5.2 Modelo de regresión lineal

Este modelo de regresión puede estudiarse como una extensión del modelo lineal simple en el que considerábamos una sola variable predictora x.

Ahora vamos a considerar que la variable de respuesta y depende de varias variables x, algunas conocidas por el investigador y otras no.

El modelo de regresión múltiple trata de estimar el efecto de las más importantes, englobando las demás en el término que denominarnos error aleatorio.

Para simplificar vamos a suponer que la variable y depende solamente de dos variables x y x y que la relación que liga a las variables sigue siendo lineal.

La ecuación de regresión es, ahora, la siguiente:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Un **ejemplo** de esta situación de investigación podría ser que y fuera la cantidad de lluvia caída en una zona en particular, x_1 la humedad del ambiente y x_2 la presión atmosférica. La variable de respuesta y depende de dos variables que llamaremos predictoras: x_1 y x_2 .

Otro ejemplo podría ser que y fuera el salario pagado por una empresa, x_1 los años de antigüedad de los empleados y x_2 la calificación anual de cada uno de ellos.

Supuestos del Modelo

Los supuestos referidos a los ε son los mismos que estudiamos en el modelo de regresión lineal simple.

5.2.1. Estimación de los parámetros del modelo

También aquí se utiliza el método de mínimos cuadrados sólo que ahora debemos derivar la función de mínimos cuadrados con respecto a α, β_1 y β_2 .

Luego de aplicar el procedimiento matemático de minimización, nos queda el siguiente sistema de 3 ecuaciones con 3 incógnitas:

$$\Sigma y_i = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$
$$\Sigma y_i x_1 = a\Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$
$$\Sigma y_i x_2 = a\Sigma x_2 + b_1 \Sigma x_2 x_1 + b_2 \Sigma x_2^2$$

donde a estima a α, b_1 a β_1 y b_2 a β_2

La resolución de este sistema de ecuaciones se complica y posiblemente no sea comprendido por todos los participantes del curso. Los cálculos que se deben realizar al efectuar un análisis de regresión múltiple a menudo consumen mucho tiempo y por ello siempre se recurre a una computadora.

Desarrollaremos el tema por medio de un ejemplo.

EJEMPLO

Los siguientes datos corresponden a la velocidad (y), la potencia (x_1) y la cantidad de cilindros (x_2) de 10 marcas de motos:

Velocidad (y)	Potencia (x_1)	Cantidad de cilindros (x_2)
160	60	2
156	26	4
193	80	6
191	67	2
200	70	2
190	70	2
194	70	2
170	36	1
132	27	1
111	17	1

Con estos datos obtenidos de una muestra de 10 motos, vamos a efectuar un análisis de regresión múltiple. Por facilidad, se utilizará la hoja de Cálculo EXCEL.

En primera medida presentamos las estimaciones de los parámetros del modelo.

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.896200002
Coeficiente de determinación R^2	0.803174443
R^2 ajustado	0.74693857
Error típico	15.05994983
Observaciones	10

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	6478.485377	3239.242689	14.28224362	0.003382869
Residuos	7	1587.614623	226.802089		
Total	9	8066.1			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	108.8659417	12.64423923	8.609924228	5.68529E-05	78.96706695	138.764816	78.96706695	138.7648164
Variable X 1	1.150233924	0.23811225	4.83063733	0.001898076	0.587187923	1.71327993	0.587187923	1.713279925
Variable X 2	0.294271352	3.518712311	0.08363041	0.935691672	-8.026161112	8.61470382	-8.026161112	8.614703816

En la tabla aparece la variable que queremos predecir por medio del análisis de la relación que liga a esta variable con sus dos predictoras x_1 y x_2.

A continuación se estiman los coeficientes que acompañan a cada una de las variables predictoras CIL x_1 (Cantidad de Cilindros) y POTE x_2 (Potencia) además de la ordenada al origen (Constant).

Las estimaciones en la siguiente sección de la tabla

	Coeficientes
Intercepción	108.8659417
Variable X 1	1.150233924
Variable X 2	0.294271352

dando los siguientes valores:

$$a = 108{,}8659$$
$$b_1 = 1{,}1502$$
$$b_2 = 0{,}2943$$

La ecuación de regresión resultante es, luego:

$$y_i = 108{,}8659 + 1{,}1502x_1 + 0{,}2943x_2$$

En la siguiente columna (Error típico) se observan las desviaciones estándares correspondientes a cada uno de estos estimadores.

Error típico
12.64423923
0.23811225
3.518712311

Así, tenemos:

$$SEa = 12{,}6442$$
$$SEb_1 = 0{,}2381$$
$$SEb_2 = 3{,}5187$$

Posteriormente se observa una columna con la letra T, la cual proporciona los valores de una variable t de Student que se utiliza para probar la significación de los estimadores en la ecuación.

Estadístico t
8.609924228
4.83063733
0.08363041

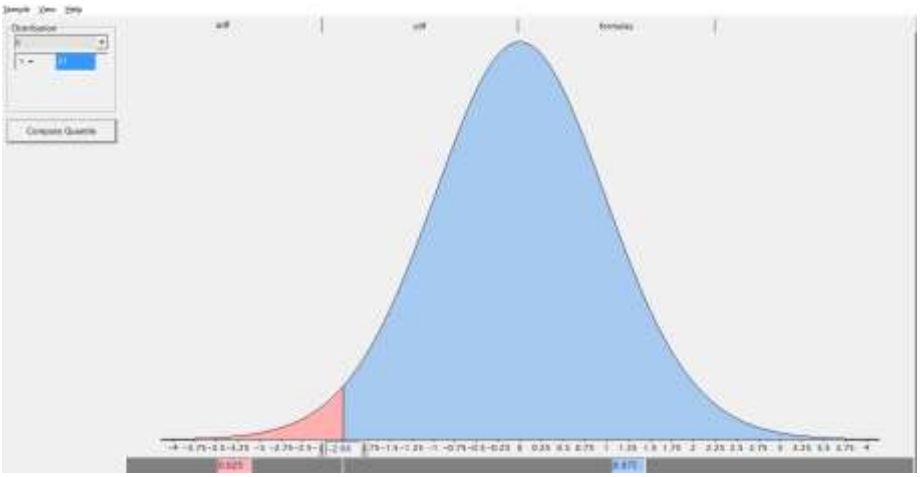
Así, para probar la hipótesis de que la ordenada al origen es igual a 0 (H_0: $\alpha = 0$, H_1: $\alpha \neq 0$), el valor de t observado es: $t = 8{,}61$

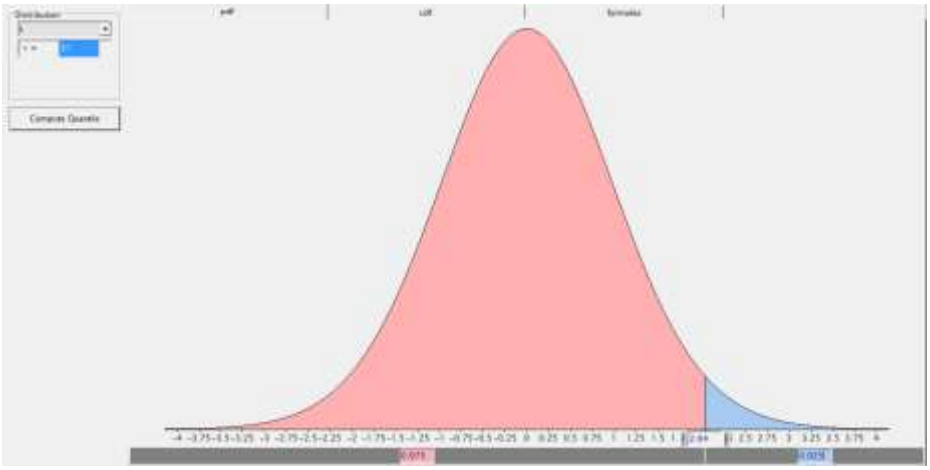
Al lado aparece el nivel de significación (0,0001) con el cual se rechaza la hipótesis nula planteada.

Es posible realizar el contraste de la hipótesis a través del valor t, o bien con el nivel de significación.

El valor crítico para se obtiene con t de $\alpha/2$ con $n-(k+1)$ grados de libertad (k indica el número de variables predictoras).

En el ejemplo:





Los valores críticos para t son -2.04 y 2.04

Analizando los valores de t, se tiene:

0.084 para CIL
4.831 para POTE
8.61 para la CONSTANTE

De lo anterior, se puede concluir que la ordenada al origen no es 0.

También podemos probar la siguiente hipótesis:
H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

El valor de t observado es t = 4.831 con un nivel de significación igual a 0.0019.

En este caso diremos que β_1 no es 0 y, en consecuencia, la variable potencia es buena predictora de la variable velocidad. Si probamos la hipótesis:

H_0: $\beta_2 = 0$
H_1: $\beta_2 \neq 0$

El valor de t observado es t = 0.084 (el cual se ubica en el área de NO RECHAZO de H0) con un nivel de significación igual a 0,9357. En esta situación, no podemos rechazar la hipótesis nula y, entonces, diremos que β_2 es igual a 0. En conclusión, la cantidad de cilindros de las motos no es una variable significativa para predecir el valor de la velocidad.

El programa de cómputo también proporciona la tabla de análisis de la varianza de la regresión:

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	6478.485377	3239.242689	14.28224362	0.003382869
Residuos	7	1587.614623	226.802089		
Total	9	8066.1			

En la primera columna aparecen las dos fuentes de variación (debida a la regresión y la residual o de error). Luego se observan los grados de libertad (DF), las sumas de cuadrados (Sum of Squares) y los cuadrados medios (Mean Square).

Después está calculado el cociente F = CMregresión/CMresidual = 14,28224 con un nivel de significación igual a 0,0034. Esta situación está indicando que la regresión es significativa.

También está calculado el coeficiente de determinación R^2 = 0,80317, lo cual indica un ajuste bastante bueno a través del modelo de regresión lineal múltiple especificado.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.896200002
Coeficiente de determinación R^2	0.803174443
R^2 ajustado	0.74693857
Error típico	15.05994983
Observaciones	10

Cuando se tiene más de una variable independiente, conviene tomar como referencia de la bondad del ajuste al R^2 ajustado, ya que el mismo está ajustado por la cantidad de parámetros que se estiman.

5.2.2. Problemas que pueden presentarse en la utilización de los modelos de regresión múltiple

A continuación vamos a tratar los problemas más importantes que se pueden presentar cuando analizamos modelos de regresión múltiples.

1) Multicolinealidad

Este problema se suscita cuando las variables predictoras están muy correlacionadas entre sí. Esta situación impide que se puedan medir aisladamente los efectos de cada una de ellas y su contribución en la ecuación de regresión como predictora de la variable de respuesta y.

En estos casos, los estimadores presentan grandes varianzas y, a menudo, ocultan contribuciones importantes de las variables predictoras.

Por este motivo, se debe tener mucho cuidado al elegir las variables predictoras y no agregar variables en la ecuación por el solo hecho de que se han medido.

La multicolinealidad puede identificarse estudiando las correlaciones entre las variables predictoras por medio del coeficiente de correlación lineal.

2) Error de especificación

Se comete un error de especificación cuando se establece una dependencia errónea de la variable de respuesta con las variables predictoras.

Este problema ocurre cuando omitimos variables predictoras importantes, introducimos variables innecesarias o suponemos que existe una relación lineal cuando en realidad la relación es curvilínea.

Cuando olvidamos incluir en el modelo variables importantes, la consecuencia suele ser la obtención de estimadores sesgados y varianzas de los estimadores más grandes.

Cuando incluimos variables innecesarias, ya vimos que se puede producir un efecto de multicolinealidad si estas variables están muy correlacionadas entre sí. Suponer una relación

lineal cuando no lo es, afecta mucho a la predicción de la variable de respuesta sobre todo si la misma se debe realizar fuera de su rango de variación.

5.2.3. Interpretación de las conclusiones de un modelo de regresión

Consideremos que se está aplicando un test para medir la habilidad de niños en escritura a partir de un texto dictado.

Vamos a simbolizar al puntaje obtenido por cada niño con y_i y con x_i al peso de cada niño.

¿Tendría algún sentido lógico encontrar una relación entre y_i y x_i?

Si se considera que en la investigación participaron niños dentro de un amplio rango de edades, por ejemplo entre 6 y 15 años, se supone que los niños de mayor edad también pesarán más. Pero, al ser mayores, también tendrán más experiencia escolar. En este caso podría decirse que habrá una fuerte relación positiva entre el peso y el puntaje de cada niño.

Este hecho no hubiera ocurrido, en cambio, si todos los niños que participaron en la experiencia tuvieran la misma edad.

Estas ilustraciones muestran que, para discutir la relación entre dos variables x e y, se necesitan especificar las circunstancias o quizás determinar muy bien la población objetivo que va a ser muestreada.

Con estas consideraciones en mente, se verán algunos conceptos y sus definiciones.

Independencia

Por ejemplo, si la experiencia referente a la lengua se llevara a cabo con niñas de 15 años, todas provenientes del mismo colegio y de hogares muy similares, se podría pensar que el peso y el puntaje en escritura no están asociados y, en consecuencia, estas variables se comportarían **independientemente**.

Dependencia

Se debe ser muy cuidadoso al emplear la palabra dependencia y sus derivados. Cuando se dice **y depende de x**, a veces se hace referencia a una dependencia exclusiva significando que, si se da un cierto valor de x, se darán determinados valores de y, generalmente apelando a la existencia de una ley.

Otras veces se utiliza la palabra dependencia como falta de independencia cuando en realidad lo que ocurre es que no se tienen en cuenta otras condiciones o variables en el estudio.

5.3 Selección de variables

En muchas situaciones se dispone de un conjunto grande de posibles variables regresoras, una primera pregunta es saber si todas las variables deben estar en el modelo de regresión y, en caso negativo, se quiere saber qué variables deben entrar y qué variables no deben estar en el modelo de regresión.

Intuitivamente parece bueno introducir en el modelo todas las variables regresoras significativas (según el contraste individual de la t) al ajustar el modelo con todas las variables posibles. Pero este procedimiento no es adecuado porque en la varianza del modelo influye el número de variables. Además puede haber problemas de multicolinealidad cuando hay muchas variables regresoras.

Para responder a estas preguntas se dispone de diferentes procedimientos estadísticos. Bajo la hipótesis de que la relación entre las variables regresoras y la variable respuesta es lineal existen procedimientos "paso a paso" (o setpwise) que permiten elegir el subconjunto de variables regresoras que deben estar en el modelo. Estos algoritmos se presentan en esta sección. También existen medidas de la bondad de ajuste de un modelo de regresión que permiten elegir entre diferentes subconjuntos de variables regresoras el "mejor" subconjunto para construir el modelo de regresión. Para la utilización de estas medidas de bondad de ajuste no es necesaria la hipótesis de linealidad. La utilización combinada de los algoritmos de selección de las variables regresoras y los criterios de bondad de ajuste permiten seleccionar adecuadamente el modelo de regresión que se debe utilizar. Estos criterios serán estudiados cuando se aborde el modulo de Análisis Multivariado.

Brevemente, se pueden mencionar los procedimientos para seleccionar las variables regresoras que deben entrar en el modelo:

- "**Eliminación progresiva**" ("***Backward Stepwise Regression***"). Este procedimiento parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la t (o de la F) hasta una cierta regla de parada. El procedimiento de eliminación progresiva tiene los inconvenientes de necesitar mucha capacidad de cálculo si k es grande y llevar a problemas de multicolinealidad si las variables están relacionadas. Tiene la ventaja de no eliminar variables significativas.

- "**Introducción progresiva**" ("***Fordward Stepwise Regression***"). Este algoritmo funciona de forma inversa que el anterior, parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa hasta una cierta regla de parada. El

procedimiento de introducción progresiva tiene la ventaja respecto al anterior de necesitar menos cálculo, pero presenta dos graves inconvenientes, el primero, que pueden aparecer errores de especificación porque las variables introducidas permanecen en el modelo aunque el algoritmo en pasos sucesivos introduzca nuevas variables que aportan la información de las primeras. Este algoritmo también falla si el contraste conjunto es significativo pero los individuales no lo son, ya que no introduce variables regresoras.

- "**Regresión paso a paso**" ("***Stepwise Regression***"). Este método es una combinación de los procedimientos anteriores, comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo.