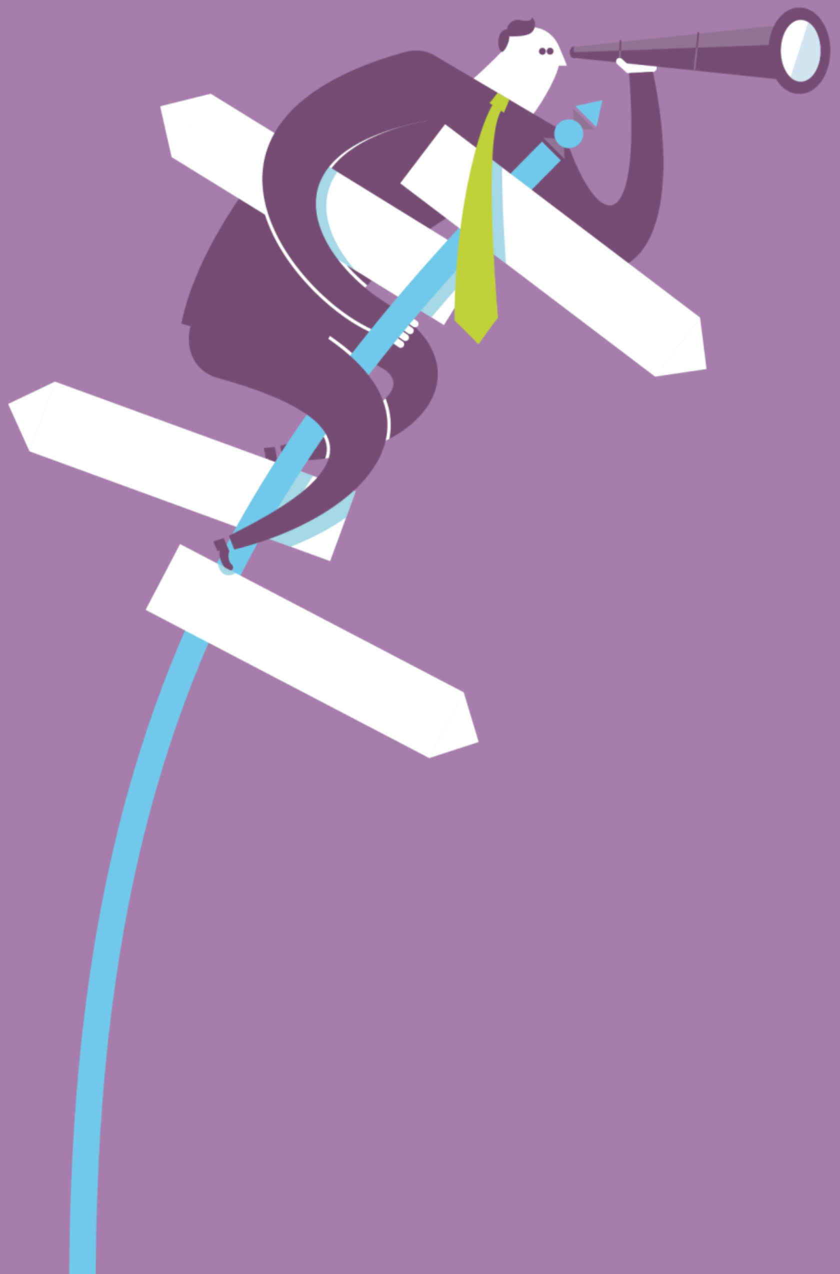

Tema 13: Inferencias en modelos de regresión múltiple y predicción.

Modelo de regresión múltiple: ejemplo.

Se realizó un experimento para determinar si el peso de un animal puede predecirse después de un periodo dado, con base en el peso inicial del animal y en la cantidad de alimento consumida por este. Se registraron los siguientes datos, medidos en kilogramos:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Peso Final Y	Peso inicial X_1	Alimentos consumidos X_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
94	38	235

Una vez estimado el modelo de regresión, se obtienen los siguientes resultados:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.9012
Coeficiente de determinación R^2	0.8121
R^2 ajustado	0.7584
Error típico	7.5797
Observaciones	10

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intersección	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

De esta regresión puede apreciarse que el modelo de regresión estimado es:

$$\hat{Y} = -22.1377 + 1.4420\ X_1 + 0.2110X_2$$

La relación entre Y (peso final) y X_1 (peso inicial) se describe por b_1= 1.4420. De este número puede decirse que en este modelo, por cada unidad adicional de peso inicial, el peso final se incrementa en 1.4420 en promedio, manteniendo constante la X_2 (alimentos consumidos).

Supóngase que se desea predecir el peso promedio final cuando X1(el peso inicial), es de 35 kilogramos y X2 (los alimentos consumidos), es igual a 280 kilogramos.

Si se utiliza la ecuación de regresión múltiple, obtenida anteriormente:

$$\hat{Y} = -22.1377 + 1.4420\ X_1 + 0.2110X_2$$

Con $X_1 = 35$ y $X_2=280$, se tiene:

$$\hat{Y} = -22.1377 + 1.4420\ (35) + 0.2110(280)$$

Y así:

$$\hat{Y} = 87.55$$

Sin embargo la regresión no termina allí …

Recordemos que: la evaluación del modelo se puede hacer en tres formas:

Analizando estas tres vertientes, junto con el análisis de los supuestos de los errores (residuos), podremos tener un buen modelo de regresión: homocedasticidad, media cero, distribución normal, no multicolinealidad.

Error estándar de la estimación

En regresión múltiple, el error estándar de la estimación se define como sigue:

$$S_{\varepsilon} = \sqrt{\frac{SCE}{n-k-1}} = \sqrt{CME}$$

En donde:

n = número de observaciones

k = número de variables independientes en la función de regresión

SCE = suma de cuadrados del error

CME = cuadrado medio del error

El número de observaciones es n=10 y el error estándar de la estimación se determina con:

$$S_\varepsilon = \sqrt{\frac{402.1607}{10-2-1}} = \sqrt{\frac{402.1607}{7}} = \sqrt{57.4515} = 7.5797$$

En este caso, el error estándar del modelo de regresión es de 7.58.

Coeficiente de determinación

El coeficiente de determinación es dado por:

$$R^2 = \frac{\text{Suma de cuadrados de regresión}}{\text{Suma de cuadrados totales}}$$

Y representa la razón de la variación de la respuesta Y explicada por su relación con las X. Para el ejemplo anterior se tiene que el coeficiente de determinación es:

$$R^2 = \frac{\text{Suma de cuadrados de regresión}}{\text{Suma de cuadrados totales}} = \frac{1738.3393}{2140.5000} = 0.8121$$

En el contexto de este problema podemos decir que el 81.21% de la variación en el peso final se explica por X_1 (peso inicial) y X_2 (alimentos consumidos). En la práctica, $0 \leq R^2 \leq 1$, y el valor de R^2 debe interpretarse en relación con los extremos, 0 y 1.

Significancia de los coeficientes de regresión

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

Para evaluar la significancia de los coeficientes de la regresión, se hacen algunas pruebas de hipótesis respecto a los coeficientes.

Pruebas de hipótesis.

$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$ (Las variables independientes no afectan a Y)

En oposición a:

$H_a : \beta_i \neq 0$ (Al menos una variable X afecta a Y)

Para **evaluar la hipótesis** se hace uso del estadístico de prueba:

$$t_{calculada} = \frac{b_i - \beta_i}{S_{b_i}}$$

Regla de decisión

$$t_{calculada} = \frac{b_i - \beta_i}{S_{b_i}} = \frac{1.4420 - 0}{0.7297} = 1.9760$$

Rechazar H_0 si $|t_{calculada}| = 1.9760$ es mayor que $t_{teórica}$.

En donde:

$$t_{teórica} = t_{\alpha/2}(n-k-1) = t_{0.05/2}(7) = t_{0.025}(7) = 2.365$$

En donde el valor de $t_{teórica}$ se obtiene de la tabla de distribución de t.

Puesto que $|t_{calculada}|$ = 1.9760 es *menor* que $t_{teórica}$ = 2.365, **no** se rechaza H_0. (Esto es, **no** existe evidencia de que el peso inicial X_1 afecte el peso final Y, o bien, la variable peso inicial X_1 no tienen efecto significativo en el peso final Y).

Intervalo de confianza

En el análisis de regresión múltiple, un intervalo de confianza para una pendiente de la población se puede estimar a partir de la siguiente expresión:

$$b_1 \pm t_{\alpha/2}(n-k-1)\, S_{b_i}$$

Para el presente ejemplo:

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

Entonces, con un 95% de confianza, se tiene que el verdadero valor β_1 se encuentra en el intervalo (-0.2837, 3.1677). Desde el punto de vista de la prueba de hipótesis, puesto que este intervalo de confianza contiene al cero, se concluye que el coeficiente de correlación β_1 no tiene efecto significativo.