Conferencia 6. Regresión Lineal Simple

Tema 4. Regresión.

Sumario:

1. Introducción.
2. Regresión Lineal Simple
3. Supuestos
4. Método de Mínimos Cuadrados.
5. Análisis de Residuos
6. Predicciones
7. Restricciones.

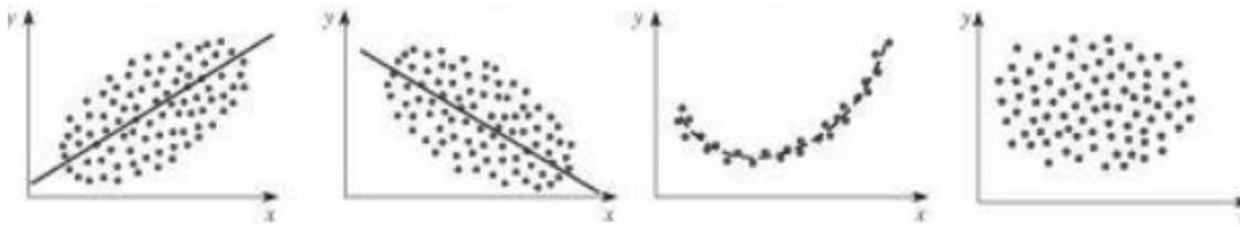
1. Introducción.

Aun cuando el coeficiente de correlación mide la fuerza de una relación lineal, no nos dice nada acerca de la relación matemática entre las dos variables. Cuando tenemos dos variables correlacionadas como Cantidad de Anuncios Promocionales y Cantidad de Ventas, el coeficiente de correlación nos dice si existe relación, que tipo y qué fuerza tiene la misma, pero no nos ayuda a pronosticar la cantidad de ventas en función de la cantidad de anuncios. El análisis de la regresión lineal consiste en encontrar la ecuación de la recta que mejor describe la relación entre las dos variables, uno de los usos de esta ecuación es hacer predicciones y hacemos uso de dichas predicciones con cierta regularidad. Por ejemplo pudiéramos predecir el éxito que un estudiante puede tener en la universidad basándonos en sus notas en el pre o predecir la distancia necesaria para que un auto se detenga si frena a determinada velocidad.

La relación entre dos variables será una expresión algebraica que describe la relación matemática entre x y y. Existen diversas relaciones posibles llamadas modelos o ecuaciones de predicción.

Lineal	$\hat{y} = b_0 + b_1 x$
Cuadrática	$\hat{y} = a + bx + cx^2$
Exponencial	$\hat{y} = a(b^x)$
Logarítmica	$\hat{y} = a \log x$

Las ecuaciones anteriores definen modelos de regresión, que no son más que patrones que siguen los datos, algunos de estos patrones graficados lucen como se muestran en la figura 1. Pero todos los datos no se comportan en formas tan “bonitas”. Muchas veces de ver el grafico no queda claro que patrón siguen los datos o siquiera si siguen un patrón especifico. Por lo que, realizando solo un análisis de correlación no podemos predecir el valor de una variable usando otras.



Regresión Lineal con Pendiente Positiva | *Regresión Lineal con Pendiente Negativa* | *Regresión Cuadrática* | *No hay Regresión*

Figura 1. 4 Tipos de Regresión

1. Regresión Lineal Simple

La idea detrás de la regresión es encontrar el modelo que mejor se ajusta a nuestros datos. Este puede ser un modelo lineal, cuadrático, logarítmico, etc. Escoger el modelo al cual nuestros datos se ajustan mejor es un área de estudio que no veremos ahora. Baste saber que para decidir si nuestros datos se ajustan al modelo de Regresión Lineal Simple es necesario que nuestra muestra cumpla una serie de suposiciones a las que se les llama supuestos del modelo. Planteemos de forma matricial la recta de la Regresión Lineal Simple.

$$Y = X\beta + e \tag{1}$$

Donde:

$Y_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ Vector del Término Dependiente. Variable Aleatoria Observada

$X_{n\times 2} = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}$ Matriz de Términos Independientes o Matriz del Diseño.

$\beta_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ Vector de los Parámetros. O Sea la pendiente (β_1) y el intercepto (β_0)

$e_{n\times 1} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ Vector de los errores. Variable Aleatoria.

En forma algebraica tendríamos que para $i = 1, \dots, n$ tenemos

$$y_i = \beta_0 + x_i\beta_1 + e_i \tag{2}$$

3. Supuestos del Modelo de Regresión Lineal Simple

Por tanto, los supuestos del modelo de regresión lineal simple son:

1. Los errores $(e_1, \dots, e_n)$ son independientes.
2. El valor esperado del error aleatorio e_i es cero $(E(e_i) = 0)$
3. La Varianza del error aleatorio es constante $(V(e_i) = \sigma^2)$. Homocedasticidad.
4. Los errores además de ser independientes son idénticamente distribuidos y siguen distribución normal con media cero y varianza constante $\left(e_i \sim N(0, \sigma^2)\right)$

Como conclusión natural de los supuestos la variable Y, que es aleatoria pero observada, sigue una distribución $N(X\beta, \sigma^2)$.

El cumplimiento de los supuestos del modelo se puede observar a través el análisis grafico de los residuos. La definición de los residuos la veremos más adelante.

Baste saber que para garantizar la veracidad de nuestro modelo necesitamos que se cumplen los supuestos.

4. Método de Mínimos Cuadrados.

El problema de la regresión es buscar la relación entre dos o más variables, en el caso de la Regresión Lineal Simple, buscamos la relación entre una variable dependiente (Y) y una variable independiente (X). Es posible que tengamos más de una relación entre dichas variables como se muestra en la figura 2, por tanto, se introduce en el modelo el termino del error. Se escoge una ecuación a partir del conjunto de posibles ecuaciones, utilizando algún método de estimación.

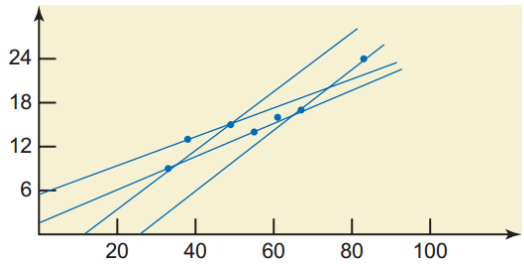


Figura 2. Posibles Rectas de Regresión.

Uno de los métodos de estimación que se pueden utilizar en la Regresión Lineal es el **método de mínimos cuadrados**. Supongamos que la ecuación (3) de una recta donde $\hat{y}$ representa el valor pronosticado de y que corresponde a un valor particular de x. El criterio de mínimos cuadrados requiere que encontremos las constantes b_0, b_1 tales que $\sum(y - \hat{y})^2$ sea lo más pequeña posible, definiendo el termino del error en la ecuación (4)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (3)$$

$$e = \hat{y} - y \quad (4)$$

La figura 3 muestra la distancia de un valor observado de y desde un valor pronosticado de $\hat{y}$. La longitud de esta distancia representa el valor $(y - \hat{y})$ (mostrados como segmentos de recta rojos en la figura 4). Note que $(y - \hat{y})$ es positiva cuando el punto (x, y) esta por encima de la recta y negativa cuando está por debajo.

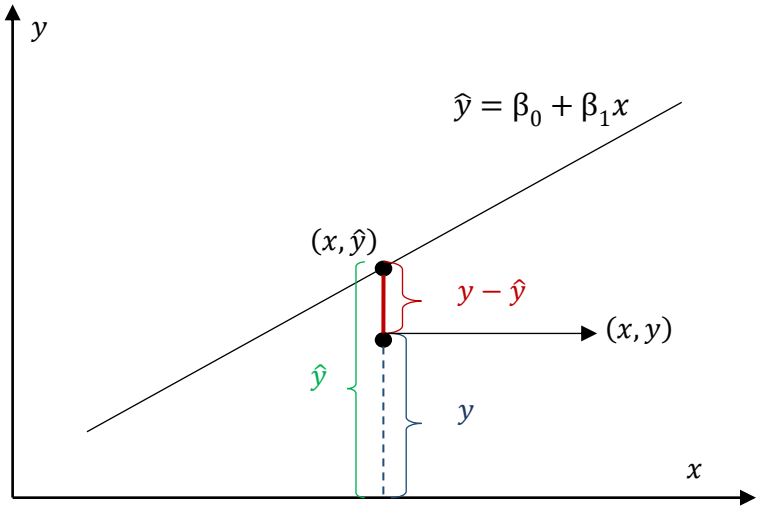
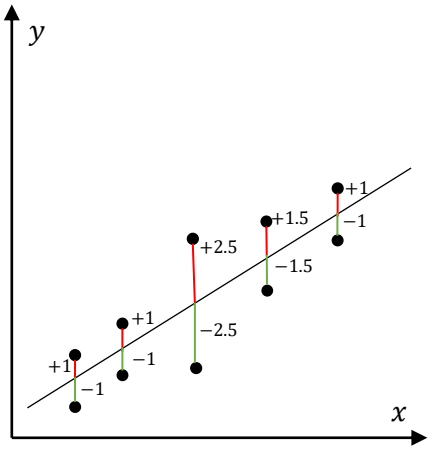
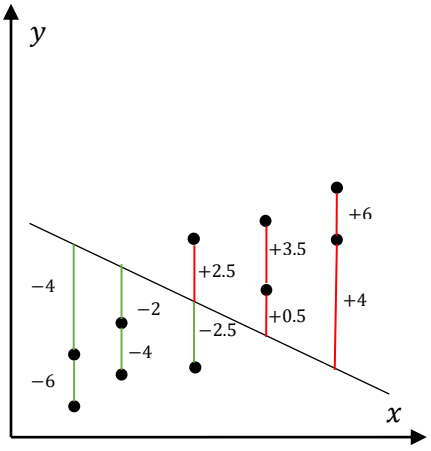


Figura 3. Regresión Lineal. Mínimos Cuadrados. Recta de Mejor Ajuste.





$$\sum (y - \hat{y})^2 = (-1)^2 + (1)^2 + (-1)^2 + (1)^2 + (-2.5)^2 + (2.5)^2 + (-1.5)^2 + (1.5)^2 + (-1)^2 + (1)^2 = 23$$

Figura 4. La Recta de Mejor Ajuste.

$$\sum (y - \hat{y})^2 = (-6)^2 + (-4)^2 + (-4)^2 + (-2)^2 + (-2.5)^2 + (2.5)^2 + (0.5)^2 + (3.5)^2 + (4)^2 + (6)^2 = 149$$

Figura 5. Una Recta que no es de Mejor Ajuste.

En la figura 4 se muestra un diagrama de dispersión con lo que parece ser la recta de mejor ajuste, junto con 10 valores individuales $(y - \hat{y})$. Los valores positivos están en rojo los negativos en verde. La suma de los cuadrados de estas diferencias se minimiza si la recta en realidad es la de mejor ajuste.

En la figura 5 se muestra un ejemplo de una recta que no se ajusta tan bien a los puntos mostrados en la figura 4. En este caso $\sum(y - \hat{y})^2 = 149$ mucho mayor que para la recta de la figura 3 donde $\sum(y - \hat{y})^2 = 23$. Toda recta diferente trazada por este conjunto de 10 puntos resultara un valor diferente para $\sum(y - \hat{y})^2$. Nuestro trabajo es hallar la recta que haga $\sum(y - \hat{y})^2$ el mínimo valor posible.

La ecuación de la recta de mejor ajuste está determinada por la pendiente $\hat{\beta}_1$ y su ordenada $\hat{\beta}_0$. Estos valores se encuentran usando la formulas siguiente:

Pendiente $$\hat{\beta}_1 = \frac{\sum(x - \hat{x})(y - \hat{y})}{\sum(x - \hat{x})^2} \quad (5)$$

Usaremos un equivalente de la fórmula matemática para la pendiente que utiliza las sumas de cuadrados encontrados en los cálculos preliminares de la correlación

Pendiente (formula computacional) $$\hat{\beta}_1 = \frac{SS(xy)}{SS(x)} \quad (6)$$

Suma de cuadrados de x y y $$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} \quad (7)$$

Suma de cuadrados de x $$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} \quad (8)$$

Ejemplo 1.

Sean las variables $X =$ Peso Corporal y $Y =$ Notas de una prueba de Eficiencia Muscular en un Preuniversitario. Tenemos las observaciones de 7 individuos y con estas obtenemos la siguiente tabla.

X	Y	$X * Y$	X^2	Y^2
120	100	12000	14400	10000
115	100	11500	13225	10000
150	75	11250	22500	5625
170	60	10200	28900	3600
120	95	11400	14400	9025
130	90	11700	16900	8100
110	100	11000	12100	10000
915	620	79050	122425	56350

Ahora deseamos hallar la recta que mejor se ajuste $\hat{y} = b_0 + b_1 x$ usando la fórmula 6 aprovechando lo calculado en la tabla de correlación y entonces:

$$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} = 79050 - \frac{915 * 620}{7} = 79050 - \frac{567300}{7}$$
$$= 79050 - 81042.86 = \mathbf{-1992.86} \quad (9)$$

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 122425 - \frac{837225}{7} = 122425 - 119603.57 = \mathbf{2821.43} \quad (10)$$

Sustituyendo 9 y 10 en 6 obtenemos el valor de la pendiente. De aquí con la fórmula 12 calculamos la ordenada que depende del valor de la pendiente y de los valores de y. Los conceptos de pendiente y ordenada son de geometría analítica, contenido anterior a esta materia. Por lo que si tiene dudas de los despejes de alguna de las formulas sugerimos consulte un libro de Geometría Analítica.

Pendiente $$\hat{\beta}_1 = \frac{SS(xy)}{SS(x)} = \frac{-1992.86}{2821.43} = \mathbf{-0.71} \quad (11)$$

Ordenada en el origen
$$\hat{\beta}_0 = \frac{\sum y - (b_1 * \sum x)}{n} = \frac{620 - (-0.71 * 915)}{7}$$
$$\hat{\beta}_0 = \frac{620 + 646.29}{7} = \mathbf{180.9} \quad (12)$$

Sustituyendo 11 y 12 en la ecuación 1 que representa el modelo de regresión lineal simple obtenemos la recta de mejor ajuste.

$$\hat{y} = 180.9 - 0.71x \quad (13)$$

Recuerde al realizar los cálculos como mínimo conservar 2 cifras significativas, para mayor exactitud.

Como tenemos el diagrama de dispersión vamos a trazar la recta de mejor ajuste para poder ver mejor la relación de la recta y los datos reales de nuestro problema. Para esto necesitamos dos puntos en el extremo dentro de los valores de nuestro problema. Tomemos entonces $x_1 = 80$ y $x_2 = 180$ que en términos de nuestro problema son valores extremos, para tener una mejor idea de la recta.

Sustituyendo x_1 en 13 $$y_1 = 180.9 - 0.71 * 80 = 124.1 \quad (14)$$

Sustituyendo x_2 en 13 $$y_2 = 180.9 - 0.71 * 180 = 53.1 \quad (15)$$

Trazando la recta de mejor ajuste utilizando los puntos $(80, 124.1)$ y $(180, 53.1)$, diferenciándose de alguna manera de los puntos de los datos, en este caso optamos por colorear los puntos de rojo, pero en muchos sistemas se usa un signo de + para diferenciarlos de los datos originales.

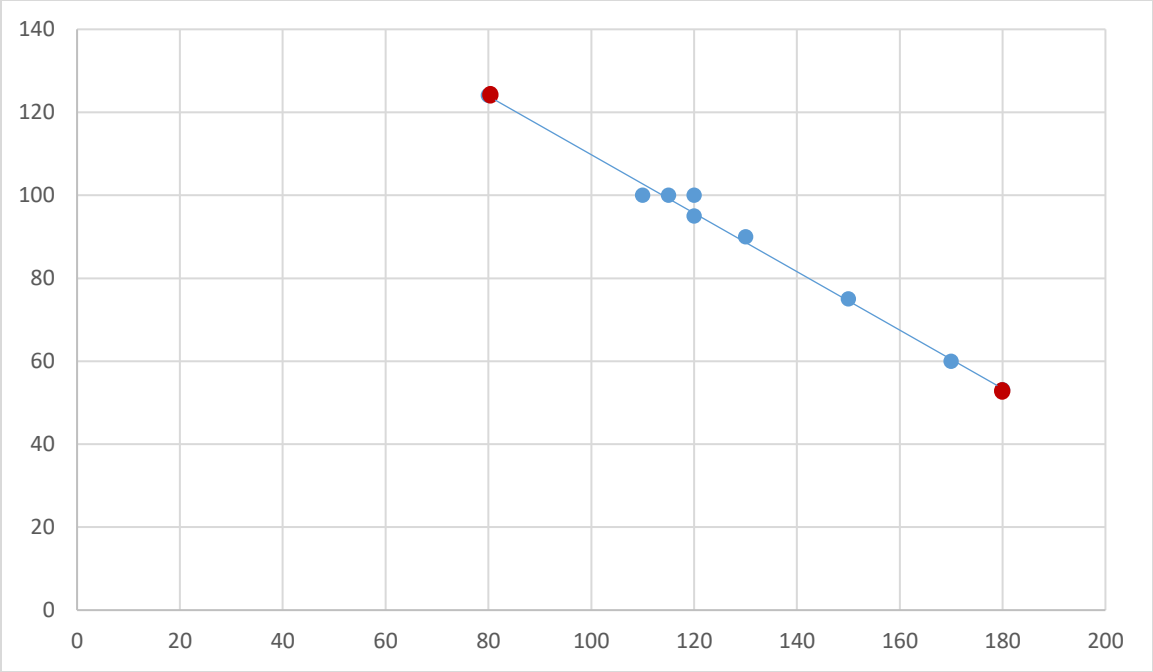


Figura 6. Diagrama de Dispersión con Recta de Mejor Ajuste.

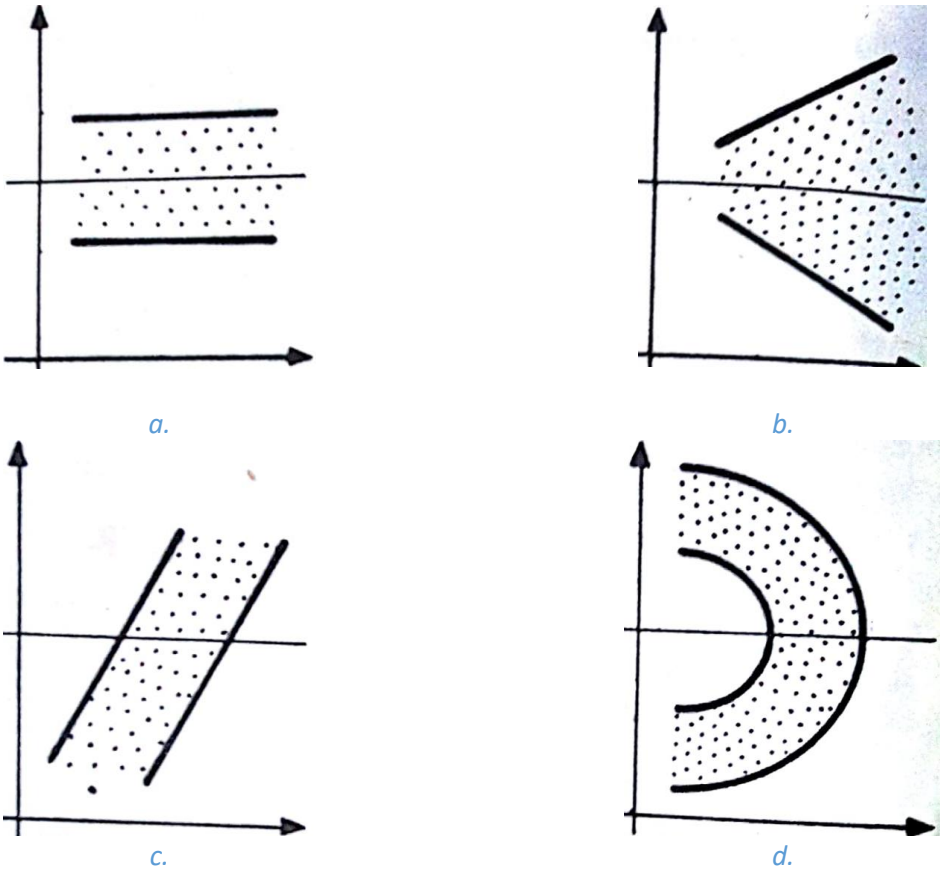
Los valores de la pendiente y la ordenada son más que simples números para trazar una recta de mejor ajuste, tienen interpretación y significado que nos permite realizar inferencias y hallar patrones en los datos.

1. La pendiente $\hat{\beta}_1$, representa el cambio pronosticado en y por un aumento o disminución unitaria en x. En el ejemplo donde $\hat{\beta}_1 = -0.71$ si un estudiante gana 15 libras de peso adicional pronosticamos que sus notas pueden caer al menos en 11 puntos $(-0.71 * 15)$. En este caso sería perder puntos pues el signo de la pendiente es negativo lo que indica una relación inversa.

2. La ordenada en el origen es el valor donde la recta de mejor ajuste cruza el eje y (si en la escala del problema tiene sentido que $x = 0$). Este no es el caso pues no puede existir nadie que pese 0 libras, por lo que en este caso $\hat{y} = 180.9$ no tiene sentido porque predecir que un estudiante obtendría 180 puntos si pesa 0 libras es absurdo. Por lo que hay que tener cuidado con la interpretación de la ordenada.

5. Análisis de Residuos.

El análisis de residuos se hace mayormente de forma gráfica, existen cuatro patrones básicos en este tipo de análisis que se mostraran en las figuras a continuación. Obteniéndose como resultado de este análisis si los supuestos han sido violados o no.



a. *b.*

c. *d.*

Figura 7. Gráficos de los Residuos.

En el caso de la figura 7.a indica que las hipótesis no fueron violadas por tanto los supuestos del modelo se cumplen. Las figuras 7.b, c, d indican que las suposiciones fueron violadas. De forma más específica en el grafico 7.b la varianza no es constante y por tanto no se cumple el supuesto de Homocedasticidad. La figura 7.c sugiere que es necesaria la inclusión de otro termino lineal. Por último la figura 7.d indica que no se cumple el supuesto de linealidad.

6. Predicciones

Uno de los objetivos de buscar la ecuación de la regresión es poder hacer predicciones una vez que logramos establecer una relación lineal (si no existe relación, no existe recta de mejor ajuste y no se pueden realizar predicciones). Una vez establecida la relación lineal y conocido el valor de la variable de entrada x, podemos predecir un valor de $y, \hat{y}$. Considere la ecuación $\hat{y} = 180.9 - 0.71x$ que relaciona el peso de un estudiante contra la nota de las pruebas de eficiencia muscular. Si un estudiante pesa 125 libras ¿Qué nota pronostica usted que obtendrá en la prueba de eficiencia muscular? En este caso el valor pronosticado ser 92.15 puntos en la prueba.

$$\hat{y} = 180.9 - 0.71x = 180.9 - 0.71 * 125 = 180.9 - 88.75 = \mathbf{92.15} \qquad (16)$$

Es importante realizar una aclaración, al realizar este pronóstico no se debe esperar que el valor pronosticado sea exacto, más bien es la nota promedio que se esperaría para todos los estudiantes que pesen 125 libras.

7. Restricciones.

1. La ecuación debe usarse para hacer predicciones solo acerca de la población de la cual se extrajo la muestra. Por ejemplo, utilizar la ecuación 13 para predecir el resultado de pruebas de eficiencia muscular en un círculo infantil, porque los resultados serían cuestionables dado las diferencias de peso.

2. La ecuación solo debe usarse dentro del dominio muestral de la variable de entrada. Porque sabemos que los datos demuestran una tendencia lineal dentro del dominio de los datos en x, pero no sabemos cuál es la tendencia fuera de este intervalo. En consecuencia los pronósticos tienen un alto riesgo fuera del dominio de los datos en x. En el ejemplo no tiene sentido predecir la nota para un peso de 0 libras. En ocasiones podría predecirse valores fuera del dominio de x, pero debe hacerse con precaución y siempre para valores cercanos al intervalo del dominio.

3. Si la muestra fue tomada en el 2000, no espere que los resultados sean válidos para muestras tomadas en 1925 o que se cumplan en 2010. Los estudiantes pueden ser muy diferentes de una década a otra.