14

Modelos de regresión: lineal simple y regresión logística

Irene Moral Peláez

14.1. Introducción

Cuando se quiere evaluar la relación entre una variable que suscita especial interés (variable dependiente que suele denominarse Y) respecto a un conjunto de variables (variables independientes, que se denominan X_1, X_2, …, X_n) las pruebas de contraste de hipótesis mostradas hasta ahora no nos aportan suficiente información sobre la relación en conjunto de todas ellas, dado que los contrastes de hipótesis que conocemos hasta ahora se basan en probar relaciones bivariantes (2 variables), en las que no se tiene en cuenta la posibilidad de que haya otras variables de interés y en las que el sentido de la relación es bidireccional. Es entonces cuando resulta adecuado y conveniente la aplicación de los modelos de regresión. Los modelos de regresión permiten evaluar la relación entre una variable (dependiente) respecto a otras variables en conjunto (independientes). Los modelos re regresión se expresan de la siguiente forma: $Y = f(x_1, x_2, \ldots) + \varepsilon$.

El objetivo principal de construir un modelo de regresión puede ser, por ejemplo, evaluar cómo afecta el cambio en unas características determinadas

(variables independientes) sobre otra característica en concreto (variable dependiente), denominado *modelo con fines explicativos;* o también nuestro objetivo podría ser intentar estimar o aproximar el valor de una característica (variable dependiente) en función de los valores que pueden tomar en conjunto otra serie de características (variables independientes), denominado entonces *modelo con fines predictivos*.

Existen varias opciones para estimar un modelo de regresión, de entre los que destacan por su facilidad de aplicación e interpretación, el modelo de regresión lineal y el modelo de regresión logística. Teniendo en cuenta el tipo de variable que deseemos estimar (variable dependiente o respuesta) aplicaremos un modelo de regresión u otro. Simplificando, cuando la variable dependiente es una variable continua, el modelo de regresión más frecuentemente utilizado es la regresión lineal, mientras que cuando la variable de interés es dicotómica (es decir, toma dos valores como sí/no, hombre/mujer) se utiliza la regresión logística. Otro tipo de modelos de regresión utilizados, aunque no tan frecuentemente, son la regresión no lineal o la regresión ordinal que estiman una serie de modelos matemáticos que pueden ajustarse mejor que un modelo lineal.

14.2. Condiciones de aplicabilidad

La regresión logística y los modelos de regresión lineal no pueden ser aplicados sobre cualquier tipo de variable. Por ejemplo, la regresión lineal no es aplicable cuando la variable de interés es categórica, dado que al estimar el modelo de regresión no se respeta la restricción de que los valores de la variable dependiente oscilan entre una serie de valores que son los permitidos o reales, siendo el resto de valores imposibles. Es por eso que resulta más conveniente utilizar en ese caso el modelo de regresión logística. Sin embargo, ambos modelos de regresión se construyen aplicando modelos matemáticos similares. Al aplicar un modelo de regresión logística, en lugar de construir un modelo de regresión para estimar los valores reales de la variable de interés, se construye una función basada en el cálculo de la probabilidad de que la variable de interés adopte el valor del evento previamente definido, de la manera siguiente:

$$Y = \ln (p / (1-p))$$

De forma que la nueva variable dependiente construida que vamos a estimar sí puede tomar cualquier valor (no está restringida a un rango de valores) y podemos recurrir a los métodos de estimación de los modelos de regresión tradicionales para construir el modelo de regresión logística.

Tras realizar una serie de transformaciones matemáticas se pude deducir que:

$$p = \frac{1}{1 + e^{-\text{modelo de regresión}}}$$

En definitiva, en el modelo de regresión logística estimaremos un modelo de regresión que en lugar de realizar estimaciones para la variable dependiente real, las realizará sobre la función de probabilidad asociada a ella, pudiendo entonces aplicar los métodos de estimación aplicables al modelo de regresión lineal, diferenciándose entonces ambos modelos únicamente en la interpretación de resultados.

Para entender en qué consiste un modelo de regresión así como para interpretar correctamente los resultados debemos relacionar dos conceptos: el coeficiente de correlación y el análisis de la varianza. Se puede demostrar que existe una relación entre el coeficiente de correlación (r) y el análisis de la varianza de la regresión, de tal forma que el cuadrado de r, llamado coeficiente de determinación, multiplicado por 100 se interpreta como el porcentaje de la varianza de la variable dependiente que queda explicada por el modelo de regresión.

14.3. Variables a introducir en un modelo de regresión

Ambos modelos de regresión permiten que las variables utilizadas para poder estimar el modelo (variables independientes) puedan ser de cualquier tipo, no existen restricciones sobre ellas. Únicamente se deben tener en cuenta una serie de consideraciones.

14.3.1. Tipos de variables a introducir en el modelo

Cuando la variable que se incluya en el modelo sea una variable continua, se introducirá la variable real o bien alguna transformación de ella (logaritmo, cuadrado, etc.), cuando sea necesario. Sin embargo cuando la variable que se quiera introducir en el modelo sea una variable categórica de más de dos categorías, se necesitará recurrir a una serie de transformaciones para que los resultados obtenidos sobre la variable en cuestión sean correctos e interpretables. En dicho caso, no se podrá introducir la variable original en el modelo, sino que si la variable tiene n categorías deberán expresarse cada una de estas categorías mediante n-1 variables *dummy*. Una variable *dummy* es una variable construida artificialmente y que únicamente puede tomar los valores 0 o 1. Es posible expresar la misma información contenida en una variable de n categorías mediante la combinación de n-1 variables *dummy*. Adicionalmente estas variables permiten realizar todas las comparaciones necesarias respecto a las n categorías de la variable original en el modelo de regresión.

Por ejemplo: Partimos con la variable tabaquismo que tiene 3 categorías y que introduciremos en el modelo mediante las variables *dummy* tabac1 y tabac2. Tabac1 nos mostrará los resultados de comparar la categoría de fumador versus las categorías de no fumador y exfumador, mientras que tabac2 nos mostrará los resultados de comparar las categorías de fumador y exfumador versus no fumador. Es condición indispensable que para interpretar los resultados obtenidos al estimar el modelo de regresión, todas las variables *dummy* que representan las categorías de una variable original han de interpretarse en conjunto, no pudiendo seleccionar únicamente las que hayan resultado estadísticamente significativas. La codificación de las variables *dummy* tabac1 y tabac2 se muestra en la Tabla 19:

Tabaquismo	Tabac1	Tabac2
Fumador	1	0
No fumador	0	1
Exfumador	0	0

Tabla 19. Ejemplo de variables *dummy* creadas para una variable con 3 categorías.

La variable Tabac1 podría interpretarse como una variable que indica si es o no fumador activo y Tabac2 como si es no fumador y no ha fumado alguna vez. Los paquetes estadísticos realizan la transformación de la variables categóricas en las variables *dummy* necesarias automáticamente y no es necesario realizar todo el proceso manualmente, únicamente debe identificarse al programa cuál es el nombre de las variables que requieren este tipo de transformación.

Es importante conocer este requisito al construir un modelo de regresión, principalmente al interpretar los resultados obtenidos, dado que en este tipo de variables, obtendremos un resultado global y otro para cada una de las variables *dummy*, debiendo conocerse qué comparación se realiza a través de cada una de las variables *dummy* que intervienen en el modelo para poder explicar los resultados correctamente.

14.3.2. Selección de las variables a introducir en el modelo

La elección de las variables que deben introducirse en el modelo no debería suponer un problema, dado que al estar el objetivo del estudio concretamente definido, automáticamente las variables que son de mayor interés también quedan identificadas. Sin embargo, no siempre queda todo tan evidente y se plantea encontrar "relaciones" posibles para evaluar una única respuesta claramente

identificada. Puede ocurrir que aunque el objetivo de nuestro estudio esté bien definido, no dispongamos de información previa o de indicio alguno que nos pueda indicar qué aspectos son los que suscitan mayor interés. Si a este hecho se le añade, el irresistible impulso de registrar más información de la que realmente necesitamos por si posteriormente pudiéramos recurrir a ella (exceso de información), construir un modelo de regresión resulta un proceso laborioso y complicado. Por lo tanto, recomendamos desde el inicio elegir con esmero las variables indispensables para su análisis estadístico.

Resulta imprescindible como primer contacto con el modelo analizar la relación bivariante entre la característica de nuestro interés y el resto de variables registradas durante el estudio una a una. Como criterios para seleccionar aquellas variables a introducir en el modelo de regresión podríamos aplicar los siguientes:

- en primera instancia y entre otros criterios igualmente válidos, introducir en el modelo aquellas variables que resultaron estadísticamente significativas en las comparaciones bivariantes realizadas previamente.
- en un segundo plano debería considerarse la conveniencia de incluir en el modelo adicionalmente aquellas variables que consideremos especialmente importantes o influyentes, como por ejemplo la edad o el género, si sospechamos que a pesar de no haber resultado estadísticamente significativas, podrían modificar o intervenir en nuestros resultados,
- otra serie de variables de las que hayamos tenido conocimiento de su influencia a través de estudios previos.

14.4. Concepto de interacción

Otro concepto importante que debe ser tenido en cuenta al construir un modelo de regresión es que pueden introducirse términos independientes únicos (una sola variable, por ejemplo efecto del tabaco) y además las interacciones entre variables de cualquier orden (efecto del tabaco según género), si consideramos que pueden ser de interés o afectar a los resultados. Al introducir los términos de interacción en un modelo de regresión es importante para la correcta estimación del modelo respetar un orden jerárquico, es decir siempre que se introduzca un término de interacción de orden superior ($x \cdot y \cdot z$), deben introducirse en el modelo los términos de interacción de orden inferior ($x \cdot y$, $x \cdot z$, $y \cdot z$) y por supuesto los términos independientes de las variables que participan en la interacción (x, y, z).

Por ejemplo, supongamos que deseamos construir un modelo de regresión para estimar la prevalencia de hipertensos en una muestra y decidimos que es

imprescindible evaluar si la interacción de las variables tabaco, género y edad es significativa o no al estimar dicha prevalencia, por lo que introduciremos el término de interacción tabaco * género * edad. Automáticamente deberían introducirse igualmente en el modelo los términos de interacción de orden inferiores, es decir, tabaco*género, tabaco*edad y género*edad, así como los términos independientes tabaco, género y edad para poder estimar el modelo correctamente. Si se introducen en un modelo de regresión términos de interacción y resultan estadísticamente significativos, no se podrán eliminar del modelo los términos de interacción de orden inferiores ni los términos independientes de las variables que participan en la interacción para simplificarlo, deben mantenerse, aunque no resulten estadísticamente significativos.

14.5. Construcción del modelo de regresión

Para construir un modelo de regresión, nos centraremos en el tipo de variables que deseamos introducir (categóricas o continuas) y posteriormente, veremos los métodos que los paquetes estadísticos nos ofrecen actualmente para obtener el modelo de regresión más fiable.

14.5.1. Selección de las variables del modelo

Existen varios métodos para construir el modelo de regresión, es decir, para seleccionar de entre todas las variables que introducimos en el modelo, cuáles son las que necesitamos para explicarlo. El modelo de regresión se puede construir utilizando las siguientes técnicas:

- Técnica de pasos hacia adelante *(Forward))*: consiste en ir introduciendo las variables en el modelo únicamente si cumplen una serie de condiciones hasta que no se pueda introducir ninguna más, hasta que ninguna cumpla la condición impuesta;
- Técnica de pasos hacia atrás *(Backward)*: se introducen en el modelo todas las variables y se van suprimiendo si cumplen una serie de condiciones definidas a priori hasta que no se pueden eliminar más, es decir ninguna variable cumpla la condición impuesta;
- Técnica por pasos *(Stepwise)*: combina los dos métodos anteriores, adelante y atrás introduciendo o eliminando variables del modelo si cumplen una serie de condiciones definidas a priori hasta que ninguna variable satisfaga ninguna de las condiciones expuestas de entrada o salida del modelo
- Técnica de introducir todas las variables obligatoriamente *(Enter):* Esta última técnica de selección de variables para construir el modelo de re-

gresión, produce que el proceso de selección de las variables sea manual, partiendo de un modelo inicial, en el que se obliga a que entren todas las variables seleccionadas, se va evaluando qué variable es la que menos participa en él y se elimina, volviendo a construir un nuevo modelo de regresión aplicando la misma técnica, pero excluyendo la variable seleccionada y aplicando el mismo proceso de selección. Este proceso se repite reiteradamente hasta que se considere que el modelo obtenido es el que mejor se ajusta a las condiciones impuestas y que no se puede eliminar ninguna variable más de las que los componen.

14.5.2. Métodos de construcción del modelo de regresión

Para evaluar la adecuación de los modelos construidos, es conveniente comenzar a evaluar el modelo saturado, es decir el modelo que contiene todas las variables de interés que queramos evaluar y todas las interacciones posibles. Progresivamente se van eliminando del modelo aquellos términos no significativos, respetando el modelo jerárquico y comenzando por los términos de interacción superiores. Como hemos dicho anteriormente, si un término de interacción es significativo, no podrán eliminarse del modelo los términos de interacción de grado inferior, ni los términos independientes de las variables que participan en la interacción. Las variables introducidas en el modelo se van eliminando progresivamente a cada nuevo modelo que se construye en base a los resultados obtenidos en el modelo anterior, y se van evaluando los nuevos modelos, de la manera que se explicará más adelante. Es importante observar que los coeficientes de las variables que permanezcan en el modelo no varían de forma exagerada tras la eliminación de alguno de los términos del modelo, dado que si así sucediera, podría tratarse de un factor de confusión y por tanto debería mantenerse la variable en cuestión en el modelo, para permitir el ajuste del resto de variables y no obtener resultados artificiales.

14.6. Obtención y validación del modelo más adecuado

Los modelos de regresión pueden ser validados en otro conjunto de datos de similares características, extraídos de la misma población, por ejemplo, con el fin de evaluar su fiabilidad. Otra posibilidad, cuando se trabaja con muestras grandes, es dividir aleatoriamente la muestra en dos grupos y utilizarlos para obtener dos modelos con el fin compararlos para comprobar si se obtienen resultados similares.

Por otro lado, identificar el modelo más adecuado consistirá en evaluar dife-

rentes parámetros de los modelos de regresión. El modelo de regresión lineal se estima mediante una técnica denominada método de los mínimos cuadrados, mientras que en la regresión logística se utiliza el método de máxima verosimilitud.

14.6.1. Modelo de regresión lineal

El método de los mínimos cuadrados, consiste en calcular la suma de las distancias al cuadrado entre los puntos reales y los puntos definidos por la recta estimada a partir de las variables introducidas en el modelo, de forma que la mejor estimación será la que minimice estas distancias. Para poder decidir qué modelo es el que mejor se adecua a los datos de los que disponemos en el modelo de regresión lineal se comparan la F parcial obtenida en cada uno de los modelos de regresión construidos. Si utilizamos cualquiera de las técnicas de selección de variables expuestas previamente, se calculará dicho coeficiente cada vez que se elimine o introduzca una variable, dado que al realizar este proceso, en realidad se están estimando nuevos modelos de regresión. En todos los casos el paquete estadístico realiza la operación automáticamente, exceptuando si utilizamos la técnica de obligar a entrar todas las variables, en cuyo caso seremos nosotros quienes vayamos estimando todos los modelos posibles manualmente, para realizar posteriormente la selección.

Otro método para validar un modelo es evaluar los residuos de la regresión, es decir la diferencia entre el valor estimado por el modelo y el valor observado y por tanto la parte que el modelo de regresión no es capaz de explicar. Si el modelo de regresión resulta adecuado para explicar nuestros datos, los residuos deberían distribuirse según una ley normal de media 0 y varianza constante. Este supuesto puede comprobarse gráficamente al representar mediante una nube de puntos cómo se distribuyen los residuos de nuestro modelo de regresión. A través de este método se puede diagnosticar la falta de linealidad o la heterocedasticidad (cuando la varianza no es constante).

El peligro de introducir en el modelo de regresión valores extremos, a pesar de ser valores realmente registrados, es que pueden provocar alteraciones importantes en los resultados de la regresión lineal, al estimarse ésta en base al método de los mínimos cuadrados, como se ha explicado y basarse en el cálculo de las distancias entre puntos. Por tanto, resulta necesario tenerlos en consideración al ajustar el modelo, estimando dos modelos: uno incluyendo y otro excluyendo dichos valores y finalmente valorando cuáles son los resultados que más se adecuan a nuestros propósitos.

14.6.2. Modelo de regresión logística

Por su parte, la identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el cociente de verosimilitud, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente al otro. La diferencia de los cocientes de verosimilitud entre dos modelos se distribuye según la ley de la Ji-cuadrado con los grados de libertad correspondientes a la diferencia en el número de variables entre ambos modelos. Si a partir de este coeficiente no se puede demostrar que un modelo resulta mejor que el otro, se considerará como el más adecuado, el más sencillo.

14.7. Factores de confusión

Durante el proceso de selección del modelo de regresión más adecuado, el que mejor se ajusta a los datos de los que disponemos, además de seguir el procedimiento explicado anteriormente, hay que considerar un último aspecto adicional, especialmente si el proceso de selección de variables se hace mediante el método manual de obligar a que todas las variables entren en el modelo y es el propio investigador el que paso a paso va construyendo el modelo de regresión más conveniente.

Durante el proceso de incorporación de variables, al eliminar una variable de uno de los modelos de regresión estimados, hay que observar si en el modelo de regresión resultante al excluir esa variable, los coeficientes asociados al resto de variables introducidas en el modelo varían significativamente respecto al modelo de regresión que sí incluía dicha variable. Si así sucede, significa que dicha variable podría ser un factor de confusión, al no mostrar una relación significativa con la variable que estamos estudiando directamente, pero sí indirectamente, al relacionarse con otras variables, que en sí mismas pueden estar significativamente relacionadas con la variable de estudio. Por lo tanto, en dicho caso, es conveniente no excluir la variable en cuestión del modelo de regresión, aunque no cumpla los requisitos para permanecer en él, obligando a que permanezca, de modo que aunque no se incluya su interpretación al evaluar los resultados del modelo, ajustemos el resultado del resto de variables seleccionadas por su posible efecto.

Por ejemplo: Al estudiar una muestra aleatoria de una población de diabéticos y analizando la posible relación lineal entre la Tensión arterial sistólica (TAS) como variable respuesta y la edad y el género de los pacientes, obtendríamos un modelo de regresión donde el género de los pacientes sería significativo, es decir, existiría una ecuación diferente de predicción para hombres y otro para

mujeres. Sin embargo, si controláramos también el índice de masa corporal (IMC) introduciéndolo en la ecuación, posiblemente la variable de género no sería significativa, mientras que pasaría a serlo el IMC. En ese caso el IMC sería un factor de confusión que deberíamos incluir en la ecuación y ello aunque su coeficiente no fuera significativo.

Aprovechamos este punto para decir que debe tenerse cuidado con los términos «relación», «correlación» o «significación» y «causalidad». Que dos factores estén relacionados no implica de ninguna manera que uno sea causa del otro. Es muy frecuente que una alta dependencia indique que las dos variables dependen de una tercera que no ha sido medida (el factor de confusión).

14.8. Interpretación de los resultados de los modelos de regresión

En el modelo de regresión lineal el resultado que obtenemos se puede interpretar como la magnitud del cambio de la variable dependiente si incrementamos en una unidad el valor de la variable independiente (en el caso de que la variable independiente sea de tipo continuo) y la magnitud del cambio en la variable dependiente si una característica determinada está o no presente (en el caso de tratar variables de tipo categórico).

Para que la aplicación de un modelo de regresión lineal resulte procedente debe cumplirse que los valores de respuesta (y) sean independientes entre sí y la relación entre las variables sea lineal de la forma:

$$Y = f(x_1, x_2, \ldots) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

Para poder interpretar el resultado del modelo de regresión logística debemos recurrir al concepto de 'odds', una de las medidas de las que se dispone para poder cuantificar el riesgo. De esta forma el o la 'odds' se define como el cociente de la probabilidad de presentar una característica y la probabilidad de no presentarla, o lo que es lo mismo el cociente del número de casos que presentan la característica entre el número de casos que no la presentan.

$$Odds = p / (1-p)$$

Se pueden comparar dos odds, por ejemplo entre los pacientes que padecen una cierta enfermedad si presentan cierta característica o no la presentan, en forma de cociente de ambas (denominada odds ratio), de manera que podamos concluir si por ejemplo la enfermedad es más frecuente entre los pacientes que presentan dicha característica o no la presentan (los términos «Odds» y «Odds-

ratio» a los que se hace referencia en este apartado y se explican con más detalle en el capítulo 15. Se puede demostrar que los coeficientes obtenidos en la regresión logística son medidas que cuantifican el riesgo de presentar cierta característica respecto a no presentarla en base a la variable de estudio, de manera que:

$$\text{Exp}\ (\beta) = \text{OR}$$

Donde β es el coeficiente resultado de la regresión logística asociado a una cierta variable participante en el modelo. Cuando la variable independiente tratada es numérica, este valor se interpreta como el cambio en el riesgo cuando se incrementa en uno el valor de la variable, mientras que el resto de variables permanecen constantes.

Siempre que se construye un modelo de regresión es fundamental, antes de pasar a extraer conclusiones, el corroborar que el modelo calculado se ajusta efectivamente a los datos usados para estimarlo. En el caso de la regresión logística una idea bastante intuitiva es calcular la probabilidad de aparición del suceso, presencia de hipertensión en nuestro caso, para todos los pacientes de la muestra. Si el ajuste es bueno, es de esperar que un valor alto de probabilidad se asocie con presencia real de hipertensión, y viceversa, si el valor de esa probabilidad calculada es bajo, cabe esperar también ausencia de hipertensión. Esta idea intuitiva se lleva a cabo formalmente mediante la prueba conocida como de Hosmer-Lemeshow (1989), que básicamente consiste en dividir el recorrido de la probabilidad en deciles de riesgo (esto es probabilidad de hipertensión ≤ 0.1, ≤ 0.2, y así hasta ≤ 1) y calcular tanto la distribución de hipertensos, como no hipertensos prevista por la ecuación y los valores realmente observados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de Ji-Cuadrado.

Finalmente, debe evitarse que en el modelo de regresión planteado pueda producirse el fenómeno de la colinealidad, que daría lugar a soluciones inestables. Se habla de colinealidad cuando dos o más variables independientes que se introducen en el modelo de regresión están altamente correlacionadas entre sí.

14.9. Ejemplos

Para conocer cómo se estima un modelo de regresión, se presentan a continuación dos ejemplos que ilustran en primer lugar, un modelo de regresión lineal con la variable dependiente continua y, en segundo lugar, un modelo de regresión logística, donde la variable respuesta es de tipo binario.

14.9.1. Ejemplo de modelo de regresión lineal

Ejemplo 1: Se desea conocer qué variables pueden estar relacionadas con las cifras de tensión arterial. Para ello, se desarrolla un estudio que incluye a 71 pacientes que acuden a consultas del médico de atención primaria. A estos pacientes se les hacen diversas mediciones. Se registra información sobre la principal variable de interés, la tensión arterial sistólica (TAS) y diversas características sociodemográficas como la edad y el género, medidas antropométricas como el peso y la altura y otras variables como tabaquismo y presencia de enfermedades concomitantes como la diabetes o la hipercolesterolemia. El objetivo del estudio sería identificar qué factores son los que se relacionan con el hecho de presentar cifras altas de TAS.

La Figura 60 y la Figura 61 muestran los pasos a seguir para la obtención del modelo anterior mediante el paquete estadístico SPSS. En primer lugar se selecciona el tipo de regresión y a continuación, se seleccionan las variables independientes y la respuesta (variable dependiente), así como aspectos técnicos del mismo, como puede ser el método a utilizar para la selección de variables.

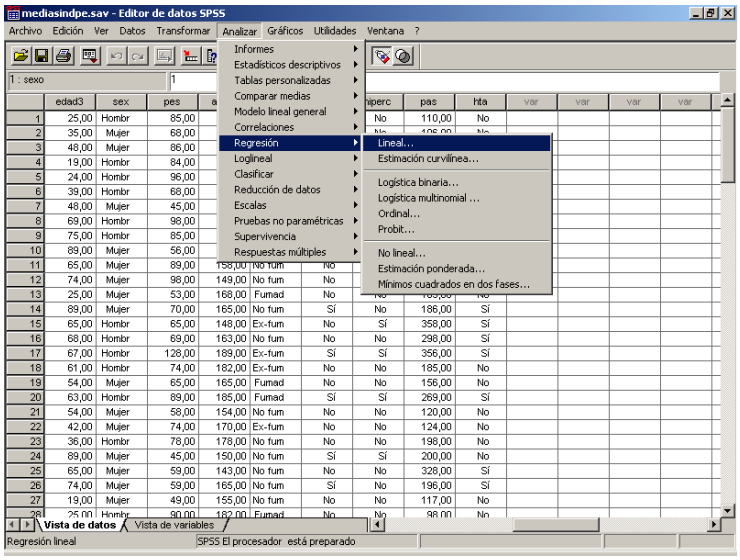

Figura 60. Obtención de un modelo de regresión lineal por menús en SPSS

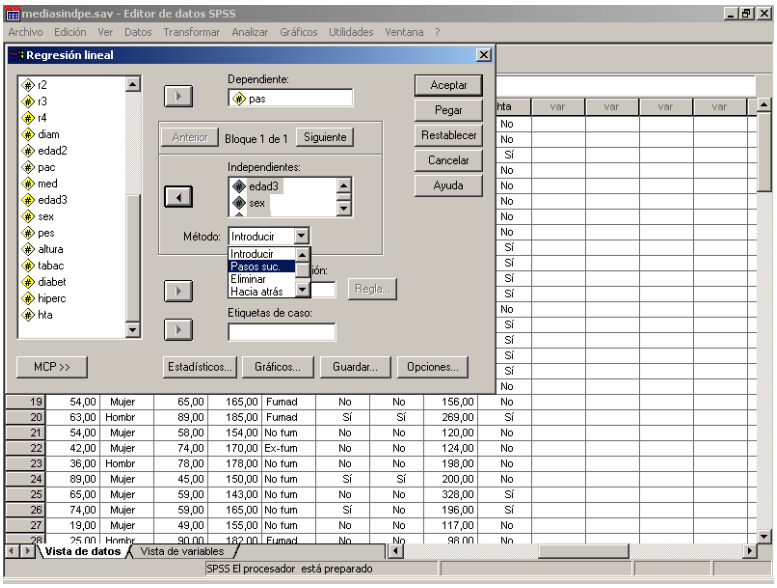

Figura 61. Obtención de un modelo de regresión lineal por menús en SPSS. Variables introducidas y método de introducción de variables

En primer lugar, se presenta un resumen de los modelos obtenidos: en nuestro caso, se han obtenido dos modelos (Figura 62). Observamos en este resumen que aparece R cuadrado (R^2), que es el cuadrado del coeficiente de correlación muestral, también denominado coeficiente de determinación. En consecuencia, este coeficiente indica la proporción de variabilidad total de la variable dependiente explicada por el modelo de regresión (recta de regresión). Por tanto, el segundo modelo, con una $R^2 = 0{,}289$, es el que mejor explica la variabilidad de la TAS.

Los resultados del análisis de la varianza (ANOVA) se proporcionan para evaluar la significación del modelo. Cuanto mayor sea el estadístico F, mejor será la predicción mediante el modelo lineal. De nuevo, si el p-valor asociado a F es menor a alfa, se rechazarán las hipótesis nulas en las que se plantea que no hay relación entre las variables. Como se muestra en la siguiente figura, ambos modelos presentan valores de F significativos, por lo tanto, la recta de regresión tendrá los coeficientes de las variables independientes distintos de cero.

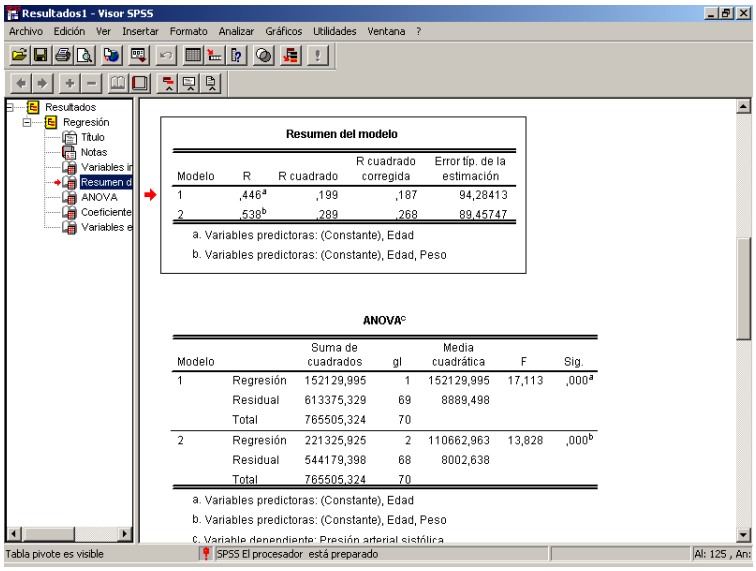

Figura 62. Resultados obtenidos en la evaluación de un modelo de regresión lineal en SPSS. R de los modelos

Se ha utilizado el método de por pasos hacia delante y atrás para seleccionar las variables en el modelo. Observamos que el modelo se ha completado en dos pasos. En el primero se introdujo la variable edad y en el segundo se volvió a introducir otra variable, el peso, finalizando en este momento el proceso de selección de variables.

El modelo de regresión que se propone en segundo lugar, explica la TAS de una persona, en función de su edad y su peso. Los coeficientes del modelo de regresión que se presentan en la figura Figura 63 como parte de los resultados obtenidos en SPSS indican los términos que "modifican" el efecto de las variables. El modelo estimado (modelo 2) sería el siguiente:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_n X_n + \varepsilon$$

$$TAS = 2{,}056 \times Edad + 1{,}7 \times Peso + \varepsilon$$

Donde ε es el error aleatorio que no se puede definir a partir de las variables incluidas en el modelo.

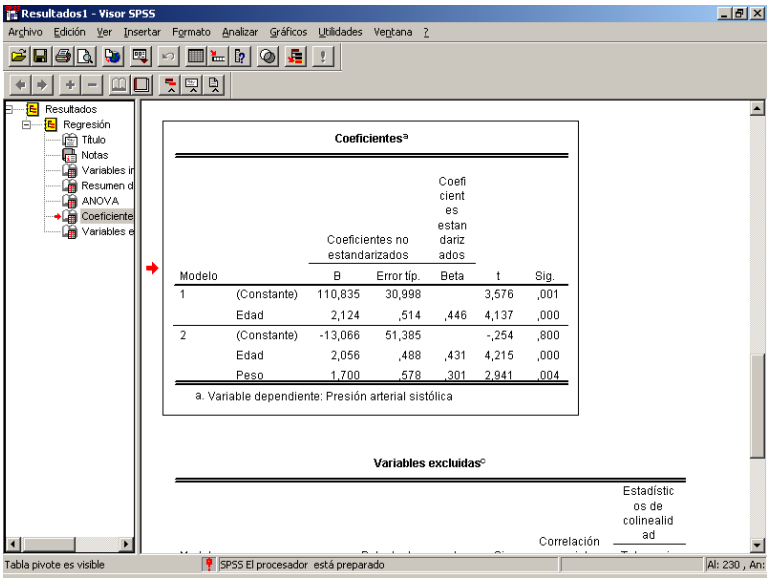

Figura 63. Resultados obtenidos en la evaluación de un modelo de regresión lineal en SPSS. Coeficientes de los modelos

A partir de los coeficientes obtenidos, podemos deducir que por cada año de más que tiene una persona, se incrementa en dos unidades la tensión arterial sistólica, mientras que por cada kilo de más que pesa una persona, la cifra de tensión arterial se incrementará en 1,7 unidades. El peso y la edad son las dos únicas variables que se relacionan con las cifras de TAS en nuestro ejemplo.

14.9.2. Ejemplo de modelo de regresión logística

Ejemplo 2: Siguiendo el ejemplo anterior, supongamos que en lugar de intentar evaluar la relación de las variables descritas con las cifras de TAS, lo que se desea es evaluar la relación entre los distintos factores expuestos y el hecho de que un paciente haya sido diagnosticado de hipertenso o no. Es en este momento cuando resulta adecuado la aplicación de un modelo de regresión logística, considerando como variable dependiente el hecho de que un paciente padezca o no hipertensión. El primer paso para ejecutar dicho modelo en SPSS es seleccionar el tipo de modelo, tal como se muestra a continuación.

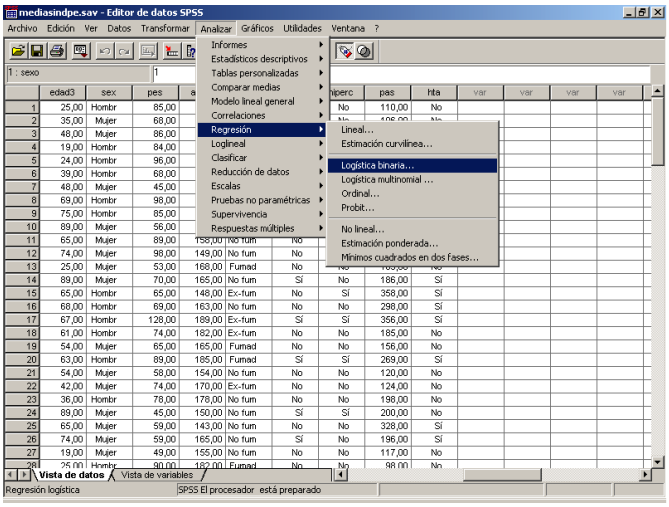

Figura 64. Obtención de un modelo de regresión logística (respuesta binaria) por menús en SPSS

A continuación (Figura 65), introducimos la variable dependiente (HTA) y las variables independientes: sexo, peso, altura, tabaquismo, diabetes e hipercolesterolemia, indicando cuáles de ellas son variables categóricas. En este ejemplo, lo son todas exceptuando el peso y la talla. El programa SPSS creará las variables *dummy* necesarias para poder crear los modelos, de esta forma deberemos indicarle a partir de qué categoría de la variable debe empezar a crearlas.

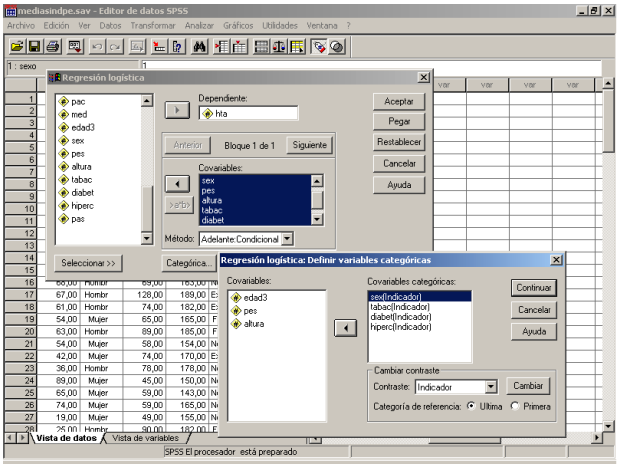

Figura 65. Obtención de un modelo de regresión logística por menús en SPSS. Variables introducidas y método de introducción de variables

Los resultados obtenidos de aplicar este modelo se muestran en la figura siguiente. En primer lugar, en la tabla de codificación de variables categóricas se muestra un resumen del formato de las variables '*dummy*' creadas para poder estimar el modelo de regresión, lo que nos dará la información necesaria para poder posteriormente interpretar los resultados obtenidos. Se le ha asignado internamente el valor 1 al hecho de que un paciente sea hipertenso, factor que también deberemos tener en cuenta al interpretar los resultados.

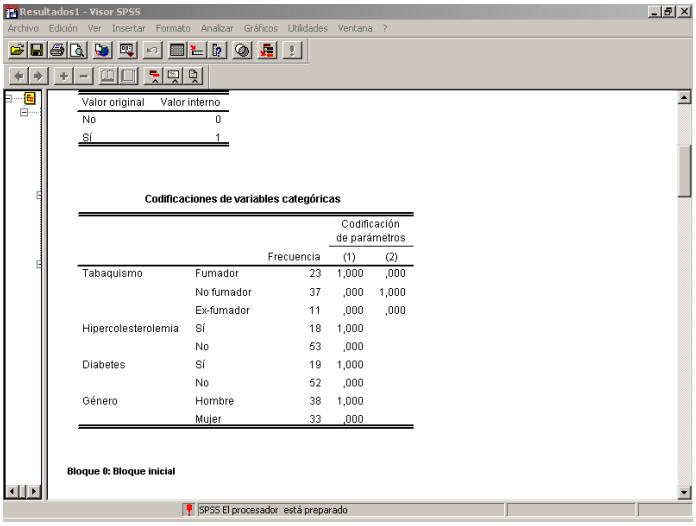

Figura 66. Resultados obtenidos en la evaluación de un modelo de regresión logística en SPSS. Correspondencia de las categorías de las variables codificadas con las variables *dummy*

Posteriormente se muestran los coeficientes de las variables introducidas en la ecuación (Figura 67). En el segundo y último modelo obtenido, se han introducido la variable edad y el género (sex), ambas variables con coeficientes estadísticamente significativos ($p<0,05$). Además de los coeficientes obtenemos información de Exp(B), que corresponde al Odds-Ratio asociado a cada factor.

Si en la ecuación de regresión tenemos un factor dicotómico, como es el género en nuestro caso, la categoría que se considera de referencia es la que se le asigna el valor '0', por lo que el OR se atribuye directamente a la otra opción de respuesta de la variable. A modo de ejemplo el OR del género se atribuye directamente al hecho de ser hombre. Es decir, que *exp(b)* es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes.

Cuando la variable es numérica, como es en este caso la edad, es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro, permaneciendo constantes el resto de variables. Así el odds ratio que supone pasar de la edad X_1 a la edad X_2, siendo *b* el coeficiente correspondiente a la edad en el modelo logístico es $OR = \exp[\, b \times (X_2 - X_1)]$. Se trata de un modelo en el que el aumento o disminución del riesgo al pasar de un valor a otro del factor es proporcional al cambio, es decir a la diferencia entre los dos valores, pero no al punto de partida, esto quiere decir que el cambio en el riesgo es el mismo cuando pasamos de 40 a 50 años que cuando pasamos de 80 a 90.

Cuando el coeficiente *b* de la variable es positivo obtendremos un odds ratio mayor que 1 y corresponde por tanto a un factor de riesgo. Por el contrario, si *b* es negativo el odds ratio será menor que 1 y se trata de un factor de protección.

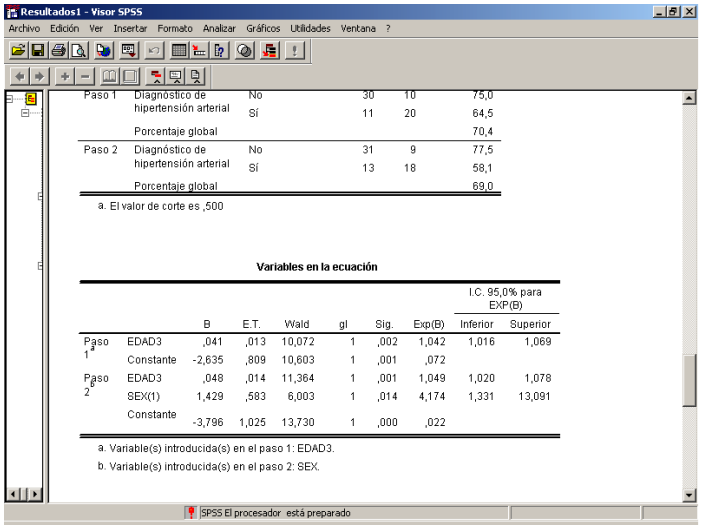

Figura 67. Resultados obtenidos en la evaluación de un modelo de regresión lineal en SPSS. Coeficientes de las variables introducidas en la ecuación.

El proceso de selección de las variables ha finalizado en 2 pasos. La primera variable introducida en el modelo fue la edad, seguida del género (sex), finalizando en este momento el proceso de selección de las variables. La variable edad se introdujo en el modelo como una variable continua, por lo que al interpretar los resultados podemos decir que conforme aumenta la edad aumenta el riesgo de padecer hipertensión arterial. Por otra parte, la variable sexo se introdujo como una variable categórica. Al observar el tipo de codificación realizada internamen-

te (mostrada en la Figura 66), podemos concluir que el hecho de ser hombre (codificado como 1), aumenta aproximadamente 4 veces (OR=4,174) el riesgo de padecer hipertensión arterial. En conclusión el hecho de ser hombre y/o de mayor edad, aumenta el riesgo de padecer hipertensión arterial.

14.10. Consideraciones importantes

Se conoce como análisis de regresión al método estadístico que permite establecer una relación matemática entre un conjunto de variables $X_1, X_2 .. X_k$ (covariantes o factores) y una variable dependiente Y. Se utiliza fundamentalmente en estudios en los que no se puede controlar por diseño los valores de las variables independientes con el fin de predecir su valor o bien, explicarlo.

Un problema fundamental que se plantea a la hora de construir un modelo multivariante es qué factores $X_1, X_2 .. X_k$ incluir en la ecuación, de tal manera que estimemos el mejor modelo posible a partir de los datos de nuestro estudio. Si buscamos un modelo predictivo será aquél que nos proporcione predicciones más fiables, más acertadas; mientras que si nuestro objetivo es construir un modelo explicativo, buscaremos que las estimaciones de los coeficientes de la ecuación sean precisas, ya que a partir de ellas vamos a efectuar nuestras deducciones. Cumplidos esos objetivos, otra característica deseable de nuestro modelo es que sea lo más sencillo posible.

La construcción de un modelo de regresión lineal es muy simple, puesto que la estimación de los coeficientes para cada variable resulta muy intuitiva. Además, el hecho de modelizar una variable numérica, implica que podamos obtener distintos valores que corresponderán a los valores estimados de la variable a predecir, siempre en función de nuestras variables dependientes.

En cuanto a la regresión logística, el objetivo primordial que resuelve esta técnica es el de cuantificar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías.

La elección de la técnica estadística apropiada debe hacerse en el momento de diseñar el estudio y en realidad debería ser el primer paso y la primera línea que se escribiera en el borrador de un protocolo.