REGRESION Y CORRELACION

Fórmulas básicas en la regresión lineal simple

Como ejemplo de análisis de regresión, describiremos el caso de Pizzería Armand, cadena de restaurantes de comida italiana. Los lugares donde sus establecimientos han tenido más éxito están cercanos a establecimientos de educación superior. Se cree que las ventas trimestrales (representadas por y) en esos restaurantes, se relacionan en forma positiva con la población estudiantil (representada por x). Es decir, que los restaurantes cercanos a centros escolares con gran población tienden a generar más ventas que los que están cerca de centros con población pequeña. Aplicando el análisis de regresión podremos plantear una ecuación que muestre cómo se relaciona la variable dependiente "*y*" con la variable independiente "x".

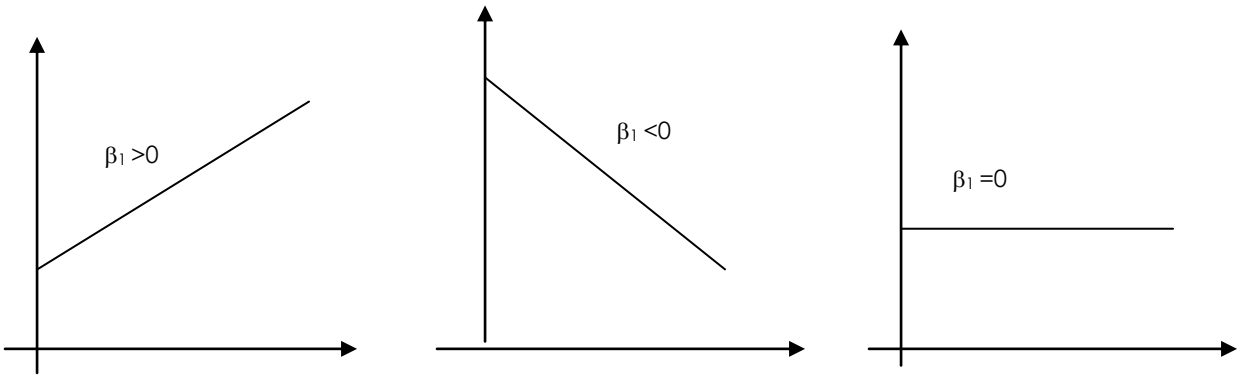
El modelo de regresión y la ecuación de regresión

En el ejemplo, cada restaurante está asociado con un valor de x (población estudiantil en miles de estudiantes) y un valor correspondiente de *y* (ventas trimestrales en miles de $). La ecuación que describe cómo se relaciona *y* con x y con un término de error se llama *modelo de regresión*. Éste usado en la regresión lineal simple es el siguiente:

Modelo de regresión lineal simple: $y = \beta_0 + \beta_1 x + \varepsilon$

β_0 y β_1 son los parámetros del modelo. ε es una variable aleatoria, llamada error, que explica la variabilidad en *y* que no se puede explicar con la relación lineal entre x y *y*.
Los errores, ε, se consideran variables aleatorias independientes distribuidas normalmente con media **cero** y desviación estándar **σ**. Esto implica que el valor medio o valor esperado de *y*, denotado por E(*Y*/x), es igual a $\beta_0 + \beta_1 x$.

Ecuación de regresión lineal simple: $E(y/x) = \beta_0 + \beta_1 x$ **(** $\mu_{Y/x}=E(Y/x)$ **)**



La ecuación estimada de regresión (lineal simple)
Los parámetros, β_0 y β_1, del modelo se estiman por los estadísticos muestrales b_0 y b_1, los cuales se calculan usando el método de **mínimos cuadrados.**

Ecuación Estimada de regresión lineal simple: $\hat{y} = b_0 + b_1 x$

En la regresión lineal simple, la gráfica de la ecuación de regresión se llama ***línea de regresión estimada***. $\hat{y}$ es el valor estimado de y para un valor específico de x.

Datos de población estudiantil y ventas trimestrales para una muestra de 10 restaurantes:

restaurante	Poblac. estudiantil (en miles) x_i	Ventas trimestrales (miles de $) y_i	
1	2	58	
2	6	105	
3	8	88	
4	8	118	
5	12	117	
6	16	137	
7	20	157	
8	20	169	
9	22	149	
10	26	202	

Diagrama de dispersión

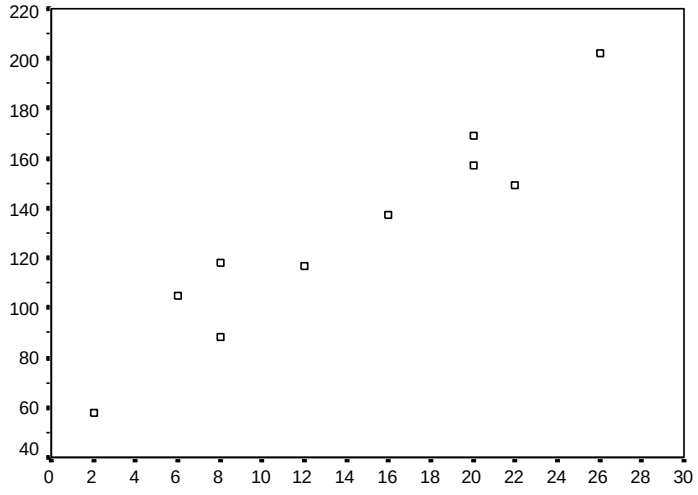
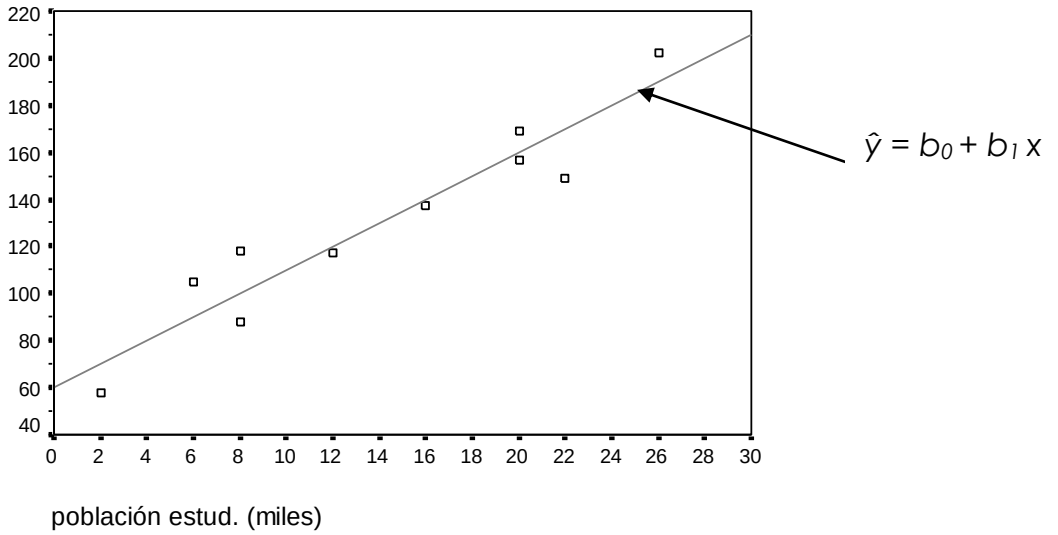


Diagrama de dispersión

(y línea de regresión estimada)



El método de mínimos cuadrados consiste en hallar los valores b_0 y b_1 que hacen mínima la *suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente*, y_i, y los valores estimados de la misma, $\hat{y}_i$. Es decir se minimiza la suma: $\Sigma(y_i - \hat{y}_i)^2$.

Al aplicar el método se llega al siguiente sistema de ecuaciones simultáneas (llamadas ecuaciones normales de la recta de regresión de y en x), cuya solución da los valores de b_0 y b_1:

$$\begin{cases} \Sigma y_i = nb_0 + (\Sigma x_i)b_1 \\ \Sigma x_i y_i = (\Sigma x_i)b_0 + (\Sigma x_i^2)b_1 \end{cases}$$

Las soluciones son las siguientes:

$$b_1 = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$ que también es $$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \frac{S_{XY}}{S_X^2}$$

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

Determine la ecuación de regresión con los datos dados.

b_1=

b_0=

$\hat{y}$ =

restaurante	x_i	y_i	$x_i y_i$	x_i^2
1	2	58		
2	6	105		
3	8	88		
4	8	118		
5	12	117		
6	16	137		
7	20	157		
8	20	169		
9	22	149		
10	26	202		
	140	1300	21040	2528

El coeficiente de determinación (r^2)

El coeficiente de determinación en la regresión lineal simple es una medida de la bondad de ajuste de la recta estimada a los datos reales.

Suma de cuadrados debida al error: $SCE = \Sigma(y_i - \hat{y}_i)^2$

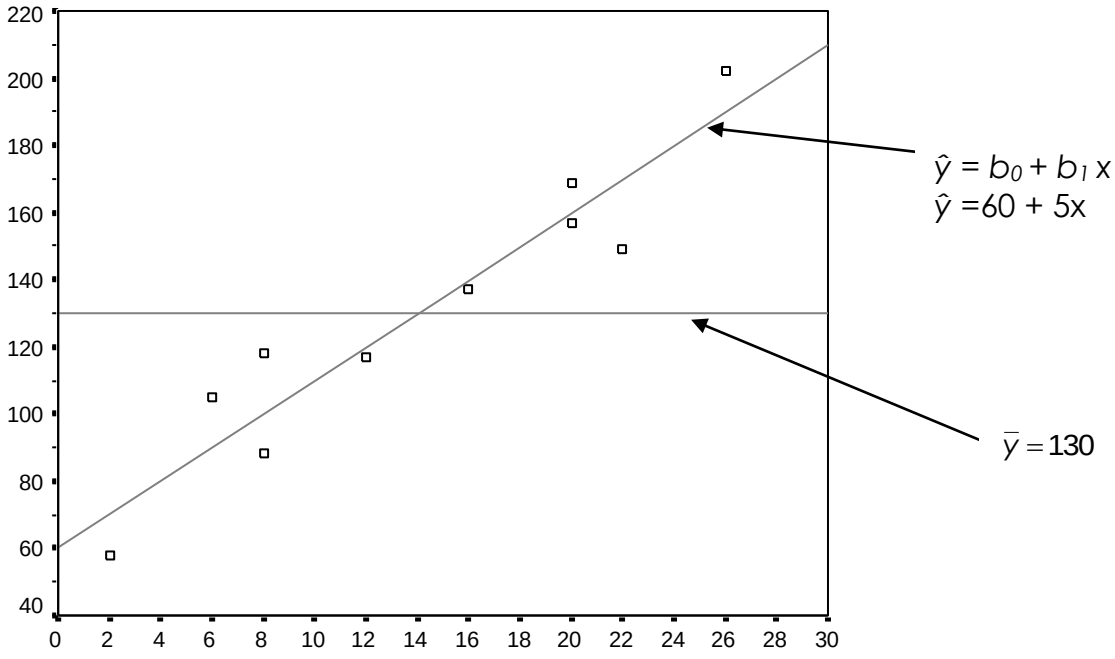
Suma de cuadrados total: $SCT = \Sigma(y_i - \bar{y})^2$

Suma de cuadrados debida a la regresión: $SCR = \Sigma(\hat{y}_i - \bar{y})^2$

Relación entre SCT, SCR y SCE: $SCT = SCR + SCE$

Coeficiente de determinación : $r^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$

Expresado **r^2** en porcentaje, se puede interpretar como **el porcentaje de la variabilidad total de "Y" que se puede explicar aplicando la ecuación de regresión**.



cálculo de SCE y SCT

restaurante	x_i *(poblac. estud)*	Y_i *(ventas trimest.)*	$\hat{y}_i = 60 + 5x_i$	Residuales $y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \overline{y}$ = $(y_i - 130)$	$(y_i - \overline{y})^2$ = $(y_i - 130)^2$
1	2	58					
2	6	105					
3	8	88					
4	8	118					
5	12	117					
6	16	137					
7	20	157					
8	20	169					
9	22	149					
10	26	202					
TOTALES	140	1,300			SCE=1,530		SCT=15,730

La suma de cuadrados debida a la regresión se calcula por diferencia:
SCR = SCT – SCE = 15,730 – 1,530 = 14,200
El coeficiente de determinación es entonces:

$$r^2 = \frac{SCR}{SCT} = 14{,}200/15{,}730 = 0.9027$$

El 90.27% de la variación en las ventas se puede explicar con la relación lineal entre la población estudiantil y las ventas.

El coeficiente de correlación lineal (r)

Es una medida descriptiva que mide la intensidad de asociación lineal entre las dos variables, *x* y *y*. Los valores del coeficiente de correlación lineal siempre están entre –1 y +1. –1 significa una relación lineal negativa perfecta, +1 significa una relación lineal positiva perfecta. Los valores cercanos a cero indican que las variables x y *y* no tiene relación lineal. El coeficiente de correlación lineal se relaciona con el coeficiente de determinación así:

$$\boxed{r = (\text{signo de } b_1)\sqrt{r^2}} \quad (1)$$

b_1 es la pendiente la recta de regresión de y en x.
El coeficiente de determinación es más general que el coeficiente de correlación lineal.

PRUEBAS DE SIGNIFICACIÓN PARA LA REGRESIÓN LINEAL

La ecuación de regresión lineal simple indica que el valor medio o valor esperado de *y* es una función lineal de x: $E(y/x) = \beta_0 + \beta_1 x$. Si $\beta_1 = 0$ entonces $E(y/x) = \beta_0$ y en este caso el valor medio no depende del valor de *x,* y concluimos que *x* y *y* no tienen relación lineal. En forma alternativa, si el valor $\beta_1 \neq 0$ llegamos a la conclusión que las dos variables se relacionan (más específicamente, que hay una componente lineal en el modelo). Existen dos pruebas, por lo menos, que se pueden utilizar para tal fin. En ambas se requiere una estimación de σ^2, la varianza de ε en el modelo de regresión.

---

(1) El coeficiente de correlación se define como $r = \frac{S_{XY}}{S_x S_Y}$; S_{XY} es la covarianza muestral y el denominador es el producto de las desviaciones típicas.

Cuadrados medios del error CME (es una estimación de σ^2)

$$S^2 = CME = SCE/(n-2)$$

n-2 son los grados de libertad asociados a SCE. 2 son los parámetros estimados en la regresión lineal (β_0 y β_1) y n es el número de pares de datos.

Error estándar de estimación (s)

Es la raíz cuadrada de s^2, $s = \sqrt{CME} = \sqrt{\frac{SCE}{n-2}}$ y es el estimador de la desviación estándar σ.

Distribución muestral de b_1

b_1 es un estadístico con distribución normal de media $\mu_{b1} = \beta_1$ y desviación estándar $\sigma_{b1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}}$. Si sustituimos σ por su estimación muestral, s, obtenemos un estimador de σ_{b1} que denotaremos por s_{b1}. $s_{b1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$. Con esta información podemos construir un estadístico t. $t = \frac{b_1 - \beta_1}{s_{b_1}}$ el cual se distribuye con ν=n-2 g.l.

Prueba t de significación en la regresión

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Estadístico de contraste bajo H_0, $\boxed{t_c = \frac{b_1 - 0}{s_{b_1}}}$

Decisión: Se rechaza H_0 en favor de H_1 si $|t_c| > t_{\alpha/2}$ o si p-valor < α

(Realizar la prueba con los datos del ejm propuesto).

Prueba de significancia usando el estadístico F (es una prueba más general)

Se usan dos estimaciones de σ^2, una basada en CME y la otra basada en CMR.

$$CME = \frac{SCE}{n-2} \quad y \quad CMR = \frac{SCR}{\text{número de variables independientes}} = \frac{SCR}{1}.$$

CME es un estimador insesgado de σ^2, mientras que CMR lo es sólo si H_0 es cierta. Si H_0 es falsa, CMR tiende a sobreestimar σ^2.

El estadístico de contraste, bajo H_0 es una F. F=CMR/CME con 1 gl en el numerador y n-2 en el denominador. Los datos se acomodan en una tabla ANOVA. Se rechaza H_0 en favor de H_1 si $F_c > F_\alpha$ o también si el p-valor correspondiente es menor que el nivel de significancia propuesto (α).

Tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	p-valor o sig.
Regresión	SCR	1	CMR	F=CMR/CME	
Error	SCE	n-2	CME		
total	SCT	n-1			

Realiza la prueba del ejemplo usando ANOVA.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	p-valor o sig.

Uso de la ecuación de regresión lineal para evaluar y predecir.

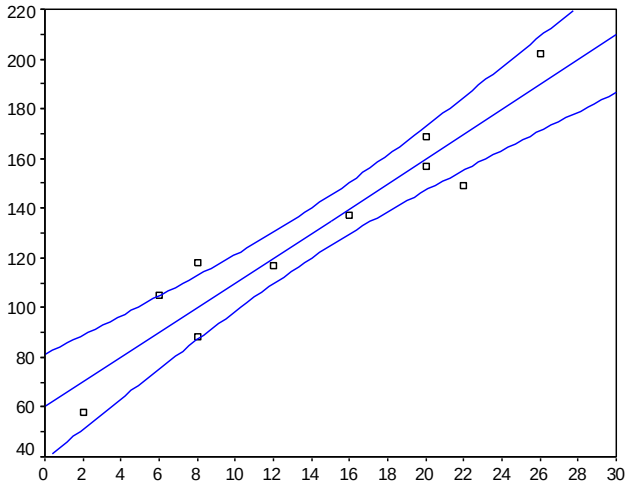
El modelo de regresión lineal simple es un supuesto acerca de la relación entre x y *y*. Si los resultados tienen una relación estadísticamente significativa entre x y *y,* y si el ajuste que proporciona la ecuación de regresión parece bueno, ésta podría utilizarse para estimaciones y predicciones.

Intervalo de confianza para estimar la media de *y* para un valor dado x_p de x.

$$\mu_{Y/X\,p} = E(y/x_p): \hat{y}_p \pm t_{\alpha/2}\ S\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

Intervalo de predicción para estimar un valor individual de Y para un valor dado x_p de x:

$$Y_p:\ \hat{y}_p \pm t_{\alpha/2}\ S\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$



Ejercicio:

a) Se desea estimar, mediante un intervalo del 95% de confianza, el promedio de venta trimestral para todos los restaurantes cercanos a centros escolares con 10,000 estudiantes:

$$\mu_Y:\ \hat{y}_p \pm t_{\alpha/2}\, S$$

$x_p = 10$; $\hat{y}_p = 60+5(10) = 110$; $\bar{x} = 140/10 = 14$; $\sum(x_i - \bar{x})^2 = 568$; $n = 10$;

$(x_p - \bar{x})^2 = (10-14)^2 = 16$; $S = \sqrt{CME} = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{1530}{8}} = 13.8293$; $t_{\alpha/2} = 2.306$

$\mu_{Y/x=10}$: 110 ± 11.415 miles de dólares.

b) Se desea predecir, mediante un intervalo del 95% de confianza, las ventas trimestrales para un restaurante que se construirá cercano a un centro estudiantil de 10,000 estudiantes :

$$Y_p:\ \hat{y}_p \pm t_{\alpha/2}\, S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Y_p: *110* ± 33.875 miles de dólares

Análisis de residuales: validación de los supuestos del modelo

Como ya se dijo, el **residual** en la observación i es la diferencia entre el valor observado de la variable dependiente (y_i) y el valor estimado de esa variable ($\hat{y}_i$).

Residual en observación i: $y_i - \hat{y}_i$

El análisis de residuales es la principal herramienta para determinar si es adecuado el modelo de regresión supuesto. $y = \beta_0 + \beta_1 x + \varepsilon$; ε es el término del error en el modelo, y se hacen los siguientes supuestos para él:

1. $E(\varepsilon) = 0$
2. La varianza de ε, representada por σ^2, es igual para todos los valores de x.
3. Los valores de ε son independientes.
4. El término del error, ε, tiene tendencia normal de probabilidad.

Estos supuestos forman la base teórica de las pruebas t y F que se usan para determinar si la relación entre x y Y es significativa, y para los estimados de intervalos de confianza y de predicción que ya se describieron.
El SPSS provee dos tipos de gráficos para determinar las característica de los residuales: Un gráfico de residuales en función de x o de $\hat{y}$, con el cual se puede analizar si la varianza es constante, y un gráfico de probabilidad normal. Generalmente se trabaja con los residuales estandarizados o tipificados.

Determinar estos gráficos para los datos del ejemplo de la pizería Armand.

Hay otros análisis para los residuales que permiten determinar **valores atípicos** y **observaciones influyentes** en los datos muestrales que por ahora no estudiaremos.

Modelos no lineales intrínsicamente lineales

Hay alguna tendencias que no son lineales pero con una adecuada transformación de variables se pueden transformar en lineales, por ejm tendencias exponenciales, potenciales, logarítmica, etc. El Spss tiene ésas y otras tendencias en el menú de regresión. Los siguientes ejercicios son de ese tipo:

1. Los siguientes datos se refieren al crecimiento de una colonia de bacterias en un medio de cultivo:

Días desde la inoculación x	Número de bacterias (en miles) y
3	115
6	147
9	239
12	356
15	579
18	864

   a) Trace $\ln(y_i)$ versus x_i para verificar que es razonable una curva exponencial.
   b) Ajuste una curva exponencial a los datos.
   c) Estime el número de bacterias al término de 20 días.

2. Los siguientes datos se refieren a la demanda de un producto (en miles de unidades) y su precio (en centavos) en cinco mercados diferentes:

Precio X	Demanda y
20	22
16	41
10	120
11	89
14	56

Ajuste una función potencial y úsela para estimar la demanda cuando el precio del producto es de 12 centavos.

3. Los siguientes datos se refieren al tiempo de secado de un cierto barniz y a la cantidad de aditivo añadido para reducir el tiempo de secado:

Cantidad de aditivo agregado (g) x	Tiempo de secado (horas) y
0	12.0
1	10.5
2	10.0
3	8.0
4	7.0
5	8.0
6	7.5
7	8.5
8	9.0

   a) Dibuje un diagrama de dispersión para verificar que es razonable suponer que la relación es parabólica.

   b) Ajuste un polinomio de segundo grado con el método de mínimos cuadrados.

Regresión múltiple

Fórmulas clave

Variables independientes $\aleph=(x_1,x_2,\ldots,x_p)$

Modelo de regresión múltiple $y=\beta_0 + \beta_1x_1 + \beta_2x_2 + \ . \ . \ . +\beta_px_p + \varepsilon$

Ecuación de regresión múltiple $\mu_{Y/\aleph}= E(y/ \aleph) =\beta_0 + \beta_1x_1 + \beta_2x_2 + \ . \ . \ . +\beta_px_p$

Ecuación de regresión múltiple estimada $\hat{Y} = b_0 + b_1x_1 + b_2x_2 + \ . \ . \ . +b_px_p$

Criterio de mínimos cuadrados $\min\Sigma(y_i - \hat{y}_i)^2$

Relación entre SCT, SCR y SCE SCT = SCR + SCE

Coeficiente de determinación múltiple $r^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$

Coeficiente de determinación múltiple ajustado $r_a^2 = 1-(1-r^2)\frac{n-1}{n-p-1}$

Cuadrado medio debido a la regresión $CMR = \frac{SCR}{p}$

Cuadrado medio del error $CME = \frac{SCE}{n-p-1}$

Estadístico de la prueba F $F = \frac{CMR}{CME}$

Estadístico de la prueba t $t = \frac{b_i}{S_{b_i}}$

Modelo de regresión múltiple

El análisis de regresión múltiple es el estudio de la forma en que una variable dependiente, **y**, se relaciona con dos o más variables independientes. En el caso general emplearemos **p** para representar la cantidad de variables independientes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ . \ . \ . + \beta_p x_p + \varepsilon$$

El término del error explica la variabilidad en "**y**" que no puede explicar las p variables independientes. El error es una variable aleatoria distribuida normalmente con media cero y varianza constante, σ^2, para todos los valores de las X_i.

Si consideramos el valor medio de la variable "**y**" dadas las variables independientes $\aleph=(x_1,x_2,\ldots,x_p)$, obtenemos la ecuación de regresión lineal

$$\mu_{Y/\aleph} = E(y/\aleph) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ . \ . \ . + \beta_p x_p$$

Utilizando los datos de una muestra de tamaño n y el método de mínimos cuadrados se determina la ecuación de regresión múltiple estimada:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ . \ . \ . + b_p x_p$$

Cada coeficiente b_i representa una estimación del cambio en "y" que corresponde a un cambio unitario en x_i cuando todas las demás variables independientes se mantienen constantes.

Coeficiente de determinación múltiple (r^2)

r^2 se interpreta como la proporción de la variabilidad de la variable dependiente que se puede explicar con la ecuación de regresión múltiple.

$$r^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

SCT: suma de cuadrados total $\Sigma(y_i - \bar{y})^2$

SCR: Suma de cuadrados debida a la regresión $\Sigma(\hat{y}_i - \bar{y})^2$

SCE: Suma de cuadrados debida al error $\Sigma(y_i - \hat{y}_i)^2$

Pruebas de significancia

Prueba F $\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1 : Uno\ o\ más\ de\ los\ parámetros\ no\ es\ cero \end{cases}$

α=0.05 es el nivel de significación de la prueba.

$F_C = \frac{CMR}{CME}$; CMR=SCE/p y CME=SCE/(n-p-1)

Se rechaza H_0 si el p-valor de F_C es menor que α.

Los resultados se acomodan en una tabla ANOVA.

Tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	p-valor o sig.
Regresión	SCR	p	CMR=(SCR/p)	F_C=CMR/CME	
Error	SCE	n-p -1	CME=(SCE/(n - p -1))		
total	SCT	n-1			

Prueba t para coeficientes individuales (β_i)

$$\begin{cases} \beta_i = 0 \\ \beta_i \neq 0 \end{cases}$$

$t_c = \frac{b_i}{S_{b_i}}$; con ν=n-p-1

Se rechaza H_0 si $|t_c| > t_{\alpha/2}$; o alternativamente, si **p-valor de t_c** es menor que α.

Multicolinealidad

En el análisis de regresión hemos empleado el término **variables independientes** para indicar cualquier variable que se usa para predecir o explicar el valor de la variable dependiente. Sin embargo, el término no indica que las variables independientes sean independientes entre sí en un sentido estadístico. Al contrario, la mayor parte de las variables independientes en un problema de correlación múltiple se correlacionan en cierto grado.
Tener un coeficiente de correlación de la muestra mayor que 0.70 o menor que -0.70 para dos variables independientes es una regla fácil para advertir la posibilidad de problemas por multicolinealidad.
Cuando las variables independientes están muy correlacionadas no es posible determinar el efecto separado de una de ellas sobre la variable dependiente.
Si es posible, se debe evitar incluir en el modelo, variables independientes que tengan mucha correlación. Sin embargo, en la práctica casi nunca es posible adherirse estrictamente a este criterio.

Empleo de la ecuación de regresión estimada para evaluar y predecir.

Podemos determinar intervalos de confianza para estimar la media de “y” e intervalos de predicción para estimar valores individuales de “y”.

Como ejemplo de análisis de regresión múltiple describiremos un problema que se presentó en la compañía Butler, una empresa dedicada a entregas de encomiendas. Para poder contar con mejores programas de trabajo, se desea estimar el tiempo diario total que viajan sus operarios. Se han considerado dos variables independientes que se cree que influyen en el tiempo diario total. A continuación se muestran los datos de una muestra de 10 recorridos:

Recorrido	millas recorridas (x1)	cantidad de entregas (x2)	tiempo de recorrido en horas (y)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Inicialmente analice el tiempo de recorrido en función de las millas recorridas y luego incorpore la cantidad de entregas en el modelo. En cada caso analice también la distribución de residuales.

¿Cuál es la ecuación de regresión estimada en cada caso?

¿Cómo interpreta los coeficientes de regresión en cada modelo?

¿Cómo interpreta el coeficiente de determinación múltiple r^2?

En general, r^2 aumenta siempre a medida que se agregan variables independientes al modelo. Hay muchas personas que prefieren ajustar r^2 de acuerdo con el número de variables independientes, para evitar una sobreestimación al agregar otras variables al modelo estudiado.

$$r_a^2 = 1-(1-r^2)\frac{n-1}{n-p-1}$$

¿Cuánto vale r_a^2 en el ejemplo?.

Adviértase que cuando r^2 es pequeño, el coeficiente ajustado puede asumir un valor negativo; en este caso el programa de computadora ajusta en cero el valor de ese coeficiente.

Estime, mediante un intervalo del 95% de confianza, la media del tiempo de viaje para todos los camiones que recorren 100 millas y hacen dos entregas.

Estime, mediante un intervalo del 95% de confianza, el tiempo de viaje para un camión que va a recorrer 100 millas y a hacer 2 entregas.

Variables independientes cualitativas

Como hemos visto, las variables involucradas en problema de regresión son todas variables numéricas tanto las independientes como la dependiente. Sin embargo, en muchas situaciones se debe incorporar al modelo variables cualitativas. El objetivo de esta sección es mostrar cómo se manejan este tipo de variables. Se crean unas variables llamadas **variables ficticias o indicadoras,** las cuales sólo pueden tomar dos valores, **0** y **1**.
Para ejemplificar el uso de estas variables consideremos el siguiente problema en la empresa **Jonson filtration,** la cual se dedica al servicio de mantenimiento de sistemas de filtrado de agua. Sus clientes se comunican solicitando servicio de mantenimiento en sus sistemas de filtrado de agua. Para estimar el tiempo y el costo de servicios, la gerencia desea predecir el tiempo necesario de reparación para cada solicitud de mantenimiento. Se cree que ese tiempo de reparación se relaciona con dos factores: la cantidad de meses transcurridos desde el último servicio y el tipo de reparación (mecánica o eléctrica). En la tabla se presentan los datos de una muestra de 10 órdenes de servicio:

órden de servicio	Meses desde el último servicio	Tipo de reparación	Tiempo de reparación (horas)
1	2	eléctrica	2.9
2	6	mecánica	3.0
3	8	eléctrica	4.8
4	3	mecánica	1.8
5	2	eléctrica	2.9
6	7	eléctrica	4.9
7	9	mecánica	4.2
8	8	mecánica	4.8
9	4	eléctrica	4.4
10	6	eléctrica	4.5

Desarrolle un modelo que explique el tiempo de reparación (Y) en función de los meses desde el último servicio (X_1) y del tipo de reparación (x_2).

$$Y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Haga un análisis de los resultados obtenidos, interprete los parámetros estimados.

Variables cualitativas más complejas

Si una variable cualitativa tiene más de dos niveles, se pueden definir varias variables indicadoras para resolver el problema. En general se necesitan k-1 variables indicadoras para incorporar una variable cualitativa con k niveles. Por ejm si una variable tiene 3 niveles o categorías (A, B y C) se pueden crear dos variables ficticias de la siguiente manera

$$x_1 = \begin{cases} 1 \text{ si es el nivel B} \\ 0 \text{ si es cualquier otro} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ si es el nivel C} \\ 0 \text{ si es cualquier otro} \end{cases}$$

Con esta definición tenemos los siguientes valores de x_1 y x_2.

categoría	x_1	x_2
A	0	0
B	1	0
C	0	1