

**Universidad de los Andes**  
**Facultad de Ciencias Económicas y Sociales**  
**Área de Métodos Cuantitativos**  
**Notas y Ejercicios de Computación I**  
**Estimación de funciones lineales. Regresión**  
**Lineal Simple**  
Gerardo Colmenares y Alexis Melo T.

### 1. INTRODUCCIÓN.

Este resumen pretende ser un documento de ayuda rápida para la formulación y evaluación superficial de modelos lineales usando el método de los mínimos cuadrados. O también conocido, el análisis de regresión lineal.

Para cualquier revisión más detallada se recomienda acudir a los textos especializados. Por ejemplo, Applied Regression Analysis de Draper y Smith, Estadística para Administración, 2ª. Edición por Mark Berenson, David Levine y Timothy Krehbiel.

### 2. DEFINICIÓN.

**Modelo lineal:** Se dice que un modelo de regresión es lineal si sus parámetros son lineales (los parámetros están elevados a la potencia 1).

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

**Modelo no lineal:** Se dice que un modelo es no lineal cuando los parámetros se encuentran elevados a una potencia diferente de 1.

$$y = b_0 + \frac{b_1x}{b_2 + x}$$

En la formulación general  $b(s)$  son los parámetros, las  $x(s)$  son las variables independientes e  $y$  es la variable dependiente.

**Método de los mínimos cuadrados ordinarios.** Consiste en estimar los valores de las  $b(s)$  que permiten minimizar  $S$ , donde:

$$S = \sum (y_i - \hat{y}_i)^2 = \text{Suma del cuadrado de los residuos.}$$

$y_i$  = valor observado  $i$  (variable dependiente)

$\hat{y}_i$  = Valor estimado de  $y_i$  por el modelo.

**El modelo lineal simple:** Se define así a un modelo de regresión lineal donde existe una sola variable independiente, siendo su fórmula  $\hat{y}_i = b_0 + b_1x_i$

### 3. COMENTARIOS ACERCA DEL ANÁLISIS DE LOS DATOS EN LOS MODELOS LINEALES.

Estos son lineamientos aproximados para el análisis. Existen otras técnicas que ayudan a la realización de análisis más exhaustivos.

#### Sugerencias:

1. Grafique individualmente cada una de las variables independientes vs. la variable y revise si algún gráfico muestra alguna tendencia particular. Por ejemplo, la

existencia de una tendencia no lineal implicará una transformación de la variable independiente hacia una relación lineal.

2. Los histogramas de todas las variables (independientes y dependiente) ayudan a determinar potenciales valores atípicos. Estos valores atípicos podrían ocasionar distorsiones en la estimación de los parámetros a causa del ruido producido ellos.
3. Las variables cuya varianza sean mayores que la varianza del error estándar del modelo podrían ser excluidas del análisis preliminar.
4. Se debe tener para evitar utilizar dos variables con medidas del mismo significado (por ejemplo ingreso y riqueza). Para ello deben hacerse una prueba de correlación entre ambas variables. Si el coeficiente de correlación es mayor de 0.95, entonces es recomendable usar solo una de ellas.
5. Tener cuidado de no incluir variables que son resultado de otras variables. Cantidades expresadas como proporciones, funciones o porcentajes podrían arrojar variables correlacionadas y de este mismo modo un análisis pobre.
6. Mientras mayor sea la cantidad de datos incluidos en el experimento, éste resulta generalmente mejor. Sin embargo, esto significaría pruebas mejor trabajadas para determinar la construcción del modelo, dando como resultado la posibilidad de incluir algunas nuevas variables.
7. Una revisión previa de las variables utilizadas en modelos semejantes que hayan sido anteriormente realizados ayudaría a establecer el modelo experimental que se desea realizar. Por ejemplo, una nueva función de demanda incluiría al menos como variable independiente al precio del bien. Otras variables que pueden incluirse en el experimento podrían ser el ingreso per cápita, el precio de bienes relacionados, etc.
8. Debe prestarse bastante atención a las conclusiones derivadas de las variables controladas.

#### 4. MEDIDAS DE IMPORTANCIA.

$$S = \sum (y_i - \hat{y}_i)^2$$

$$F_{calc.} = \frac{SSR}{SSE}$$

SSR = Suma de los cuadrados de la Regresión

SSE = Suma de los cuadrados de los Residuos

#### Prueba de que todos los parámetros sean cero:

Si  $F_{calc.} > F_{tabulado}$ , entonces la hipótesis nula de que todos los parámetros son simultáneamente cero se rechaza y, desde luego, se acepta el modelo. Las pruebas débiles no son muy útiles para cuando los valores de  $F_{calc.}$  y  $F_{tabulado}$  sean similares. Sin embargo, hay un medio heurístico conocido como la regla del dedo para la evaluación de  $F_{calc.}$

En Excel puede observar en los resultados del Análisis de Varianza este valor de F, la regla mnemotécnica que se sigue regularmente es la que se muestra en el cuadro que sigue.

Valores de $F_{calc.}$	Observación
3-10	Puede ser considerado un modelo
10-100	Puede ser un modelo interesante
100-1000	Es un buen modelo

> 1000	Es un modelo increíble
--------	------------------------

### ANÁLISIS DE VARIANZA

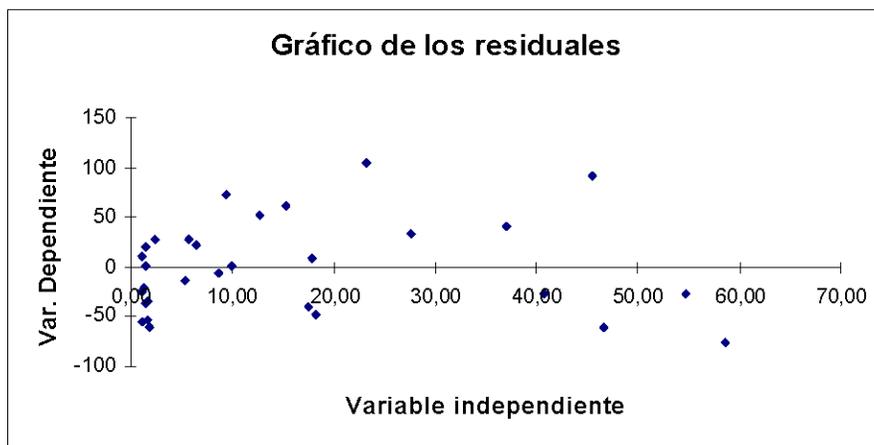
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	171485,629	171485,629	74,7835831	2,14561E-09
Residuos	28	64206,5734	2293,09191		
Total	29	235692,203			

### 5. RESIDUOS.

Los residuos o residuales son los valores obtenidos al hacer la diferencia entre los valores observados y los estimados. Es decir  $\text{residual} = y_{\text{observado}} - y_{\text{estimado}}$ .

Observación	Valor estimado de la variable respuesta	Residuos
1	81,32087792	-24,1408779
2	81,4508658	-55,2808658
3	81,66751227	11,1224877
4	81,97081732	-20,3708173
5	82,7940739	-36,2940739
6	83,18403754	1,87596246
7	83,18403754	20,3359625
8	83,87730624	-52,9973062
9	84,09395271	-34,8039527
10	84,22394059	-60,0839406

Cuando se grafica, aparentemente sería algo como el siguiente gráfico.



1. Los residuos deben tener una media cero, varianza constante y seguir una distribución normal.
2. El gráfico de los residuos vs. los valores estimados de la variable dependiente no deben mostrar signos de una variación sistemática en el que se puede observar un patrón de comportamiento. Si se observa la relación sistemática, entonces una transformación de  $y$  (no lineal) podría ser de gran ayuda.

Ejemplos:

Función	Transforma a
$y^k$	$y^{1-k}$
$y$	$\log y$

3. El gráfico de los residuos no debe mostrar una variación sistemática. Si se observa una variación de este estilo, se recomienda una transformación de los datos de tal forma que las unidades de medida conserven una misma equivalencia (por ejemplo, todas las variables en kilogramos, en millardos, en toneladas, etc.).
4. Si el gráfico de los residuos vs. cualquier variable independiente relevante (para los modelos de regresión múltiple) muestra alguna variación sistemática, entonces estaría violándose alguno de los supuestos de normalidad y en consecuencia la variable independiente en cuestión debería ser revisada. Por ejemplo, una presencia de multicolinealidad con alguna o algunas otras variables independientes o la presencia de no linealidad en la variable. Estos fenómenos podrían generar heterocedasticidad o autocorrelación. Algunos patrones de observación de los residuos podrían ser los siguientes: una figura de trompeta (la varianza es creciente), un patrón no lineal (cuadrático, exponencial, rectas con pendientes positivas o negativas), un patrón siguiendo la figura dentada (la varianza no es constante).

## Algunos ejercicios de Regresión Lineal Simple

Ejercicio 44. El consumo de energía eléctrica en una casa totalmente de componentes eléctricos, ubicada en Apartaderos fue registrada durante veinticuatro meses. Los datos son los siguientes:

Mes	Consumo en Kilvartios	Temperatura promedio	Mes	Consumo en Kilvartios	Temperatura promedio
1	126	-1	13	123	-2,5
2	132	-3,5	14	121	0,5
3	114	-1,5	15	138	-2
4	87	5	16	99	3,5
5	67	8	17	64	7,5
6	50	14,5	18	52	15,5
7	39	18,5	19	49	18,5
8	45	23	20	41	20,5
9	39	20	21	44	19
10	43	15	22	53	16
11	61	6,5	23	59	10,5
12	92	2	24	118	-2,5

- Establezca el diagrama de dispersión y responda cuál es el patrón que sigue, hay alguna asociación entre las dos variables.
- Determine cuánto es el nivel de asociación lineal.
- Si hay fuerte asociación, construya la función lineal mediante el método de los mínimos cuadrados, que aproxime mejor estos datos. Identifique las variables independiente (explicativa) y dependiente (respuesta)
- Indique la expresión que aproxima estos datos:
- Haga inferencia con un grupo de 10 nuevos valores para la variable independiente y pronostique cual será su respuesta para la variable dependiente siguiendo el modelo estimado.

**Ejercicio 45.** Una compañía que tiene los derechos de distribución de las ventas de videos desea estimar la cantidad que puede esperar vender. Se dispone de los datos de 30 películas que indican la ganancia bruta en taquilla y el número de unidades de video vendidas.

Película	Ganancia Bruta en taquilla (Millones de Bs)	Videos vendidos (en miles)	Película	Ganancia Bruta en taquilla (Millones de Bs)	Videos vendidos (en miles)
1	1,10	57,18	16	9,36	190,80
2	1,13	26,17	17	9,89	121,57
3	1,18	92,79	18	12,66	183,80
4	1,25	61,60	19	15,35	204,72
5	1,44	46,50	20	17,55	112,47
6	1,53	85,06	21	17,91	162,95
7	1,53	103,52	22	18,25	109,20
8	1,69	30,88	23	23,13	280,79
9	1,74	49,29	24	27,62	229,51
10	1,77	24,14	25	37,09	277,68
11	2,42	115,31	26	40,73	226,73
12	5,34	87,04	27	45,55	365,14
13	5,70	128,45	28	46,62	218,64
14	6,43	126,64	29	54,70	286,31
15	8,59	107,28	30	58,51	254,58

- Establezca el diagrama de dispersión y responda cuál es el patrón que sigue, hay alguna asociación entre las dos variables.
- Determine cuánto es el nivel de asociación lineal.
- Si hay fuerte asociación, construya la función lineal mediante el método de los mínimos cuadrados, que aproxime mejor estos datos. Identifique las variables independiente (explicativa) y dependiente (respuesta)
- Indique la expresión que aproxima estos datos:
- Haga inferencia con un grupo de 10 nuevos valores para la variable independiente y pronostique cual será su respuesta para la variable dependiente siguiendo el modelo estimado.
- Pronostique el número promedio de unidades de video que se venderán para una película que tuvo una ganancia en taquilla de 20 millones de bolívares.

**Ejercicio 46.** Un analista del departamento de costos está interesado en usar el cambio que se produce en los precios del azúcar para pronosticar los cambios que podría tener los precios del caramelo durante cualquier mes. El observó la siguiente *secuencia presentada por filas* durante trece meses consecutivos.

Mes	Precios												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Azúcar	80	82	85	81	80	80	80	84	88	89	90	88	84
Caramelos	105	100	105	114	107	105	104	105	110	117	120	121	118

Mediante Excel en la Hoja 1 y teniendo en cuenta que los valores observados están dispuestos por filas, responda en la misma Hoja 1 pero identifíquela como EJERCICIO 35.

- Identifíquese en la primera línea del ejercicio. (Nombre y Apellido), e indique la fecha automática de hoy en una celda adicional mediante las funciones disponibles.
- Grafique el diagrama de dispersión. Observa algún patrón lineal en los datos observados? Si hay alguna relación lineal, cuál es la medida que le permite conocer la fuerza de asociación lineal y que sentido tiene?.
- Identifique las dos variables del modelo e indique cuál será la variable respuesta (dependiente) y cuál la variable explicativa (independiente).
- Se podrá construir una función lineal entre las variables seleccionadas? Cuál es la herramienta que se puede emplear?. Si es procedente, haga los cálculos que se requieran para construir tal función.
- De ser posible la estimación de la función lineal, cuál son los coeficientes de la función  $Y=ax+b$ :  
 $a=$ ,  $b=$ , quién es  $Y$ : y quién es  $x$ :.
- Escriba la expresión resultante: .
- De acuerdo a su criterio y al valor observado de la F en los resultados del cálculo de la estimación, cree que la función construida es un: (mal, normal, buen, excelente) **modelo de pronóstico**. Cómo lo hizo?
- Si concluyó que se puede pronosticar con el modelo, haga las siguientes estimaciones de pronóstico:
  - Que pasaría con el precio de azúcar, si el precio del caramelo sufre un aumento a 140.
  - Si el precio del azúcar disminuye a 10, cuánto sería el precio del caramelo?
- Escriba una estructura de la expresión lineal construida en la hoja anterior tal que se coloquen los *coeficientes a y b* automáticamente vinculándolos desde la hoja EJERCICIO, siguiendo el patrón  $y=ax+b$ , dedicándole una celda a cada uno de los términos o coeficientes de la expresión.
- Haga tres pruebas de predicción, dándoles valores a X, y luego coloque los resultados en una tabla preparada para ello en la misma plantilla diseñada.
- Guardar los resultados con el nombre *Ejercicio46*.

**Ejercicio 47.** Un proyectil es lanzado al espacio y luego regresa a tierra. Los valores observados que se registraron fueron la altura y el tiempo en que alcanzaba esa altura. Se quiere **construir una función lineal** a partir de estos valores para estimar la altura a un tiempo dado. Los datos observados están en la siguiente tabla.

Valores observados	Tiempo en alcanzar la altura	Altura
1	0	0
2	1	112
3	2	192
4	3	240
5	4	256
6	5	240
7	6	192
8	7	112
9	8	0

Mediante Excel en la Hoja 1, identificada como EJERCICIO 35 y respondiendo a las preguntas en la misma hoja de Excel,

- Identifíquese en la primera línea del ejercicio. (Nombre y Apellido), e indique la fecha automática de hoy en una celda adicional usando las funciones disponibles.
- Grafique el diagrama de dispersión. Observa algún patrón lineal en los datos observados? Si hay alguna relación lineal, cuál es la medida que le permite concluir que existe una asociación lineal y que sentido tiene. Cuál es el valor?. Que le quiere decir este valor?
- De existir un modelo lineal, identifique las dos variables del modelo e indique cuál sería la variable respuesta (dependiente) y cuál la variable explicativa (independiente).
- Se podrá construir una función lineal entre las variables seleccionadas? Cuál es la herramienta que se podría emplear?. Si es procedente, haga los cálculos que se requieran para construir tal función.

**Ejercicio 48.** Mediante Excel en la Hoja 2, identificada como EJERCICIO 48.

- Construya, con los valores de rendimiento en kilómetros recorridos a diferentes velocidades, un modelo de regresión de la forma  $Y=ax+b$ , para estimar los valores de  $a$  y  $b$  y así poder hacer los pronósticos que se solicitan en el punto siguiente. Recuerde que los datos están organizados en filas.

Velocidad en (Km)	30,00	37,50	45,00	52,50	60,00	67,50	75,00	82,50	90,00
Km por litro	8,68	8,29	7,89	9,08	7,50	7,11	6,32	5,53	4,34

- En una fila nueva escriba la expresión lineal e indique quienes son los coeficientes de la expresión lineal, sabiendo que es de la forma  $y= ax+b$ .
- Haga tres pruebas de predicción, dándoles valores la variable independiente, y luego coloque los resultados en una tabla preparada para ello en la misma plantilla diseñada.
- Guardar los dos ejercicios con el nombre *Ejercs47-48*.

**Ejercicio 49.** Para evaluar el impacto social que produce el desempleo en la conducta de los individuos, un investigador obtiene las tasas de desempleo y suicidio de ocho áreas urbanas. Los resultados de tales mediciones están en la tabla más abajo. Los valores de las tasas, han sido registradas de la siguiente manera: la de desempleo es de un desempleado por cada 100 y la de suicidio es de un suicidio por cada 10.000 habitantes.

Areas Urbanas	Desempleo	Suicidio
1	8,6	19,1
2	6,1	12
3	6,3	10,2
4	8,4	27,5
5	14,9	29
6	12,1	31,4
7	12,7	40,4
8	9,3	23,7

Mediante Excel en la Hoja 1, responda en la misma Hoja 1 pero identifíquela como EJERCICIO 38.

- Identifíquese en la primera línea del ejercicio. (Nombre y Apellido), e indique la fecha de hoy en una celda adicional, usando las funciones disponibles.
- Grafique el diagrama de dispersión. Observa algún patrón lineal en los datos observados? Si hay alguna relación lineal, cuál es la medida que le permite conocer la fuerza de asociación lineal. Sigue algún sentido (positivo o negativo) y que significa eso?
- Identifique las dos variables del modelo e indique cuál será la variable respuesta (dependiente) y cuál la variable explicativa (independiente).
- Se podrá construir una función lineal entre las variables seleccionadas? Cuál es la herramienta que se puede emplear?. Si es procedente, haga los cálculos que se requieran para construir tal función.
- De ser posible la estimación de la función lineal, cuál son los coeficientes de la función  $Y=ax+b$ :  
 $a=$ ,  $b=$ , quién es  
 $Y$ : y quién es  $x$ :.
- Escriba la expresión resultante: .
- De acuerdo a su criterio y al valor observado de la F en los resultados del cálculo de la estimación, cree que la función construida es un: **modelo de pronóstico**. Cómo lo hizo?
- Si concluyó que se puede pronosticar con el modelo, haga las siguientes estimaciones de pronóstico:
  - Cómo se afectaría la inclinación al suicidio si el desempleo llega a los valores de 18.8.
  - Como profilaxis social y mediante ayudas psicológicas, el suicidio bajó a valores de 9,1 por cada 10.000. Bajaría o aumentaría el desempleo?. A cuánto:?
- Haga tres pruebas de predicción, dándoles valores a X, y luego coloque los resultados en una tabla preparada para ello en la misma plantilla diseñada.
- Guardar los con el nombre *Ejercicio49*.

**Ejercicio 50.** La casa de bolsa de valores de Caracas desea predecir el número de transacciones por día y ha decidido usar el número de llamadas que llegan como una variable de predicción. Los datos se recolectaron durante 35 días con los resultados mostrados en la siguiente tabla.

Día	No. De Llamadas	Transacciones	Día	No. De Llamadas	Transacciones
1	2591	417	18	2237	397
2	2146	321	19	2328	365
3	2185	362	20	2078	330
4	2245	364	21	2134	312
5	2600	442	22	2192	340
6	2510	386	23	1965	339
7	2394	370	24	2147	364
8	2486	376	25	2015	295
9	2483	463	26	2046	292
10	2297	389	27	2073	379
11	2106	302	28	2032	294
12	2035	266	29	2108	329
13	1936	339	30	1923	274
14	1951	369	31	2069	326
15	2292	403	32	2061	306
16	2094	319	33	2010	352
17	1897	306	34	1913	290
			35	1904	283

Mediante Excel en la Hoja 1, identificada como EJERCICIO 39 y respondiendo a las preguntas en la misma hoja de Excel,

- Identifíquese en la primera línea del ejercicio. (Nombre y Apellido), e indique la fecha automática de hoy en una celda adicional usando las funciones disponibles.
- Construya un diagrama de dispersión. Qué patrón observa? Marca alguna tendencia los datos? Cuál es la medida de asociación y cuánto es?
- En el supuesto de que exista asociación lineal, utilice el método de los mínimos cuadrados para encontrar los coeficientes de la recta de regresión.
- Indique la expresión que establece la ecuación de regresión en una línea donde se incluya una estructura de la expresión lineal tal que se coloquen los *coeficientes a y b*, siguiendo el patrón  $y=ax+b$ , dedicándole una celda a cada uno de los términos o coeficientes de la expresión.
- Pronostique el número promedio de transacciones en un día que hubo 2000 llamadas.
- Si las transacciones fueron de 500 operaciones en la bolsa, cuántas llamadas cree que se hicieron durante ese día.
- Grafique la recta de regresión estimada. Cómo lo hace?
- De acuerdo a su criterio y al valor observado de la F en los resultados del cálculo de la estimación, cree que la función construida es un: **modelo de pronóstico**. Cómo lo hizo?
- Haga seis pruebas de predicción, dándoles valores a X, y luego coloque los resultados en una tabla preparada para ello. Grafique el resultado mediante una línea
- Guarde ambos archivos de Excel como *Ejercicio50*