UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ECONOMÍA

TRABAJO DE APLICACIÓN DE ESTADÍSTICA
(PONENCIA)

PRESENTADO POR: **MTRO. SERGIO HORACIO NUÑEZ MEDINA**

MEXICO DF CD UNIVERSITARIA 10 DE FEBRERO DE 2012

MODELO DE REGRESION LINEAL SIMPLE

Sea un modelo con una sola variable explicatorio x, con una variable de respuesta y, tal que la relación entre ambas es una línea recta. Este modelo de regresión lineal simple es

$$y = \beta_o + \beta_1 x + \varepsilon \qquad (4.1)$$

en donde la ordenada al origen β_o y la pendiente β_1 son constantes desconocidas y ε es un componente aleatorio de error. Se supone que los errores tiene un valor esperado de cero y una varianza σ^2 desconocida. Se supone que los errores no están correlacionados. Es decir, el valor de un error no depende del valor de cualquier otro error, $E(\varepsilon_i\ \varepsilon_j) = 0$, con $i \neq j$. Se considera a x como variable controlada para el análisis de datos, mientras que y, la variable de respuesta, es aleatoria debido a que depende del componente aleatorio de error. Así existe una distribución de probabilidad de y para cada valor de x. La media de esta distribución es

$$E(y / x) = \beta_o + \beta_1 x \qquad (4.2)$$

y la varianza es

$$Var(y / x) = Var(\beta_o + \beta_1 x + \varepsilon) = \sigma^2 \qquad (4.3)$$

Así, la media de *y* es una función lineal de x, aunque la varianza de *y* no depende del valor de x. Además, dado que los errores no están correlacionados, las respuestas diversas de *y* tampoco lo están. A los parámetros β_o y β_1 se les suele llamar coeficientes de regresión. Estos tienen una interpretación simple y frecuentemente útil. La pendiente β_1 es el cambio de la media de la distribución de y producido por un cambio unitario en x. La ordenada al origen β_o (cuando esta es distinta de cero) indica que a pesar de que x descienda a cero, $y > 0$.

Estamos ahora en posición de enlistar los supuestos que especifican completamente el modelo de regresión lineal simple (o múltiple)

1.- La relación entre *y* y x es lineal como se describe en la ecuación (4.1).
2.- Las x son variables no aleatorias o adoptan valores fijos.
3(a) El término de error tiene un valor esperado de cero y varianza constante para todas las observaciones; esto es, $E(\varepsilon_i) = 0$ y $E(\varepsilon^2_i) = \sigma^2$.
3(b) Las variables aleatorias ε_i son estadísticamente independientes. Así, $E(\varepsilon_i\ \varepsilon_j) = 0$, para todo $i \neq j$.
3(c) El término de error se distribuye normalmente.

Esta lista de supuestos, excluyendo 3(c) constituyen el modelo de regresión lineal clásico.

Teorema de Gauss-Markov.
Dados los supuestos 1, 2, 3(a) y 3(b), los estimadores $\tilde{\beta}_o$ y $\tilde{\beta}_1$ son los mejores y más eficientes estimadores insesgados de β_o y β_1 en el sentido de que tienen varianza mínima de entre todos los estimadores insesgados.

Estimación de los parámetros por mínimos cuadrados.

Los parámetros β_o y β_1 son desconocidos y se deben estimar con los datos de la muestra. Suponemos que existen n pares de datos: (x_1, y_1), (x_2, y_2), …,(x_n, y_n) que pueden obtenerse en un experimento controlado, diseñado en forma especifica para recolectarlos, en un estudio observacional o a partir de registros históricos disponibles.

Estimación de β_o y β_1

Para estimar β_o y β_1 se usa el método de mínimos cuadrados. Esto es, se estiman β_o y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima. Según la ecuación (4.1), se puede escribir

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (4.4)$$

Se puede considerar que la ecuación (4.1) es un **modelo poblacional de regresión**, mientras que la ecuación (4.4) es un **modelo muestral de regresión,** escritos en términos de los n pares de datos (x_i, y_i) (i = 1, 2, …,n). Así, el criterio de mínimos cuadrados es

$$S(\beta_o, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i)^2 \qquad (4.5)$$

Los estimadores, por mínimos cuadrados, de β_o y β_1, que se designarán por $\tilde{\beta}_o$ y $\tilde{\beta}_1$ deben satisfacer

$$(\partial S) / (\partial \beta_o) = -2 \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i) = 0$$

y

$$(\partial S) / (\partial \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i) x_i = 0$$

Se simplifican estas dos ecuaciones y se obtiene

$$n\beta_o + \beta_1 \sum x_i = \sum y_i$$
$$\beta_o \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \qquad (4.6)$$

Las ecuaciones (4.6) son llamadas ecuaciones normales de mínimos cuadrados. Su solución es la siguiente

$$\tilde{\beta}_o = y_p - \tilde{\beta}_1 x \qquad (4.7)$$

y

$$\tilde{\beta}_1 = [\sum x_i y_i - (\sum x_i \sum y_i)/n] / [\sum x^2_i - (\sum x_i)^2 / n] \qquad (4.8)$$

en donde

$$y_p = (\sum y_i) / n \quad y \quad x_p = (\sum x_i) / n$$

son los promedios de y_i y x_i, respectivamente. Por consiguiente, $\tilde{\beta}_o$ y $\tilde{\beta}_1$ en las ecuaciones (4.7) y (4.8) son los **estimadores por mínimos cuadrados** de la ordenada al origen y la pendiente, respectivamente. El modelo ajustado de regresión lineal simple es, entonces,

$$\tilde{y} = \tilde{\beta}_o + \tilde{\beta}_1 x \qquad (4.9)$$

La ecuación (4.9) produce un estimado puntual de la media de y para una determinada x.

Como el denominador de la ecuación (4.8) es la suma corregida de cuadrados de las x_i y el numerador es la suma corregida de los productos cruzados de x_i y y_i, estas ecuaciones se pueden escribir en una forma más compacta como sigue:

$$S_{xx} = \sum^n_{i=1} x^2_i - (\sum^n_{i=1} x_i)^2 / n = \sum^n_{i=1} (x_i - x_p)^2 \qquad (4.10)$$

y

$$S_{xy} = \sum^n_{i=1} x_i y_i - (\sum^n_{i=1} x_i)(\sum^n_{i=i} y_i) / n = \sum^n_{i=1} y_i (x_i - x_p) \qquad (4.11)$$

Entonces, una forma cómoda de escribir la ecuación (4.8) es

$$\tilde{\beta}_1 = S_{xy} / S_{xx} \qquad (4.12)$$

La diferencia entre el valor observado y_i y el valor ajustado correspondiente y^*_i se llama residual. Matemáticamente, el i-ésimo residual es

$$e_i = y_i - \tilde{y}_i = y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i), \quad i = 1, 2, \ldots, n \tag{4.13}$$

Varianzas y desviaciones estándar de los estimadores de mínimos cuadrados.

De las ecuaciones (4.7) y (4.12) se infiere que los estimadores de mínimos cuadrados están en función de la muestra. Sin embargo, como los datos pueden cambiar fácilmente de una muestra a otra, de igual manera cambiarán las estimaciones que se obtengan. Por consiguiente, se hace necesario encontrar alguna medida de la confiabilidad o precisión de los estimadores $\tilde{\beta}_0$ y $\tilde{\beta}_1$. En estadística, la precisión de un estimador se mide por medio de su error estándar (se $\tilde{\beta}_i$)[1]. Dados los supuestos del modelo de regresión lineal, los errores estándar de los estimadores de MCO pueden obtenerse mediante:

$$\text{Var}(\tilde{\beta}_1) = \sigma^2 / \Sigma x^2_i \tag{4.14}$$

donde $\Sigma x^2_i = \Sigma(X_i - \tilde{X})^2$

$$\text{Se}(\tilde{\beta}_1) = \sigma / \sqrt{\Sigma x^2_i} \tag{4.15}$$

$$\text{Var}(\tilde{\beta}_0) = \Sigma X_i^2 \, \sigma^2 / n \, \Sigma x^2_i \tag{4.16}$$

$$\text{Se}(\tilde{\beta}_0) = (\sqrt{\Sigma X_i^2} / n \, \Sigma x^2_i) \, \sigma \tag{4.17}$$

Todas las cantidades que aparecen en estas últimas cuatro ecuaciones pueden estimarse a partir de los datos de la muestra con excepción de la varianza σ^2 de e_i. Dicho parámetro se estima mediante la fórmula siguiente:

$$\tilde{\sigma}^2 = \Sigma e^2_i / (n - 2) \tag{4.18}$$

donde $\tilde{\sigma}^2$ es el estimador de MCO del verdadero pero desconocido σ^2 y el denominador $n - 2$ se conoce como el número de grados de libertad y Σe^2_i es la suma de los residuos al cuadrado (SRC).

[1] El error estándar no es otra cosa que la desviación estándar de la distribución muestral del estimador, siendo ésta última una distribución de frecuencias o de probabilidades de un estimador.

La demanda del TAP en función del precio de traslado origen-destino.

Sea la siguiente función de demanda de TAP, como una especificación no lineal que explica los cambios de la variable de estudio: el total de pasajeros transportados según el precio de traslado origen-destino.

$$\text{Pasajet} = \beta_o - \beta_1\text{preciod2} + \beta_2\text{preciod3} - \beta_3\text{preciorc} \qquad (1)$$

Donde, la variable de estudio: pasajet indica la cantidad total de pasajeros transportados (vía aérea) en el periodo: 2003 a 2009, en las rutas: Monterrey-México, Cancún-México y México-Guadalajara. La variable independiente preciod2 indica el precio de traslado origen-destino retrasado dos periodos (años), la variable preciod3 indica el precio retrasado tres periodos (años) y la variable preciorc señala la raíz cuadrada del vector precio inicial.

Así pues se presenta el cuadro siguiente que resume los resultados de la estimación puntual del modelo (1)

Cuadro 1

Variable dependiente	**pasajet**
Método	Mínimos cuadrados
Muestra ajustada	4 a 21
Observaciones incluidas	18 (después de ajustes)
Software empleado	E-views 5.0

Variable	**coeficiente**	**Std error**	**t-statiscs**	**Prob**
C	2252.473	519.3936	4.336736	0.0007
preciorc	-55.95953	12.36369	-4.526119	0.0005
Preciod3	0.767364	0.130619	5.874819	0.0000
Preciod2	-0.049220	0.081060	-0.607208	0.5534

R^2	**0.713893**	**Meandependvar**	**1134.611**
RA^2	0.652585	SD depend var	423.2371
SEof regresion	249.4641	Akaike info criter	14.06964
SRC	871252.9	Schwarz criterion	14.26750
Loglikelihood	-122.6267	F statistics	11.64427
DWstat	1.399083	Prob(F statistics)	0.000427

De acuerdo con los resultados presentados por la función de demanda (1) en el cuadro 1, por cada unidad porcentual que aumenta la variable independiente: preciod3, la variable de estudio: pasajet aumenta un 76.7

por ciento aproximadamente (es decir, su tasa de crecimiento anual es del 76.7 por ciento), considerando constantes a las otras variables de dicha función, lo que refleja el impacto importante de la variable: preciod3 (es decir, el precio por traslado origen-destino, rezagado tres periodos). A su vez, por cada unidad porcentual que aumenta la variable independiente: preciorc, la variable de estudio: pasajet disminuye un 5595 por ciento, aproximadamente (es decir, su tasa de crecimiento anual es de -5595 por ciento), considerando constantes las otras variables de la función de demanda. Este porcentaje de cambio en la variable pasajet resulta muy alto si consideramos los otros porcentajes asociados con las otras variables de la función de demanda (1) y sugiere la pregunta siguiente: ¿cómo interpretar económicamente el efecto de la variable preciorc en los cambios de la variable de estudio: pasajet?. Una respuesta a esta interrogante está en la sensibilidad del mercado interno del TAP ante cambios no constantes del precio de traslado aéreo origen-destino. Recuérdese que el mercado doméstico de aviación es un mercado oligopólico y está concentrado en algunas rutas donde se concentra la actividad económica nacional como la Ciudad de México, Guadalajara, Monterrey, Cancún, Acapulco, Veracruz, Tijuana, entre otros. Además de que dicho mercado interno es relativamente reducido, en comparación con el mercado interno de E.U. o de Canadá, dado que tal característica es consecuencia de las caídas de la tasa de crecimiento anual (TCRA3, TCRA4) de la demanda de TAP en México y por las caídas en el ritmo de expansión de la actividad económica nacional e ilustrado por la tasa de crecimiento del PIB y el PIBP, en por lo menos los últimos veinte años.

Por otro lado, las variables independientes: preciod3 y preciorc de la función (2) son significativas estadísticamente puesto que sus estadísticos $t_c = 5.8464 > t_{\alpha/2} = 2.11$ (para el caso de la variable: preciod3) dado un nivel de significancia $\alpha = 0.05$, tal que se rechaza la $H_o:\beta_2 = 0$ frente a la $H_a:\beta_2 \neq 0$. De manera análoga, el estadístico $t_c = /-4.526/ > t_{\alpha/2} = 2.11$ (para el caso de la variable: preciorc) dado un nivel de significancia $\alpha = 0.05$, tal que se rechaza la $H_o:\beta_1 = 0$ frente a la $H_a:\beta_1 \neq 0$. Observe que la variable: preciod2, no es estadísticamente significativa, debido a que su estadístico t_c permite no rechazar la $H_o:\beta_3 = 0$ frente a la $H_a:\beta_3 \neq 0$, aunque se incluyo en la función por motivos estadísticos (la presencia de dicha variable facilita el que la función (2) no presente problemas de autocorrelacion de primer y segundo orden). Observe que la bondad del ajuste de esta función correspondiente a su coeficiente de determinación (R^2) es del 0.7138, lo que indica que la función (1) explica el 71.3 por ciento aproximadamente de los cambios de la variable de estudio: pasajet. Observe que el modelo es significativo estadísticamente en su conjunto puesto que $F_c = 11.6442 > F_\alpha$

= 3.16, dado un $\alpha = 0.05$ y así, se rechaza la H_o: $\sum\beta_i = 0$ frente a la H_a: $\sum\beta_i \neq 0$.

Los datos que forman la muestra para estimar la función (2) se presentan a continuación, que incluyen a las variables precio, pasajet , preciorc (raíz cuadrada del precio de traslado), preciod2 (precio retrasado dos periodos) y preciod3 (precio retrasado tres periodos).

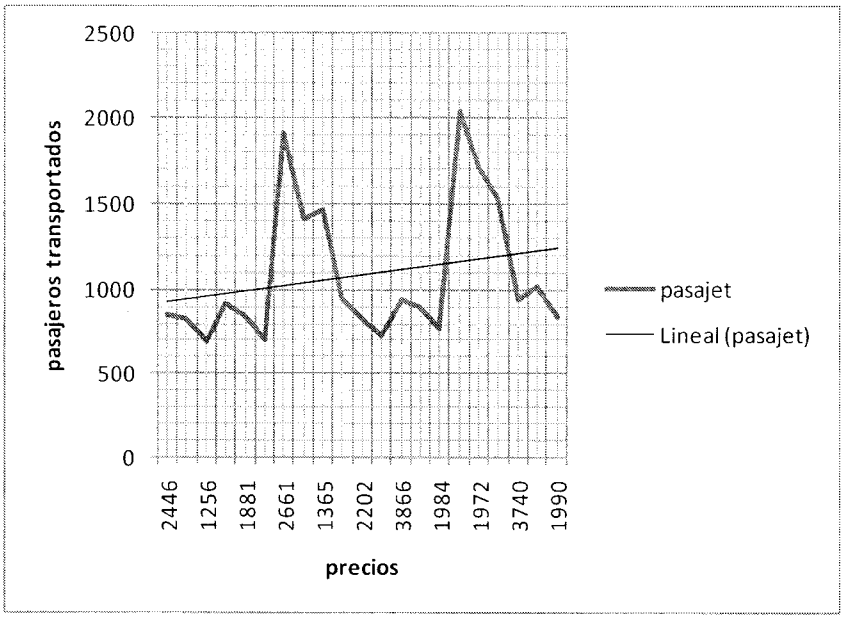
Cuadro 2. Precios y pasajeros transportados en las rutas Monterrey-México, Cancún-México y México-Guadalajara.

Precio	pasajet	preciorc	Preciod2	Preciod3
2646	848	51.4392		
1715	829	41.4125		
1256	689	35.4400	2446	
2683	918	51.7976	1715	2446
1881	842	43.3705	1256	1715
1378	705	37.1214	2683	1256
2661	1912	51.5848	1881	2683
1863	1404	43.1624	1378	1881
1365	1460	36.9460	2661	1378
3146	955	56.0892	1863	2661
2202	829	46.9254	1365	1863
1614	723	40.1746	3146	1365
3866	939	62.1771	2202	3146
2706	900	52.0192	1614	2202
1984	770	44.5421	3866	1614
2770	2031	52.6307	2706	3866
1972	1708	44.4072	1984	2706
1474	1528	38.3927	2770	1984
3740	938	61.1555	1972	2770
2662	1017	51.5945	1474	1972
1990	844	44.6094	3740	1474

Fuente: SCT. Estadística operacional origen-destino en servicio regular nacional. Periodo: 2003, 2009.

Ahora es conveniente presentar una grafica que ilustre la relación entre los precios de traslado y los pasajeros según los datos del cuadro 4.

Grafica 1.



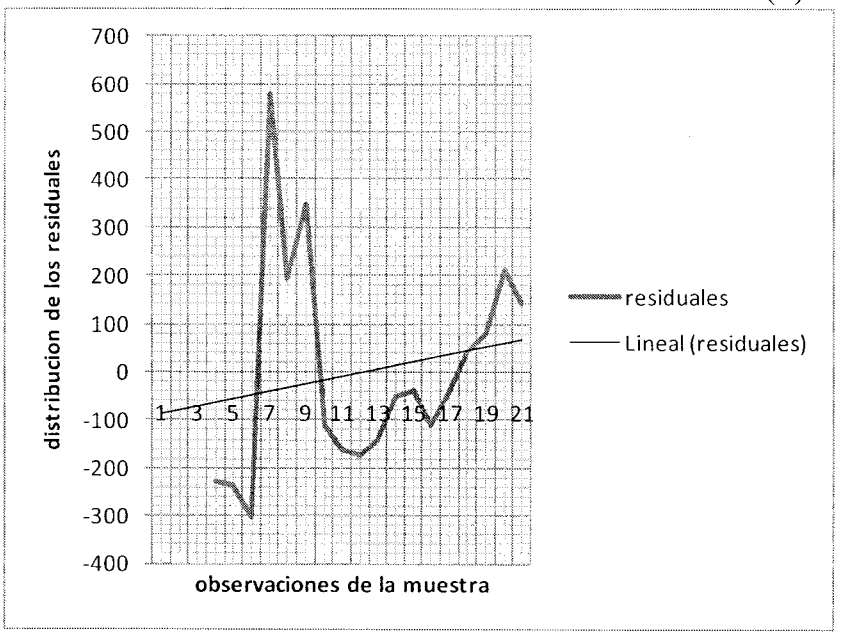
Note la tendencia positiva de la variable pasajet a lo largo del periodo: 2003 a 2009. Esto se debe a la concentración de la actividad económica nacional que se concentra en los destinos de las rutas comprendidas según el cuadro 4.

Tabla ANOVA de la función (1)

Modelo	sumadecuadrados	gdel	mediadecuadrados	F	sig
Regresión	2173951.406	3	724650.469	11.644	0.000
Residual	871252.872	14	62232.348		
total	3045204.278	17			

Observe que los datos de la tabla ANOVA confirman los comentarios presentados para la función estadística (1) y los resultados del cuadro 1. Esto puede comprobarse con el valor de $F_c = 11.644 > F_\alpha = 3.16$, dado un $\alpha = 0.05$ y así, se rechaza la H_o: $\sum\beta_i = 0$ frente a la H_a: $\sum\beta_i \neq 0$.
SRC = 871252.872 < 2173951.406 = SEC y con R^2 = SEC/STC=0.713893.

Grafica 2. Distribución de los residuales de la función (1)



Para el modelo 1. Software: E-views 5.
Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.457792	Probability	0.643284
Obs*R-squared	1.276018	Probability	0.528343

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 05/30/11 Time: 16:28
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.632040	543.0366	-0.015896	0.9876
PRECIOD3	0.016093	0.137662	0.116904	0.9089
PRECIOD2	0.004714	0.084681	0.055670	0.9565
PRECIORC	-0.713230	12.89736	-0.055300	0.9568
RESID(-1)	0.261993	0.290926	0.900549	0.3855
RESID(-2)	0.035497	0.303576	0.116929	0.9089

R-squared	0.070890	Mean dependent var	-7.26E-13
Adjusted R-squared	-0.316239	S.D. dependent var	226.3850
S.E. of regression	259.7258	Akaike info criterion	14.21833
Sum squared resid	809489.8	Schwarz criterion	14.51512
Log likelihood	-121.9650	F-statistic	0.183117
Durbin-Watson stat	1.888085	Prob(F-statistic)	0.963594

Resultados y comentarios

La función estadística (1) presenta buenos resultados de estimación de la variable de estudio: pasajeros transportados vía aérea en rutas nacionales y por servicio regular (pasajet) debido a que dicha función se fundamenta en la significancia estadística de las variables que explican los cambios de la variable de estudio. Tales variables son: preciorc, preciod3, cuyo significado está ya definido. Así pues, la demanda de transportación aérea aumenta un 76.7 por ciento por cada unidad porcentual que aumenta la variable: preciod3, si los demás variables de la función (1) son constantes, lo que refleja el impacto importante de la variable: preciod3 (es decir, el precio por traslado origen-destino, rezagado tres periodos). A su vez, por cada unidad porcentual que aumenta la variable independiente: preciorc, la variable de estudio: pasajet disminuye un 5595 por ciento, aproximadamente (es decir, su tasa de crecimiento anual es de -5595 por ciento), considerando constantes las otras variables de la función de demanda. Este porcentaje de cambio en la variable pasajet resulta muy alto si consideramos los otros porcentajes asociados con las otras variables de la función de demanda (1) y sugiere la pregunta siguiente: ¿cómo interpretar económicamente el efecto de la variable preciorc en los cambios de la variable de estudio: pasajet?. Una respuesta a esta interrogante está en la sensibilidad del mercado interno del TAP ante cambios no constantes del precio de traslado aéreo origen-destino. Recuérdese que el mercado doméstico de aviación es un mercado oligopólico y está concentrado en algunas rutas donde se concentra la actividad económica nacional como la Ciudad de México, Guadalajara, Monterrey, Cancún, Acapulco, Veracruz, Tijuana, entre otros.

Observe que la variable independiente: preciod2 es estadísticamente no significativa tal como lo indica su estadístico $t_c = -0.6072 < t_{\alpha/2} = 2.11$ dado un nivel de significancia $\alpha = 0.05$, tal que no se rechaza la $H_o: \beta_1 = 0$ frente a la $H_a: \beta_1 \neq 0$. Observe que la bondad del ajuste de esta función correspondiente a su coeficiente de determinación (R^2) es del 0.7138, lo que indica que la función (1) explica el 71.3 por ciento aproximadamente de los cambios de la variable de estudio: pasajet. Observe que el modelo es significativo estadísticamente en su conjunto puesto que $F_c = 11.6442 > F_\alpha = 3.16$, dado un $\alpha = 0.05$ y así, se rechaza la $H_o: \sum\beta_i = 0$ frente a la $H_a: \sum\beta_i \neq 0$. Finalmente, la prueba Breusch-Godfrey de autocorrelacion serial de los residuales permite afirmar que la función (1) está libre de este problema, dado que los resid(-1) y resid(-2) no son estadísticamente significativos.