

Estadística Inferencial  
Apuntes de clase  
Universidad Autónoma de Aguascalientes  
Lic. en Turismo

Paul Ramírez De la Cruz

Enero - Junio 2008



# Índice general

---

<b>1. Elementos de probabilidad</b>	<b>1</b>
1.1. Introducción . . . . .	1
1.2. Conceptos básicos de la teoría de conjuntos . . . . .	2
1.2.1. Subconjuntos y conjuntos notables . . . . .	4
1.3. Conceptos básicos de probabilidad . . . . .	9
1.4. Enfoques de la probabilidad . . . . .	12
1.5. Funciones de probabilidad . . . . .	15
1.6. Ejercicios . . . . .	17
1.7. Probabilidad condicional e independencia de eventos . . . . .	20
<b>2. Variables aleatorias</b>	<b>23</b>
2.1. Introducción . . . . .	23
2.2. Distribución de probabilidad . . . . .	25
2.2.1. Propiedades de una función de distribución de probabilidad discreta . . . . .	26
2.2.2. Propiedades de una función de probabilidad continua . . . . .	27
2.3. Función de distribución acumulada discreta . . . . .	28
2.4. Valor esperado y varianza de una variable aleatoria discreta . . . . .	29
<b>3. Algunas distribuciones de probabilidad</b>	<b>33</b>
3.1. Distribuciones discretas . . . . .	33
3.1.1. Distribución de probabilidad binomial . . . . .	33
3.1.2. Distribución de probabilidad geométrica . . . . .	35
3.2. Distribuciones continuas . . . . .	37
3.2.1. Distribución de probabilidad normal . . . . .	37
<b>4. Elementos de estadística inferencial</b>	<b>43</b>
4.1. Distribuciones muestrales . . . . .	43
4.1.1. Introducción . . . . .	43
4.1.2. Distribución t de Student . . . . .	45
4.1.3. Distribución ji-cuadrada . . . . .	45
4.1.4. Distribución F de Fisher . . . . .	45
4.1.5. Distribución muestral de la media . . . . .	45



# 1. ANÁLISIS DE EXPERIMENTOS

---

## 1.1. Motivación

¿Se puede ahorrar dando a los empleados gratificaciones en efectivo o certificados de regalo de la principal zona comercial?<sup>1 2</sup>

Suponga que la administradora del Hospital A, Helen Alston, desea examinar una iniciativa que busca disminuir de manera importante el número de días de ausentismo laboral, dirigida a los empleados de limpieza, abastecimiento y cuidados. A estos trabajadores que tienen registros de asistencia ininterrumpida por seis meses se les haría una gratificación en efectivo en su sueldo, o bien se les daría certificados de regalo de la principal zona comercial de la ciudad para que los utilicen en donde deseen como recompensa por simplemente asistir a su trabajo.

Helen considera que el interés de los empleados por la gratificación podría reducir los costos incurridos por contratar más personal, además de evitar las molestias de tener que hacerlo de última hora.

¿Cómo podría Helen obtener evidencia de su suposición? El diseño de experimentos nos permite proporcionar una solución estadística a este problema.

## 1.2. Conceptos básicos

El diseño de experimentos se refiere a un conjunto de métodos estadísticos que permite obtener observaciones mediante las cuales contrastar una hipótesis sobre un parámetro. Introduciremos los conceptos básicos del diseño de experimentos a partir del ejemplo del Hospital A. Para que Helen pueda establecer si la propuesta efectivamente reduciría costos necesita un punto de comparación, es decir, debe medir el costo incurrido por aplicar la nueva medida A (bonificación) o la nueva medida B (certificados de regalo) en comparación con el costo de seguir como antes. Para ello se requiere de tres muestras de empleados: a unos se les aplicará la medida A, a otros la medida B, y a los otros se les dejará con el esquema actual. En este caso se dice que existe un único factor con tres niveles o tratamientos (la aplicación de la nueva medida A, la B, o el

---

<sup>1</sup>En este documento las cantidades numéricas están expresadas en “estilo europeo”, por lo cual el separador de miles es un punto (.), mientras que el separador decimal es una coma (,).

<sup>2</sup>Adaptado de [?]

esquema actual).

Consideremos las siguientes definiciones <sup>3</sup>:

**Definición 1 (Experimento)** *Prueba o examen práctico que se realiza para probar la eficacia de una cosa o examinar sus propiedades.*

**Definición 2 (Unidad experimental)** *Es un individuo o cosa al cual se aplica el experimento. Se requiere de varias unidades experimentales para conducir un buen experimento. Tratándose de personas, a las unidades experimentales se les llama también sujetos.*

**Definición 3 (Factor)** *Un factor de un experimento es una variable independiente controlada, cuyos distintos valores o niveles los determina el experimentador.*

**Definición 4 (Tratamiento)** *Es cualquier cosa que el experimentador administra o aplica a las unidades experimentales. Los tratamientos se dan a las unidades experimentales en diferentes niveles, donde un nivel implica una cantidad o magnitud.*

*Un tratamiento es la combinación de una o más variables independientes o factores.*

En nuestro ejemplo hay un tratamiento con tres niveles: dar bonificación en efectivo, dar certificados de regalo, o bien, continuar como hasta ahora (no dar certificados de regalo). Las variables independientes producen un resultado o respuesta, que es la variable que se mide para establecer el efecto de los tratamientos.

Diseñar experimentos adecuados implica estar al tanto de todos los posibles factores que pueden influir en el resultado de modo que se pueda introducir al experimento los que son de interés al tiempo que se controla los que no lo son para que no afecten el resultado del experimento.

Al diseñar experimentos en donde intervienen personas, debe considerarse la complejidad del comportamiento humano además de tener en cuenta cuestiones éticas. Al conducir experimentos también debe tenerse cuidado con la falta de realismo, es decir, con la sobresimplificación de los posibles factores que pueden causar la respuesta.

Los principios básicos en el diseño de experimentos son el control, la aleatorización y la replicación.

### 1.2.1. Control

Los experimentos permiten estudiar los efectos de los tratamientos de interés al tiempo que se controla el medio ambiente de las unidades experimentales de modo que se puede mantener constantes los efectos causados por otros factores que no son de interés.

---

<sup>3</sup>Definiciones tomadas de [?] y [?]

La forma más simple de control consiste en comparar los tratamientos, lo cual incluye comparar tratados contra no tratados.

Las dos formas básicas de realizar la comparación son:

- Aplicación de distintos tratamientos a distintos grupos de unidades experimentales o comparación entre sujetos. Esto implica realizar la comparación de tres o más muestras independientes. Debe tratarse de que los distintos grupos sean tan parecidos entre sí como sea posible en todos sentidos, excepto por el tratamiento que reciben. Debe existir un grupo que no reciba tratamiento al cual se le conoce como grupo de control.
- Aplicación de los distintos tratamientos a las mismas unidades experimentales o comparación dentro de los sujetos. A cada unidad experimental se le hace la aplicación de todos los tratamientos. Esto implica realizar la comparación de muestras pareadas.

### 1.2.2. Aleatorización

Los resultados de comparar los efectos de distintos tratamientos son confiables solamente si todos los tratamientos se aplicaron a grupos similares de unidades experimentales.

La manera más simple de lograr esto es hacer intervenir la aleatoriedad en la asignación de las unidades experimentales a los distintos tratamientos. A esto se le llama aleatorización.

### 1.2.3. Replicación

Un experimento que funciona en un individuo o un objeto no es prueba suficiente de que funcione otra vez. Se requiere de repetir o replicar el experimento varias veces. Es necesario utilizar una cantidad suficiente de unidades experimentales a fin de que sea posible sacar conclusiones, desde el punto de vista estadístico, con respecto a los resultados observados.

### 1.2.4. Variables ocultas y de confusión

En una situación ideal, los resultados de un experimento dependen únicamente de las variables que se se está probando. En la práctica, sin embargo hay dos tipos de variables que afectan los resultados:

**Definición 5 (Variable oculta)** *Es aquella que tiene un efecto importante en los resultados, pero de la cual no se tiene conocimiento o bien se le conoce pero se le ha considerado irrelevante para el experimento, por lo cual no se le incluyó entre las variables a observar.*

**Ejemplo 6** *Supongamos que la administradora del Hospital A, Helen Alston comienza su programa de bonificaciones, pero al mismo tiempo comienza un sistema de rotación de turnos que permite al personal tener tiempo libre cuando más lo requiere y conocer su rol de horarios mensual por anticipado.*

*El sistema de rotación es una variable oculta porque puede afectar los resultados observados. Si el personal comienza a asistir con mayor regularidad a sus labores será difícil determinar qué tanto esto fue por el sistema de bonificaciones y qué tanto por el sistema de rotación de turnos.*

**Definición 7 (Variable de confusión)** *Es una variable cuya inclusión en el experimento distorsiona la medida de asociación entre otras dos variables. La presencia de una variable de confusión puede causar:*

- **Confusión positiva.** *Situación en que se observa un efecto en donde realmente no lo existe, o bien se exagera una asociación real entre variables.*
- **Confusión negativa.** *Ocurre cuando se atenúa una asociación real, o incluso se invierte el sentido de una asociación real entre variables.*

**Definición 8 (Diseño completamente aleatorizado)** *Es un experimento en el cual se asigna los tratamientos a las unidades experimentales de manera aleatoria.*

**Ejemplo 9** *Supongamos que la administradora del hospital tiene la nómina a la mano y de allí aleatoriamente selecciona un grupo de personas para aplicar el programa de bonificaciones A, un grupo para las bonificaciones B y un grupo que no recibirá gratificaciones.*

### 1.3. Análisis de varianza unifactorial para un diseño completamente aleatorizado

**Definición 10 (Análisis de varianza)** *El análisis de varianza (llamado ANOVA, por ANalysis Of Variance; o ANVA, por ANálisis de VArianza, como la castellanización del anterior) es un método estadístico para analizar los efectos de una o más variables independientes (que pueden medirse en escalas nominal ordinal o dicotómica) sobre una variable dependiente que se distribuye Normal (y por tanto es continua), así como los efectos entre las variables independientes.*

A continuación estableceremos los elementos básicos del ANVA.

En una situación general, se tiene  $k$  poblaciones. Las poblaciones están definidas por grupos existentes en las unidades experimentales, o bien por el tratamiento que se les asigna en el experimento. De cada población se toma una muestra aleatoria de tamaño  $n$ . Se asume que las  $k$  poblaciones son independientes y tienen distribución Gaussiana o Normal con medias  $\mu_1, \mu_2, \dots, \mu_k$  y una varianza común  $\sigma^2$ . La aleatorización permite suponer que estos requisitos se cumplen.

Nos interesa contrastar las hipótesis:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_a : \text{Al menos dos de las medias no son iguales}$$

**Notación 11** Utilizaremos las siguientes literales:

- $y_{ij}$  es la  $j$ -ésima observación del  $i$ -ésimo tratamiento (por ejemplo,  $y_{21}$  es la primera observación del tratamiento número dos,  $y_{32}$  es la segunda observación del tercer tratamiento, etc.)
- $\bar{y}_i$  es la media de las observaciones de la muestra en el  $i$ -ésimo tratamiento.
- $T_i$  es el valor total de las observaciones en la muestra correspondientes al  $i$ -ésimo tratamiento.
- $T_{..}$  es el valor total de todas las observaciones.
- $\bar{y}$  es la media de todas las observaciones.

Entonces podemos hacer un resumen de las observaciones como se indica en la tabla a continuación:

Obs.	Tratamiento						Total	
	1	2	...	$i$	...	$k$		
1	$y_{11}$	$y_{21}$	...	$y_{i1}$	...	$y_{k1}$		
2	$y_{12}$	$y_{22}$	...	$y_{i2}$	...	$y_{k2}$		
⋮	⋮	⋮	...	⋮	...	⋮		
$n$	$y_{1n}$	$y_{2n}$	...	$y_{in}$	...	$y_{kn}$		
Media	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_i$	...	$\bar{y}_k$		$\bar{y}_{..}$
Total	$T_1$	$T_2$	...	$T_i$	...	$T_k$		$T_{..}$

Cada observación puede escribirse de la forma

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

donde  $\varepsilon_{ij}$  es la desviación de la observación  $j$ -ésima de la  $i$ -ésima muestra de la correspondiente media del tratamiento.

El contraste de hipótesis se basa en una comparación de dos estimaciones independientes de la varianza poblacional común,  $\sigma^2$ . Tales estimaciones se obtienen de separar la variabilidad total de los datos en dos componentes.

### 1.3.1. Identidad de suma de cuadrados

La variabilidad total en las observaciones se puede separar en dos partes, una que representa la variabilidad debida a los tratamientos y otra que representa la variabilidad que no es debida a los tratamientos (Véase [?]).

- A la variabilidad total se le llama Suma de Cuadrados Total, se le denota por  $SCT$  y, cuando todas las muestras son de tamaño  $n$ , se calcula mediante la expresión

$$SCT = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

- La variabilidad debida a los tratamientos se conoce como Suma de Cuadrados de los Tratamientos, se denota por  $SCA$  y, cuando todas las muestras son de tamaño  $n$ , está dada por

$$SCA = n \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2$$

- A la diferencia entre  $SCT$  y  $SCA$  se le considera como la parte de la variación presente en los datos que no se explica por la aplicación de los tratamientos. Se le llama Suma de Cuadrados del Error, se representa por  $SCE$  y se calcula como

$$SCE = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2,$$

también suponiendo que todas las muestras son del mismo tamaño,  $n$ .

- La identidad de la suma de cuadrados establece que

$$SCT = SCA + SCE.$$

### 1.3.2. Fórmulas simplificadas para el cálculo de las sumas de cuadrados

El análisis estadístico de datos ha tenido un gran desarrollo a partir de la segunda mitad del siglo XX con el uso de sistemas de cómputo y programas cada vez más potentes; sin embargo antes de dichos avances la mayoría de los cálculos se realizaba a mano. Por ello es común que los textos de referencia ofrezcan expresiones alternativas que facilitan la obtención de dichos resultados cuando no se dispone de equipo de cómputo. Las presentamos a continuación. En todas las expresiones que siguen, consideramos que  $N = n_1 + n_2 + \dots + n_k$ :

1. Cálculo simplificado o alternativo de la Suma de Cuadrados Total:

a) Cuando las muestras son de tamaños  $n_1, n_2, \dots, n_k$ , respectivamente,

$$SCT = \sum_{i=1}^k \sum_{j=1}^n y_{ij}^2 - \frac{T_{..}^2}{N}$$

b) Si todas las muestras son de tamaño  $n$ , la expresión anterior se simplifica a

$$SCT = \sum_{i=1}^k \sum_{j=1}^n y_{ij}^2 - \frac{T_{..}^2}{nk}$$

2. Cálculo simplificado o alternativo de la Suma de Cuadrados de los Tratamientos:

a) Si las muestras son de tamaños  $n_1, n_2, \dots, n_k$ , respectivamente,

$$SCA = \sum_{i=1}^k \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$$

b) Cuando todas las muestras son de tamaño  $n$ , también se puede utilizar

$$SCA = \frac{1}{n} \sum_{i=1}^k T_{i.}^2 - \frac{T_{..}^2}{nk}$$

3. Cuando hemos calculado  $SCT$  y  $SCA$ , el cálculo alternativo de la Suma de Cuadrados del Error es muy simple:

$$SCE = SCT - SCA.$$

Notemos que, desde luego, este cálculo es el mismo ya sea que todas las muestras sean del mismo tamaño o no.

### 1.3.3. Cuadrados medios

Se le llama Cuadrado Medio a cada una de las dos estimaciones de  $\sigma^2$  que se calcula a partir de los datos:

▪ Cuadrado medio del tratamiento

$$s_1^2 = \frac{SCA}{k-1}$$

A  $s_1^2$  también se le denota por  $CMA$ . Cuando  $H_0$  es verdadera,  $s_1^2$  es un estimado insesgado de  $\sigma^2$ . Cuando  $H_0$  es falsa,  $s_1^2$  sobreestima el valor de  $\sigma^2$ .

■ Cuadrado medio del error

- Cuando las muestras son de tamaños  $n_1, n_2, \dots, n_k$  y  $N = n_1 + n_2 + \dots + n_k$ ,

$$s^2 = \frac{SCE}{N - k}$$

- Cuando cada una de las muestras es de tamaño  $n$ , también podemos calcularlo como

$$s^2 = \frac{SCE}{k(n - 1)}.$$

A  $s^2$  se le representa también como *CME*. El estadístico  $s^2$  siempre es un estimador insesgado de  $\sigma^2$ , aunque  $H_0$  no sea verdadera.

### 1.3.4. Estadístico de prueba del análisis de varianza

Cuando  $H_0$  es verdadera, el estadístico

$$F_{Calc} = \frac{s_1^2}{s^2}$$

sigue la distribución  $F$  de Fisher-Snedecor con

- $k - 1$  g.l. en el numerador y  $N - k$  g.l. en el denominador, si las muestras son de tamaño  $n_1, n_2, \dots, n_k$  y  $N = n_1 + n_2 + \dots + n_k$  (o bien,  $k - 1$  grados de libertad (g.l.) en el numerador y  $k(n - 1)$  g.l. en el denominador, si cada muestra es de tamaño  $n$ ).
  - Recordemos que cuando  $H_0$  es falsa,  $s_1^2$  sobreestima a  $\sigma^2$ , mientras que  $s^2$  sigue siendo un estimador insesgado de  $\sigma^2$ . Esto implica que si  $H_0$  es falsa el cociente  $\frac{s_1^2}{s^2}$  aumenta de valor. Por tanto, mientras mayor sea el valor de  $F_{Calc}$ , más dudaremos de la veracidad de  $H_0$ , y en consecuencia:
1. Considerando el enfoque de contraste de hipótesis de Neyman-Pearson, rechazaremos  $H_0$  al nivel de significancia  $\alpha$  si

$$F_{Calc} > F_{k-1, k(n-1), \alpha}$$

donde  $F_{k-1, k(n-1), \alpha}$  es el cuantil que deja una probabilidad igual a  $\alpha$  en la cola derecha de la distribución  $F$  con  $k - 1$  y  $k(n - 1)$  g.l. (valor de tablas)

**Observación 12 (¿Cómo elegir el valor de  $\alpha$ ?)** Recordemos que el valor de  $\alpha$  representa la probabilidad de cometer un Error Tipo I (rechazar una hipótesis nula verdadera) y este es un riesgo que queremos mantener tan pequeño como sea posible. Por tanto,  $\alpha = 0,05$  y  $\alpha = 0,01$  son

valores de uso convencional. En caso de que la aplicación no especifique el valor de  $\alpha$ , se puede elegir cualquiera de los dos anteriores, haciendo la aclaración al momento de calcular la región de rechazo y el valor de tablas de  $F$ .

2. Con el enfoque de contraste de hipótesis de Fisher, podemos calcular el Valor  $- p$  asociado al valor del estadístico:

$$\text{Valor} - p = P(F_{k-1, k(n-1)} \geq F_{Calc})$$

y entonces decir que existe evidencia estadística significativa contra  $H_0$  si  $\text{Valor} - p \leq 0,5$ , o bien, que existe evidencia estadística altamente significativa contra  $H_0$  si  $\text{Valor} - p \leq 0,1$ .

### 1.3.5. Tabla resumen del análisis de varianza para un solo criterio de clasificación

De acuerdo con lo expuesto anteriormente, podemos resumir el procedimiento de análisis de varianza cuando se tiene un solo criterio de clasificación (factor) y todas las muestras son del mismo tamaño en una tabla como la siguiente:

Cuando se realiza análisis de varianza de un solo factor con tamaños de muestra  $n_1, n_2, \dots, n_k$  y  $N = n_1 + n_2 + \dots + n_k$ , tenemos:

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Tratamientos	$SCA$	$k - 1$	$s_1^2 = \frac{SCA}{k-1}$	$F_{Calc} = \frac{s_1^2}{s^2}$
Error	$SCE$	$N - k$	$s^2 = \frac{SCE}{N-k}$	
Total	$SCT$	$N - 1$		

Si cada muestra tiene  $n$  observaciones, entonces también podemos utilizar la siguiente versión de la tabla de ANVA.

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Tratamientos	$SCA$	$k - 1$	$s_1^2 = \frac{SCA}{k-1}$	$F_{Calc} = \frac{s_1^2}{s^2}$
Error	$SCE$	$k(n - 1)$	$s^2 = \frac{SCE}{k(n-1)}$	
Total	$SCT$	$nk - 1$		

**Ejemplo 13** Con el fin de investigar el efecto de la altura de los estantes en un supermercado sobre las ventas de los alimentos para perro Arf, se llevó a cabo un experimento consistente en utilizar tres niveles para el estante: a la rodilla, a la cintura y a los ojos. Durante un periodo de 8 días, la altura del estante se cambió aleatoriamente en tres ocasiones cada día. A las secciones restantes de la góndola que contenía la marca de interés se les llenó con latas de otras marcas de alimento para perro que eran familiares y no familiares para los compradores de la región <sup>4</sup>.

<sup>4</sup>Tomado de [7]

La tabla siguiente muestra las ventas, en cientos de dólares, para cada nivel del estante en cada uno de los ocho días. ¿Existe evidencia estadística contra la suposición de que los valores promedio de venta son iguales para los distintos niveles en los que se colocó el estante?

Observación	Altura del estante		
	Rodilla	Cintura	Ojos
1	77	88	85
2	82	94	85
3	86	93	87
4	78	90	81
5	81	91	80
6	86	94	79
7	77	90	87
8	81	87	93

**Solución 14** Podemos ver que la tabla de análisis de varianza resultante es<sup>5</sup>:

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Tratamientos	399,25	2	199,625	$F_{Calc} = \frac{199,625}{13,75} = 14,5182$
Error	288,75	21	13,75	
Total	688	23		

A partir de la tabla anterior, podemos ver que  $Valor - p = 0,000110$ , por tanto existe evidencia estadística en contra de la igualdad de los promedios de venta para los tres niveles en que se colocó el estante.

**Ejercicio 15** Se dividió un grupo de doce parcelas en tres grupos. A los dos primeros grupos se les aplicó los fertilizantes A y B, en tanto que el tercero es un grupo de control al que no se le aplica ningún fertilizante. Las producciones fueron como se presenta a continuación [?]:

A	B	C
75	74	60
70	78	64
66	72	65
69	68	55

**Ejercicio 16** Supóngase que una muestra aleatoria de las utilidades anuales de 5 franquicias, llevada a cabo en varias ciudades en 1985, produjo los siguientes precios (en miles de dólares) [?, modificado]:

Ciudad	Observación				
	1	2	3	4	5
Boston	110	160	93	206	171
Indianápolis	72	38	45	108	42
Rochester	88	66	112	47	52
San Diego	57	81	181	165	106

<sup>5</sup>Para realizar los cálculos se utilizó la aplicación que se encuentra en <http://faculty.vassar.edu/lowry/anova1u.html>

### 1.3.6. Inferencias acerca de pares de valores medios de tratamiento

El contraste de hipótesis basado en el análisis de varianza de una vía nos indica, en el caso de que se rechace la hipótesis nula, que al menos dos de los tratamientos proporcionan valores distintos para el promedio de la variable de interés; pero no nos dice cuáles son las medias que producen la diferencia. Una forma de establecer cuáles medias son distintas entre sí consiste en calcular un intervalo de confianza para la diferencia de medias.

Si comparamos la media del tratamiento  $i$  con la del tratamiento  $j$ , la expresión del intervalo de confianza de  $100(1 - \alpha)\%$  para la verdadera diferencia entre las medias poblacionales de los tratamientos,  $\mu_i - \mu_j$ , es

$$(\bar{x}_i - \bar{x}_j) \pm t_{N-k, \alpha/2} \sqrt{CME \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

donde

$\bar{x}_i$  = la media del tratamiento  $i$

$\bar{x}_j$  = la muestra del tratamiento  $j$

$k$  = Número de tratamientos

$N = n_1 + n_2 + \dots + n_k$  = Número total de observaciones

$n_i$  = Número de unidades experimentales al que se le aplicó el tratamiento  $i$

$n_j$  = Número de unidades experimentales al que se le aplicó el tratamiento  $j$

$t_{N-k, \alpha/2}$  = Valor del cuantil que deja una probabilidad de  $\alpha/2$  en la cola derecha de la distribución  $t$  de Student con  $N - k$  grados de libertad

$CME = \frac{SCE}{N-k}$  = Cuadrado medio del error.

La conclusión sobre si existe diferencia entre las medias de los tratamientos comparados, se formula con base en lo siguiente:

- Si el intervalo de confianza calculado incluye al cero, entonces decimos que no hay evidencia de diferencia entre las medias de los grupos considerados.
- Si el intervalo de confianza no incluye al cero, entonces decimos que existe evidencia, al nivel  $\alpha$ , de que las medias de los tratamientos considerados son distintas.

**Ejemplo 17** *Los siguientes datos representan el costo de colegiaturas (en miles de dólares) de una muestra de universidades privadas en diversas regiones de Estados Unidos de América.*

1. *Al nivel de significancia de 0.05, ¿puede concluirse que existe alguna diferencia en el costo promedio de las colegiaturas?*
2. *Obtenga un intervalo de 95 por ciento de confianza para la diferencia entre las medias de las regiones noreste y oeste. ¿Son diferentes las medias?*

Noreste	Sureste	Oeste
10	8	7
11	9	8
12	10	6
10	8	7
12		6

**Solución 18** Las hipótesis son:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_a : \text{Al menos dos medias son distintas entre sí}$$

La región de rechazo la forman los valores del estadístico de prueba que son mayores que  $F_{k-1, N-k, \alpha} = F_{2, 11, 0, 05} = 3,98$  (de tablas). Por tanto, rechazaremos  $H_0$  si el valor del estadístico de prueba,  $F_{Calc} = \frac{CMA}{CME}$ , es mayor que el valor de tablas, 3,98.

La tabla de análisis de varianza es la siguiente:

(Coloque la tabla de análisis de varianza)

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Región (tratamientos)	44,164	2	22,082	$F_{Calc} = \frac{22,082}{0,868} = 25,44$
Error	9,550	11	0,868	
Total	53,714	13		

El valor de tablas es  $F_{2, 11, 0, 05} = 3,982$ . Luego, como  $F_{Calc} = 25,44 > 3,982 = F_{2, 11, 0, 05}$ , rechazamos  $H_0$ . Por tanto existe evidencia estadística, al nivel  $\alpha = 0,05$ , de que los costos promedios de las colegiaturas son distintos en las diferentes regiones examinadas.

Ahora examinaremos si existe diferencia entre los promedios de las regiones noroeste y oeste mediante un intervalo de confianza. Para ello recordamos que la expresión es

$$(\bar{x}_i - \bar{x}_j) \pm t_{N-k, \alpha/2} \sqrt{CME \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

de donde tenemos que los límites del intervalo de confianza son:

$$\begin{aligned} \text{Límite Inferior de Confianza} &= LIC = (\bar{x}_i - \bar{x}_j) - t_{N-k, \alpha/2} \sqrt{CME \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \\ &= (11 - 6,8) - 1,796 \sqrt{0,868 \left( \frac{1}{5} + \frac{1}{5} \right)} \\ &= 3,1417. \end{aligned}$$

y

$$\begin{aligned} \text{Límite Superior de Confianza} &= LSC = (\bar{x}_i - \bar{x}_j) + t_{N-k, \alpha/2} \sqrt{CME \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \\ &= (11 - 6,8) + 1,796 \sqrt{0,868 \left( \frac{1}{5} + \frac{1}{5} \right)} \\ &= 5,2583. \end{aligned}$$

Luego el intervalo de 95 % de confianza para la verdadera diferencia entre los costos promedio en las regiones noroeste y oeste es (3,1417, 5,2583), y como dicho intervalo no contiene al cero, existe evidencia estadística, al nivel de significación de 0.05, de que los costos promedio en las regiones noroeste y oeste son distintos.

**Ejemplo 19** Una egresada de contaduría tiene ofertas de trabajo de cuatro empresas. Para examinar un poco más las propuestas, solicitó a una muestra de personas de nuevo ingreso decirle cuántos meses trabajó cada una para su compañía, antes de recibir un aumento de sueldo. La información muestral es:

<i>CPA, Inc.</i>	<i>AB Intl.</i>	<i>Acct Ltd</i>	<i>Pfisters</i>
12	14	18	12
10	12	12	14
14	10	16	16
12	10		

La tabla de análisis de varianza de una vía resultante es  
(Coloque Tabla de ANVA)

de donde se concluye que existe diferencia entre los tiempos promedios de promoción laboral. ¿Se puede decir que existe diferencia entre los tiempos promedio de CPA Inc. y Acct Ltd.? Base su respuesta en el cálculo de un IC de 99 % para la diferencia de medias.

### 1.3.7. Ejercicios

Los siguientes ejercicios fueron tomados de [?].

1. Durante un determinado curso se somete a cuatro grupos de estudiantes a diferentes técnicas de enseñanza y se les examina al final del curso obteniéndose los siguientes rendimientos:

A	B	C	D
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

La tabla de resumen del ANVA es:

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Técnicas (tratamientos)	712.6	3	237.53	$F_{Calc} = \frac{237.53}{62.98} = 3,77$
Error	1196.6	19	62.98	
Total	1909.2	22		

A partir de la información anterior:

- Elabore una gráfica de cintas para los datos.
- Realice el contraste de hipótesis y responda: ¿Existe evidencia que indique una diferencia significativa en el rendimiento medio de las cuatro técnicas de enseñanza? Utilice  $\alpha = 0,05$ .
- Encuentre un intervalo de confianza del 95 % para la diferencia entre los rendimientos medios de las técnicas C y D. ¿Puede decirse que son distintos dichos rendimientos medios?

## 1.4. Análisis de varianza para un diseño aleatorizado por bloques

Aquí faltan las notas sobre este tema.

## 1.5. Análisis de varianza para un diseño con dos factores

En la sección anterior, conocimos los conceptos básicos del diseño de experimentos con un solo factor. Hay ocasiones en las que se requiere controlar más de un factor, con lo cual, además, se puede obtener mayor información del fenómeno. En esta sección hablaremos sobre los diseños experimentales en los

que el experimentador controla dos variables independientes o factores, por lo cual se les llama diseños bifactoriales.

Consideremos el siguiente ejemplo<sup>6</sup>, en el que un fabricante de ropa que suministra uniformes a un hotel debe cortar faldas, camisas y pantalones en varias tallas, de rollos de tela. El desperdicio de tela tiene un efecto importante en las utilidades. El fabricante tiene que elegir entre tres máquinas cortadoras asistidas por computadora. Para ello, decide conducir un experimento en donde controla dos variables o factores: la máquina y el tipo de prenda. En dicho experimento, hace que cada máquina corte varios lotes de faldas, otros de camisas y varios más de pantalones.

Supongamos que los resultados del experimento produjeron los siguientes porcentajes de desperdicio promedio:

Factor 1 (máquina)	Factor 2 (Tipo de prenda)			
	Faldas	Camisas	Pantalones	Promedio
<b>A</b>	7.6	9.1	7.3	<b>8.0</b>
<b>B</b>	6.5	8.0	6.2	<b>6.9</b>
<b>C</b>	5.1	6.6	4.8	<b>5.5</b>
<b>Promedio</b>	<b>6.4</b>	<b>7.9</b>	<b>6.1</b>	<b>6.8</b>

Notemos que hay un patrón persistente en esta tabla. La máquina A (la peor) siempre tiene más desperdicio que la máquina B (1.1 puntos porcentuales por arriba), sin importar qué tipo de prenda corte. La máquina B, a su vez, siempre tiene 1.4 puntos porcentuales de desperdicio más que la máquina C.

(Falta parte del ejemplo)

### 1.5.1. Identidad de suma de cuadrados

En un experimento con dos factores, se supone que la variación total en las observaciones (Suma de Cuadrados Total, SCT) se divide en tres partes:

- La variación debida al primer factor (también se le llama factor en los renglones),  $SCA$
- La variación debida al segundo factor (también se le llama factor en las columnas),  $SCB$
- La variación debida a la interacción entre los factores 1 y 2,  $SCAB$
- La variación debida al error,  $SCE$

La identidad de suma de cuadrados es ahora

$$SCT = SCA + SCB + SCAB + SCE$$

---

<sup>6</sup>Tomado de [?, pp. 520 - 521] (modificado)

1.5.2. Expresiones para las sumas de cuadrados en un diseño con dos factores con interacción

Las expresiones para las sumas de cuadrados ahora son:

- Suma de cuadrados total

$$SCT = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}^2 - \frac{T_{...}^2}{abr}$$

donde

$y_{ijk}$  =  $k$  – *ésima* réplica del  $i$  – *ésimo* nivel del factor 1 y el  $j$  – *ésimo* nivel del factor 2

$T_{...}$  = Total global

$a$  = Número de niveles del factor 1

$b$  = Número de niveles del factor 2

$r$  = Número de réplicas (observaciones) en cada tratamiento

- Suma de cuadrados del tratamiento 1

$$SCA = \frac{1}{br} \sum_{i=1}^a T_{i..}^2 - \frac{T_{...}^2}{abr}$$

donde

$T_{i..}$  = Total del  $i$  – *ésimo* nivel del factor 1

- Suma de cuadrados del tratamiento 2

$$SCB = \frac{1}{ar} \sum_{j=1}^b T_{.j.}^2 - \frac{T_{...}^2}{abr}$$

donde

$T_{.j.}$  = Total del  $j$  – *ésimo* nivel del factor 2

- Suma de cuadrados de la interacción

$$SCAB = \frac{1}{r} \sum_{i=1}^a \sum_{j=1}^b T_{ij.}^2 - \frac{1}{br} \sum_{i=1}^a T_{i..}^2 - \frac{1}{ar} \sum_{j=1}^b T_{.j.}^2 - \frac{T_{...}^2}{abr}$$

donde

$T_{ij.}$  = Total del  $ij$  – *ésimo* tratamiento (es decir, aquel en donde el factor 1 está en el nivel  $i$  y el factor 2 en el nivel  $j$ )

Fuente de variación	gl	SC	CM	F
Factor 1	$a - 1$	SCA	$CM_A = \frac{SCA}{a-1}$	$F_{Calc1} = \frac{CM_A}{CME}$
Factor 2	$b - 1$	SCB	$CM_B = \frac{SCB}{b-1}$	$F_{Calc2} = \frac{CM_B}{CME}$
Interacción	$(a - 1)(b - 1)$	SCAB	$CM_{AB} = \frac{SCAB}{(a-1)(b-1)}$	$F_{Calc3} = \frac{CM_{AB}}{CME}$
Error	$ab(r - 1)$	SCE	$CME = \frac{CME}{ab(r-1)}$	
Total	$abr - 1$	SCT		

**Ejemplo 20** <sup>7</sup> Con objeto de averiguar la estabilidad de la vitamina C en concentrado de jugo de naranja congelado reconstituido que se almacena en un refrigerador en un periodo de hasta una semana, en 1975 se llevó a cabo el estudio “Vitamin C retention in reconstituted frozen orange juice” a cargo del Departamento de Alimentos y nutrición humana en el Instituto Politécnico y Universidad Estatal de Virginia. Se probaron tres tipos de concentrado de jugo de naranja congelado utilizando tres periodos diferentes de tiempo. Estos últimos se refieren al número de días que transcurren desde que el jugo de naranja se mezcla hasta que se somete a la prueba. Los resultados, en miligramos de ácido ascórbico por litro, se registraron de la siguiente manera, según se indica en la tabla. Utilice un nivel de significancia de 0.01 para contrastar las hipótesis de que:

- No existe diferencia en los contenidos de ácido ascórbico entre las diferentes marcas de concentrado de jugo de naranja.
- No existe diferencia entre los contenidos de ácido ascórbico debido a los diferentes periodos de tiempo.
- Las marcas de concentrado de jugo de naranja y el número de días que transcurre desde que el jugo se mezcla hasta que se somete a la prueba no interactúan.

Marca	Tiempo (días)					
	0	3	7	7	7	7
Richfood	52.6	54.2	49.4	49.2	42.7	48.8
	49.8	46.5	42.8	53.2	40.4	47.6
Sealed-Sweet	56.0	48.0	48.8	44.0	49.2	44.0
	49.6	48.4	44.0	42.4	42.0	43.2
Minute Maid	52.5	52.0	48.0	47.0	48.5	43.4
	51.8	53.6	48.2	49.6	45.2	47.6

**Solución 21** La salida de Minitab© para los datos en Análisis de Varianza Bifactorial, sin considerar un modelo aditivo, es:

Fuente	gl	SC	CM	F	Valor-p
Marca	2	32.962	16.481	1.75	0.193
Tiempo	2	226.676	113.338	12.04	0.000
Interacción	4	17.301	4.325	0.46	0.765
Error	27	254.140	9.413		
Total	35	531.079			

<sup>7</sup> Tomado de [4], con modificaciones

Recordemos que ahora tenemos tres juegos de hipótesis a contrastar:  
a) Igualdad de medias entre las marcas (factor 1):

$$H_{1,0} : \mu_R = \mu_S = \mu_M$$

$H_{1,a}$  : Al menos dos de las marcas tienen medias distintas entre sí

El valor del estadístico de prueba es  $F_{Calc1} = 1.75$ , y el valor de tablas es  $F_{Tabla1} = F_{2,27,0.01} = 5.49$ . Entonces, como  $F_{Calc1}$  no es mayor que  $F_{Tabla1}$ , no se rechaza  $H_{1,0}$ , es decir, no hay evidencia de que la concentración de vitamina C varíe entre las distintas marcas.

b) Igualdad de medias entre los tiempos (factor 2)

$$H_{2,0} : \mu_0 = \mu_3 = \mu_7$$

$H_{2,a}$  : Al menos dos de los tiempos tienen medias distintas entre sí

El valor del estadístico de prueba es  $F_{Calc2} = 12.04$ , y el valor de tablas es  $F_{Tabla2} = F_{2,27,0.01} = 5.49$ . Entonces,  $F_{Calc2} > F_{Tabla2}$ , por tanto, se rechaza  $H_{2,0}$ , es decir, existe evidencia de que la concentración de vitamina C varía si se utilizan distintos tiempos.

c) Existencia de interacción entre las marcas (factor 1) y los tiempos (factor 2):

$H_{3,0}$  : No existe interacción entre las marcas y los tiempos

$H_{3,a}$  : Existe interacción entre las marcas y los tiempos

El valor del estadístico de prueba es  $F_{Calc3} = 0.46$ , y el valor de tablas es  $F_{Tabla3} = F_{4,27,0.01} = 4.11$ . Entonces, como  $F_{Calc3}$  no es mayor que  $F_{Tabla3}$ , no se rechaza  $H_{3,0}$ , es decir, no hay evidencia de que la combinación de marca con tiempo haga variar la concentración de vitamina C.

**Ejemplo 22** *Ganancias de joyerías. Resultados de SPSS para las variaciones porcentuales en las ganancias debidas a la localidad y el incremento aplicado al precio*

(Coloque los resultados de SPSS)

Las diferencias entre los distintos aumentos son aditivas. El aumento  $A_1$  produce mayores ganancias. El aumento  $A_1$  produce mayores ganancias en ambos tipos de ciudades. Le siguen el aumento  $A_2$  y  $A_3$ . En ciudades pequeñas, el aumento  $A_3$  produce mayor disminución en el valor medio.

## 2. REGRESIÓN LINEAL SIMPLE

---

### 2.1. Introducción

En esta unidad estaremos interesados en establecer, estadísticamente, si existe una relación entre dos variables y, en caso afirmativo, en calcular una ecuación que plasme dicha relación. Por ejemplo, podemos hacernos preguntas como las siguientes [?]:

1. ¿Existe una relación entre lo que gasta una empresa y sus ventas durante un año?
2. Con base en el área que tiene una casa habitación, ¿se puede calcular el costo de la calefacción doméstica?
3. ¿Hay una relación entre la antigüedad en el trabajo de un empleado de producción y el número de unidades que elabora?

Nuestro interés en establecer si existen dichas relaciones se debe a una de las siguientes razones:

- Existe una variable, a la que etiquetaremos como  $Y$ , cuyos valores requerimos conocer, pero que resulta difícil medir directamente, por lo cual intentaremos hacerlo de manera indirecta a través de una variable fácil de medir, a la que llamaremos  $X$ , mediante la relación entre ambas variables.
- Existe una variable,  $Y$ , cuyos valores pueden explicarse, al menos de manera parcial, por los valores de otra variable,  $X$ , a través de una relación entre ambas variables.

A los métodos estadísticos que nos permiten predecir valores para una variable,  $Y$ , con base en valores de otra,  $X$ , se les llama **métodos de regresión**.

A la variable cuyos valores son de nuestro interés,  $Y$ , le llamamos **variable dependiente o de respuesta**. A la variable  $X$  cuyos valores nos ayudan a calcular los valores de la primera, le llamamos **variable independiente o variable explicativa**. En un esquema general, los valores de  $X$  no son aleatorios, sino que los fija el experimentador, mientras que los valores de  $Y$  son aleatorios y su promedio depende del valor de  $X$ .

Hemos dicho que nos interesa establecer una relación entre  $X$  y  $Y$ . De el álgebra y la geometría analítica sabemos que entre dos variables puede establecerse distintas relaciones, por ejemplo:

(Coloque una gráfica con distintos ejemplos de relaciones entre variables)

La figura anterior ilustra cuatro tipos de relación entre las variables  $X$  y  $Y$ , las cuales poseen características diferentes. Todos estos tipos de relación son comúnmente utilizados por los métodos de regresión; sin embargo sabemos que la relación más simple entre dos variables está dada por una línea recta, del tipo de la primera mostrada en la gráfica anterior.

Si la relación entre  $X$  y  $Y$  fuera perfectamente lineal, entonces podríamos expresarla como sigue:

$$Y = \beta_0 + \beta_1 X$$

donde a  $\beta_0$  le llamamos ordenada al origen, porque es el punto en que la recta corta al eje  $Y$ , y  $\beta_1$  es la pendiente de la recta. Sin embargo, veremos que en general las relaciones lineales que establezcamos no serán exactas o determinísticas, sino que existirá un componente de aleatoriedad, razón por la cual, necesitamos de la estadística y la probabilidad para establecer la relación entre las variables dependiente e independiente.

Para entrar en materia, considérese el siguiente

**Ejemplo 23** *Suponga que se desea conocer la relación que existe entre la cantidad de fertilizante utilizada en una parcela y la producción de trigo obtenida de dicha superficie. Supóngase que se dispone de fondos para efectuar únicamente siete observaciones. Así, el valor de  $X$  (cantidad de fertilizante) se establece en siete diferentes niveles, con una observación de  $Y$  (producción) en cada caso, como se presenta en la tabla a continuación<sup>1</sup>:*

*(Coloque la tabla de  $X =$  Fertilizante y  $Y =$  Producción)*

*Podemos representar gráficamente los datos de ambas variables en un gráfico llamado **de dispersión**, como el que se muestra a continuación:*

*(Coloque el gráfico de dispersión)*

*Observamos que los datos muestran que existe cierta relación lineal entre las variables  $X$  y  $Y$  (a saber, notamos que mientras más fertilizante se utilice, la producción será mayor) pero que sin embargo, esta no es perfecta. En particular, podemos observar en la gráfica que no todos los puntos caen exactamente sobre la misma recta. ¿A qué se debe esto?*

*En un modelo de regresión lineal simple, asumimos que la relación entre las variables es esencialmente lineal, sin embargo, el valor final de  $Y$  puede estar afectado (y generalmente lo está) por muchas otras variables que no estamos tomando en cuenta. Por ejemplo, puede ser que la cantidad producida en una cierta parcela dependa, además de la cantidad de fertilizante, de factores como:*

- *Quién realizó la siembra*
- *Si hay más de una máquina para sembrar o cosechar, cuál de ellas se utilizó*

---

<sup>1</sup> 1 bushel = 8 galones = 35.24 litros

1 acre = 43,560 pie<sup>2</sup> = 4046.86 metros<sup>2</sup>

- *Quién realizó la cosecha*
- *Qué día se llevó a cabo la siembra*
- *Qué día se realizó la cosecha*
- *Cuál es la inclinación del terreno*
- *Etcétera*

*Como nuestro modelo no incorpora ninguno de esos factores, lo que a fin de cuentas observamos es que en promedio  $Y$  está dada por  $\beta_0 + \beta_1 X$ , pero tiene un cierto componente aleatorio que hace que en ocasiones se observe valores que estarían por abajo, y otras veces, valores que estarían por arriba de dicha recta. Estas variaciones las incorporamos a nuestro modelo en forma de una variable aleatoria que denominamos “error” y representamos por la letra griega épsilon:  $\varepsilon$ . Entonces nuestro modelo completo es:*

$$Y = \beta_0 + \beta_1 X + \varepsilon. \quad (2.1)$$

*Dado que tendremos a nuestra disposición solamente una muestra de observaciones, no podemos saber el valor exacto de  $\beta_0$  y  $\beta_1$ , pero sí valores estimados de ellos a los cuales llamaremos  $\hat{\beta}_0$  y  $\hat{\beta}_1$ , donde el circunflejo ( $\hat{\phantom{x}}$ ), como de costumbre, indica que se trata de una estimación. Por tanto, la recta que estime la relación entre  $X$  y  $Y$  dada en [??] tendrá la forma*

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (2.2)$$

*Después veremos que nuestra suposición de que  $\varepsilon$  es un error aleatorio tiene otras implicaciones, por ahora requerimos encontrar una manera de calcular los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ . Notemos que, en la gráfica de dispersión de los datos de nuestro ejemplo, podríamos dibujar más de una línea que pasara por algunos puntos o que de algún modo estuviera “cerca de ellos”:*

*(Coloque la gráfica que muestra que hay distintas rectas posibles)*

*Ahora nuestro problema consiste en definir la manera de elegir una recta. Eso lo veremos en el desarrollo que se presenta a continuación.*

## 2.2. Estimadores de mínimos cuadrados

Para establecer cuál sería una recta adecuada, consideramos el criterio de **mínimos cuadrados**: Los coeficientes de “la mejor recta” son aquellos que minimizan la suma de los cuadrados de los errores, es decir, la suma de las distancias entre cada punto observado y la recta,  $e_i = y_i - \hat{y}_i$ , cada una elevada al cuadrado. Por ello, a los estimadores de los coeficientes de la recta de regresión lineal se les llama “estimadores de mínimos cuadrados”.

Para aproximar los coeficientes de la recta  $Y = \beta_0 + \beta_1 X + \varepsilon$ , el criterio que se usa es el de encontrar los estimadores de  $\beta_0$  y  $\beta_1$  que produzcan el menor valor de la suma de cuadrados del error, la cual definimos como:

$$SCE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (2.3)$$

Aplicando cálculo vectorial<sup>2</sup>, se puede ver que los coeficientes de la recta que minimizan la suma de cuadrados de los errores, son

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (2.4)$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (2.5)$$

## 2.3. Supuestos básicos del modelo de regresión lineal simple

Un modelo es una representación simplificada de la realidad. Dicha simplificación implica establecer ciertos supuestos con respecto al comportamiento general del fenómeno, los cuales nos permiten obviar muchos detalles de la realidad. Cuando se busca explicar cualquier fenómeno de la naturaleza mediante un modelo, el qué tan bien esta representación permita describir el fenómeno depende de que los supuestos básicos en los que se apoya el modelo se cumplan.

En el caso particular de la regresión lineal simple, las suposiciones que hacemos las podemos resumir de la siguiente forma:

1. Se puede representar la respuesta  $Y$  mediante un modelo lineal que depende de  $X$ :

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Es decir que suponemos que efectivamente existe una relación lineal entre  $X$  y  $Y$ , y además existe un componente de aleatoriedad,  $\varepsilon$ , lo cual hará que observemos valores de  $Y$  a veces por arriba y otras por debajo del promedio  $\beta_0 + \beta_1 X$ .

2.  $X$  se mide sin error, es decir, no es aleatoria.

---

<sup>2</sup>El desarrollo que se requiere para deducir los estimadores no es del interés de este curso, en cambio sí lo son las expresiones finales que aquí presentamos

3.  $\varepsilon$  es un error aleatorio tal que, para un valor dado de  $X$ ,

$$E(\varepsilon) = 0, \quad \sigma_\varepsilon^2 = \sigma^2,$$

con lo que decimos que, dado un valor de  $X$ , los errores son cero en promedio y tienen siempre la misma varianza, desconocida pero fija,  $\sigma^2$ .

Además, los  $\varepsilon$  son independientes entre sí.

4. Los errores  $\varepsilon$  se distribuyen según una distribución normal, es decir,

$$\varepsilon \sim N(0, \sigma^2).$$

## 2.4. Inferencias sobre los parámetros en un modelo de regresión lineal simple

### 2.4.1. Estimador de $\sigma^2$

Definamos las siguientes sumas:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \\ S_{xy} &= \sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) \\ S_{yy} &= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2 \end{aligned}$$

Notemos que, con estas sumas, podemos definir las estimaciones de mínimos cuadrados de  $\beta_0$  y  $\beta_1$  como:

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Además, puede verse que es posible escribir la suma de cuadrados del error,  $SCE$ , como

$$SCE = S_{yy} - \hat{\beta}_1 S_{xy}.$$

A partir de esto, puede demostrarse que un estimador insesgado de  $\sigma^2$ , la varianza que comparten todos los errores  $\varepsilon$ , es

$$\begin{aligned} s^2 &= \frac{SSE}{n-2} \\ s^2 &= \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}. \end{aligned} \tag{2.6}$$

A  $s = \sqrt{s^2}$  se le llama el error estándar de estimación y es una medida de la dispersión de los valores observados con respecto a la línea de regresión.

#### 2.4.2. Intervalo de confianza para $\beta_1$

Recordemos que los valores que obtenemos para los coeficientes de la recta de regresión son estimaciones puntuales de los parámetros basadas en la muestra observada, por tanto, resulta conveniente hacer otras inferencias con respecto a los parámetros, en particular, calcular estimaciones por intervalo o hacer contrastes de hipótesis.

Un intervalo de confianza de  $100(1 - \alpha)\%$  para  $\beta_1$  está dado por:

$$\begin{aligned} LIC &= \hat{\beta}_1 - t_{n-2, \alpha/2} \frac{s}{\sqrt{S_{xx}}} \\ LSC &= \hat{\beta}_1 + t_{n-2, \alpha/2} \frac{s}{\sqrt{S_{xx}}} \end{aligned} \quad (2.7)$$

Es decir que el intervalo

$$\left( \hat{\beta}_1 - t_{n-2, \alpha/2} \frac{s}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{n-2, \alpha/2} \frac{s}{\sqrt{S_{xx}}} \right). \quad (2.8)$$

contendrá al verdadero valor de la pendiente de la recta de regresión,  $\beta_1$ , con una confianza de  $100(1 - \alpha)\%$ .

#### 2.4.3. Intervalo de confianza para $\beta_0$

Un intervalo de confianza de  $100(1 - \alpha)\%$  para  $\alpha$  se puede calcular mediante las siguientes expresiones:

$$\begin{aligned} LIC &= \hat{\beta}_0 - t_{n-2, \alpha/2} \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} \\ LSC &= \hat{\beta}_0 + t_{n-2, \alpha/2} \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} \end{aligned} \quad (2.9)$$

Por tanto, el intervalo

$$\left( \hat{\beta}_0 - t_{n-2, \alpha/2} \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}, \hat{\beta}_0 + t_{n-2, \alpha/2} \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} \right). \quad (2.10)$$

contendrá al verdadero valor de la ordenada al origen de la recta de regresión con una confianza de  $100(1 - \alpha)\%$ .

#### 2.4.4. Contraste de hipótesis sobre $\beta_1$

El contraste de hipótesis de mayor interés sobre  $\beta_1$  es el que nos permita decir si el valor de  $\beta_1$  es distinto de cero. ¿Por qué nos interesa particularmente hacer esta comparación? Recordemos que  $\beta_1$  representa la pendiente de la recta (es el coeficiente de la  $X$ ). Si resulta que el verdadero valor de  $\beta_1$  es igual a cero, entonces eso indica que el valor de la variable (supuestamente) dependiente  $Y = \beta_0 + \beta_1 X + \varepsilon$  es en realidad  $Y = \beta_0 + 0 \cdot X + \varepsilon = \beta_0 + \varepsilon$ , es decir que  $Y$  es simplemente un valor aleatorio alrededor de la constante  $\beta_0$  y no tiene relación con  $X$ , y en tal caso,  $X$  no serviría para predecir los valores de  $Y$ , lo cual era lo que buscábamos desde un inicio. En resumen, este contraste de hipótesis es importante, porque nos dice si en realidad tiene sentido el modelo de regresión que propusimos.

El contraste de hipótesis sobre  $\beta_1$ , en particular para determinar si su valor es distinto de cero, es:

1.  $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

2. El estadístico de la prueba es

$$T_{Calc} = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$$

el cual, cuando  $H_0$  es verdadera, sigue la distribución  $T$  de Student con  $n - 2$  grados de libertad.

3. La región de rechazo la constituyen los valores del estadístico de la prueba tales que  $|T_{Calc}| > t_{n-2, \alpha/2}$ , por tanto

4. Rechazamos  $H_0$  si  $T_{Calc} > t_{n-2, \alpha/2}$ , o si  $T_{Calc} < -t_{n-2, \alpha/2}$ .

#### 2.4.5. Contraste de hipótesis sobre $\beta_0$

Si los datos indican que en realidad  $\beta_0 = 0$ , entonces tendríamos que la recta de regresión pasa por el origen. Esto no es igual de grave que el caso para  $\beta_1$ , porque simplemente significa que cuando  $X = 0$ ,  $Y$  también es cero. Por ejemplo, si  $X$  fuera el número de días que un artículo está en exhibición y  $Y$  indicara las ventas obtenidas durante ese periodo, pues resulta claro que si  $X = 0$  (el producto se exhibió ningún día) esperaríamos que también  $Y = 0$  (no hubo ventas). En general, los contrastes de hipótesis sobre el valor de  $\beta_0$  son menos utilizados que aquellos para  $\beta_1$ , pero los incluimos aquí para presentar una visión más completa de las inferencias con respecto a los parámetros de la recta de regresión lineal simple.

Un contraste de hipótesis sobre  $\beta_0$ , la ordenada al origen de la recta de regresión, tiene los siguientes componentes:

1.  $H_0 : \beta_0 = b_0$

$H_a : \beta_0 \neq b_0$

donde  $b_0$  es un valor particular contra el cual nos interesa comparar  $\beta_0$ .

2. El estadístico de la prueba es

$$T_{Calc} = \frac{\widehat{\beta}_0 - b_0}{\frac{s}{nS_{xx}} \sqrt{\sum_{i=1}^n x_i^2}}$$

el cual, cuando  $H_0$  es verdadera, sigue la distribución  $T$  de Student con  $n - 2$  grados de libertad.

3. La región de rechazo la constituyen los valores del estadístico de la prueba tales que  $|T_{Calc}| > t_{n-2, \alpha/2}$ , por tanto

4. Rechazamos  $H_0$  si  $T_{Calc} > t_{n-2, \alpha/2}$ , o si  $T_{Calc} < -t_{n-2, \alpha/2}$ .

Si la hipótesis alternativa que interesa es que  $\beta_0$  es menor (o mayor) que  $b_0$ , entonces se debe modificar de manera correspondiente la región de rechazo. Se deja dicha modificación como ejercicio al lector.

## 2.5. Correlación

### 2.5.1. Coeficiente de correlación

Entre los supuestos básicos del modelo de regresión lineal simple está el que establece que la variable  $X$  no es aleatoria debido a que se mide sin error, o bien con un error muy pequeño. En la práctica suele ocurrir que  $X$  también es aleatoria. En esta situación, estamos interesados en obtener una medida de la relación lineal entre  $X$  y  $Y$ . Dicha medida recibe el nombre de **coeficiente de correlación**. Si suponemos que  $X$  se distribuye normal, el coeficiente de correlación poblacional entre  $X$  y  $Y$  está dado por:

$$\rho = \sqrt{\frac{\sigma_Y^2 - \sigma^2}{\sigma_Y^2}},$$

que podemos estimar mediante la siguiente expresión:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

El valor de  $r$  (al igual que el de  $\rho$ ) está siempre entre -1 y 1. Los distintos valores de  $r$ , los podemos interpretar de la siguiente manera:

- Un valor de  $r = -1$  indica una perfecta relación lineal negativa entre  $X$  y  $Y$ , es decir que todos los puntos caen sobre la recta de regresión, y dicha recta tiene una pendiente negativa.

(Coloque una gráfica que muestre una relación negativa perfecta entre  $X$  y  $Y$ )

- Un valor de  $r = 0$  indica que no existe relación lineal entre  $X$  y  $Y$ , lo cual implica que, o bien no existe ninguna relación entre las variables, o bien existe alguna relación, pero esta no es lineal.

(Coloque una gráfica que muestre una relación no lineal entre  $X$  y  $Y$ )

(Coloque una gráfica que muestre ausencia de relación lineal entre  $X$  y  $Y$ )

- Un valor de  $r = 1$  indica una relación lineal positiva perfecta entre  $X$  y  $Y$ , lo cual significa que todos los puntos caen sobre la recta de regresión y esta tiene pendiente positiva.

(Coloque una gráfica que muestre una relación positiva perfecta entre  $X$  y  $Y$ )

Un valor cualquiera de  $r$  negativo indicará que existe algún grado de asociación lineal entre las variables y que esta asociación es negativa, es decir, que cuando  $X$  crece,  $Y$  decrece y viceversa.

Cualquier valor de  $r$  positivo indicará que existe algún grado de asociación lineal positivo entre  $X$  y  $Y$ , es decir que cuando la variable independiente aumenta su valor, también lo hace la variable dependiente.

Es difícil establecer un valor para el cual pueda decirse que la relación es “fuerte”, ya que tal aseveración dependerá del contexto del problema, pero como una regla general para problemas **económicos, administrativos y sociales**, podría decirse que:

- Valores de  $r$  entre -0.3 y 0.3 indican que se tiene **escasa** relación lineal entre las variables.
- Si  $r$  está entre -0.5 y -0.3, o entre 0.3 y 0.5, diremos que existe una **moderada** relación lineal.
- Cuando  $r$  toma valores entre -0.8 y -0.5, o entre 0.5 y 0.8, consideraremos que hay una relación lineal **alta** entre  $X$  y  $Y$ .
- Si observamos valores de  $r$  entre -0.9 y -0.8 o entre 0.8 y 0.9, diremos que estos indican que hay una relación lineal **muy alta** entre las variables.
- Valores de  $r$  menores que -0.9 o mayores que 0.9 indican que existe una relación lineal **extremadamente alta** entre las variables.

Lo recién expresado, puede resultar más claro si lo observamos en la siguiente representación gráfica:

(Coloque la gráfica de una posible descripción detallada del valor de  $r$ )

Si se tratara de un fenómeno físico o químico en el cual se busca confirmar relaciones o características en dichos ámbitos no sería extraño que el índice de correlación fuera extremadamente alto (abajo de -0.9 o arriba de 0.9) puesto que se trata de características de la materia que son constantes universales; pero en el ámbito de las ciencias sociales, es decir, en donde se encuentra involucrado el comportamiento de seres humanos, la obtención de un índice de correlación de estas magnitudes debería hacernos sospechar que algo anda mal, o bien que los datos han sido manipulados. Un índice de correlación cercano a 0.5 (o a -0.5) puede considerarse como bueno.

### 2.5.2. Coeficiente de determinación

Otra medida que sirve para indicar el grado de relación lineal entre las variables dependiente e independiente es el coeficiente de determinación,  $r^2$ , que es precisamente el cuadrado del coeficiente de correlación.

El coeficiente de determinación indica la proporción de la variación en  $Y$  que el modelo explica. Como esta medida se obtiene de elevar al cuadrado el coeficiente de correlación, entonces siempre su valor está entre 0 y 1, y por tanto se acostumbra proporcionar su valor como un porcentaje. Mientras más cercano a 100 % esté el valor de  $r^2$ , esto implicará que un porcentaje importante de la variación en  $Y$  está explicada por el modelo lineal que se propuso.

El valor de  $r^2$  lo podemos calcular con la expresión

$$r^2 = 1 - \frac{SCE}{S_{yy}}$$

o bien, con

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

## 2.6. Análisis de varianza para el modelo de regresión

Recordemos que cuando trabajamos con experimentos unifactoriales, el análisis de varianza se basaba en dividir la variación total en las observaciones (SCT) en dos partes: una debida a los tratamientos del factor (SCA) y otra debida al error aleatorio (SCE). En el análisis de regresión se tiene algo similar:

la variación total de los datos (SCT) se divide en una parte debida al modelo, llamada suma de cuadrados de regresión (SCR) y otra debida al error aleatorio, llamada suma de cuadrados del error (SCE).

Entonces tenemos que

$$SCT = SCR + SCE$$

donde

$$SCT = S_{yy} \quad (2.11)$$

$$SCE = S_{yy} - \hat{\beta}_1 S_{xy} \quad (2.12)$$

y

$$SCR = SCT - SCE$$

$$SCR = \hat{\beta}_1 S_{xy}$$

La tabla de análisis de varianza para un modelo de regresión lineal simple tiene la siguiente forma:

Fuente de variación	Suma de cuadrados	g.l.	Cuadrado medio	Estadístico
Regresión	$SCR$	1	$CMR = \frac{SCR}{1}$	$F_{Calc} = \frac{CMR}{CME}$
Error	$SCE$	$n - 2$	$CME = \frac{SCE}{n-2}$	
Total	$SCT$	$n - 1$		

Las hipótesis que se contrastan son con respecto a  $\beta_1$ :

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

El estadístico de prueba es

$$F_{Calc} = \frac{CMR}{CME},$$

mientras que la regla de conclusión es: Rechace  $H_0$  al nivel  $\alpha$  si  $F_{Calc} > F_{1,n-2,\alpha}$ .



# 3. REGRESIÓN LINEAL MÚLTIPLE

---

## 3.1. Introducción

En el capítulo anterior estudiamos el análisis de regresión lineal simple y vimos que es un procedimiento que nos permite pronosticar los valores de la variable difícil de medir, a la cual llamamos *respuesta* o *dependiente* ( $Y$ ) con base en los valores de una variable que es fácil de medir, a la cual llamamos *explicativa* o *independiente* ( $X$ ). Sin embargo, también vimos que hay ocasiones en las que el modelar a  $Y$  solamente en función de  $X$  no proporciona buenas estimaciones. Esto nos hace pensar que tal vez si incluyéramos más de una variable explicativa, podríamos tener un modelo más adecuado para pronosticar los valores de la variable respuesta. El análisis de regresión lineal múltiple nos proporciona, precisamente, métodos para explicar a  $Y$  con base en dos o más  $X$ 's.

Por ejemplo, si estamos interesados en pronosticar el monto de las ventas del siguiente mes ( $Y$ ), entonces podríamos basarnos en los gastos durante este mes en publicidad en periódicos ( $X_1$ ), publicidad en radio ( $X_2$ ), número de agentes de venta contratados ( $X_3$ ), y tal vez otras variables.

## 3.2. Modelo de regresión lineal múltiple

El modelo de regresión múltiple es una extensión del modelo de regresión lineal simple, la diferencia es que ahora tenemos más de una variable independiente:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon. \quad (3.1)$$

Aquí también suponemos que las  $X_i$ ,  $i = 1, \dots, k$  explican la variabilidad en  $Y$ , aunque puede haber otras características o circunstancias que no se está tomando en cuenta, y a todas ellas se les engloba en el error  $\varepsilon$ .

Los supuestos del modelo de regresión lineal múltiple son los siguientes:

1. La relación entre  $Y$  y las  $X_i$ ,  $i = 1, 2, \dots, k$  es lineal.
2. Las  $X_i$ ,  $i = 1, 2, \dots, k$  se miden sin error, es decir, no son aleatorias, o bien, si son aleatorias, son independientes de los errores  $\varepsilon$  y son independientes entre sí. Cuando dos o más de las variables  $X_i$  no son independientes entre sí, se dice que existe *multicolinealidad*.

3. Para todos los niveles de las  $X_i$ ,  $i = 1, 2, \dots, k$ , los errores  $\varepsilon$  se distribuyen normal, con media cero y la misma varianza  $\sigma^2$ , es decir,

$$\varepsilon \sim N(0, \sigma^2).$$

A la igualdad de varianzas de los errores se le llama *homoscedasticidad*. Si este supuesto no se cumple, se dice que existe *heteroscedasticidad* ( $\equiv$  varianzas distintas) en los errores.

4. Los errores  $\varepsilon$  son independientes entre sí.  
 5. Las observaciones sucesivas de la variable dependiente no están correlacionadas entre sí.

A fin de calcular los valores de las  $\beta_i$ ,  $i = 0, 1, \dots, k$ , se toma un conjunto de observaciones. Dado que los valores de las  $\beta_j$  que se obtiene a partir de dicha muestra son estimaciones, entonces escribimos:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k \quad (3.2)$$

donde los  $\hat{\beta}_i$  son los estimadores de mínimos cuadrados de los  $\beta_i$ ;  $i = 0, 1, \dots, k$ .

### 3.3. Tabla de análisis de varianza para un modelo de regresión múltiple

En la tabla de análisis de varianza de un modelo de regresión múltiple se divide la varianza en dos partes: la explicada por el modelo y el error o variación aleatoria.

La tabla de ANVA en este caso es la siguiente:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	$F_{Calc}$
Regresión	SCR	$k$	$CMR = \frac{SCR}{k}$	$\frac{CMR}{CME}$
Error	SCE	$n - k - 1$	$CME = \frac{SCE}{n - k - 1}$	
Total	SCT	$n - 1$		

Las hipótesis que se contrasta mediante esta tabla de análisis de varianza son las siguientes:

$H_0$  : Todos los coeficientes  $\beta_j$  de las variables independientes son iguales a cero

$H_a$  : Al menos uno de los coeficientes  $\beta_j$  es distinto de cero.

El valor de  $F_{Calc}$  se compara contra los cuantiles de una distribución  $F$  con  $k$  gl en el numerador y  $n - (k + 1)$  gl en el denominador.

Al igual que en el caso de la regresión lineal simple, en caso de no rechazar  $H_0$ , tenemos que el modelo no es adecuado para explicar la variación en  $Y$ .

Por otro lado, si se rechaza  $H_0$ , entonces se ha determinado, al nivel de significación empleado, que al menos uno de los coeficientes  $\beta_j$ ;  $j = 1, 2, \dots, k$  es distinto de cero, lo cual implica que la  $X_j$  a la que pertenece dicho coeficiente resulta de utilidad para pronosticar el valor de  $Y$ .

## 3.4. Inferencias sobre los coeficientes del modelo

### 3.4.1. Contraste de hipótesis sobre $\beta_j$

Para establecer cuál o cuáles de los coeficientes de las variables independientes es significativamente distinto de cero, se realiza el siguiente contraste de hipótesis para cada  $j = 1, 2, \dots, k$ :

$$H_0 : \beta_j = 0$$

$$H_a : \beta_j \neq 0$$

El estadístico de prueba es

$$T_{Calc} = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}},$$

donde  $s_{\widehat{\beta}_j}$  es la desviación estándar del estimador  $\widehat{\beta}_j$ . Dicho estadístico sigue la distribución  $t$  de Student con  $n - k - 1$  grados de libertad. De momento no especificaremos la forma de calcular la desviación estándar de los coeficientes de la regresión múltiple, pero todos los paquetes de cómputo estadístico que permitan trabajar con regresión lineal múltiple proporcionan una tabla que resume los resultados de dichos contrastes de hipótesis, similar a la siguiente:

Predictor	Coficiente	Err.Std. Coef	$T_{Calc}$	Valor - $p$
$X_1$	$\widehat{\beta}_1$	$s_{\widehat{\beta}_1}$	$T_1$	$p_1$
$X_2$	$\widehat{\beta}_2$	$s_{\widehat{\beta}_2}$	$T_2$	$p_2$
...	...	...	...	...
$X_k$	$\widehat{\beta}_k$	$s_{\widehat{\beta}_k}$	$T_k$	$p_k$

Recordemos que el *Valor -  $p$*  representa una forma distinta de establecer si se rechaza o no una hipótesis. Mientras más pequeño sea el *Valor -  $p$* , mayor evidencia presentará la muestra en contra de la hipótesis en cuestión.

A partir de la tabla de ANVA podemos calcular una estimación para el valor de la desviación estándar de los errores,  $\sigma = \sqrt{\sigma^2}$ . A dicha estimación se le llama *error estándar múltiple de la estimación*; representa la variabilidad de las estimaciones que se obtiene para  $Y$  a partir de la ecuación de regresión lineal múltiple. Al error estándar múltiple de la estimación se le representa con  $s$  y se calcula como sigue:

$$s = \sqrt{\frac{SCE}{n - k - 1}}.$$

Se tomó los siguientes dos ejercicios de [[?]] pp. 505, 506 y 513.

**Ejercicio 24** Refiérase a la siguiente tabla de ANVA.

Fuente	SC	gl	CM	$F_{Calc}$
Modelo	21	3	7.0	2.33
Error	45	15	3.0	
Total	66	18		

- ¿De qué tamaño es la muestra?
- ¿Cuántas variables independientes hay?
- Calcule el coeficiente de correlación múltiple.
- Determine el error estándar múltiple de estimación.

**Ejercicio 25** Refiérase a la siguiente tabla de ANVA.

Fuente	SC	gl	CM	$F_{Calc}$
Modelo	60	5	12	1.714
Error	140	20	7	
Total	200	25		

- ¿De qué tamaño es la muestra?
- ¿Cuántas variables independientes hay?
- Calcule el coeficiente de correlación múltiple.
- Determine el error estándar múltiple de estimación.

**Ejercicio 26** La empresa Salsberry Realty vende casas en la costa este de Estados Unidos. Una de las preguntas que los posibles compradores hacen con frecuencia es: si adquirimos esta casa, ¿cuánto tendremos que pagar por la calefacción en invierno? ... Se consideró que el costo (en dólares) incluye tres variables (1) la temperatura media diaria en el exterior (en grados F), (2) el espesor del material de aislamiento térmico que se coloca en el desván (el pulgadas), y (3) la antigüedad del calefactor (en años). Para realizar esta investigación, ... (se) seleccionó una muestra aleatoria de 20 casas vendidas recientemente. ... se presenta la información muestral.

(Coloque una tabla con los datos de la muestra)

Se introdujo esta información en un proyecto de Minitab y se obtuvo los resultados que se muestra a continuación.

a) Contraste la hipótesis de que el modelo es adecuado para explicar  $Y$  mediante la tabla de ANVA.

b) Determine cuáles variables resultan importantes para describir el costo de la calefacción realizando un contraste de hipótesis sobre el coeficiente de cada variable independiente del modelo.

**Resultados de la regresión:**

La ecuación de regresión es

$$\text{Costo} = 427 - 4,58\text{Temperatura} - 14,8\text{Aislante} + 6,10\text{Calefactor}$$

La tabla de análisis de varianza es la siguiente:

Fuente	SC	gl	CM	$F_{Calc}$	Valor - p
Regresión	71,220	3	1 57073	21.90	0.000
Error	41,695	16	2606		
Total	112,915	19			

La tabla que resume los contrastes de hipótesis individuales, se presenta a continuación:

Predictor	Coficiente	Err. Std.	$T_{Calc}$	Valor - p
Constante	427.19	59.60	7.17	0.000
Temperatura	-4.5827	0.7723	-5.93	0.000
Aislante	-14.831	4.754	-3.12	0.007
Calefactor	6.101	4.012	1.52	0.148

### 3.4.2. Intervalo de confianza para $\beta_j$

Un intervalo de  $100(1 - \alpha)\%$  de confianza para  $\beta_j$ ,  $j = 0, 1, \dots, k$ , está dado por:

$$\begin{aligned} \text{Límite Inferior de Confianza} &= LIC = \hat{\beta}_j - t_{n-k-1, \alpha/2} \times s_{\hat{\beta}_j} \\ \text{Límite Superior de Confianza} &= LSC = \hat{\beta}_j + t_{n-k-1, \alpha/2} \times s_{\hat{\beta}_j} \end{aligned}$$

**Ejemplo 27** Suponga que se tiene un conjunto de 15 datos sobre la calificación obtenida por un egresado en el examen GPA y la indicación de si egresó de la carrera de administración o de alguna otra para pronosticar su salario inicial al comenzar a trabajar. Con base en tales datos, se obtuvo el siguiente modelo de regresión lineal múltiple:

$$\text{Salario} = 23,4 + 2,77\text{GPA} + 1,31\text{Administracion}.$$

Además, se cuenta con la siguiente tabla sobre los coeficientes del modelo.

Predictor	Coficiente	SE Coef
Constante	23.447	3.490
GPA	2.775	1.107
Administracion	1.3071	0.4660

a) Calcule un intervalo de confianza de 95 % para el coeficiente de la variable GPA.

b) Calcule un intervalo de confianza de 95 % para el coeficiente de la variable Administración.

## 3.5. Correlación

### 3.5.1. Coeficiente de determinación múltiple

A partir de la tabla de ANVA, podemos calcular el valor del coeficiente de determinación múltiple,  $R^2$ :

$$R^2 = \frac{SCR}{SCT}$$

este se interpreta, como en el caso de la regresión lineal simple, como la proporción de la variabilidad en  $Y$  que explica el modelo. Como  $SCT = SCR + SCE$ , entonces  $SCR = SCT - SCE$ , y por tanto también se puede calcular el coeficiente de determinación múltiple como

$$R^2 = 1 - \frac{SCE}{SCT}.$$

### 3.5.2. Coeficiente de determinación múltiple ajustado

Aunque el coeficiente de determinación múltiple es un indicativo de qué tan bien explica el modelo el comportamiento de la variable dependiente, debe considerársele con precaución ya que el valor de  $R^2$  aumentará siempre que agreguemos una variable al modelo, aunque dicha variable no contribuya de manera notoria a la predicción de  $Y$ .

Para evitar esta dificultad, se utiliza también el coeficiente de correlación múltiple ajustado,  $R_a^2$ , el cual se define por

$$R_a^2 = 1 - \frac{n-1}{n-k-1} \frac{SCE}{SCT}.$$

La ventaja de  $R_a^2$  sobre  $R^2$  es que aquél toma en cuenta el número de datos y el número de parámetros  $\beta$  en el modelo (incluyendo  $\beta_0$ ). Esto tiene como consecuencia que no se puede hacer que  $R_a^2$  se acerque a 1 simplemente agregando variables al modelo.

### 3.5.3. Matriz de correlación

Es una matriz que contiene los coeficientes de correlación entre todos los pares de variables independientes. También se puede incluir en ella a la variable independiente. Se le representa con la letra griega sigma mayúscula,  $\Sigma$ .

Por ejemplo, si se tiene tres variables independientes  $X_1, X_2$  y  $X_3$ , la matriz de correlación entre ellas tendría la forma siguiente:

$$\Sigma = \begin{bmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{12} & \rho_{22} & \rho_{23} \\ \rho_{13} & \rho_{23} & \rho_{33} \end{bmatrix}$$

donde  $\rho_{ij}$  (rho) representa el índice de correlación entre  $X_i$  y  $X_j$ , que, recordemos, se define como

$$\rho_{ij} = \sqrt{\frac{Cov(X_i, X_j)}{Var(X_i) Var(X_j)}}$$

### 3.5.4. Multicolinealidad

Se refiere a la existencia de una alta correlación lineal entre dos variables independientes. Existe un contraste de hipótesis específico sobre multicolinealidad, pero de manera general podemos decir que entre dos variables independientes puede existir multicolinealidad si el coeficiente de correlación entre ellas es menor que  $-0,7$  o mayor que  $0,7$ .

Si se sospecha de la existencia de multicolinealidad entre dos variables independientes, debe eliminarse una de ellas del modelo. La razón para esto es que si las variables tienen una correlación muy alta, podríamos pensar que ambas miden casi la misma característica, por lo cual no

Tras la eliminación de la variable, se debe volver a calcular los coeficientes de este, junto con todas las otras medidas que nos sirven para evaluar el modelo ( $R$ ,  $R^2$ , la tabla de ANVA para el modelo completo, los contrastes de hipótesis individuales para cada coeficiente, etc.).

## 3.6. Métodos secuenciales para la selección del modelo

Cuando existe multicolinealidad, debe eliminarse una de las variables del modelo. Cuando no resulta obvio cuál de las dos variables...



# Bibliografía

---

- [1] Lee J. Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. PWS-Kent Publishing Company, Estados Unidos de América, 1992.
- [2] Wayne W. Daniel. *Estadística con Aplicaciones a Las Ciencias Sociales Y la Educación*. Mc Graw-Hill Interamericana de México, S.A. de C.V., México, 1988.
- [3] A. Efímov, A. Karakulin, P. Pospélov, A. Teréschenko, E. Vukólov, V. Zemskov, and Yu. Zolotarev. *Problemas de Las Matemáticas Superiores III*. Editorial Mir, Moscú, URSS, 1986.
- [4] John E. Freund and Ronald E. Walpole. *Estadística Matemática con Aplicaciones*. Prentice - Hall Hispanoamericana, S. A., México, 1990.
- [5] Paul G. Hoel. *Introduction to Mathematical Statistics*. John Wiley and Sons, Estados Unidos de América, 1984.
- [6] Robert Johnson. *Estadística Elemental*. Grupo Editorial Iberoamérica, México, 1993.
- [7] William Mendenhall. *Estadística Para Administradores*. Grupo Editorial Iberoamérica, México, 1990.
- [8] Ronald E. Walpole, Raymond H. Myers, and Myers Sharon L. *Probabilidad Y Estadística Para Ingenieros*. Prentice Hall Hispanoamericana, S.A., México, 1999.