confianza, para cada uno de los parámetros del modelo, y la tercera, el extremo derecho. Así los intervalos de confianza para los distintos parámetros son:

Parámetro	Intervalo de confianza al 95 %
β_0	$(16.7169, 19.1401)$
β_1	$(-0.5775, 0.06401)$
β_2	$(-0.0971, -0.0768)$
β_3	$(0.0075, 0.0109)$
β_4	$(-0.0008, 0.0001)$

7.2.3. Regresión curvilínea

En algunos casos encontramos modelos que, aunque no corresponden de manera exacta al esquema descrito para los modelos lineales, pueden transformarse en modelos lineales aplicando transformaciones sobre las variables.

Un ejemplo de este tipo de ajustes son los ajustes polinómicos, en los que el modelo es de la forma $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$.

Otros ejemplos son los modelos: $y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$ ó $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + \varepsilon$.

Nótese que en todos los casos la variable dependiente se escribe como un término independiente sumado a una combinación lineal de las variables regresoras o de funciones de cada una de las variables regresoras.

Ejemplo 2: Ajustar una parábola a los datos

x	-3	-2	-1	0	1	2	3
y	1	0	0	-1	-1	0	0

y responder a las siguientes cuestiones:

a) Dibujar el diagrama de dispersión que se obtiene al considerar x como variable independiente e y como variable dependiente. En vista del gráfico obtenido, ¿tiene sentido ajustar una parábola a los datos?

b) Determinar e interpretar el coeficiente de determinación, R^2, que se obtiene para el ajuste parabólico.

c) Se desea estudiar si, efectivamente, existe relación parabólica entre las variables regresoras y la variable dependiente. ¿Qué conclusión se obtiene para un nivel de significación de 0.05?

d) ¿Puede asumirse nulo alguno de los coeficientes del modelo para una significación de 0.05?

e) Expresar el modelo tras obtener una estimación de los valores de los coeficientes.

f) Realizar también un ajuste de regresión lineal simple.

g) Teniendo en cuenta los valores obtenidos de los coeficientes de determinación corregidos, ¿qué modelo elegirías?

Comenzamos por introducir los datos correspondientes a las variables x e y en el editor de datos. A continuación, construimos los datos corresondientes a x^2. Para ello[2] seleccionamos el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Calcular una nueva variable`. En la ventana emergente, como nombre de la nueva variable escribimos, por ejemplo, $x2$ y en la expresión a calcular escribimos x^2. Finalmente, pulsamos en el botón `Aceptar`.

a) Seleccionamos el menú `Gráficas`→`Diagrama de dispersión`. En la ventana emergente seleccionamos la variable x en el espacio `variable x` y la variable y en el espacio `variable y`; a continuación pulsamos en el botón `Aceptar`. Observando el gráfico podemos ver que la nube de puntos se asemeja a una parábola. En consecuencia, tiene sentido ajustar una parábola a los datos.

b) Tenemos que estudiar el modelo $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Para ello realizamos un análisis de regresión lineal entre las variables y, x y $x2$. La variable y es la `Variable explicada` y las variables x y $x2$ serán las `Variables explicativas`. Indicamos como nombre del modelo, *Modelo2*, por ejemplo.

   En la salida observamos que el valor del coeficiente R^2 es 0.8, lo que significa que el modelo explica el 80 % de la variabilidad de la variable y.

c) Para responder a la pregunta tenemos que resolver el contraste

$$\begin{cases} H_0: & \beta_1 = \beta_2 = 0 \\ H_1: & \beta_i \neq 0 \text{ para algún } i \end{cases}$$

   En la salida del análisis de regresión observamos que el p-valor correspondiente a este contraste es igual a 0.04. Puesto que es menor que la significación del contraste tenemos evidencia para rechazar la hipótesis nula y podemos afirmar que existe relación parabólica entre la variable y y la variable x.

d) Tenemos que resolver los contrastes sobre los parámetros:

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

   para cada $i = 0, 1, 2$, con nivel de significación $\alpha = 0.05$.

[2] Al tener el ejemplo los datos de, únicamente, 7 casos podríamos calcular e introducir manualmente los valores de x^2 pero no será esta la situación habitual.

En la salida observamos que el p-valor del contraste sobre β_0 es 0.0307; el p-valor del contraste sobre β_1 es 0.1161 y el p-valor del contraste sobre β_2 es 0.0257. Por tanto, teniendo en cuenta que el nivel de significación indicado es $\alpha = 0.05$, el único coeficiente para el que no tenemos evidencia para afirmar que sea distinto de 0, es el coeficiente β_1.

e) Teniendo en cuenta la estimación de los coeficientes, el modelo ajustado quedaría:

$$y = -071429 - 0.14286x + 0.14286x^2$$

f) Realizamos ahora un ajuste de regresión lineal entre las variables y y x. Obtenemos que el modelo ajustado es $y = -0.1429 - 0.1429x$.

e) En el ajuste parabólico obtuvimos un valor del coeficiente R^2-corregido (`Adjusted R-squared`) igual a 0.7. En el caso del ajuste de regresión lineal simple, hemos obtenido un valor del coeficiente R^2-corregido igual a 0.04. Por tanto, si tenemos que elegir entre los dos modelos, elegiríamos el modelo parabólico por ser el que presenta un mayor valor del coeficiente R^2-corregido.

7.3. Regresión no lineal

En numerosas ocasiones el conocimiento previo de la situación experimental sugiere el uso de modelos que no son lineales. Los siguientes son ejemplos de modelos no lineales:

$$y = \alpha e^{\beta x} + \varepsilon \qquad\qquad y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \varepsilon$$

A veces es posible obtener una aproximación del problema "linealizando" el modelo mediante un cambio de variable aunque, eso sí, a costa de perder información en lo que respecta a la estructura del mismo. En cualquier caso, la solución obtenida mediante este procedimiento sería simplemente una aproximación a la solución óptima que se obtendría mediante el método de mínimos cuadrados.

Ejemplo 3: Se están estudiando dos magnitudes X e Y que se ajustan a un modelo del tipo $y = \alpha x^{\beta} + \varepsilon$.

Estimar los parámetros del modelo a partir de la siguiente muestra utilizando para ello el método de mínimos cuadrados:

x	2.718	7.389	20.086	54.598	148.41
y	7.389	20.086	403.429	2980.96	8103.08

El procedimiento utilizado por R para aplicar el método de mínimos cuadrados a estos modelos requiere conocer unos valores iniciales de los parámetros. Comenzaremos pues por determinar dichos valores.

En el caso del modelo propuesto podemos linealizarlo aplicando logaritmos a ambos lados de la igualdad. Utilizaremos este procedimiento para obtener un valor inicial de los parámetros. Al aplicar logaritmo neperiano a ambos miembros de la igualdad resulta un modelo lineal de la forma $lny = ln\alpha + \beta lnx + \varepsilon^*$ donde ε^* es una variable aleatoria de distribución desconocida.

Nótese que, aunque el error de este nuevo modelo siga una distribución desconocida, el método de mínimos cuadrados no hace ningún tipo de hipótesis sobre el error del modelo por lo que podemos aplicarlo para obtener un valor aproximado de los parámetros.

Comenzamos entonces por crear un nuevo conjunto de datos, de nombre *nolin*, con los valores de las variables x e y. A continuación, usando el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Calcular una nueva variable`, construimos las variables $ln\,x$ y $ln\,y$, como el logaritmo neperiano de las variables x e y, respectivamente. El logaritmo neperiano de una variable se calcula mediante la función $log()$.

El siguiente paso consiste en realizar un ajuste de regresión lineal entre las variables *lnx* (variable explicativa) y la variable *lny* (variable explicada). Obtenemos el modelo ajustado $lny = -0.09986 + 1.89997 \cdot lnx$. Identificando coeficientes, resulta que $ln\alpha = -0.09986$ y que $\beta = 1.89997$, por lo que los valores estimados de los parámetros son $\hat{\alpha} = e^{-0.09986} = 0.9049641$ y $\hat{\beta} = 1.89997$.

Ahora vamos a realizar un ajuste no lineal utilizando como valores iniciales de los parámetros las estimaciones obtenidas anteriormente. Para ello usaremos la instrucción *nls(ecuación,data=nombre del conjunto de datos,start=list(valores iniciales de los parámetros))*. Nótese que las variables utilizadas en la ecuación deben formar parte del conjunto de datos. En nuestro caso, puesto que el nombre que hemos dado al conjunto de datos es *nolin*, tenemos que utilizar el comando:

*nls(y∼alfa*x^beta,data=nolin,start=list(alfa=0.9049641,beta=1.89997))*

Introducimos esta instrucción en la ventana de comandos y, tras pulsar el botón `Ejecutar`, obtenemos como estimación de α y β los valores 27.536 y 1.138, respectivamente, En consecuencia el modelo ajustado sería $y = 27.536x^{1.138}$.