UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

E.A.P. DE ESTADÍSTICA

"REGRESIÓN NO PARAMÉTRICA UTILIZANDO SPLINE PARA LA SUAVIZACIÓN DE LA ESTRUCTURA DE LA MORTALIDAD EN EL PERÚ"

TESIS

Tesis para optar al Título Profesional de

Licenciado en Estadística

AUTOR

Luis Alberto Meza Santa Cruz

ASESOR

Mg. Rosa Ysabel Adriazola Cruz

Lima-Perú

2013

A la memoria de mi madre Leonidas Santa Cruz Ponce Vda. de Meza y de mis abuelos Agustín Santa Cruz San Miguel y Venancia Ponce Pardavé de Santa Cruz

Reconocimiento y agradecimiento

Mi reconocimiento y agradecimiento eterno a mi Profesora Asesora Magister Rosa Ysabel Adriazola Cruz, por haber creído en mis capacidades y haberme enseñado, orientado y ayudado incansablemente y con mucha paciencia durante los 2 años que llevó preparar esta Tesis, primero como mi Profesora de los cursos Tesis I y Tesis II, y luego como mi Profesora Asesora en la continuación y consumación de mi Tesis. Un millón de Gracias por sus constantes palabras de aliento y por su apoyo incondicional. Sin su ayuda no hubiera sido posible elaborar esta Tesis.

Muchas gracias a los profesores de Matemáticas Mg. Víctor Osorio Vidal y Mg. Jorge Ícaro Condado Jaúregui, por su valioso tiempo empleado en enseñarme la parte matemática del desarrollo de los polinomios en general y al entendimiento del funcionamiento matemático del polinomio Spline en particular.

Muchas gracias a las Profesoras Dra. Doris A. Gómez Ticerán, Mg. Ana María Cárdenas Rojas y Mg. Olga Lidia Solano Dávila, por sus palabras de aliento constante.

Muchas gracias al Profesor Mg. Antonio Bravo Quiroz, por su apoyo académico y amical incondicional.

Muchas gracias a los estadísticos Johnny Iván Bravo Alonso y Mixsi Joanne Casas Bendezú, por su apoyo desinteresado en la parte computacional del modelo Spline.

Mi reconocimiento a mi madre Leonidas Santa Cruz Ponce Vda. de Meza y mis abuelos Agustín Santa Cruz San Miguel y Venancia Ponce Pardavé de Santa Cruz, cuya honradez, dedicación y enseñanzas hicieron posible el forjarme con los Valores Morales que hoy poseo.

Mi reconocimiento a mi padre el Eco. Baldomero Meza Sánchez por su constancia en aprender cada día más, y su honradez, que al igual que mi madre y mis abuelos maternos, contribuyeron en mi formación como persona.

REGRESIÓN NO PARAMÉTRICA UTILIZANDO SPLINE PARA LA SUAVIZACIÓN DE LA ESTRUCTURA DE LA MORTALIDAD EN EL PERÚ

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para optar el Título Profesional de Licenciado en Estadística

Aprobada por:

Mg. Ana María Cárdenas Rojas _____
Presidenta

Lic. Grabiela Montes Quintana _____
Miembro

Mg. Rosa Ysabel Adriazola Cruz _____
Miembro - Asesor

Lima – Perú
Diciembre 2013

FICHA CATALOGRÁFICA

MEZA SANTA CRUZ, LUIS ALBERTO

Regresión no paramétrica utilizando Spline para la suavización de la estructura de la mortalidad en el Perú, (Lima) 2013.

106, 30 cm. (UNMSM, Licenciado en Estadística, 2013)
Tesis, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas. Estadística.
I. UNMSM / F. de C.M. II. Regresión no paramétrica utilizando Spline para la suavización de la estructura de la mortalidad en el Perú.

ÍNDICE

RESUMEN

REGRESIÓN NO PARAMÉTRICA UTILIZANDO SPLINE PARA LA SUAVIZACIÓN DE LA ESTRUCTURA DE LA MORTALIDAD EN EL PERÚ

Presentado por: Bachiller MEZA SANTA CRUZ, Luis Alberto
Profesora Asesora: Magister ADRIAZOLA CRUZ, Rosa Ysabel

DICIEMBRE 2013

En esta investigación se hace un estudio preliminar de la regresión en general y sus tipos para luego centrarse en el estudio teórico del Modelo de Regresión no Paramétrico Spline, que es un polinomio cúbico por secciones o trozos, demostrándose sus bondades y ductilidad con respecto a los polinomios en general.

Al unir dos polinomios para obtener un polinomio por secciones mayormente el punto de unión no es suave o simplemente no se unen, lo que deriva en cambios bruscos, pero si se utilizan polinomios Spline que es un caso particular de los polinomios por secciones, y que tiene como una de sus propiedades que la primera derivada de la función Spline hace que la unión no sea brusca y la segunda derivada permite la concavidad al unir dos polinomios, lográndose una curva suavizada de tendencia continua.

En las últimas décadas investigadores están utilizando modelos de regresión no paramétricos para suavizar curvas correspondientes a un conjunto de pares de datos, en demografía para hacer aproximaciones de la tendencia de los componentes demográficos tales como la fecundidad, mortalidad y la población propiamente dicha, por ello en la presente investigación se aplica el modelo de regresión no paramétrico Spline en la suavización la curva correspondiente a la estructura de mortalidad por sexo y edad utilizando datos de las defunciones de

las Estadísticas Vitales del año 2007 y Censo Nacional de Población del 2007, correspondientes al departamento de Lima.

Se concluye que la suavización de la estructura de la mortalidad con el Spline es adecuado y se sugiere su utilización como una forma alternativa de suavizamiento de dicha estructura.

Palabras clave

Spline, regresión no paramétrica, polinomio cúbico, estructura de muertes, interpolación.

ABSTRACT

NONPARAMETRIC REGRESSION WITH SPLINE SMOOTHING FOR MORTALITY STRUCTURE IN PERU

By: Bachiller MEZA SANTA CRUZ, Luis Alberto

Assessor: Magister ADRIAZOLA CRUZ, Rosa Ysabel

DECEMBER 2013

This research is a preliminary study of regression in general and their types, and then focus on the theoretical study of non-Parametric Regression Spline, which is a cubic polynomial in sections or pieces, demonstrating its benefits and ductility compared to general polynomials.

When joining two polynomials to obtain a polynomial by sections mostly the junction is not smooth or just do not join, resulting in abrupt changes, but using spline polynomials which is a particular case of polynomials in sections, and wich has as one of its properties that the first derivative of the spline function makes the union is not sharp and the second derivative allows the socket to join two polynomials, achieving a continuous trend smoothed curve.

In recent decades, researchers are using nonparametric regression models to smooth curves corresponding to a set of pairs of data, demographic approaches to the trend of demographic components such as fertility, mortality and population properly speaking, for that in this research I applies the nonparametric regression smoothing spline on the curve corresponding to the structure of mortality by sex and age using data from the deaths of Vital Statistics 2007 and National Population Census 2007, for the department of Lima.

I conclude that the smoothing structure Spline mortality is suitable, for this reason I suggests its use as an alternative form of smoothing the structure.

Key words

Spline, Nonparametric regression, Polynomial cubic, Structure of mortality, Interpolation.

CAPÍTULO 1: INTRODUCCIÓN

Si bien es cierto que los métodos estadístico-matemáticos son herramientas muy poderosas, estos realmente llegan a ser útiles sólo cuando contribuyen a solucionar los problemas numéricos de diversa índole que se presentan en la vida cotidiana.

En particular, la regresión paramétrica permite modelar conjuntos de datos correspondientes a una variable dependiente y una independiente.

Generalmente estos modelos siguen los patrones rígidos de la distribución normal, esto quiere decir que muchos eventos de la vida real quedarían fuera del alcance de poder ser modelados con los métodos estadístico-matemáticos, lo cual negaría la posibilidad de dar validez científica a un modelo diferente, que se ajuste a un conjunto de datos.

Como alternativa, el Modelo de Regresión no Paramétrico Spline (MRNPS) permite la generación de modelos sin la rigidez de la regresión paramétrica que requiere del supuesto de normalidad, es decir una vez ordenados en algún sentido el conjunto de datos pertenecientes a la variable en estudio se tiene la libertad de adaptar el modelo a los comportamientos algunas veces sinuosos correspondientes a la data que provienen de hechos reales. Se denomina no paramétrico porque no tiene un parámetro o parámetros que especifiquen un modelo de probabilidad.

El MRNPS, es una función polinómica mayormente de tercer grado, que permite suavizar los datos correspondientes a la variable en estudio. Los polinomios desde hace siglos tienen un papel importante en la teoría de la aproximación y el análisis numérico.

El MRNPS tiene la particularidad de permitir unir polinomios por secciones mediante puntos llamados nodos, de tal forma que la unión de las pequeñas secciones no presentarán cambios bruscos en la curva del modelo; sino más bien se obtendrá una continuidad suave, esta es una característica resaltante

de los polinomios Spline. La continuidad en los nodos son iguales en sus primera y segunda derivadas. Es decir, no se comporta como un polinomio en general, que si puede presentar cambios bruscos en la unión de polinomios por secciones.

Los espacios de los polinomios en general, de orden m, tienen entre sus características: son espacios lineales finito dimensionales; son funciones de suavizamiento; la derivada y antiderivada de un polinomio es también un polinomio; dada cualquier función continua sobre un intervalo $[a, b]$, existe un polinomio que es uniformemente cerrado en dicho intervalo; los valores de convergencia exactos pueden ser dados por aproximación de funciones de suavizamiento por polinomios; también muchos procesos de aproximación de polinomios generan curvas que oscilan bruscamente.

El mayor problema en los espacios de polinomios en general es que son relativamente inflexibles. Pareciera que con un mayor número de intervalos se lograría una mejor aproximación, pero no ocurre así, particularmente si el orden del polinomio es mayor que 3 o 4.

Se conoce que los polinomios de bajo grado divididos en pequeñas secciones llegan a tener una mejor aproximación y suavizamiento. Estas son características de los espacios de polinomios Spline de orden m (siendo $m \leq 3$) y con nodos en los puntos $x_1, x_2, \ldots, x_k$, los que se encuentran dentro del intervalo $[a, b]$.

Los espacios de polinomios Spline tienen entre otras las siguientes características importantes: son lineales finito dimensionales; son funciones de suavizamiento; las derivadas y antiderivadas de los polinomios Spline también son polinomios Spline; cada función continua en el intervalo $[a, b]$ puede ser arbitraria por el polinomio Spline de orden m fijado, generando tantos nodos como sea permitido; los valores de convergencia pueden ser dados por aproximación de funciones de suavizamiento por el MRNPS, así sean de orden

alto sus derivadas continuarán siendo buenas; Los polinomios Spline de bajo orden son muy flexibles y no presentan las oscilaciones bruscas que generan los polinomios en general.

1.1 Situación problemática

Siempre ha tenido importancia el modelamiento de eventos socio-económicos y demográficos utilizando métodos estadístico-matemáticos, con la finalidad de simular procesos, y obtener funciones que representen un comportamiento adecuado de los eventos de la vida real.

Montgomery y col.[1]: "El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. … De hecho, puede ser que el análisis de regresión sea la técnica estadística más usada".

1.2 Formulación del problema

Cuando se inicia un estudio sobre un evento o suceso en particular, se parte de un conjunto de pares de datos correspondientes a dicho evento, los cuáles se grafican en un diagrama de dispersión y se obtiene una nube de puntos, a través de la cual se observará la relación que hay entre las variables x e y, que generarán la ecuación

$$y = \beta_0 + \beta_1 x$$

donde:
β_0 es la ordenada al origen, y β_1 es la pendiente.

Wayne W. Daniel[2] dice que la típica introducción a los cursos de estadística analiza en principio los procedimientos estadísticos paramétricos que incluyen las pruebas basadas en la distribución t de Student, análisis de varianza,

[1] Montgomery, Douglas C., Peck, Elizabeth A. y Vining, G. Geoffrey. Introducción al Análisis de Regresión Lineal. Compañía Editorial Continental. Primera reimpresión, México, 2004.
[2] Daniel, Wayne W. Applied Nonparametric Statistics. Houghton Mifflin Company, Boston, 1974.

análisis de correlación y análisis de regresión. Una característica de estos procedimientos paramétricos es el hecho de que para que sean útiles se deben cumplir ciertos supuestos. Los procedimientos inferenciales en análisis de varianza, por ejemplo, se basan en el supuesto que las muestras provienen de poblaciones distribuidas normalmente y con varianzas homogéneas.

Dado que las poblaciones en estudio no siempre satisfacen las pruebas paramétricas, frecuentemente se necesitan supuestos inferenciales cuya validez no dependa de la rigidez paramétrica.

Los procedimientos de la estadística no paramétrica satisfacen muchas veces esta necesidad, desde que son validados bajo supuestos generales, y por ende ayudan a solucionar los problemas que se presentan a los investigadores.

1.3 Justificación teórica

En particular el Modelo de Regresión No Paramétrico Spline (MRNPS) permite modelar un conjunto de datos correspondientes a un evento u objeto estudiado, cuya función de distribución no es conocida y en los que la identificación de la forma de la curva no es fácil resolver con la regresión paramétrica, teniendo la ventaja de amoldarse a una nube de puntos con alguna tendencia y especialmente permite la combinación de curvas y rectas.

Al tener estas características, es muy útil su aplicación en los diversos campos de las ciencias e ingenierías, ya que permite modelar curvas que no pueden ser ajustadas con los modelos de regresión paramétricos usuales.

Los conjuntos de datos correspondientes a diversos eventos, generalmente no tienen un comportamiento con distribución normal o de cualquiera de las distribuciones paramétricas conocidas, muchos sucesos importantes de la vida real necesitan ser modelados adecuadamente para determinar sus tendencias, a la vez que tengan representatividad.

El MRNPS es muy flexible en la determinación de curvas no lineales a través de las funciones cuadráticas o cúbicas que se genera, donde se toman en cuenta todos o la gran mayoría de puntos correspondientes a un conjunto de datos.

Su utilidad se ha ido incrementando a lo largo de las décadas, ya que su formulación data de aproximadamente la segunda década del siglo XX. En los años 70 del siglo pasado con la aparición de la computadora y luego la microcomputadora, muy generalizada en la actualidad, se pudieron resolver las ecuaciones que no eran posible solucionar por otros medios más sencillos, y así como en otros campos de las ciencias básicas, en nuestro caso las matemáticas, se pudo resolver innumerables ecuaciones complejas como las del Spline, que no hubieran sido posible cristalizarse sin la gran ayuda de la computadora.

En Demografía, se tienen problemas con estas características, por ejemplo en la determinación de una buena estructura de la mortalidad por sexo y edad.

En la elaboración de las proyecciones de población la componente mortalidad se obtiene de la estructura de las defunciones por sexo y edad, para ello se requiere de una población y su correspondiente número de defunciones por sexo y edad.

Además en el análisis demográfico se puede utilizar este modelo para determinar tendencias de los indicadores de fecundidad, mortalidad, modelos de las estructuras de la fecundidad y mortalidad, entre otros.

1.4 Justificación práctica

El MRNPS es ampliamente utilizado en las ingenierías para determinar por ejemplo contornos de piezas de variadas maquinarias y equipos, en cambio en las Ciencias Sociales su aplicación ha sido limitada y siendo este un modelo adecuado para la suavización de conjuntos de datos con tendencia no lineal, amerita su aplicación (ver apéndice A).

En la literatura mundial se encuentran muy pocas investigaciones o estudios relacionadas con la aplicación del MRNPS específicamente a Demografía para determinar por ejemplo las estructuras de mortalidad, a pesar de ello es una buena herramienta estadístico-matemática y se utiliza en el presente trabajo, para determinar una estructura de mortalidad suavizada y representativa de la mortalidad en un departamento del Perú, que no será cuestionada por investigadores propios y de otras disciplinas afines, como cuando se utilizan métodos demográficos.

En cuanto la estructura de mortalidad por sexo y edad de un área determinada del Perú sea representativa, en el presente caso el departamento de Lima, mejor será la representatividad de la mortalidad. Si bien es cierto que en los diversos departamentos del Perú se carece de una cobertura completa de formularios de las estadísticas vitales de defunciones (hay una omisión de alrededor del 40 a 50 por ciento), el conjunto de datos básicos con el que se cuenta para elaborar la estructura de defunciones es adecuada.

No habiendo modelos de mortalidad que se adapten o representen adecuadamente la tendencia de la mortalidad en los departamentos del Perú, se ha utilizado el MRNPS para obtener estructuras de mortalidad por sexo y edad, que reflejan la situación de la mortalidad peruana, teniendo en cuenta que hasta la actualidad no existen en nuestro país, trabajos preliminares respecto a la utilización de este modelo. He aquí el gran valor de este esfuerzo contributivo para el desarrollo de nuestro país, y por extensión método que puede ser utilizado en otros países de Latinoamérica.

A finales de la década de los años 60 del siglo pasado se construyeron tablas de mortalidad para los departamentos del Perú, utilizando el modelo logito; posteriormente, en los años 80, se utilizaron las tablas modelo de Coale & Demeny, que reemplazaban a la real estructura de la mortalidad peruana.

Las proyecciones de población que desde las últimas décadas del siglo pasado viene adquiriendo una mayor demanda, a nivel desagregado de departamentos, provincias y distritos, son el resultado de haber realizado una

investigación empírica respecto a los llamados componentes demográficos y a través de los cuales se obtienen las tendencias de la fecundidad, mortalidad y migración. Cada uno de estos componentes requiere de una investigación profunda ya que son fenómenos en los que se encuentran inmerso la población humana.

Las tablas modelo de mortalidad a los que se hace mención fueron elaborados mayormente con información de defunciones correspondientes a países desarrollados correspondientes al siglo XIX y comienzos del siglo XX, que representan una mortalidad propia de los países de donde provienen y no necesariamente representan la mortalidad de los países latinoamericanos.

1.5 Objetivos

1.5.1 Objetivo general

Estudiar la regresión no paramétrica con énfasis en el modelo Spline, determinando sus bondades en la suavización de curvas, y aplicación a la estructura de mortalidad en Perú.

1.5.2 Objetivos específicos

1.5.2.1 Estudiar sobre la evolución de los polinomios en general a los polinomos Spline en particular.

1.5.2.2 Aplicar del polinomio Spline a la estructura de la mortalidad de la Región de Lima, Perú.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Marco filosófico o epistemológico de la investigación

Francis Galton en el siglo XIX estudia la altura de los padres e hijos utilizando más de mil registros de familias, al comparar las estaturas de los padres altos con sus hijos encontró que los hijos eran algo menores de talla es decir volvían a un "promedio o media", y los hijos de padres más bajos eran más altos, también estos tendían a un "promedio o media", o sea había una tendencia a "regresar" entonces le dio el nombre de "ley de la regresión universal" en su libro publicado en 1889 llamado *"Natural inheritance",* y que fue confirmada por su amigo Karl Pearson, es así como nace el término *regresión* en estadística

Al tener un conjunto de pares de datos (x,y) correspondientes a un suceso, éstos se grafican el *diagrama de dispersión* correspondiente que mostrará la relación que hay entre las variables *x* e *y*, y como resultado se determinará una función

$$y = \beta_0 + \beta_1 x,$$

donde β_0 es la ordenada al origen y β_1 es la pendiente.

El modelamiento de los fenómenos económicos, sociales y demográficos es posible si se utilizan modelos estadístico-matemáticos, con la finalidad de suavizar la gráfica de la nube de puntos, de tal forma que la representen adecuadamente.

Montgomery (2004): "De hecho, da la impresión que los datos, caen, en general, pero no exactamente, en una línea recta. … por lo que se debe modificar la ecuación … para tomar en cuenta esto. Sea la diferencia entre el valor observado de "y" y el de la línea recta ($\beta_0 + \beta_1 x$), un error (o ruido) ε. Conviene imaginar que ε es un error estadístico, esto es, que es una variable aleatoria que explica, porqué el modelo no se ajusta exactamente a los datos".

Entonces el modelo final:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Es llamado modelo de regresión lineal. A "X" se le llama variable independiente y a "Y" variable dependiente. "Sin embargo, eso causa confusión con el concepto de independencia estadística, así que llamaremos a X la variable predictora o regresora y a Y la variable de respuesta. Como la última ecuación sólo tiene una variable regresora, es llamada *modelo de regresión lineal simple*"[3].

Montgomery (2004): "Para comprender mejor el modelo de regresión lineal, supongamos que se puede fijar el valor de la variable regresora X para observar el valor correspondiente a la variable respuesta Y. Ahora, si X está fijo, el componente aleatorio ε del lado derecho de la ecuación mencionada … determina las propiedades de la variable Y. Supongamos que el promedio y la varianza de ε son 0 y σ^2 , respectivamente. Entonces, la respuesta media en cualquier valor de la variable regresora es

$$E(y|x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x$$

La varianza de Y para cualquier valor dado de X es:

$$Var(y|x) = \sigma^2_{y|x} = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

Así el verdadero modelo de regresión $\mu_{y|x} = \beta_0 + \beta_1 x$ es una línea recta de valores promedios, esto es, la altura de la línea de regresión en cualquier valor de X no es más que el valor esperado de Y para esa X. Se puede interpretar

[3] Montgomery, D.C., Peck, E.A. y Vining, G.G. Introducción al Análisis de Regresión Lineal. Compañía Editorial Continental, primera reimpresión, México, 2004.

que la pendiente β_1 es el cambio de la media de Y para un cambio unitario de X".

Montgomery (2004): "En general, las ecuaciones de regresión sólo son válidas dentro del rango de las variables regresoras contenidas en los datos observados. Por ejemplo, …, supongamos que se reunieron datos de las variables X e Y en el intervalo (rectilíneo) $x_1 \leq X \leq x_2$. En este intervalo, la ecuación de regresión lineal … es una buena aproximación de la verdadera relación. Sin embargo, supongamos que se usara esta ecuación para calcular valores de Y con valores de la variable regresora en la región (curvada) $x_2 \leq X \leq x_3$. Es claro que no funciona bien el modelo de regresión lineal dentro de este intervalo de X, porque hay error de modelo o error de ecuación.

En general, la variable de respuesta Y se puede relacionar con k regresores x_1, x_2, …, x_k, de modo que

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

A esto se le llama *modelo de regresión lineal múltiple*, ya que implica a más de un regresor. El adjetivo lineal es para indicar que el modelo es lineal respecto a los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ y no porque sea una función lineal de la X. … muchos modelos en los que Y se relaciona con las X en forma no lineal, se puede seguir manejando como modelos de regresión lineal, siempre y cuando la ecuación sea lineal en las β".

Montgomery (2004): "El modelo de regresión lineal $Y = X\beta + \varepsilon$ es un modelo general de ajuste de toda relación que sea lineal en los parámetros desconocidos β. Entre las relaciones está incluida la clase importante de los ***modelos polinomiales de regresión***. Por ejemplo, el polinomio de segundo orden de una variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

y el polinomio de segundo orden de dos variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

son modelos de regresión lineal.

Los polinomios se usan mucho en casos en los que la respuesta es curvilínea, y aún las relaciones no lineales complejas se pueden modelar en forma adecuada con polinomios dentro de límites razonablemente pequeños de las x.

Al polinomio de segundo orden también se le llama modelo polinomial de segundo orden o modelo cuadrático donde a "β_1 se le llama parámetro de efecto lineal y a β_2 parámetro de efecto cuadrático". El parámetro β_0 es el promedio de *y* cuando $x = 0$, si el rango de los datos incluye a $x = 0$. En caso contrario, β_0 no tiene interpretación física.

En general, el modelo polinomial de k-ésimo orden en una variable es:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$$

Si se define $x_j = x^j$, $j = 1, 2, \ldots, k$, la ecuación se transforma en un ***modelo de regresión lineal múltiple*** con los k regresores $x_1, x_2, \ldots, x_k$. Así, un modelo polinomial de orden k se puede ajustar con"[4] los mismos métodos del modelo de regresión lineal.

Continúa Montgomery (2004): "Los modelos polinomiales son útiles en los que se sabe que hay efectos curvilíneos presentes en la función verdadera de respuesta. También son útiles como funciones de aproximación a relaciones no lineales, desconocidas y posiblemente muy complejas. En este sentido, el modelo polinomial es simplemente el desarrollo en serie de Taylor de la función desconocida. Esta clase de aplicaciones parece presentarse con la mayor frecuencia en la práctica".

[4] Montgomery, D.C., Peck, E.A. y Vining, G.G. Introducción al Análisis de Regresión Lineal. Compañía Editorial Continental, primera reimpresión, México, 2004.

Respecto a las curvas Spline, Montgomery y col. (2004): “A veces se ve que un polinomio de bajo orden proporciona mal ajuste a los datos, y que al aumentar en forma modesta, el orden del polinomio no mejora mucho la situación. Los indicios de esto son que no se estabiliza la suma de cuadrados de residuales, o que las gráficas de residuales muestran una estructura raramente inexplicable. Este problema se puede presentar cuando la función se comporta en forma distinta en diferentes partes del rango de X. A veces, las transformaciones de X y/o Y eliminan ese problema; sin embargo, el método acostumbrado es dividir el rango de X en segmentos y ajustar la curva adecuada en cada segmento. Las funciones SPLINE ofrecen una forma útil de hacer este tipo de ajuste polinomial por segmentos.

Los Splines son polinomios de orden k por segmentos. Los puntos de unión de los segmentos se suelen llamar nudos. Por lo general, se requiere que los valores de la función y de las primeras $k-1$ derivadas concuerden con los nodos, para que la función Spline sea continua con $k-1$ derivadas. El Spline cúbico es adecuado para la mayor parte de los problemas prácticos”.

Dado que las poblaciones en estudio no siempre satisfacen las pruebas paramétricas supuestas líneas arriba, frecuentemente se necesita supuesto inferenciales cuya validez no dependa de supuestos rígidos.

Los procedimientos de la estadística no paramétrica satisface muchas veces esta necesidad, desde que son validados bajo supuestos generales, y por ende ayuda a los requerimientos de este tipo por parte de los investigadores.

Por convención, dos tipos de procedimientos estadísticos son tratados como no paramétricos: los procedimientos realmente no paramétricos y los procedimientos de distribución libre.

Estrictamente hablando los procedimientos no paramétricos no se refieren a parámetros de la población. Por ejemplo, la prueba de bondad de ajuste y de aleatoriedad tiene que ver con alguna otra característica diferente al valor del

parámetro de la población. La validez de los procedimientos de libre distribución no depende de la forma de la función de la población de la cual proviene la muestra"[5] .

2.2 Antecedentes de investigación

El estudio de las formas curvadas se viene dando desde la época de los romanos en que se construían curvas de forma libre en la construcción naval. "Las costillas (estructura base o armazón) de una nave producidas con tablones de madera que emanan de la quilla debían producirse con plantillas que pudieran usarse muchas veces.

Esto llevó al empleo de técnicas que fueron perfeccionadas por los venecianos (siglos XIII al XVI). La forma de las costillas se definió en términos de arcos circulares tangentes continuos. El casco de la nave se obtuvo variando la forma de las costillas a lo largo de la quilla (una manifestación temprana de las superficies definidas producto tensorial en la actualidad). Hasta esta época no existió ninguna representación gráfica para definir el casco de una nave, estos se hicieron populares en Inglaterra allá por el año 1600 y probablemente en aquella época se inventó el clásico "Spline" definido como una tira flexible de madera o caucho usada para dibujar curvas lisas.

En general, las curvas fueron empleadas para diseñar durante siglos: la mayoría de éstas eran círculos, pero algunas fueron de "forma libre". El empleo de las mismas parte desde el diseño de naves hasta llegar a la arquitectura.

Cuando las curvas tuvieron que ser dibujadas exactamente, la herramienta comúnmente usada fue un juego de plantillas conocidas como "curvas francesas" que consisten en porciones cónicas y espirales. Otra herramienta mecánica utilizada fue el llamado Spline, tira de madera a la que se le daba cierta forma y para mantener esa forma se utilizaban pesas de metal conocidas como ducks. La contraparte matemática de un Spline mecánico es una curva

[5] Daniel, Wayne W. Applied Nonparametric Statistics. Houghton Mifflin Company, Boston, 1974.

spline definida en forma no paramétrica y que al igual que un Spline mecánico se utilizó para diseñar.

En los años cincuenta, dos matemáticos franceses, Paul de Faget de Casteljau y Pierre Bézier trabajaron en forma independiente y llegaron a resultados similares descubriendo así las hoy conocidas curvas Bézier. El descubrimiento de estas curvas fue de tal trascendencia que su uso en el diseño se adoptó a nivel mundial.

Posteriormente se mejoran los resultados obtenidos con las curvas de Bézier al descubrirse su generalización: Las curvas de B-Spline (Basis Spline).

Alrededor de los años 60 del siglo pasado, Carl de Boor empezó a trabajar para los laboratorios de investigación de la General Motors usando en este trabajo los B-Spline para efectuar representaciones geométricas. Posteriormente se vuelve uno de los más arduos propulsores de los B-Spline en la teoría de la aproximación. La evaluación recursiva de las curvas B-Spline se debe a él y en la actualidad se conoce como el algoritmo de Boor. Esta evaluación recursiva fue descubierta en forma independiente por de Boor, L. Mansfield y M. Cox. Gracias a esta evaluación recursiva los B-Spline se convierten en una herramienta viable en el Computer Aided Geometric Design (CAGD), ya que antes del descubrimiento del algoritmo de Boor los B-Spline fueron definidos usando un tedioso método de diferencias divididas que era muy inestable.

La programación de los B-Spline en lenguajes tradicionales como Fortran, Pascal, C, etc. mediante la evaluación recursiva propuesta por de Boor está sumamente difundida"[6].

[6] B-splines con Mathematica 5.1, Ipanaqué Chero, R., Urbina Guzmán, R.T. y Correa Erazo, S.B. Universidad de Piura, 1998.

2.3 Bases teóricas

2.3.1 Regresión paramétrica

Por lo general el análisis de regresión conlleva estimar coeficientes de regresión, obtener gráficos de residuos, etc., los que han sido obtenidos mediante el ajuste a un modelo lineal de un conjunto de datos. Esto es un aspecto del análisis de regresión, sin embargo, hay muchos otros tópicos que forman parte del análisis de regresión, uno de ellos es la regresión no paramétrica.

Para empezar nuestro estudio es importante centrar nuestra atención en un modelo de regresión simple que proporcione un marco adecuado para la discusión de diferentes enfoques del análisis de regresión. Específicamente vamos a suponer que las observaciones son tomadas de una variable aleatoria continua *"Y"* correspondiente a n valores predeterminados de una variable continua independiente *"t"*. Sean (t_i, y_i), para $i = 1, \ldots, n$, valores de t e Y los cuáles son obtenidos de una muestra y se asume que t_i e y_i están relacionados por el modelo de regresión

$$y_i = \mu(t_i) + \varepsilon, \quad i = 1, \ldots, n$$

donde la sumatoria de los ε_i tiene media cero, es incorrelacionada, la variable aleatoria tiene varianza σ^2 *y los* $\mu(t_i)$ son valores de una función desconocida μ en los puntos $t_1, \ldots, t_n$. *Se supone que* $0 \le t_1 \le \cdots \le t_n \le 1$.

La función μ de la ecuación es llamada función de regresión o curva de regresión. El análisis de regresión se refiere a métodos de inferencia estadística acerca de una función de regresión.

La determinación de una adecuada metodología inferencial para el modelo permitirá que sus supuestos se cumplan para μ. Un modelo de regresión

paramétrico presupone que la forma de μ es conocida excepto para un número finito de parámetros.

Más específicamente se supone que existe un vector de parámetros

$$\beta = (\beta_1, \ldots, \beta_p)^T \sim \beta, \text{ subconjunto de } R^p,$$

y una función conocida $\mu(.;\beta)$ tal que $\mu(.) = \mu(.;\beta)$.

Por lo tanto es claro que para un modelo de regresión paramétrico la inferencia sobre μ es equivalente a la inferencia sobre β.

Los modelos paramétricos pueden tener forma lineal o no lineal. Por ejemplo, las funciones de regresión tales como que

$$\mu(t) = \beta_1 t^{\beta 2} \text{ y } \mu(t) = \beta_1 + \beta_2 exp\{-\beta_3 t^{\beta 4}\}$$

muestran casos donde los parámetros están entre un modelo lineal y un modelo no lineal. La primera curva de regresión $\mu(t) = \beta_1 t^{\beta 2}$ es lineal en β_1 y no lineal en β_2, mientras que la segunda ecuación es lineal en β_1 y β_2 y no lineal en β_3 y β_4. En contraste, para un modelo lineal puro están las funciones conocidas $x_1, \ldots, x_p$ tal que,

$$\mu(t) = \sum_{j=1}^{p} \beta_j x_j(t)$$

Se Infiere acerca de los parámetros cuando el dominio está referido a un típico análisis de regresión lineal.

Los métodos de análisis de regresión para modelos paramétricos representan una aproximación que lleva a inferir acerca de μ. Usando una metodología de estimación apropiada tal como mínimos cuadrados, posibilita utilizar la data

para estimar parámetros y de ese modo estimar μ. El resultado es una curva apropiada que fue seleccionada de una familia de curvas que permite al modelo ajustarse a los datos.

2.3.2 Regresión no paramétrica

En los estudios que se realizan para determinar si un conjunto de datos se aproxima o ajusta a alguna de las funciones de distribución conocidas, de no conocer bien la función adecuada, utilizar métodos paramétricos para su solución puede ser desacertado.

En tales casos, la mayor parte de la información acerca de la exactitud de μ es mejor que lo explique el que conduce el estudio experimental. En conformidad, parece ser más razonable usar las técnicas inferenciales con las cuales se ajusten mejor los datos. Ello es la razón de las técnicas de regresión no paramétricas que son ideales para ajustarse a problemas de inferencia cuando el conocimiento disponible acerca de μ es limitado por naturaleza.

Los métodos de regresión no paramétricas superan las dificultades inherentes de las técnicas paramétricas, es decir el conocimiento de la forma funcional de μ pero los estimadores no paramétricos son menos precisos cuando se quiere sustituir los modelos paramétricos por no paramétricos.

Para la mayoría de los estimadores paramétricos el riesgo o error al cuadrado esperado de la estimación, converge a cero a una tasa de n^{-1}. La tasa correspondiente a los estimadores no paramétricos es en general n^{-5} para algún número δ perteneciente a $(0,1)$ el cual depende de la suavidad de μ. Por ejemplo, si μ es dos veces diferenciable $n^{-4/5}$ es una relación con frecuencia citada (cotizada). Así, las técnicas de regresión no paramétricas adolecen de una eficiencia cuando son comparados con los métodos paramétricos.

Sin embargo, esta comparación no es enteramente imparcial ya que si se conociera la forma de μ se volvería a utilizar una aproximación paramétrica

para estimar en la mayoría de los casos. Si los estimadores no paramétricos se convirtieran en candidatos para la estimación de μ solo cuando hay un problema acerca de una forma paramétrica apropiada para μ. En ese caso, si un modelo paramétrico incorrecto es usado la tasa n^{-1} no converge a cero y las técnicas de regresión no paramétricas serán mucho mejor que los métodos paramétricos.

El resultado de un análisis de regresión no paramétrico es una curva ajustada a un conjunto de datos. Desde esta curva se produce sin suponer una forma paramétrica para μ, habrá alguna pérdida en la interpretación de los estimadores obtenidos de este modo en que no dejará de ser cantidades tales como coeficientes de regresión estimados para ser interpretada. Sin embargo, la curva ajustada es en sí misma una estimación del parámetro μ y cualquier funcional de μ es también un parámetro que se puede estimar utilizando la estimación de la curva de regresión (ver apéndice B).

2.3.3 Polinomios

Desde hace mucho tiempo desempeñan un papel importante en la teoría de la aproximación y el análisis numérico.

Para resaltar el poder de los mismos se denota el espacio

$$\mathcal{P}_m = \left\{ p(x) : p(x) = \sum_{i=1}^{m} c_i x^{i-1}, c_1 \ldots c_m \, , x \in \Re \right\}$$

Como un espacio de polinomios de orden m y que tiene las siguientes ventajas:

1. $\mathcal{P}_m$ es un espacio lineal finito dimensional sobre una base conveniente;
2. Los polinomios son funciones de suavizamiento;
3. Los polinomios son fáciles de almacenar, manipular y evaluar con el apoyo de una computadora;

4. La derivada y antiderivada de un polinomio son polinomios cuyos coeficientes se determinan algebraicamente;
5. El número de ceros de un polinomio de orden m no puede exceder a $m-1$;
6. Diversas matrices (obtenidas en interpolaciones y aproximaciones por polinomios) son siempre no singulares y tienen fuertemente las propiedades de regularidad del signo;
7. La estructura del signo y la forma de un polinomio está íntimamente relacionada con la estructura del signo de los coeficientes;
8. Dada cualquier función continua sobre un intervalo *[a,b]*, existe un polinomio el cual es uniformemente cerrado en él;
9. Los valores de convergencia exactos pueden ser dados por aproximación de funciones de suavizamiento por polinomios;
10. Muchos procesos de aproximación de polinomios traen consigo la tendencia a producir aproximaciones de polinomios que oscilan bruscamente (ver apéndice C).

2.3.4 Polinomios por secciones

El mayor problema de los espacios de polinomios $\mathcal{P}_m$ para propósitos de aproximación es que son relativamente inflexibles.

Con un número suficiente de intervalos pequeños parece que se corregiría este defecto pero cuando son muchos intervalos, severas oscilaciones pueden aparecer, particularmente si m es mayor que 3 o 4.

Esta observación sugiere que el orden logra una clase de aproximación a funciones con gran flexibilidad, cuando se trabaja con polinomios de relativo bajo grado, y dividiendo los intervalos en pequeñas secciones.

En base a todo esto se define:
Sea

$$a = x_0 < x_1 < x_2 < ... < x_k < x_{k+1} = b$$

y se escribe

$$\Delta = \{x_i\}_0^{k+1}$$

Entonces Δ es la partición del intervalo $[a,b]$ con $k+1$ subintervalos,

$$I_i = [x_i, x_{i+1}), i = 0,1, ..., k-1 \quad e \quad I_k = [x_k, x_{k+1}]$$

Dado un entero positivo m *se tiene*

f : existen polinomios $\mathcal{PP}_m(\Delta) = p_0, p_1, \ldots, p_k$ en $\mathcal{P}_m$

con $f(x) = p_i(x)$, para x perteneciente a $I_i, i = 0,1, \ldots, k$

Se llama a $\mathcal{PP}_m(\Delta)$ el espacio de polinomios por secciones de orden m con nodos $x_1, \ldots, x_k$.

2.3.5 Funciones Spline

En orden a mantener la flexibilidad de los polinomios por secciones mientras que al mismo tiempo se logra algunos grados de suavización global, se puede definir la siguiente clase de funciones:

Sea Δ una partición del intervalo $[a,b]$ y sea m un entero positivo

Sea

$$\delta_m(\Delta) = \mathcal{PP}_m(\Delta) \cap C^{m-2}[a,b]$$

donde $\mathcal{PP}_m(\Delta)$ es el espacio de polinomios por secciones definidos en la sección anterior.

Llamamos a $\delta_m(\Delta)$ el espacio de polinomios Spline de orden m con nodos simples en los puntos $x_1, \ldots, x_k$.

Los polinomios Spline poseen las siguientes características:

1. Los espacios de polinomios Spline son espacios lineales finito dimensionales con unas bases muy convenientes;
2. Los polinomios Spline son relativamente funciones de suavizamiento;
3. Los polinomios Spline son fáciles de almacenar, manipular y evaluar en una computadora;
4. Las derivadas y antiderivadas de los polinomios Spline son también polinomios Spline, cuyos desarrollos pueden ser encontrados con una computadora;
5. Los polinomios Spline poseen buenas propiedades del cero análogas a la de los polinomios en general;
6. Varias matrices derivadas naturalmente del uso del Spline en teoría de aproximación y análisis numérico tienen un conveniente signo y propiedades determinantes;
7. La estructura del signo y la forma de un polinomio Spline está relacionada con la estructura del signo de los coeficientes;
8. Cada función continua en el intervalo $[a,b]$ puede ser aproximadamente arbitraria por el polinomio Spline de orden m fijado, procurando un número suficiente de nodos como pueda permitirse;
9. Los valores de convergencia pueden ser dados por aproximación de funciones de suavizamiento por Spline no solamente las funciones se aproximan a un orden alto, sino que sus derivadas son simultánea y aproximadamente buenas;
10. Los Spline de bajo orden son muy flexibles, y no presentan las oscilaciones generalmente asociadas con los polinomios.

Estas propiedades de los polinomios Spline son compartidas por una amplia variedad de otros espacios por secciones.

2.3.6 Espacio de los polinomios Spline

2.3.6.1 Propiedades básicas

1)

$Sea\ [a,b]\ un\ \text{int}ervalo\ cerrado\ de\ extremos\ finitos,$
$y\ sea\ \Delta = \{x_1, x_2, ..., x_k\}\ con\ a = x_0 < x_1 < x_2 < ... < x_k < x_{k+1} = b$
$una\ partición\ de\ k+1\ sub\text{int}ervalos$
$I_i = [x_i, x_{i+1}),\ donde\ \ i = 0, 1, 2, ..., k-1 \quad e \quad I_k = [x_k, x_{k+1}]$

2)

Sea m un entero positivo, y sea $\mathcal{M} = (m_1, m_2, \ldots, m_k)$ un vector de enteros (exponente de los términos del polinomio) con $1 \leq m_i \leq m,\ i = 1,2, \ldots, k$

2.3.6.2 Definición

Al espacio

$\delta(\mathcal{P}_m; \mathcal{M}; \Delta)$ = {s: existen polinomios $s_0, \ldots, s_k$ en $\mathcal{P}_m$ tal que $s(x) = s_i(x)$

para x perteneciente a I_i, $i = 0,1, \ldots, k$, y

$D^j s_{i-1}(x_i) = D^j s_i(x_i)$ para $j = 0, 1, \ldots, m-1-m_i$, $i = 1, \ldots, k$}

Llamamos **espacio de polinomios Spline** de orden m con nodos $x_1, x_2, \ldots, x_k$ de multiplicidad (número de veces que se repite) $m_1, m_2, \ldots, m_k$.

$\mathcal{M}$ es el vector multiplicidad, que controla la naturaleza del espacio $\delta_m(\mathcal{P}_m; \mathcal{M}; \Delta)$ mediante el suavizamiento de los Spline en los nodos (intersecciones).

Si $m_i = m$ el promedio de las dos secciones de polinomios adyacentes s_{i-1} y s_i en los intervalos adyacentes al nodo x_i son inconexos y pueden dar un salto discontinuo en x_i.

Si $m_i < m$ se logra que las dos secciones de polinomios se unan suavemente, de tal forma que el Spline s y sus primeras $m\text{-}1\text{-}m_i$ derivadas también son continuas en dicho nodo.

Por ejemplo, sea el espacio de polinomios de orden m

$$\mathcal{P}_m = \left\{ p(x) : p(x) = \sum_{i=1}^{m} c_i x^{i-1},\ c_1 \ldots c_m ,\ x \in \Re \right\}$$

Para $m=4$

$$p(x) = \sum_{i=1}^{4} c_i x^{i-1} = c_1 x^{1-1} + c_2 x^{2-1} + c_3 x^{3-1} + c_4 x^{4-1} =$$

$$= c_1 x^0 + c_2 x^1 + c_3 x^2 + c_4 x^3$$

Según la segunda propiedad básica del espacio de polinomios Spline

$\mathcal{M} = (m_1, m_2, \ldots, m_k) = (m_1, m_2, m_3, m_4) = (1, 2, 3, 4)$

$\Rightarrow\ k = 1, 2, 3, 4$

$1 \le m_i \le m,\ i = 1, 2, \ldots, k,$ que en este caso $k = 4$

$1 \le m_1 = 1 \le m = 4 \quad \Rightarrow \quad 1 \le 1 \le 4$

$1 \le m_2 = 2 \le m = 4 \quad \Rightarrow \quad 1 \le 2 \le 4$

$1 \le m_3 = 3 \le m = 4 \quad \Rightarrow \quad 1 \le 3 \le 4$

$1 \le m_4 = 4 \le m = 4 \quad \Rightarrow \quad 1 \le 4 \le 4$

Por la definición del espacio de polinomios Spline, si $m = 4$

$\delta(\mathcal{P}_4; \mathcal{M}; \Delta)$ = {s: existen polinomios $s_0, \ldots, s_k$ en $\mathcal{P}_4$ tal que $s(x) = s_i(x)$

para x perteneciente a I_i, $i = 0,1, \ldots, k$, y

$D^j s_{i-1}(x_i) = D^j s_i(x_i)$ para $j = 0, 1, \ldots, m\text{-}1\text{-}m_i$, $i = 1, \ldots, k$}

Como $\mathcal{M} = (m_1, m_2, \ldots, m_k) = (m_1, m_2, m_3, m_4) = (1, 2, 3, 4) \Rightarrow i = 1, 2, 3, 4$

Si $j = m - 1 - m_i \quad con \quad i = 1, 2, 3, 4$

Para $i = 1 \Rightarrow j = 4 - 1 - m_1 = 4 - 1 - 1 = 4 - 2 = 2$

Para $i = 2 \Rightarrow j = 4 - 1 - m_2 = 4 - 1 - 2 = 4 - 3 = 1$

Para $i = 3 \Rightarrow j = 4 - 1 - m_3 = 4 - 1 - 3 = 4 - 4 = 0$

Para $i = 4 \Rightarrow j = 4 - 1 - m_4 = 4 - 1 - 4 = 4 - 5 = -1$

$\Rightarrow D^0 s_{i-1}(x_i) = D^0 s_i(x_i)$, la derivada en 0 es la misma función pero continua.

$\Rightarrow D^1 s_{i-1}(x_i) = D^1 s_i(x_i)$, la primera derivada suaviza la función, no puede hacer ángulo.

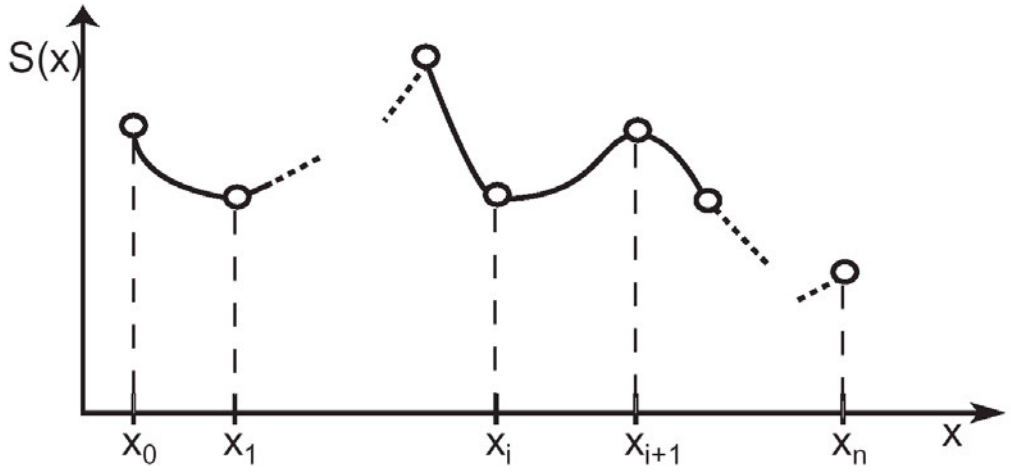
$\Rightarrow D^2 s_{i-1}(x_i) = D^2 s_i(x_i)$, en la segunda derivada la función mantiene el sentido de la concavidad.

Además por la propiedad 1) el polinomio Spline está definido en cada uno de los intervalos.

Todo ello permite controlar el suavizamiento en los nodos (intersecciones), siendo esta las propiedades básicas del polinomio Spline, y que sirven para suavizar las curvas no lineales.

2.3.6.3 El polinomio Spline

Dados $n+1$ puntos $(x_0;f(x_0)),(x_1;f(x_1)), \ldots, (x_n;f(x_n))$



Con $x_0 < x_1 < \ldots < x_n$ y la función f, definidos en el intervalo $[a, b]$ que tiene nodos $x_0, x_1, \ldots, x_n$, se aproximará la función f en cada subintervalo $[x_j; x_{j+1}]$, $j = 0, 1, \ldots, n-1$, siendo el polinomio correspondiente

$$P_j(x) = c_1 + c_2(x - x_j) + c_3(x - x_j)^2 + c_4(x - x_j)^3,$$

con $j = 0, 1, \ldots, n-1$

2.3.7 El modelo de regresión no paramétrico "Spline"

Eubank (1999) menciona que los Spline "modernos iniciales" fueron propuestos por Whittaker (1923), mientras que su formulación moderna fue propuesta por Schoenberg (1964) y Reinsche (1967); finalmente su implementación en estadística se da gracias a Grace Wahba (1990).

"El análisis de regresión permite construir modelos matemáticos que estudian la relación existente entre una variable dependiente y una o más variables independientes. Estos modelos se utilizan para estimar respuesta de valores futuros no observados de la o las variables independientes.

En el caso simple cuando ambas variable la dependiente Y, y la independiente X, son escalares, dadas las observaciones (x_i,y_i) para $i = 1, \ldots, n$, se relaciona un modelo de regresión para dichas variables de la siguiente forma:

$$y_i = f(x_i) + \varepsilon_i, \quad para\ i = 1, \cdots, n \tag{2.1}$$

Donde f es la función de regresión y ε_i son los errores aleatorios independientes con media cero y varianza común σ^2. El objetivo del análisis de regresión es construir un modelo para f y realizar la estimación en base a los datos observados.

A menudo f es no lineal en x. Un enfoque frecuente es tratarla como una relación no lineal aproximando f a un polinomio de orden m

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_{m-1} x^{m-1} \tag{2.2}$$

En general, **un modelo de regresión paramétrico** asume que la forma de f es conocida excepto para un número finito de parámetros desconocidos. La forma específica de f puede provenir de la teoría conocida y/o aproximaciones a mecanismos bajo algunos supuestos simplificados. Los supuestos pueden ser demasiado restrictivos y la aproximación puede ser demasiado bruta para algunas aplicaciones. Un modelo inapropiado conduce a un sesgo sistemático y conclusiones engañosas. En la práctica, siempre se debe comprobar la forma supuesta para la función f.

Esto a menudo es dificultoso, a veces imposible, de obtener una forma funcional para f.

Un **modelo de regresión no paramétrico** no supone una forma predeterminada. En lugar de ello se hace supuestos de las propiedades cualitativas de f. Por ejemplo, uno puede estar dispuesto a asumir que f se suaviza y que no se reducirá a una forma específica con un número finito de

parámetros. Más bien, en general conduce a algunos grupos de funciones infinito dimensionales.

La idea básica de la regresión no paramétrica es dejar que los datos hablen por sí mismos. Es dejar que la data decida que función se ajusta mejor sin imponer una forma específica de f. Por consiguiente los métodos no paramétricos son en general más flexibles. Ellos pueden descubrir estructuras en los datos que de lo contrario se perderían.

La técnica de regresión no paramétrica puede ser aplicada en diferentes partes del análisis de regresión: exploración de datos, construcción de modelos, pruebas de modelos paramétricos, y diagnósticos. De hecho, la suavización Spline es una herramienta potente y versátil para la construcción de modelos estadísticos en la explotación de estructuras de datos.

El polinomio (2.2) es un modelo global que suaviza las variaciones puntuales de un conjunto de datos. Las observaciones individuales pueden tener influencia negativa sobre los extremos del conjunto de datos.

Estas variaciones lleva a oscilaciones en ambos extremos del rango en el polinomio de ajuste. Una solución para superar esta limitación es la utilización de polinomios por tramos o trozos, que es la idea central de los polinomios Spline.

Sea $a < t_1 < \ldots < t_k < b$ puntos fijos llamados nodos. Sea $t_0 = a$ y $t_{k+1} = b$. en términos generales los polinomios Spline son polinomios por trozos unidos suavemente en sus nodos o extremos. Formalmente, un polinomio Spline de orden r es una función real valorada sobre $[a,b]$, $f(t)$, de tal manera que:

(i) f es un polinomio por trozos de orden r sobre $[t_i,t_{i+1}]$, $i = 0, 1, \ldots, k$;

(ii) f tiene $r-2$ derivadas continuas y la derivada $r-1$ (segunda derivada si $r=3$) es una función escalonada con saltos en los nudos.

Ahora consideremos a las órdenes representadas como $r = 2m$. La función f es un polinomio Spline natural de orden $2m$ si, en adición a (i) y (ii), satisface las condiciones naturales límite,

(iii) $f^{(j)}(a) = f^{(j)}(b) = 0, \quad j = m, \cdots; 2m-1$

Las condiciones naturales de contorno implican que f es un polinomio de orden m en los sub-intervalos extremos $[a,t_1]$ y $[t_k,b]$, denotando la función del espacio de polinomios naturales Spline de orden $2m$ con nodos $t_1, \ldots, t_k$ como

$$N\ S^{2m}(t_1, ..., t_k)$$

Una aproximación conocida como regresión Spline, es acercar f utilizando un polinomio Spline o un polinomio natural Spline. Para conseguir una buena aproximación, se necesita decidir el número y ubicación de los nodos. Acá se cubre un enfoque diferente conocido como suavización Spline. Se empieza con un espacio modelo definido para f y se introduce una penalidad para evitar el exceso de ajuste. Luego se describe esta aproximación para los polinomios Spline.

Consideremos el modelo de regresión (2.1). Suponemos que f es un modelo suavizado. Específicamente, se asume que

$$f \in W_2^m[a,b]$$

donde el espacio Sobolev[7]

$$W_2^m[a,b] = \left\{ f : f, f', \cdots, f^{(m-1)} \text{ son absolutamente continuas}, \int_a^b (f^{(m)})^2 dx < \infty \right\} \quad (2.3)$$

Para cualquier $a \le x \le b$, el *Teorema de Taylor* afirma que

[7] Los espacios de Sobolev hacen el papel de derivadas fraccionarias.

$$f(x) = \underbrace{\sum_{v=0}^{m-1} \frac{f^{(v)}(a)}{v!}(x-a)^v}_{\text{polinomio de orden } m} + \underbrace{\int_a^x \frac{(x-u)^{m-1}}{(m-1)!} f^{(m)}(u)du}_{\text{Re}\, m(x)} \qquad (2.4)$$

Está claro que el modelo de regresión polinomial (2.2) ignora el término del resto del polinomio Rem(x) suponiendo que es insignificante. A menudo, en la práctica es difícil verificar esta hipótesis. La idea detrás del suavizamiento Spline está en dejar que la data genere que tan grande debe ser $Rem(x)$. Desde que $W_2^m[a,b]$ es un espacio finito dimensional, un ajuste directo de f minimizando por mínimos cuadrados (LS)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \qquad (2.5)$$

conduce a la interpolación. Por lo tanto, es necesario cierto control sobre Rem(x). Un enfoque básico es controlar hasta qué punto se puede permitir a f alejarse del modelo del polinomio.

Bajo reglas apropiadas una medida de distancia entre f y los polinomios es $\int_a^b (f^{(m)})^2 dx$. Entonces es razonable estimar f minimizando los mínimos cuadrados (LS) bajo la restricción

$$\int_a^b (f^{(m)})^2 dx \le \rho \qquad (2.6)$$

para una constante ρ. Introduciendo el multiplicador Lagrange, el problema de minimización restringida (2.5) y (2.6) son equivalentes a minimizar los mínimos cuadrados penalizados (PLS):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int_a^b (f^{(m)})^2 dx \qquad (2.7)$$

O sea el polinomio Spline se referirá a la solución del PLS (2.7) en el espacio modelo $W_2^m[a,b]$, un Spline cúbico es un caso especial de los polinomios Spline

con $m=2$. La medida de la aspereza de la función f, $\int_a^b (f^{(m)}))^2 dx$ está referida como una penalidad de aspereza. Es obvio que no hay penalidad para polinomios de orden menor o igual que m. El parámetro de suavización ʎ equilibra el intercambio entre la bondad de ajuste por el LS y la medida de aspereza por $\int_a^b (f^{(m)}))^2 dx$ "[8].

2.3.8 El parámetro ʎ

Si hacemos [a,b] = [0,1] podemos reemplazar 2.7 por:

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_0^1 f^{(m)}(t)^2 dt, \quad \lambda > 0, \quad f \in W_2^m[0,1] \qquad (2.8)$$

donde el parámetro λ en la ecuación regula el equilibrio entre el suavizamiento y la bondad de ajuste y, por esta razón se suele hacer referencia como el parámetro de suavizamiento. Cuando el valor de λ es grande (es decir cercana a 1) el suavizamiento es casi una recta, y los estimadores potenciales con m derivadas muy grandes son penalizados. En el caso límite de

$$\lambda = \infty \ (o \ q = 1)$$

se produce un polinomio de regresión de orden m o m-1 grados. A la inversa, cuando el valor de λ es pequeño, le corresponde una mejor bondad de ajuste, con λ = q = 0 se tiene un estimador que interpola todos los datos.

Se quieres ajustar un conjunto de datos a una función que refleje la indispensable característica de los datos pero conservando algún grado de suavizamiento.

Una medida natural de suavizamiento asociado con una función

$$f \in W_2^m[0,1] \quad es \quad \int_0^1 f^{(m)}(t)^2 dt$$

[8] Wang, Yuedong. Smoothing Splines, Methods and Applications. Monographs on Statistics and Applied Probability 121. Taylor & Francis Group, LLC 2011.

Mientras que una medida estándar de bondad de ajuste a los datos es la media residual de la suma de los n^{-1} cuadrados

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2$$

Así una valoración de la calidad de un candidato estimador f es proporcionado por la suma convexa

$$(1-q)\, n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + q\int_0^1 f^{(m)}(t)^2\, dt$$

Para algunos $0 < q < 1$, un estimador óptimo podría ser obtenido reduciendo al mínimo la función

$$W_2^m[0,1]$$

al establecer

$$\lambda = \frac{q}{(1-q)}$$

Esta se vuelve equivalente a la estimación de μ por la función μ_λ que la minimiza.

2.3.9 Determinación de los coeficientes de la función Spline

El término "Spline" se refiere a un conjunto de funciones de diversa graduación, todas ellas utilizadas para la interpolación de datos o suavizamiento de curvas. Existen varios tipos de funciones "Spline" entre ellos los lineales, cuadráticos, cúbicos, y de mayor grado. Pero los que según la experiencia demostrada en las diferentes investigaciones y aplicaciones que se realizan y para los que sirven con un alto grado de suavización son los "Spline" cúbicos o de tercer grado.

"Las funciones para la interpolación por Spline normalmente se determinan como minimizadores de la aspereza sometidas a una serie de restricciones".

También se la identifica como una herramienta para la interpolación polinómicas por trozos. Donde dados $n+1$ puntos $(x_0;f(x_o))$, $(x_1;f(x_1))$, … , $(x_n;f(x_n))$ se puede lograr una curva suavizada que se adapte al conjunto de dichos pares.

La función Spline en cada subintervalo $(x_i;x_{i+1})$, es la siguiente

$$p_k(x) = c_{1k} + c_{2k}(x-x_k) + c_{3k}(x-x_k)^2 + c_{4k}(x-x_k)^3 \; ; \; k = 0;1;\ldots;n-1$$

conteniendo cuatro coeficientes c_{1k}, c_{2k}, c_{3k}, $c_{4k.}$
Estos coeficientes son despejados, de los n polinomios de grado menor o igual que tres.

Un método de interpolación cuyo proceso se ha generalizado con el soporte computacional (aunque anteriormente se realizaba en forma manual aunque tediosa), es el de la interpolación Spline cúbico. Al igual que en otros métodos de interpolación este se ajusta a un polinomio cúbico por secciones, con algunas ventajas respecto a las otras, y cuya forma es:

$$y = c_1 + c_2 x + c_3 x^2 + c_4 x^3$$

aplicado a una sección de los datos.

Sin embargo, con los Spline cúbicos, uno logra engranar la relación de una sección del polinomio Spline cúbico con su siguiente sección, de tal forma que la pendiente del límite superior del primer polinomio debe coincidir con la pendiente del límite inferior del siguiente polinomio, además de tener una suavización que es característica de los polinomios Spline.

Esta propiedad de los polinomios Spline por secciones nos permite encontrar un sistema lineal de ecuaciones que tienen solución, lo que nos procura hallar

un conjunto de coeficientes para cada una de las secciones del polinomio Spline.

Se comienza por suponer que se tiene una colección de puntos x_1, x_2, ..., x_n ordenados y a lo largo de una curva continua. A cada uno de estos puntos se asocia algún $y_i = f(x_i)$. A raíz de las derivaciones hechas por Johnson y Percy (2000), y, Burden y Faires (1993) se divide esta secuencia continua en "i" intervalos.

En cada intervalo el objetivo es ajustar un polinomio cúbico, sea $h_i = x_{i+1} - x_i$, es decir, h_i es la diferencia entre dos puntos x_i sucesivos perteneciente a dos intervalos sucesivos. En el i-ésimo intervalo, se desea ajustar un polinomio de la forma

$$y = c_{4i}(x - x_i)^3 + c_{3i}(x - x_i)^2 + c_{2i}(x - x_i) + c_{1i}$$

donde x_i es el primer valor de x en el i-ésimo intervalo. Recordemos que, para ajustarse a un polinomio de tercer orden, el intervalo debe contener al menos cuatro puntos.

El objetivo en este punto es encontrar soluciones para c_{1i}, c_{2i}, c_{3i}, c_{4i}, en el i-ésimo intervalo. Se procederá a desarrollar estas soluciones, la escritura de cada coeficiente, tanto como sea posible, en términos de valores observados x_i y y_i.

En el extremo inferior del intervalo, el polinomio es simple, es sólo

$$y = c_{4i}(x_i - x_i)^3 + c_{3i}(x_i - x_i)^2 + c_{2i}(x_i - x_i) + c_{1i} \quad \Rightarrow y = c_{1i}$$

En el extremo superior del intervalo, el polinomio es

$$y = c_{4i}(x - x_i)^3 + c_{3i}(x - x_i)^2 + c_{2i}(x - x_i) + c_{1i}, \quad sea\ h_i = x - x_i \Rightarrow$$

$$y = c_{4i}(h_i)^3 + c_{3i}(h_i)^2 + c_{2i}(h_{i)}) + c_{1i}$$

Tomamos la primera y segunda derivada de este polinomio, y obtenemos

$$\frac{dy}{dx} = 3c_{4i}(h_i)^2 + 2c_{3i}h_i + c_{2i}$$

y

$$\frac{d^2y}{dx^2} = 6c_{4i}(h_i) + 2c_{3i}$$

Una vez más después de las derivaciones hechas por Jhonson y Percy, y, Burden y Faires, se escriben los coeficientes en términos de la segunda derivada en los extremos del intervalo. Así, en el extremo inferior del intervalo de orden i,

$$S_i = \left(\frac{d^2y}{dx^2}\right)_i = 6c_{4i}(x_i - x_i) + 2c_{3i} = 2c_{3i} \quad \Rightarrow \quad c_{3i} = \frac{S_i}{2}$$

y en el extremo superior del intervalo de orden i,

$$S_{i+1} = \left(\frac{d^2y}{dx_2}\right)_{i+1} = 6c_{4i}(x_{i+1} - x_i) + 2c_{3i} = 6c_{4i}h_i + 2c_{3i}$$

como, $S_i = 2c_{3i}$, se sustituye en la ecuación del extremo superior del intervalo la ecuación correspondiente al extremo inferior del intervalo, y se tiene:

$$S_{i+1} = 6c_{4i}h_i + S_i$$

Se resuelve para c_{4i} y se obtiene:

$$c_{4i} = \frac{S_{i+1} - S_i}{6h_i}$$

Ahora se sustituye c_{4i}, c_{3i}, c_{1i} en la ecuación del extremo superior del intervalo, y se obtiene:

$$y_{i+1} = \frac{S_{i+1} - S_i}{6h_i}(h_i)^3 + \frac{S_i}{2}h_i^2 + c_{2i}h_i + y_i$$

Por último, se despeja para c_{2i}:

$$y_{i+1} - \left[\frac{S_{i+1} - S_i}{6h_i}h_i^3\right] - \left[\frac{S_i}{2}h_i^2\right] - y_i = c_{2i}h_i$$

$$\left[\frac{1}{h_i}y_{i+1} - y_i\right] - \left\{\left[\frac{1}{h_i}\frac{S_{i+1} - S_i}{6h_i}h_i^3\right] - \left[\frac{1}{h_i}\frac{S_i}{2}h_i^2\right]\right\} = c_{2i}$$

$$\left[\frac{y_{i+1} - y_i}{h_i}\right] - \left\{\left[\frac{S_{i+1} - S_i}{6h_i^2}h_i^3\right] + \left[\frac{S_i}{2h_i}h_i^2\right]\right\} = c_{2i}$$

$$\left[\frac{y_{i+1} - y_i}{h_i}\right] - \left[\frac{S_{i+1}h_i - S_ih_i}{6} + \frac{S_i}{2}h_i\right] = c_{2i}$$

se toma m. c. m.

$$\left[\frac{y_{i+1} - y_i}{h_i}\right] - \left[\frac{S_{i+1}h_i - S_ih_i + 3S_ih_i}{6}\right] = c_{2i}$$

$$\left[\frac{y_{i+1} - y_i}{h_i}\right] - \left[\frac{S_{i+1}h_i + 2S_ih_i}{6}\right] = c_{2i}$$

$$\left[\frac{y_{i+1} - y_i}{h_i}\right] - \left[\frac{2h_iS_i + h_iS_{i+1}}{6}\right] = c_{2i}$$

$$c_{2i} = \frac{y_{i+1} - y_i}{h_i} - \frac{2h_iS_i + h_iS_{i+1}}{6}$$

Estas sustituciones han dado ecuaciones para c_{1i}, c_{2i}, c_{3i}, c_{4i} y, en el intervalo de orden i, en el que estas constantes se expresan en términos de valores conocidos $(y_i, y_{i+1}, y\ h_i)$ y aún desconocidos como la primera derivada $(S_i's)$.

Para encontrar las primeras derivadas, se utiliza la condición que las pendientes de dos polinomios sucesivos son iguales en su punto común. Usando la definición de la derivada

$$\left(\frac{dy}{dx}\right)_i = c_{2i} \qquad y$$

$$\left(\frac{dy}{dx}\right)_{i-1} = 3c_{4i}h^2{}_{i-1} + 2c_{3i}h_{i-1} + c_{2(i-1)}$$

Se igualan

$$c_{2i} = 3c_{4(i-1)}h^2{}_{i-1} + 2c_{3(i-1)}h_{i-1} + c_{2(i-1)}$$

Ahora se sustituye para los valores c_{2i}, c_{4i}, c_{3i} y $c_{2(i-1)}$, y se resuelve la ecuación

$$h_{i-1}S_{i-1} + (2h_i + 2h_i)S_i + h_iS_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right)$$

Esta relación tiene valores conocidos y_i, y_{i+1}, h_i, h_{i-1}, y valores desconocidos S_i, S_{i-1}, y S_{i+1}.

Ahora bien, esta relación contiene valores conocidos y_i, y_{i+1}, h_i, h_{i-1}, y las incógnitas de S_i, S_{i-1} y S_{i+1}. Mediante la combinación de todas las ecuaciones implícitas, para todos los intervalos i, se construye el sistema de la siguiente forma:

$$\begin{bmatrix}
h_1 & 2(h_1+h_2) & h_2 & & & & \\
 & h_2 & 2(h_2+h_3) & h_3 & & & \\
 & & h_3 & 2(h_3+h_4) & h_4 & & \\
 & & & \ddots & \ddots & \ddots & \\
 & & & & \ddots & \ddots & \\
 & & & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1}
\end{bmatrix}$$

$$\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ \vdots \\ S_{n-1} \\ S_n \end{bmatrix} = 6 \begin{bmatrix} \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ \frac{y_5 - y_4}{h_4} - \frac{y_4 - y_3}{h_3} \\ \vdots \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \end{bmatrix}$$

El sistema lineal en la ecuación anterior contiene $n\text{-}2$ ecuaciones y n incógnitas. Dos ecuaciones se necesitan más para hacer única la solución. Si se aplican los valores finales $S_1 = S_n = 0$ (lo que implica que el polinomio es plano en los extremos inferior y superior), se puede resolver este sistema de ecuaciones para todos los S_i's. La aplicación de estas dos condiciones de contorno elimina efectivamente dos columnas, la primera y la última, en la matriz, y se crea el sistema:

$$\begin{bmatrix} 2(h_1 + h_2) & h_2 & & & & & \\ h_2 & 2(h_2 + h_3) & h_3 & & & & \\ & h_3 & 2(h_3 + h_4) & h_4 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & & \\ & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & & \end{bmatrix} \begin{bmatrix} S_2 \\ S_3 \\ \vdots \\ \vdots \\ S_{n-1} \end{bmatrix}$$

$$= 6\begin{bmatrix} \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ \frac{y_5 - y_4}{h_4} - \frac{y_4 - y_3}{h_3} \\ \vdots \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \end{bmatrix}$$

Esta última ecuación es el sistema solución para resolver los valores desconocidos $S_2 \ldots S_{n-1}$.

Recapitulando, luego de estudiar la diferenciación y bondades de los dos tipos importantes de regresión: la paramétrica y la no paramétrica, se pasó al estudio de los polinomios en general; de los polinomios por secciones; para luego centrarnos en el tema de esta tesis: las funciones Spline, que como se ha visto es un polinomio por secciones con propiedades especiales que la diferencian del resto de polinomios. Se hace un desarrollo de las funciones Spline para luego pasar al modelo Spline y finalmente concluir con la determinación de sus coeficientes, cabe aclarar que cada sección del modelo Spline es un polinomio, por consiguiente cada sección tendrá sus propios coeficientes.

CAPÍTULO 3: MATERIALES Y MÉTODOS

El conocimiento e interés por las estructuras de la mortalidad en las poblaciones humanas, se remonta al siglo XVII.

Fue John Graunt quién en base a los Bills of Mortality, que eran boletines que se publicaban semanalmente en la ciudad de Londres desde comienzos del siglo XVII, y que contenía la relación de las defunciones (a veces los nacimientos) registrados en las diferentes Parroquias de la ciudad de Londres, publica en el año de 1662 *"Natural and political observations mentioned in a following index, and made upon the Bills of mortality, with reference to the government, religión, trade, growth, air, diseases and the several changes of the said city"*, y en el que se encuentran las primeras estructuras de mortalidad por sexo y edad. Se tiene conocimiento que fue el primero que aplicó el Análisis Exploratorio de Datos y calcula de forma simple las tasas de mortalidad por sexo y edad.

Posteriormente en el siglo XIX, exactamente en 1825, Gompertz construyó funciones matemáticas para las tasas de mortalidad en las edades adultas (de 45 a más años), para dicha elaboración se basó en funciones exponenciales.

En 1955, por iniciativa de V.G. Valaoras se publican las primeras series modernas de Tablas Modelo de Mortalidad de las Naciones Unidas, "se basa en un conjunto de 158 tablas de vida observadas para cada sexo"…"se construyeron bajo el supuesto que el valor de cada ${}_5q_x$ [9] es una función cuadrática del valor q anterior"…"Como los coeficientes de las ecuaciones cuadráticas que relacionan cada valor ${}_5q_x$ con su predecesor no se conocían a priori, tuvieron que estimarse en base a datos observados. Se recurrió a la regresión para estimar esos coeficientes con los 158 patrones de mortalidad disponibles para cada sexo"[10].

[9] Probabilidad de morir entre la edad x y la x+5 en una tabla de vida.
[10] Manual X Técnicas Indirectas de Estimación Demográfica. Departamento de Asuntos Económicos y sociales Internacionales, Estudios de Población, Nº 81. Naciones Unidas, Nueva York, 1986.

En 1959 Ledermann y Breas mediante el Análisis Factorial determinan 5 factores que explicaban la variabilidad en 154 tablas de mortalidad observadas. "El primero y más importante se refiere al nivel general de la mortalidad; el segundo refleja la relación entre la mortalidad en la niñez y adulta; el tercero está relacionado con el patrón de la mortalidad en las edades avanzadas; mientras que el cuarto va asociado con los patrones de mortalidad por debajo de los cinco años y, por último, el quinto refleja las diferencias entre la mortalidad masculina y la femenina en las edades comprendidas entre los 5 y 70 años."[11]

Coale y Demeny, en 1966, presenta sus cuatro familias de Tablas Modelo de Mortalidad (Norte, Sur, Este y Oeste), obtenidas en base a "coeficientes de las ecuaciones lineales que relacionaban los valores $_nq_x$ con e_{10} esperanza de vida a los 10 años, y de aquellas que relacionaban los valores de $\log_{10}({}_nq_x)$ con e_{10}, se estimaron utilizando la regresión por mínimos cuadrados"[12] que realizan sobre 192 tablas de mortalidad observadas (seleccionadas de un total de 326); 39 de ellas correspondían al siglo XIX, y 69 a después de la Segunda Guerra Mundial. Encontrándose sobre-representada la experiencia occidental (Europa, América del Norte, Australia y Nueva Zelanda con un total de 176 tablas; 3 de Israel, 6 de Japón, 3 de Taiwán y 4 de la población blanca de Sudáfrica.

En 1968 Brass y colegas, obtienen "un modelo que brinda un mayor grado de flexibilidad"…"mejor conocido como sistema logito. Brass intentó relacionar matemáticamente dos tablas de vida diferentes. Descubrió que una determinada transformación de las probabilidades de sobrevivir hasta la edad x (valores de l(x) en términos de la tabla de vida) hacía que la relación entre las correspondientes probabilidades de las distintas tablas de vida resultase aproximadamente lineal"[13].

[11] NN UU. Manual X . Op. Cit.
[12] NN UU. Manual X, Op. Cit.
[13] NN UU. Manual X. Op. Cit.

En el Perú no se cuenta con modelos teóricos que representen adecuadamente la tendencia de la mortalidad en nuestro país, dada las bondades del MRNPS amerita utilizar para obtener estructuras de mortalidad por sexo y edad que reflejen la realidad de la mortalidad peruana, teniendo en cuenta que hasta la actualidad no existen trabajos preliminares en Perú respecto a la utilización de este modelo específicamente. A finales de la década de los años 60 del siglo pasado se construyeron tablas de mortalidad para el Perú utilizando el modelo logito, posteriormente se utilizaron las tablas modelo de Coale y Demeny.

A nivel mundial los modelos de mortalidad elaborados son de la década de los años 60 del siglo pasado y para lo cual utilizaron el modelo logito en base a un modelo estándar de mortalidad, o regresión por mínimos cuadrados, obteniéndose en ambos casos modelo de mortalidad con información del siglo XIX y primera mitad del siglo XX, y que no representan la mortalidad actual, ya que además de no pertenecer a nuestra región, no reflejan la situación de salud que en su época si predominaba.

Las estructuras deben reflejar una mortalidad que sea representativa de la realidad del país y en base a la data de la que se dispone especialmente las defunciones de las estadísticas vitales, porque al no considerar con precisión las muertes, se puede modifica sustancialmente las proyecciones de población.

El MRNPS es muy flexible en la determinación de curvas no lineales a través de funciones cuadráticas o cúbicas que se generan para un conjunto de datos del que se disponga.

Para la aplicación del MRNPS se dispone de información proveniente de las estadísticas vitales de defunciones del Ministerio de Salud y los datos de población correspondientes al último Censo Nacional de Población 2007 del INEI.

3.1 Datos y Software

Los datos de Perú con los que se trabaja en la presente Tesis corresponden a una parte del mismo, concretamente al departamento o región de Lima, como un ejemplo de cómo se puede estructurar la mortalidad de las áreas mayores o departamentos del país.

El insumo que se necesita para la aplicación del MRNPS es la tasa de mortalidad por sexo y edades simples, la cual es el resultado de dividir las defunciones de cada una de las edades entre la población total de la misma edad.

Se ha tomado las bases de datos de las defunciones ocurridas en el departamento de Lima por sexo y edades simples (0 a 95 y más años de edad), para los años 2006 y 2008, información proporcionada por el Ministerio de Salud (MINSA) y que en el caso del departamento de Lima reúne a un buen volumen (aproximadamente el 25 por ciento) de las defunciones ocurridas en el Perú.

Asimismo, se ha tomado la información de la población del departamento de Lima por sexo y edades simples (0 a 95 y más años de edad) correspondiente al Censo Nacional de Población del año 2007 publicado por el Instituto Nacional de Estadística e Informática (INEI).

Para viabilizar la aplicación del modelo Spline se ha hecho uso de la versión 10 de prueba por 30 días del software John's Macintosh Project (JMP) del grupo SAS, que es un programa que ayuda a hacer pruebas y análisis estadísticos; la cual tiene entre otras opciones las bondades de poder modelar rápidamente con ayuda del mouse, la curva que más se ajusta al conjunto de datos con los que se está trabajando, a su vez se va mostrando el coeficiente de determinación R^2, el valor del parámetro de suavización λ, así como la suma de cuadrados residuales.

Una vez que se tiene las defunciones y la población censada, ambas por sexo y edades simples, en el caso de las defunciones se procede a calcular un promedio de las defunciones del 2006 y 2008, para centrarlas en el 2007, en el caso del Censo de Población del 2007, como este se realizó el 21 de octubre del 2007, con una tasa de crecimiento intercensal 1993-2007, se procede a retroceder la población al 30 de junio del 2007, para que de esta manera quede centrada la población correspondiente al año 2007.

A continuación se presenta la información cruda inicial correspondiente a las defunciones y a la población censada, y las primeras operaciones que se han realizado en ella.

CUADRO Nº 3.1

DEPARTAMENTO DE LIMA: DEFUNCIONES 2006, 2008 Y PROMEDIO CENTRADO AL 2007

Edad	Defunciones 2006			Defunciones 2008			Defunciones promedio 2007		
	Total	Hombre	Mujer	Total	Hombre	Mujer	Total	Hombre	Mujer
Total	**21551**	**11446**	**10105**	**28714**	**15012**	**13702**	**25183**	**13254**	**11929**
0	951	560	391	1004	547	457	978	554	424
1	94	48	46	92	48	44	93	48	45
2	41	21	20	55	26	29	49	24	25
3	23	14	9	39	25	14	32	20	12
4	22	16	6	32	21	11	28	19	9
5	19	12	7	26	14	12	23	13	10
6	24	14	10	22	14	8	23	14	9
7	16	7	9	23	14	9	20	11	9
8	29	16	13	33	15	18	32	16	16
9	23	17	6	24	17	7	24	17	7
10	21	16	5	25	14	11	23	15	8
11	13	6	7	27	16	11	20	11	9
12	20	13	7	25	15	10	23	14	9
13	23	14	9	28	17	11	26	16	10
14	28	16	12	36	23	13	33	20	13
15	24	17	7	38	20	18	32	19	13
16	40	27	13	46	28	18	44	28	16
17	39	19	20	58	35	23	49	27	22
18	54	33	21	58	26	32	57	30	27
19	37	25	12	78	46	32	58	36	22
20	62	34	28	97	58	39	80	46	34
21	62	42	20	84	50	34	73	46	27
22	70	46	24	76	52	24	73	49	24
23	65	46	19	76	51	25	71	49	22
24	78	45	33	100	60	40	90	53	37
25	78	50	28	122	93	29	101	72	29
26	63	41	22	106	64	42	85	53	32
27	76	49	27	118	83	35	97	66	31
28	86	59	27	100	69	31	93	64	29
29	99	62	37	102	62	40	101	62	39
30	82	62	20	128	90	38	105	76	29
31	90	56	34	109	78	31	100	67	33
32	92	69	23	93	58	35	93	64	29
33	87	53	34	122	82	40	105	68	37
34	109	80	29	118	71	47	114	76	38
35	107	67	40	132	86	46	120	77	43
36	104	68	36	120	74	46	112	71	41
37	82	52	30	126	81	45	105	67	38
38	103	66	37	142	73	69	123	70	53
39	101	67	34	166	93	73	134	80	54
40	117	72	45	170	96	74	144	84	60
41	118	64	54	137	71	66	128	68	60
42	107	56	51	192	103	89	150	80	70
43	121	62	59	163	87	76	143	75	68
44	121	72	49	181	102	79	151	87	64
45	141	75	66	200	112	88	171	94	77

Continúa ...

CUADRO Nº 3.1

DEPARTAMENTO DE LIMA: DEFUNCIONES 2006, 2008 Y PROMEDIO CENTRADO AL 2007

Conclusión.

Edad	Defunciones 2006			Defunciones 2008			Defunciones promedio 2007		
	Total	Hombre	Mujer	Total	Hombre	Mujer	Total	Hombre	Mujer
46	139	72	67	179	89	90	160	81	79
47	140	77	63	181	94	87	161	86	75
48	134	71	63	206	112	94	171	92	79
49	160	83	77	226	123	103	193	103	90
50	194	102	92	274	146	128	234	124	110
51	186	95	91	239	119	120	213	107	106
52	187	100	87	279	145	134	234	123	111
53	182	101	81	274	157	117	228	129	99
54	209	110	99	221	105	116	216	108	108
55	190	120	70	242	131	111	217	126	91
56	215	110	105	273	141	132	245	126	119
57	211	106	105	300	150	150	256	128	128
58	228	137	91	278	127	151	253	132	121
59	251	148	103	309	171	138	281	160	121
60	306	174	132	371	198	173	339	186	153
61	236	117	119	351	178	173	294	148	146
62	250	120	130	369	190	179	310	155	155
63	260	136	124	341	180	161	301	158	143
64	283	155	128	381	219	162	332	187	145
65	374	193	181	390	213	177	382	203	179
66	307	174	133	361	221	140	335	198	137
67	306	184	122	436	268	168	371	226	145
68	377	214	163	448	250	198	413	232	181
69	384	225	159	500	279	221	442	252	190
70	449	257	192	577	314	263	514	286	228
71	403	222	181	525	278	247	464	250	214
72	445	240	205	554	312	242	500	276	224
73	430	224	206	561	299	262	496	262	234
74	448	240	208	613	342	271	531	291	240
75	477	267	210	660	357	303	569	312	257
76	521	266	255	604	333	271	563	300	263
77	493	282	211	667	350	317	580	316	264
78	583	319	264	753	403	350	668	361	307
79	485	264	221	768	389	379	627	327	300
80	591	308	283	731	372	359	661	340	321
81	532	275	257	765	412	353	649	344	305
82	564	312	252	750	395	355	658	354	304
83	535	272	263	680	359	321	608	316	292
84	516	246	270	715	360	355	616	303	313
85	563	267	296	724	376	348	644	322	322
86	508	246	262	656	335	321	583	291	292
87	402	193	209	695	347	348	549	270	279
88	401	183	218	603	270	333	503	227	276
89	371	172	199	518	244	274	445	208	237
90	349	154	195	497	224	273	423	189	234
91	325	143	182	456	188	268	391	166	225
92	283	104	179	397	171	226	341	138	203
93	252	94	158	407	142	265	330	118	212
94	212	85	127	277	93	184	245	89	156
95y+	742	261	481	1083	361	722	915	312	603

CUADRO Nº 3.2

DEPARTAMENTO DE LIMA: POBLACIÓN CENSADA EL 21 DE OCTUBRE DEL 2007 Y POBLACIÓN CENSADA RETROCEDIDA AL 30 DE JUNIO DEL 2007

Edad	Población censada el 21 de octubre del 2007			Población censada retrocedida al 30 de junio del 2007			Edad	Población censada el 21 de octubre del 2007			Población censada retrocedida al 30 de junio del 2007		
	Total	Hombre	Mujer	Total	Hombre	Mujer		Total	Hombre	Mujer	Total	Hombre	Mujer
Total	**8445211**	**4139686**	**4305525**	**8394200**	**4114683**	**4279517**	**46**	88035	42208	45827	87503	41953	45550
							47	101166	48862	52304	100555	48567	51988
0	132652	67663	64989	131850	67254	64596	**48**	89718	42608	47110	89176	42351	46825
1	137083	70271	66812	136255	69847	66408	**49**	80595	38390	42205	80108	38158	41950
2	151267	76778	74489	150353	76314	74039	**50**	91904	42708	49196	91349	42450	48899
3	150196	76635	73561	149289	76172	73117	**51**	67598	31891	35707	67189	31698	35491
4	142609	73147	69462	141747	72705	69042	**52**	84974	40595	44379	84461	40350	44111
5	135737	69455	66282	134917	69035	65882	**53**	79509	37523	41986	79028	37296	41732
6	133386	68022	65364	132580	67611	64969	**54**	75243	35369	39874	74788	35155	39633
7	139824	71624	68200	138979	71191	67788	**55**	67186	32016	35170	66781	31823	34958
8	141106	72079	69027	140254	71644	68610	**56**	63342	30622	32720	62959	30437	32522
9	135291	69230	66061	134474	68812	65662	**57**	67119	32094	35025	66713	31900	34813
10	148102	75333	72769	147207	74878	72329	**58**	57907	27768	30139	57557	27600	29957
11	146706	74551	72155	145820	74101	71719	**59**	52550	25618	26932	52232	25463	26769
12	152154	77192	74962	151235	76726	74509	**60**	63542	30101	33441	63158	29919	33239
13	147214	74318	72896	146325	73869	72456	**61**	40494	20083	20411	40250	19962	20288
14	152495	76430	76065	151574	75968	75606	**62**	50576	24629	25947	50270	24480	25790
15	156445	77060	79385	155500	76595	78905	**63**	47192	22792	24400	46907	22654	24253
16	145761	71290	74471	144880	70859	74021	**64**	43293	21145	22148	43031	21017	22014
17	154252	76437	77815	153320	75975	77345	**65**	48687	23505	25182	48393	23363	25030
18	169019	84247	84772	167998	83738	84260	**66**	34887	17007	17880	34676	16904	17772
19	173687	85141	88546	172640	84628	88012	**67**	41664	20020	21644	41412	19899	21513
20	178023	87446	90577	176950	86919	90031	**68**	34593	16283	18310	34384	16185	18199
21	155328	76143	79185	154390	75683	78707	**69**	30258	15228	15030	30075	15136	14939
22	167932	83445	84487	166918	82941	83977	**70**	38095	17936	20159	37865	17828	20037
23	166517	81450	85067	165511	80958	84553	**71**	24418	12196	12222	24270	12122	12148
24	169238	83295	85943	168216	82792	85424	**72**	31330	14881	16449	31141	14791	16350
25	167682	82049	85633	166669	81553	85116	**73**	27554	13260	14294	27388	13180	14208
26	153880	74731	79149	152951	74280	78671	**74**	26812	12433	14379	26650	12358	14292
27	163787	81015	82772	162798	80526	82272	**75**	28671	13336	15335	28497	13255	15242
28	152011	74553	77458	151093	74103	76990	**76**	22133	10655	11478	22000	10591	11409
29	144113	69639	74474	143242	69218	74024	**77**	23270	11045	12225	23129	10978	12151
30	166588	81102	85486	165582	80612	84970	**78**	21811	10177	11634	21680	10116	11564
31	133281	64399	68882	132476	64010	68466	**79**	17145	8230	8915	17041	8180	8861
32	147552	71523	76029	146661	71091	75570	**80**	20338	8965	11373	20215	8911	11304
33	143455	70647	72808	142588	70220	72368	**81**	12316	5801	6515	12242	5766	6476
34	130233	63408	66825	129446	63025	66421	**82**	14105	6374	7731	14019	6335	7684
35	131965	63584	68381	131168	63200	67968	**83**	11640	5218	6422	11569	5186	6383
36	122011	58324	63687	121274	57972	63302	**84**	11062	4894	6168	10995	4864	6131
37	133981	64377	69604	133172	63988	69184	**85**	10774	4513	6261	10709	4486	6223
38	123238	58862	64376	122493	58506	63987	**86**	8731	3824	4907	8678	3801	4877
39	115609	55660	59949	114911	55324	59587	**87**	8471	3513	4958	8420	3492	4928
40	128881	61631	67250	128103	61259	66844	**88**	5643	2438	3205	5609	2423	3186
41	96264	46500	49764	95682	46219	49463	**89**	4941	2062	2879	4912	2050	2862
42	121072	58795	62277	120341	58440	61901	**90**	4619	1734	2885	4592	1724	2868
43	103858	49712	54146	103231	49412	53819	**91**	2476	1033	1443	2461	1027	1434
44	96374	46877	49497	95792	46594	49198	**92**	2812	1061	1751	2795	1055	1740
45	95597	45217	50380	95020	44944	50076	**93**	2344	849	1495	2330	844	1486
							94	1887	708	1179	1876	704	1172
							95y+	6325	2198	4127	6287	2185	4102

Tanto las defunciones como la población censada adolecen de una omisión, sea de registro y/o envío a las oficinas del Ministerio de Salud en todo el país en el caso de las defunciones, sea de no empadronamiento de las personas en el caso del Censo Nacional de Población.

El demógrafo matemático William Brass creó una metodología para poder calcular el volumen de las defunciones omitidas, para la que usó la estructura de la población censada por sexo y edad. El porcentaje de omisión obtenido de la aplicación de esta metodología, sirve para que se complete las defunciones faltantes (ver apéndice D y anexo 1).

Una vez obtenido el porcentaje de omisión en las defunciones, se procede a aplicarse a las defunciones registradas por edad y sexo, para obtener el total de las defunciones, las cuáles serán el numerador para el cálculo de las tasas de mortalidad por sexo y edad (ver anexo 2).

Con respecto a la población censada, paralelamente al Censo o dentro de los días siguientes al mismo se realiza una encuesta post-censal cuya finalidad es calcular la población y viviendas, que no han sido empadronadas o registradas el día del Censo. Esta información conduce a la obtención de un porcentaje de omisión de la población que no fue censada. Dicho porcentaje asume que no hay diferenciales respecto al sexo y edad, con dicho porcentaje se calcula la población omitida el día del Censo, que para el caso del departamento de Lima fue 1.4%. (ver anexo 3).

Finalmente la data estará lista para proceder al cálculo de la tasa de mortalidad por sexo y para cada una de las edades comprendidas entre 0 y 95 años (ver anexo 4), utilizando la siguiente fórmula:

$$Tasa\ de\ mortalidad\ de\ la\ edad\ x_i = \frac{Número\ de\ defunciones, de\ la\ edad\ x}{Población\ \ total\ a\ mitad\ de\ año, de\ la\ edad\ x}$$

$$donde\ x = 0, 1, 2, \cdots, 95$$

Se procede a importar los logaritmos (ver anexo 5) de las tasas de mortalidad por sexo y edad, al software JMP versión 10, luego de ingresar al módulo: Spline cúbico, se obtiene una gráfica bivariada de la Edad y Tasa de Mortalidad por Edad. También se muestra su correspondiente λ, R^2 y lasuma de errores al cuadrado.

El valor de λ es mayor que cero y puede variar de acuerdo a la amplitud entre el mínimo y máximo valor en estudio. Cuando el valor de λ se aproxima a cero el ajuste se hace más flexible y curvado, cuando λ aumenta el ajuste se hace más rígido (menos curvado) aproximándose a una línea recta, cuando llega al límite superior cercano al valor máximo del estudio.

El software permite determinar el mejor ajuste para la curva del diagrama de dispersión, moviendo el mouse en la barra deslizadora que se encuentra debajo de la gráfica y del valor de lambda (λ). Se pueda observar a simple vista la mejor gráfica para los valores esperados de Y en X.

El informe que se guarda, correspondiente a cada suavizamiento del Spline (ver anexo 6), contiene no sólo los coeficientes correspondientes al polinomio de cada sección del Spline, sino también el R^2 y la suma de cuadrados residuales de todo el Spline. Se pueden realizar varias pruebas hasta lograr la mejor suavización de la curva.

Luego se hacen las comparaciones tanto del R^2 como de la suma de cuadrados residuales, para decidir por la curva con el mejor ajuste Spline.

Los resultados de la aplicación del modelo Spline a los logaritmos de las tasas de mortalidad por sexo y edad se presentan en el siguiente cuadro, seguido de sus gráficas correspondientes, finalmente se presentan los cuadros de los antilogaritmos de las tasas trabajadas, para tener las tasas de mortalidad por sexo y edad ajustadas con el Spline.

CUADRO Nº 3.3

DEPARTAMENTO DE LIMA: RESULTADOS DE APLICAR EL MODELO SPLINE, LN DE TASA DE MORTALIDAD, RESIDUALES Y PREDICTORES SPLINE POR EDAD, HOMBRES 2007

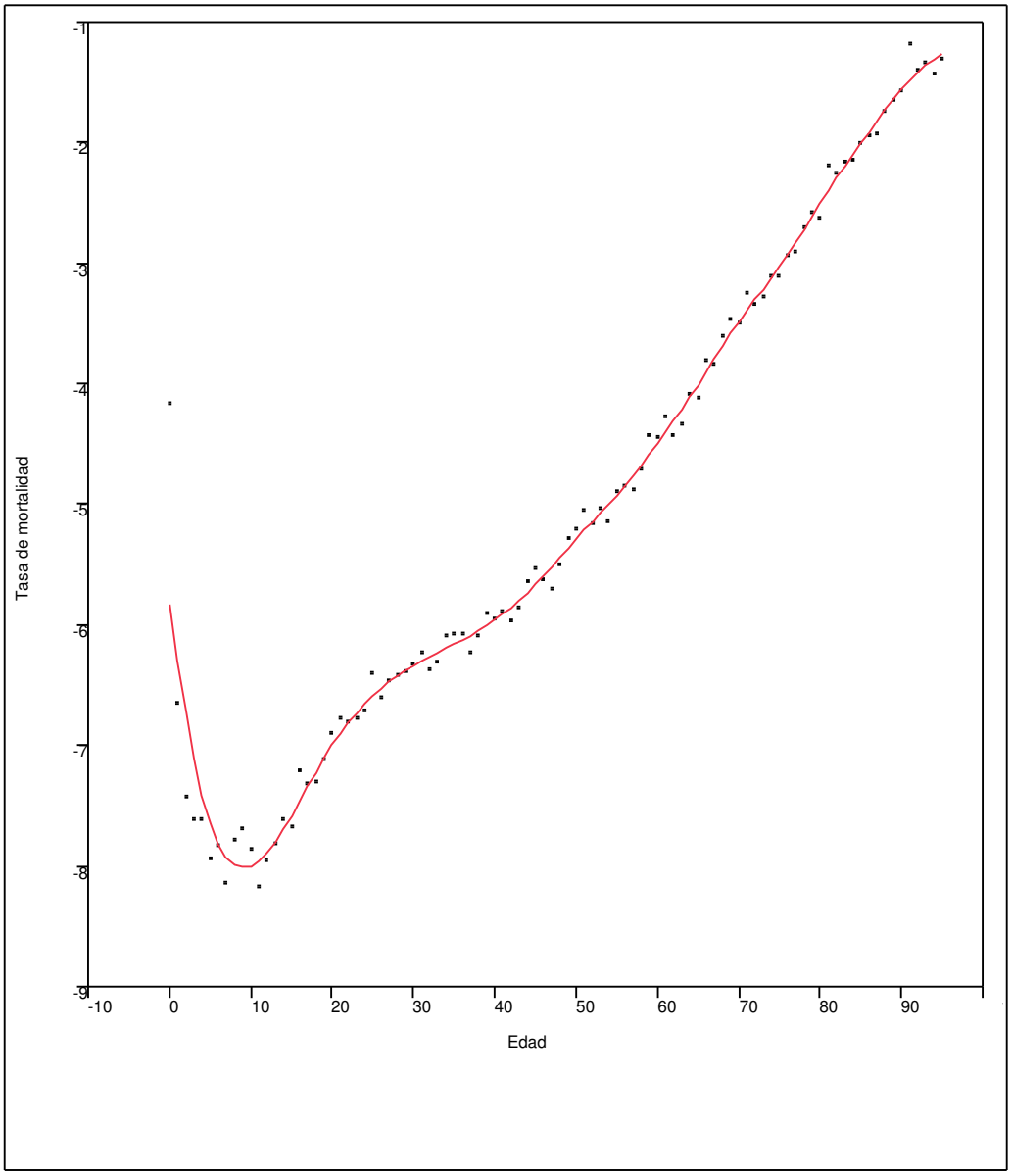
Edad	ln de tasa de mortalidad	Predictor	Residual	Predictor Spline	Edad	ln de tasa de mortalidad	Predictor	Residual	Predictor Spline
0	-4,1690	-5,8233	1,6543	-5,8233	**46**	-5,6215	-5,5943	-0,0271	-5,5943
1	-6,6573	-6,2903	-0,3670	-6,2903	**47**	-5,7049	-5,5216	-0,1833	-5,5216
2	-7,4281	-6,7255	-0,7026	-6,7255	**48**	-5,5031	-5,4448	-0,0582	-5,4448
3	-7,6173	-7,1041	-0,5131	-7,1041	**49**	-5,2855	-5,3660	0,0805	-5,3660
4	-7,6247	-7,4140	-0,2108	-7,4140	**50**	-5,2064	-5,2884	0,0820	-5,2884
5	-7,9376	-7,6525	-0,2851	-7,6525	**51**	-5,0600	-5,2139	0,1540	-5,2139
6	-7,8398	-7,8229	-0,0169	-7,8229	**52**	-5,1641	-5,1428	-0,0213	-5,1428
7	-8,1427	-7,9326	-0,2101	-7,9326	**53**	-5,0354	-5,0729	0,0375	-5,0729
8	-7,7924	-7,9911	0,1988	-7,9911	**54**	-5,1537	-5,0018	-0,1519	-5,0018
9	-7,6875	-8,0102	0,3227	-8,0102	**55**	-4,9014	-4,9270	0,0256	-4,9270
10	-7,8704	-7,9983	0,1279	-7,9983	**56**	-4,8569	-4,8480	-0,0089	-4,8480
11	-8,1828	-7,9589	-0,2239	-7,9589	**57**	-4,8873	-4,7644	-0,1229	-4,7644
12	-7,9663	-7,8930	-0,0733	-7,8930	**58**	-4,7142	-4,6762	-0,0380	-4,6762
13	-7,8230	-7,8047	-0,0182	-7,8047	**59**	-4,4388	-4,5854	0,1466	-4,5854
14	-7,6146	-7,6999	0,0853	-7,6999	**60**	-4,4511	-4,4942	0,0431	-4,4942
15	-7,6769	-7,5844	-0,0925	-7,5844	**61**	-4,2738	-4,4029	0,1291	-4,4029
16	-7,2123	-7,4635	0,2513	-7,4635	**62**	-4,4328	-4,3103	-0,1225	-4,3103
17	-7,3204	-7,3427	0,0223	-7,3427	**63**	-4,3351	-4,2138	-0,1213	-4,2138
18	-7,3065	-7,2240	-0,0824	-7,2240	**64**	-4,0922	-4,1123	0,0201	-4,1123
19	-7,1260	-7,1090	-0,0170	-7,1090	**65**	-4,1146	-4,0069	-0,1077	-4,0069
20	-6,9095	-6,9999	0,0904	-6,9999	**66**	-3,8172	-3,8988	0,0816	-3,8988
21	-6,7711	-6,8996	0,1285	-6,8996	**67**	-3,8488	-3,7907	-0,0581	-3,7907
22	-6,8074	-6,8090	0,0016	-6,8090	**68**	-3,6146	-3,6844	0,0698	-3,6844
23	-6,7832	-6,7271	-0,0560	-6,7271	**69**	-3,4652	-3,5823	0,1172	-3,5823
24	-6,7231	-6,6529	-0,0701	-6,6529	**70**	-3,5018	-3,4855	-0,0163	-3,4855
25	-6,4031	-6,5861	0,1830	-6,5861	**71**	-3,2514	-3,3934	0,1419	-3,3934
26	-6,6146	-6,5269	-0,0877	-6,5269	**72**	-3,3506	-3,3046	-0,0460	-3,3046
27	-6,4741	-6,4737	-0,0005	-6,4737	**73**	-3,2879	-3,2163	-0,0716	-3,2163
28	-6,4233	-6,4254	0,0021	-6,4254	**74**	-3,1190	-3,1259	0,0070	-3,1259
29	-6,3884	-6,3813	-0,0071	-6,3813	**75**	-3,1193	-3,0319	-0,0874	-3,0319
30	-6,3348	-6,3406	0,0058	-6,3406	**76**	-2,9344	-2,9333	-0,0011	-2,9333
31	-6,2289	-6,3025	0,0737	-6,3025	**77**	-2,9175	-2,8301	-0,0874	-2,8301
32	-6,3818	-6,2661	-0,1157	-6,2661	**78**	-2,7022	-2,7231	0,0209	-2,7231
33	-6,3137	-6,2297	-0,0840	-6,2297	**79**	-2,5890	-2,6140	0,0251	-2,6140
34	-6,0887	-6,1932	0,1045	-6,1932	**80**	-2,6368	-2,5047	-0,1322	-2,5047
35	-6,0777	-6,1578	0,0800	-6,1578	**81**	-2,1892	-2,3967	0,2075	-2,3967
36	-6,0765	-6,1231	0,0466	-6,1231	**82**	-2,2532	-2,2930	0,0397	-2,2930
37	-6,2285	-6,0874	-0,1411	-6,0874	**83**	-2,1676	-2,1937	0,0261	-2,1937
38	-6,1006	-6,0484	-0,0523	-6,0484	**84**	-2,1459	-2,0979	-0,0480	-2,0979
39	-5,9112	-6,0056	0,0944	-6,0056	**85**	-2,0045	-2,0042	-0,0003	-2,0042
40	-5,9618	-5,9597	-0,0021	-5,9597	**86**	-1,9400	-1,9119	-0,0281	-1,9119
41	-5,8954	-5,9099	0,0145	-5,9099	**87**	-1,9302	-1,8204	-0,1098	-1,8204
42	-5,9660	-5,8554	-0,1106	-5,8554	**88**	-1,7383	-1,7299	-0,0083	-1,7299
43	-5,8592	-5,7955	-0,0637	-5,7955	**89**	-1,6582	-1,6423	-0,0159	-1,6423
44	-5,6513	-5,7310	0,0796	-5,7310	**90**	-1,5801	-1,5599	-0,0202	-1,5599
45	-5,5399	-5,6637	0,1238	-5,6637	**91**	-1,1932	-1,4853	0,2921	-1,4853
					92	-1,4033	-1,4205	0,0172	-1,4205
					93	-1,3362	-1,3636	0,0274	-1,3636
					94	-1,4410	-1,3115	-0,1295	-1,3115
					95y+	-1,3166	-1,2609	-0,0557	-1,2609

Cuadro Nº 3.4

DEPARTAMENTO DE LIMA: RESULTADOS DE APLICAR EL MODELO SPLINE, LN DE TASA DE MORTALIDAD, RESIDUALES Y PREDICTORES SPLINE POR EDAD, MUJERES 2007

Edad	ln de tasa de mortalidad	Predictor	Residual	Predictor Spline	Edad	ln de tasa de mortalidad	Predictor	Residual	Predictor Spline
0	-4,3261	-5,8906	1,5644	-5,8906	**46**	-5,6593	-5,7449	0,0857	-5,7449
1	-6,5959	-6,4005	-0,1953	-6,4005	**47**	-5,8424	-5,6669	-0,1755	-5,6669
2	-7,2946	-6,8799	-0,4147	-6,8799	**48**	-5,6869	-5,5882	-0,0987	-5,5882
3	-7,9950	-7,3019	-0,6931	-7,3019	**49**	-5,4434	-5,5094	0,0660	-5,5094
4	-8,2662	-7,6485	-0,6176	-7,6485	**50**	-5,3955	-5,4326	0,0371	-5,4326
5	-8,1140	-7,9142	-0,1998	-7,9142	**51**	-5,1159	-5,3594	0,2435	-5,3594
6	-8,2054	-8,1047	-0,1007	-8,1047	**52**	-5,2836	-5,2898	0,0062	-5,2898
7	-8,2479	-8,2308	-0,0171	-8,2308	**53**	-5,3449	-5,2203	-0,1246	-5,2203
8	-7,6538	-8,3054	0,6516	-8,3054	**54**	-5,2034	-5,1471	-0,0563	-5,1471
9	-8,4673	-8,3397	-0,1277	-8,3397	**55**	-5,2502	-5,0681	-0,1821	-5,0681
10	-8,4305	-8,3367	-0,0938	-8,3367	**56**	-4,9107	-4,9831	0,0723	-4,9831
11	-8,3042	-8,2993	-0,0049	-8,2993	**57**	-4,9073	-4,8940	-0,0133	-4,8940
12	-8,3424	-8,2322	-0,1102	-8,2322	**58**	-4,8122	-4,8028	-0,0094	-4,8028
13	-8,2091	-8,1406	-0,0685	-8,1406	**59**	-4,6997	-4,7111	0,0114	-4,7111
14	-7,9516	-8,0316	0,0801	-8,0316	**60**	-4,6826	-4,6209	-0,0617	-4,6209
15	-7,9943	-7,9132	-0,0810	-7,9132	**61**	-4,2348	-4,5341	0,2993	-4,5341
16	-7,7297	-7,7930	0,0633	-7,7930	**62**	-4,4129	-4,4524	0,0394	-4,4524
17	-7,4635	-7,6789	0,2154	-7,6789	**63**	-4,4336	-4,3736	-0,0600	-4,3736
18	-7,3484	-7,5778	0,2294	-7,5778	**64**	-4,3232	-4,2944	-0,0288	-4,2944
19	-7,5927	-7,4926	-0,1000	-7,4926	**65**	-4,2393	-4,2119	-0,0274	-4,2119
20	-7,1879	-7,4229	0,2350	-7,4229	**66**	-4,1647	-4,1241	-0,0406	-4,1241
21	-7,2803	-7,3678	0,0876	-7,3678	**67**	-4,3001	-4,0295	-0,2706	-4,0295
22	-7,4606	-7,3237	-0,1368	-7,3237	**68**	-3,9097	-3,9283	0,0186	-3,9283
23	-7,5526	-7,2852	-0,2674	-7,2852	**69**	-3,6648	-3,8242	0,1594	-3,8242
24	-7,0387	-7,2494	0,2107	-7,2494	**70**	-3,7753	-3,7210	-0,0542	-3,7210
25	-7,2883	-7,2168	-0,0715	-7,2168	**71**	-3,3390	-3,6208	0,2817	-3,6208
26	-7,1127	-7,1861	0,0734	-7,1861	**72**	-3,5914	-3,5245	-0,0669	-3,5245
27	-7,1887	-7,1561	-0,0326	-7,1561	**73**	-3,4061	-3,4300	0,0240	-3,4300
28	-7,1880	-7,1250	-0,0630	-7,1250	**74**	-3,3872	-3,3350	-0,0522	-3,3350
29	-6,8442	-7,0912	0,2470	-7,0912	**75**	-3,3825	-3,2371	-0,1454	-3,2371
30	-7,2866	-7,0535	-0,2331	-7,0535	**76**	-3,0703	-3,1352	0,0649	-3,1352
31	-6,9435	-7,0082	0,0647	-7,0082	**77**	-3,1296	-3,0300	-0,0995	-3,0300
32	-7,1694	-6,9538	-0,2156	-6,9538	**78**	-2,9288	-2,9223	-0,0065	-2,9223
33	-6,8729	-6,8892	0,0163	-6,8892	**79**	-2,6852	-2,8137	0,1286	-2,8137
34	-6,7612	-6,8161	0,0550	-6,8161	**80**	-2,8609	-2,7062	-0,1547	-2,7062
35	-6,6636	-6,7366	0,0731	-6,7366	**81**	-2,3554	-2,6003	0,2449	-2,6003
36	-6,6390	-6,6516	0,0127	-6,6516	**82**	-2,5297	-2,4975	-0,0322	-2,4975
37	-6,8019	-6,5609	-0,2410	-6,5609	**83**	-2,3853	-2,3966	0,0112	-2,3966
38	-6,3984	-6,4645	0,0661	-6,4645	**84**	-2,2769	-2,2959	0,0190	-2,2959
39	-6,3088	-6,3656	0,0568	-6,3656	**85**	-2,2624	-2,1937	-0,0687	-2,1937
40	-6,3120	-6,2670	-0,0451	-6,2670	**86**	-2,1161	-2,0883	-0,0278	-2,0883
41	-6,0109	-6,1707	0,1598	-6,1707	**87**	-2,1712	-1,9787	-0,1924	-1,9787
42	-6,0845	-6,0785	-0,0060	-6,0785	**88**	-1,7455	-1,8655	0,1200	-1,8655
43	-5,9730	-5,9905	0,0175	-5,9905	**89**	-1,7914	-1,7511	-0,0403	-1,7511
44	-5,9425	-5,9060	-0,0366	-5,9060	**90**	-1,8060	-1,6375	-0,1684	-1,6375
45	-5,7792	-5,8242	0,0451	-5,8242	**91**	-1,1530	-1,5272	0,3742	-1,5272
					92	-1,4476	-1,4238	-0,0238	-1,4238
					93	-1,2471	-1,3267	0,0795	-1,3267
					94	-1,3157	-1,2339	-0,0818	-1,2339
					95y+	-1,2177	-1,1429	-0,0748	-1,1429

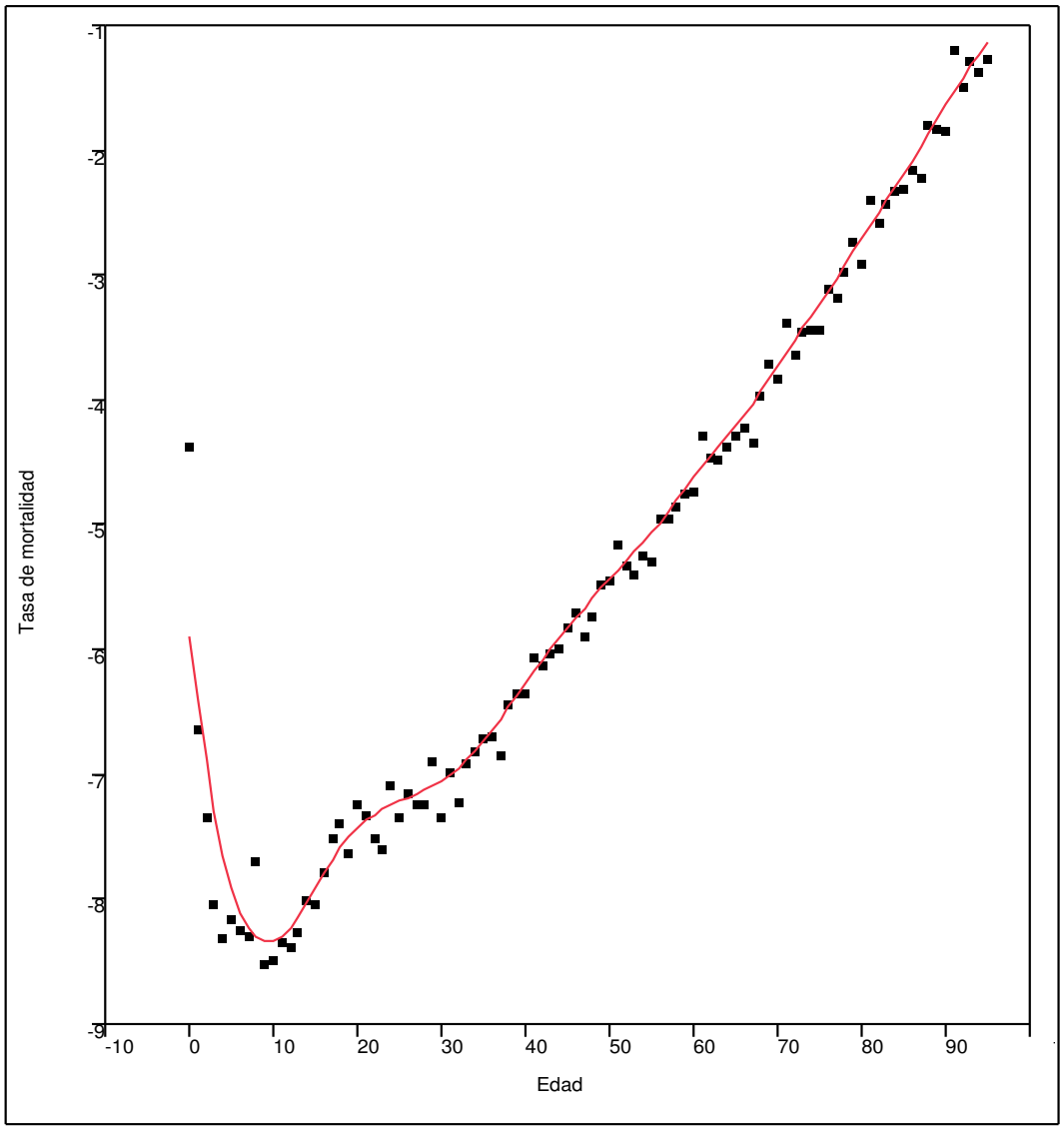
GRÁFICO Nº 3.1
DEPARTAMENTO DE LIMA: TASAS DE MORTALIDAD MASCULINAS SUAVIZADAS CON SPLINE, 2007



Smoothing Spline Fit, lambda=100

R-Square	0,987954
Sum of Squares Error	4,746199

GRÁFICO Nº 3.2
DEPARTAMENTO DE LIMA: TASAS DE MORTALIDAD FEMENINAS SUAVIZADAS CON SPLINE, 2007



Smoothing Spline Fit, lambda=100

R-Square	0,987526
Sum of Squares Error	5,491795

Cuadro Nº 3.5

DEPARTAMENTO DE LIMA: TASAS DE MORTALIDAD SUAVIZADAS CON SPLINE POR SEXO Y EDAD, 2007

Edad	Hombre	Mujer	Edad	Hombre	Mujer
0	0,0030	0,0028	**46**	0,0037	0,0032
1	0,0019	0,0017	**47**	0,0040	0,0035
2	0,0012	0,0010	**48**	0,0043	0,0037
3	0,0008	0,0007	**49**	0,0047	0,0040
4	0,0006	0,0005	**50**	0,0050	0,0044
5	0,0005	0,0004	**51**	0,0054	0,0047
6	0,0004	0,0003	**52**	0,0058	0,0050
7	0,0004	0,0003	**53**	0,0063	0,0054
8	0,0003	0,0002	**54**	0,0067	0,0058
9	0,0003	0,0002	**55**	0,0072	0,0063
10	0,0003	0,0002	**56**	0,0078	0,0069
11	0,0003	0,0002	**57**	0,0085	0,0075
12	0,0004	0,0003	**58**	0,0093	0,0082
13	0,0004	0,0003	**59**	0,0102	0,0090
14	0,0005	0,0003	**60**	0,0112	0,0098
15	0,0005	0,0004	**61**	0,0122	0,0107
16	0,0006	0,0004	**62**	0,0134	0,0117
17	0,0006	0,0005	**63**	0,0148	0,0126
18	0,0007	0,0005	**64**	0,0164	0,0136
19	0,0008	0,0006	**65**	0,0182	0,0148
20	0,0009	0,0006	**66**	0,0203	0,0162
21	0,0010	0,0006	**67**	0,0226	0,0178
22	0,0011	0,0007	**68**	0,0251	0,0197
23	0,0012	0,0007	**69**	0,0278	0,0218
24	0,0013	0,0007	**70**	0,0306	0,0242
25	0,0014	0,0007	**71**	0,0336	0,0268
26	0,0015	0,0008	**72**	0,0367	0,0295
27	0,0015	0,0008	**73**	0,0401	0,0324
28	0,0016	0,0008	**74**	0,0439	0,0356
29	0,0017	0,0008	**75**	0,0482	0,0393
30	0,0018	0,0009	**76**	0,0532	0,0435
31	0,0018	0,0009	**77**	0,0590	0,0483
32	0,0019	0,0010	**78**	0,0657	0,0538
33	0,0020	0,0010	**79**	0,0732	0,0600
34	0,0020	0,0011	**80**	0,0817	0,0668
35	0,0021	0,0012	**81**	0,0910	0,0743
36	0,0022	0,0013	**82**	0,1010	0,0823
37	0,0023	0,0014	**83**	0,1115	0,0910
38	0,0024	0,0016	**84**	0,1227	0,1007
39	0,0025	0,0017	**85**	0,1348	0,1115
40	0,0026	0,0019	**86**	0,1478	0,1239
41	0,0027	0,0021	**87**	0,1620	0,1382
42	0,0029	0,0023	**88**	0,1773	0,1548
43	0,0030	0,0025	**89**	0,1935	0,1736
44	0,0032	0,0027	**90**	0,2102	0,1945
45	0,0035	0,0030	**91**	0,2264	0,2171
			92	0,2416	0,2408
			93	0,2557	0,2654
			94	0,2694	0,2912
			95y+	0,2834	0,3189

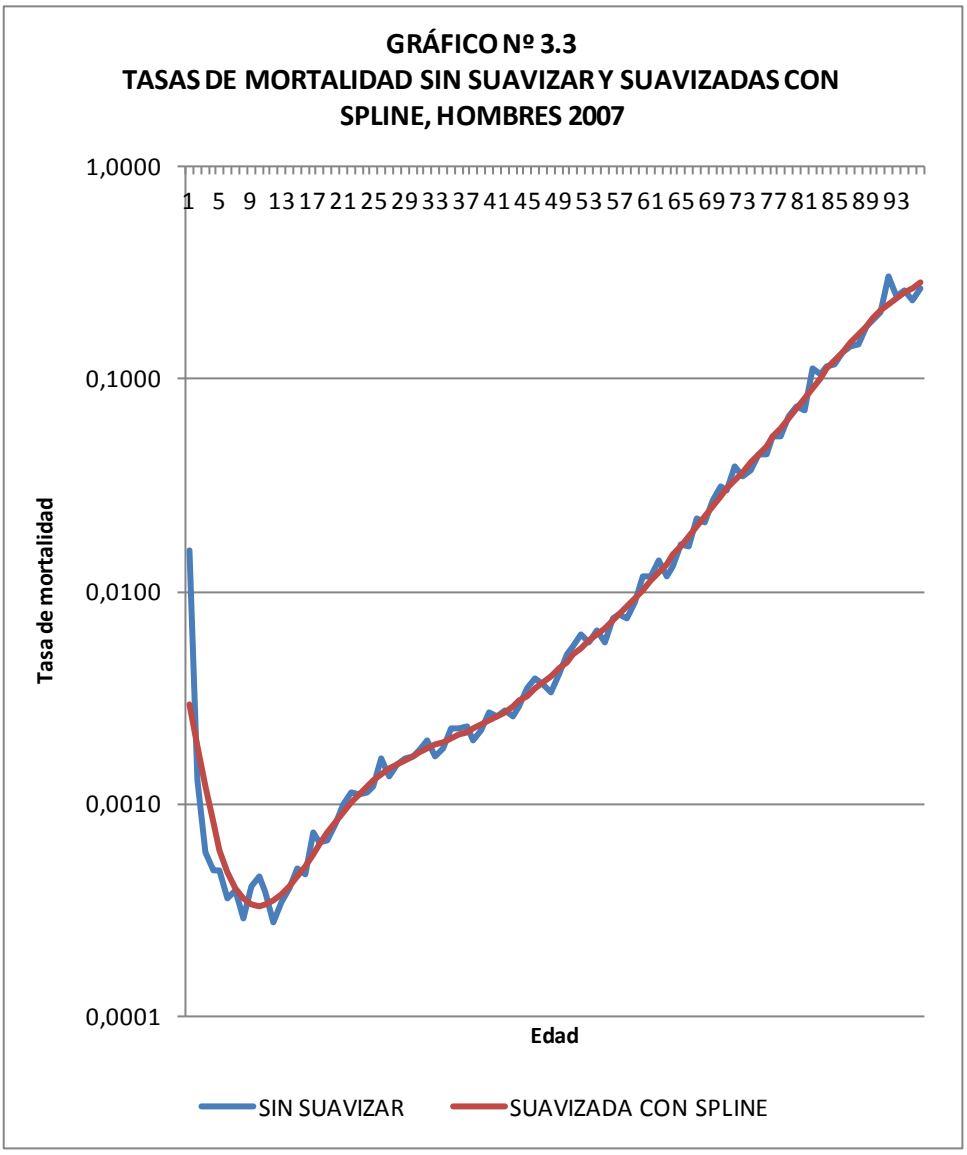
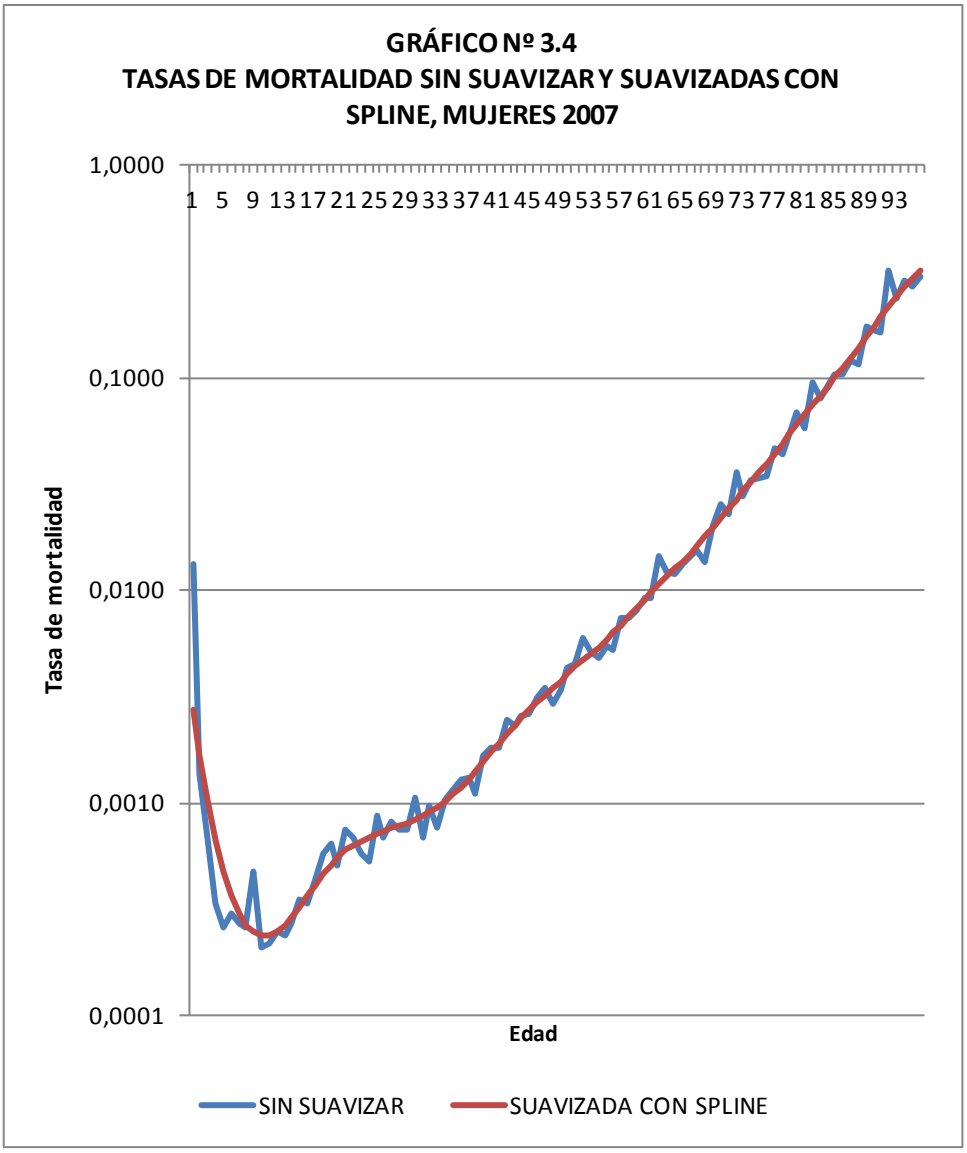


GRÁFICO Nº 3.3
TASAS DE MORTALIDAD SIN SUAVIZAR Y SUAVIZADAS CON SPLINE, HOMBRES 2007
1,0000
0,1000
0,0100
0,0010
0,0001
Tasa de mortalidad
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93
Edad
SIN SUAVIZAR
SUAVIZADA CON SPLINE

GRÁFICO Nº 3.4
TASAS DE MORTALIDAD SIN SUAVIZAR Y SUAVIZADAS CON SPLINE, MUJERES 2007



Como puede apreciarse en las gráficas correspondientes a hombres y mujeres, tanto las tasas de mortalidad por edad sin suavizar, como las tasas de mortalidad por edad suavizadas con el MRNPS (curva modelo), muestran que el ajuste es bastante bueno con un R cuadrado muy próximo a 1 y suma de cuadrados de los errores 4.75 en los hombres y 5.49 en las mujeres.

Podemos asegurar que los modelos obtenidos para los dos conjuntos de datos se ajustan a su correspondiente nube de puntos.

3.2 Discusión: ¿Se logra un mejor ajuste con la aplicación del MRNPS?

En base a lo observado en las primeras gráficas, podemos afirmar que las estructuras de muerte obtenidas por el MRNPS, son representativas del comportamiento de la estructura de la mortalidad, con lo cual queda confirmado que se está utilizando un modelo estadístico-matemático que permite obtener estructuras de muerte suavizadas y más robustas.

3.3 Conclusiones

1. La estadística matemática tiene una serie de técnicas y métodos que permiten precisar más adecuadamente los comportamientos de los fenómenos demográficos, esto conlleva el conocimiento demográfico adecuado de las poblaciones, para poder adaptar los diversos modelos a las realidades concretas.

2. El MRNPS, conocido también por algunos investigadores como polinomio cúbico, interpolador cúbico, polinomio por segmentos, aunque como se demostró el desarrollo de su teoría no es tan simple como el polinomio en sí, permite modelar las estructuras de mortalidad por edades simples, curva que una vez suavizada es la base para la construcción de las tablas de mortalidad o tablas de vida demográficas.

3. En vista de todo lo realizado en el presente estudio se recomienda la utilización del MRNPS para la suavización de la estructura de mortalidad

por edades simples y sexo, a partir de los 5 hasta los 95 y más años de edad, por la precisión de los métodos estadístico-matemáticos; de la edad 0 a los 4 años todavía será necesario la utilización de los métodos demográficos específicos.

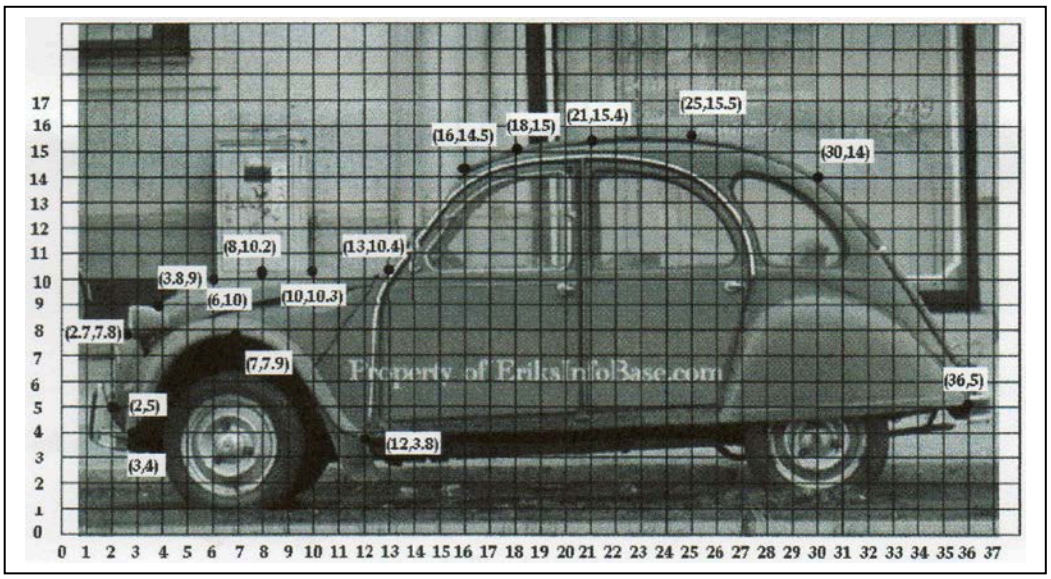
APÉNDICES

Apéndice A

Aplicaciones prácticas

Por ejemplo, con el Spline se puede reproducir con precisión la figura de un auto, superponiendo una cuadrícula a la lámina, para obtener una serie de puntos del contorno de dicho auto. Luego se aplica Spline y se puede modificar el contorno del auto para una nueva producción de automóviles.

FIGURA Nº A.1



Tomado de: Hernández, G. y col. Auxiliar 6: Interpolación mediante Spline cúbicos. Curso de Cálculo Numérico, Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Semenstre Otoño 2007.

Igualmente, si se desea hacer estudios sobre la composición de las rocas de determinada montaña (o calcular la velocidad de descongelamiento de los glaciales en un periodo determinado), se procede de la forma que se explica en el párrafo anterior y luego a través del Spline se hacen los cálculos respectivos.

FIGURA Nº A.2

Tomado de: Hernández, G. y col. Auxiliar 6: Interpolación mediante Spline cúbicos. Curso de Cálculo Numérico, Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Semenstre Otoño 2007.

Apéndice B

En la estimación no paramétrica los métodos de suavización conforman una colección de métodos que permiten estimar una curva ajustada a una distribución de probabilidad que correspondan a los datos observados, bajo mínimas suposiciones.

Pedro Delicado[14]: "Las funciones que habitualmente se estiman son:

- La función de densidad, sus derivadas o su integral (*función de distribución*):

$$X \sim f(x)$$
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$
$$f(x) = F'(x)$$

- La función de regresión o sus derivadas:

$$(X, Y) \sim F(x, y)$$
$$m(x) = E(Y \mid X = x)$$

- La función de riesgo, sus derivadas o su integral (*función de riesgo acumulada*):

$$X \sim f(x)$$
$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$
$$\Lambda(x) = \int_{-\infty}^{x} \lambda(u)du$$
$$\lambda(u) = \Lambda'(x)$$

- La curva principal, que es una versión no lineal de la primera componente principal", es decir se obtiene una curva que representa al conjunto de datos observados, de forma suavizada.

Se presenta ejemplos de las dos primeras funciones, en las que se pueden comparar las diferencias entre la regresión paramétrica y no paramétrica.

[14] Delicado, Pedro. Curso de Modelos no Paramétricos. Cap. 2, págs. 29-36. Departamento de Estadística e Investigación Operativa, Universidad Politécnica de Catalunya. 14 de setiembre de 2008.

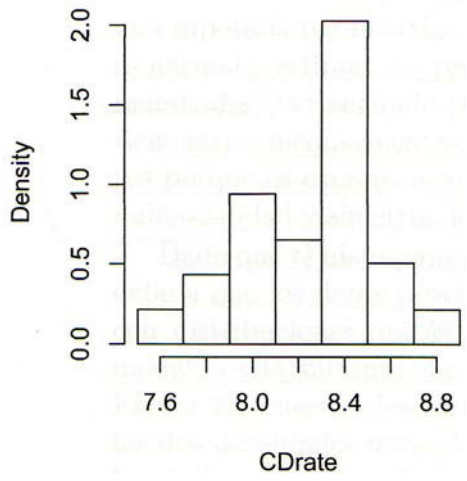
Aplicación

A través de los siguientes ejemplos se mostrará la diferencia y bondades entre la regresión paramétrica y la no paramétrica.

En el Capítulo 1 de Simonoff (1996) se presenta un ejemplo que se refiere a información correspondiente a los intereses que abonan 69 entidades financieras para su producto llamado Certificado de Depósito. La variable en estudio es la CDrate, considerando los bancos y cajas de ahorros.

Con el histograma de los datos observados (Figura B.1), se aprecia bimodalidad, pues se ve que dos barras del histograma tienen más frecuencia. El histograma es un estimador no paramétrico de la función de densidad.

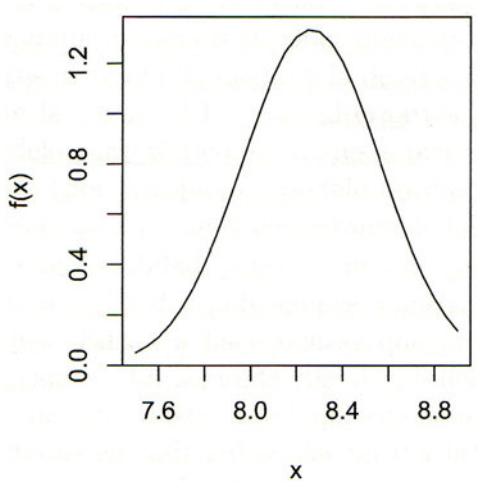
FIGURA Nº B.1
HISTOGRAMA DE LA VARIABLE CDRATE



Pero el histograma representa una función poco suave, característica que no es la que generalmente tiene la función de densidad.

Si se graficara la distribución normal que le corresponde a este conjunto de datos se obtiene la grafica mostrada en la Figura B.2.

FIGURA Nº B.2
CURVA NORMAL DE LA VARIABLE CDRATE

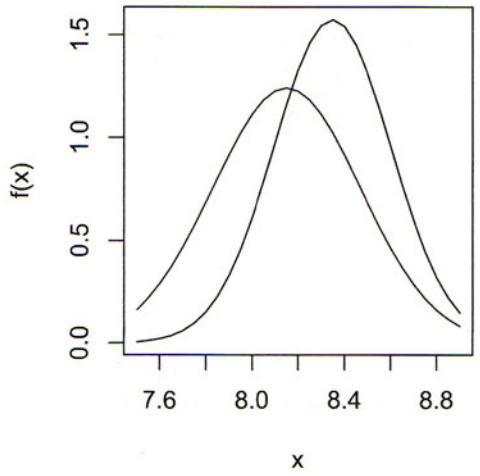


Se observa que cuando se hace una hipótesis paramétrica para obtener un estimador de la densidad de la variable CDrate, se puede suponer que la variable CDrate tiene distribución normal y que sus parámetros μ y σ^2 son estimados puntualmente por la media muestral y desviación estándar muestra, respectivamente.

Esta aplicación nos lleva a resultados no satisfactorios: un modelo paramétrico no se ajusta bien a los datos de la variable en estudio, porque los supuestos del modelo son muy rígidos (el modelo normal impone unimodalidad y simetría, lo que no se observa en el histograma de datos de la variable CDrate).

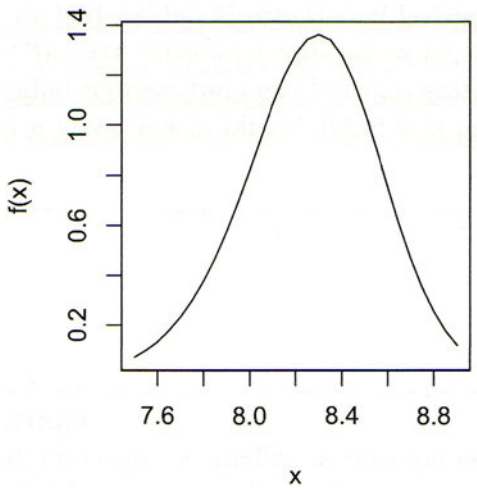
Como en el histograma se observa bimodalidad, esto puede sugerir que los datos provienen de mezclar dos poblaciones: bancos y cajas, con posibles distribuciones diferentes. Ello hace pensar que un posible modelo sería una mixtura de dos distribuciones normales. La Figura B.3 muestra las dos densidades normales ajustadas para cada población.

FIGURA Nº B.3
AJUSTE NORMAL PARA CADA SUBPOBLACIÓN



En la Figura B.4 se observa la mixtura de ambas normales, se corrige la falta de simetría ajustando a una única normal.

FIGURA Nº B.4
MIXTURA DE LAS DOS NORMALES



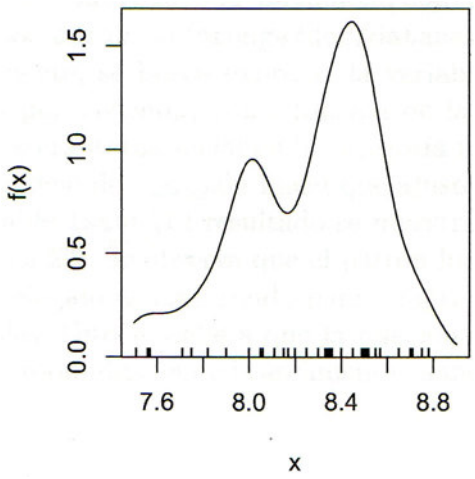
Un estimador no paramétrico alternativo de la densidad asociada al histograma de frecuencias es el estimador núcleo. Glivenko (1934) demostró la convergencia del histograma de frecuencias a la densidad, los estudios de estimación no paramétricos de esta función han sido numerosos durante los años ’80 del siglo pasado. Entre los estimadores no paramétricos de las funciones de densidad están los basados en la definición de una función núcleo o “Kernel” [Rosenblatt (1956), Parzen (1962), Nadaraya (1989)] que son los que más se han estudiado y para los que existe un gran número de

aplicaciones a datos reales. Para ello es necesario elegir tanto el "núcleo" como un valor del parámetro de alisado, ambos conducirán a una función de densidad estimada. El núcleo es una función *K(x),* a partir de la cual se puede establecer el estimador no paramétrico de cualquier función de densidad *f(x)* (Rosenblatt (1956):

$$f_n(x, h_n) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_.}{h_n}\right)$$

donde h_n es el parámetro de alisado y $X_1, \ldots, X_n$ son los datos observados.

FIGURA Nº B.5
AJUSTE NO PARAMÉTRICO



La aplicación de este estimador a datos de la variable CDrate, además de suavizarla, respeta la bimodalidad y asimetría correspondiente, y se logra la gráfica mostrada en la Figura B.5.

Regresión con respuesta continua.

Considerando los datos del Boston Housing Data del año 1978 http://lib.stat.cmu.edu/datasets/boston_corrected.txt, o http://www.ailab.si/orange/doc/datasets/housing.htm), y se analiza la relación entre dos de sus variables correspondientes a las viviendas de 506 barrios de Boston.

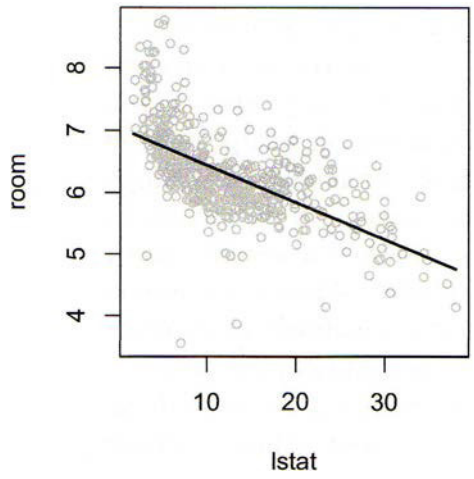
Las variables seleccionadas son:

- *room* (número medio de habitaciones por vivienda)
- *lstat* (porcentaje de población con estatus social en la categoría inferior).

Se utiliza un modelo de regresión lineal para ajustar la variable *room* como función de la variable *lstat*.

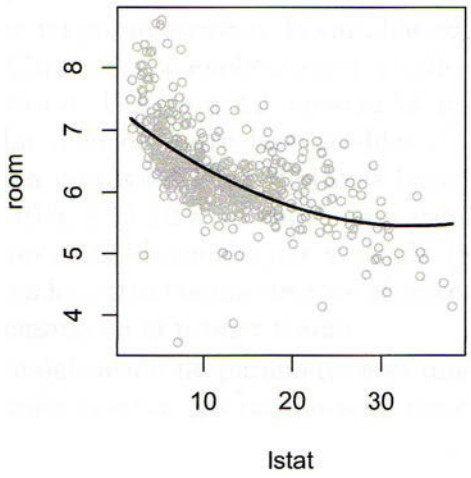
En la Figura B.6 se observa que el ajuste lineal obtenido por el modelo paramétrico es burdo, ya que no se adapta a la tendencia que siguen los pares ordenados correspondientes a las variables estudiadas, lo que estaría indicando que la relación entre las dos variables no es lineal, mientras la variable *room* (número medio de habitaciones por vivienda) desciende bruscamente cuando la variable *lstat* (porcentaje de población con estatus social en la categoría inferior) pasa del 0% al 15%, para el resto del rango de lstat, la relación se mantiene casi lineal.

FIGURA Nº B.6
AJUSTE LINEAL



La solución podría ser considerar una ecuación cuadrática (el cuadrado de l*stat*), el resultado se ve en la figura B.7. Si bien es cierto que se ajusta aproximadamente entre el 5% y 15%, para el resto del rango de lstat (porcentaje de población con estatus social en la categoría inferior) la curva del modelo paramétrico cuadrático no es representativo de la nube de puntos correspondiente, no consiguiendo adaptarse en su gran mayoría a la data.

FIGURA Nº B.7
AJUSTE CUADRÁTICO

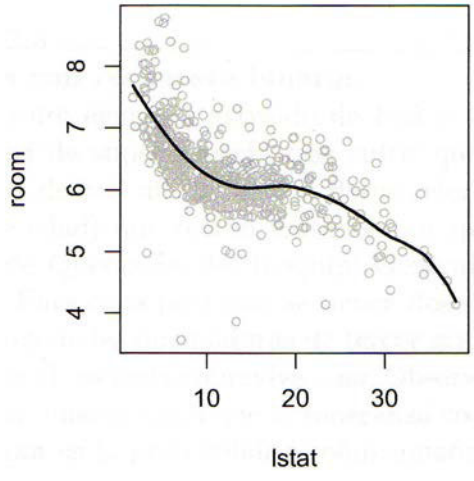


Entonces se procede a realizar un ajuste no paramétrico de la variable *room* (número medio de habitaciones por vivienda) como función de la variable *lstat* (porcentaje de población con estatus social en la categoría inferior).

La figura B.8 muestra la gráfica resultante, se observa que la relación entre las dos variables varía según el porcentaje de población de extracción social baja (*lstat*) sea inferior a 10%, esté entre el 10% y 20%, o supere este valor.

En el tramo intermedio, el número medio de habitaciones por vivienda (*room*) se mantiene casi constante, mientras que en los otros tramos decrece al crecer *lstat* (porcentaje de población con estatus social en la categoría inferior).

FIGURA Nº B.8
AJUSTE NO PARAMÉTRICO



Estos dos ejemplos muestran que la modelización no paramétrica es mucho más flexible que la paramétrica y permite resaltar los patrones de dependencia correspondientes a los datos estudiados.

Dada la problemática observada en la aplicación de los modelos de regresión paramétrica, y conociendo que los polinomios son una gran ayuda en la modelización de las curvas de muchos eventos de la vida cotidiana y de las diferentes disciplinas, pues permiten la funcionalidad de los modelos no paramétricos, es importante estudiar detalladamente las características y bondades de los polinomios en general y del Spline en particular, que siendo un polinomio de tercer grado con algunas propiedades específicas, permite el ajuste adecuado de muchos conjuntos de datos, a los cuáles la modelización paramétrica no permite una solución adecuada, dada la rigidez de los supuestos paramétricos.

Apéndice C

La inflexibilidad de los polinomios

El mayor defecto de los polinomios para propósitos de aproximación es su relativa inflexibilidad.

Por ejemplo, supongamos que queremos aproximar por polinomios la función

$$f(x) = \frac{1}{1+x^2}$$

En el intervalo [-5,5]

Una aproximación natural se logra seleccionando *m* puntos igualmente espaciados y se interpola en esos puntos, esto es:

$$t_i = -5 + 10 * \frac{(i-1)}{m-1}, \quad i = 1, 2, ..., \quad m$$

luego por el teorema de la interpolación de Hermite (teorema 3.6)[15] existe un polinomio único

$$L_m f \quad en \quad P_m \quad que \operatorname{int} erpola \ f \ en los \ puntos \ \{t_i\}_1^m \ ,$$

Al graficar *f* y $L_m f$ para *m* = 5 y *m* = 15 (ver figuras) se observa que el polinomio $L_{15}\ f$ responde mejor en la parte central del intervalo [-5,5] pero desmejora en los extremos.

Ello parece razonable esperar, como quiera que si m se incrementa con finitos puntos de modo que coincidan $L_m f$ *y f* , $L_m f$ *se aproximará a f* en el intervalo

[15] Schumaker, Larry L. Spline functions: Basic Theory. Third Edition, Cambridge University Press, 2007. (600 pág.)

[-5,5]. El teorema de Runge confirma que esto no tendrá una buena aproximación.

Sea

$$f(x) = \frac{1}{1+x^2}$$

En el intervalo [-5,5] y $L_m f$ un polinomio de orden m que interpola f en m en puntos igualmente espaciados como en la ecuación

$$t_i = -5 + 10 * \frac{(i-1)}{m-1}, \quad i = 1, 2, ..., \quad m$$

Entonces

$$\overline{\lim} \left| f(x) - Lmf(x) \right| \to \infty \text{ , } como\ m \to \infty\ para\ |x| > 3{,}64$$

Se puede observar que la no convergencia no es debido a la falta de suavización de f , desde que f es a menudo infinitamente diferenciable sobre $\Re$.

El teorema de Runge muestra que la secuencia de interpolación de polinomios en puntos igualmente espaciados puede no converger.

La convergencia puede no estar garantizada por cualquier secuencia predeterminada de puntos a interpolar.

Apéndice D

La ecuación del equilibrio de William Brass

Como la información correspondiente a las estadísticas vitales de defunciones es incompleta, se cuenta con un método ideado por Brass, llamado ecuación del equilibrio, para poder completar el total de defunciones de un área determinada. Para ello se necesita como insumos:

- Defunciones promedio, por edad y sexo, correspondiente a tres años consecutivos (el promediar tiene como objetivo disminuir las distorsiones anuales en la estructura de muertes), centrándose esta información en el año intermedio.
- Población total por edad y sexo correspondiente al Censo más cercano a la fecha para las que se quiere calcular la estructura de la mortalidad. La información censal es recalculada al 30 de junio del año en estudio.

Este método utilizando el análisis de regresión lineal, compara el número de las defunciones con su correspondiente población censada, por sexo y edad, y la población correspondiente, para de esta manera poder determinar el porcentaje faltante de las defunciones que corresponderían a la población censada del estudio.

La relación lineal a la que llega Brass es la siguiente:

$$\frac{N(x)}{N(x+)} = r + K\frac{D(x+)}{N(x+)}$$

donde: ***N(x),*** número de personas de edad x.

N(x+), número total de personas de x y más años de edad.

r, tasa de crecimiento en una población estable, que es igual para todas las edades.

K, factor de ajuste, pendiente de la línea definida por los puntos [(D(x+)/N(x+),N(x)/N(x+)].

D(x+), número total de defunciones que se producen en personas de x y más años de edad.

"En la práctica los puntos [(D(x+)/N(x+),N(x)/N(x+)] rara vez corresponderán exactamente a una línea recta, y K se obtendrá seleccionando la línea de regresión que mejor se ajuste a los puntos observados. En algunos casos, sin embargo, las desviaciones de esos puntos respecto a una tendencia lineal serán tan pronunciadas que no se justificará el uso de este método de estimación"[16].

"Casi todas las poblaciones humanas han visto aumentar sus probabilidades de supervivencia en las últimas décadas y no pueden, por tanto ser estables. Sin embargo, simulaciones que se han realizado han demostrado que cuando las poblaciones estables se desestabilizan por cambios lentos pero prolongados de la mortalidad [que es lo que está ocurriendo en las últimas décadas], el sesgo introducido en la estimación de K por esa falta de estabilidad es relativamente pequeño. Sólo cuando ocurren cambios abruptos el sesgo adquiere importancia. Como es lógico, las simulaciones mencionadas se han hecho usando datos que, aparte de representar una población no estable, son perfectas en lo demás"[17].

Obtención del total de defunciones por edad y sexo aplicando la ecuación de equilibrio de Brass a las estadísticas vitales de defunciones registradas

Un ejemplo detallado con datos de Perú, para el año 2004

Se promediaron las defunciones por sexo y edad de los 3 años consecutivos (2003, 2004 y 2005), centrándose en el 2004 (año intermedio).

La población del Censo del 2007, por sexo y edad se recalculó al 30 de junio del 2004.

El presente ejercicio será con los datos de la población masculina.

[16] NNUU. Manual X. Técnicas Indirectas de Estimación Demográfica. Estudios de Población, Nº 81. Departamento de Asuntos Económicos y Sociales Internacionales. Nueva York, 1986. Pág. 150.
[17] Ibid, pág. 150.

Paso 1, cálculo de la población en una edad exacta. El valor de N(x), número de personas de edad exacta x, se obtiene sumando el número de personas en dos grupos contiguos de edad y dividiendo por el número de años abarcados por ambos grupos (generalmente 10).

N(5) = N(00 a 09) = (1320875 + 1299732)/10 = 262061
N(10) = N(05 a 14) = (1299732 + 1429348)/10 = 272908
. .
. .
N(50) = N(45 a 54) = (638759 + 533420)/10 = 117218
N(65) = N(60 a 69) = (342439 + 270579)/10 = 61302
. .
. .
N(80) = N(75 a 80 y más) = (159198 + 163960)/10 = 32316

Paso 2, cálculo de la población por encima de una edad exacta. El número de personas de edad x y más [N(x+)] viene dado por la acumulación de la población en cada grupo de edad, empezando por la última fila (de la columna de población censada) y sumando hacia arriba hasta la edad x.

N(50+) = 163960 + 159198 + 209621 + 270579 + 342439 + 417169 + 533420 = 2096386
N(65+) = 163960 + 159198 + 209621 + 270579 = 803358

Paso 3, cálculo del número de defunciones a partir de una edad exacta. El número de defunciones correspondientes a la edad x y más [D(x+)] viene dado por la acumulación de las defunciones en cada grupo de edad, empezando por la última fila (de la columna de defunciones) y sumando hacia arriba hasta la edad x.

D(50+) = 11686 + 5036 + 4482 + 3571 + 2862 + 2353 + 2073 = 32063
D(65+) = 11686 + 5036 + 4482 + 3571 = 24775

CUADRO Nº D.1

PERÚ: POBLACIÓN, DEFUNCIONES, N(x), N(x+) Y D(x+), HOMBRES, 2004.

Edad	Población	Defunciones	N(x)	N(x+)	D(x+)
Total	12952191	48136			
00-04	1320875	4856			48136
05-09	1299732	483	262061	11631316	43280
10-14	1429348	469	272908	10331584	42797
15-19	1305782	826	273513	8902236	42328
20-24	1193943	1290	249973	7596454	41502
25-29	1072135	1486	226608	6402511	40212
30-34	965670	1465	203781	5330376	38726
35-39	861468	1580	182714	4364706	37261
40-44	768093	1736	162956	3503238	35681
45-49	638759	1882	140685	2735145	33945
50-54	533420	2073	117218	2096386	32063
55-59	417169	2353	95059	1562966	29990
60-64	342439	2862	75961	1145797	27637
65-69	270579	3571	61302	803358	24775
70-74	209621	4482	48020	532779	21204
75-79	159198	5036	36882	323158	16722
80 y +	163960	11686	32316	163960	11686

Paso 4, cálculo de los puntos definidos por las tasas parciales de natalidad y mortalidad. Utilizando los valores de N(x), N(x+) y D(x+) obtenidos en los pasos anteriores se calculan los cocientes N(x)/N(x+) y D(x+)/N(x+)

D(50+)/N(50+) = 32063/2096386 = 0.0153

D(65+)/N(65+) = 24775/803358 = 0.0368

N(50)/N(50+) = 117218/2096386 = 0.0559

N(65)/N(65+) = 61302/803358 = 0.0763

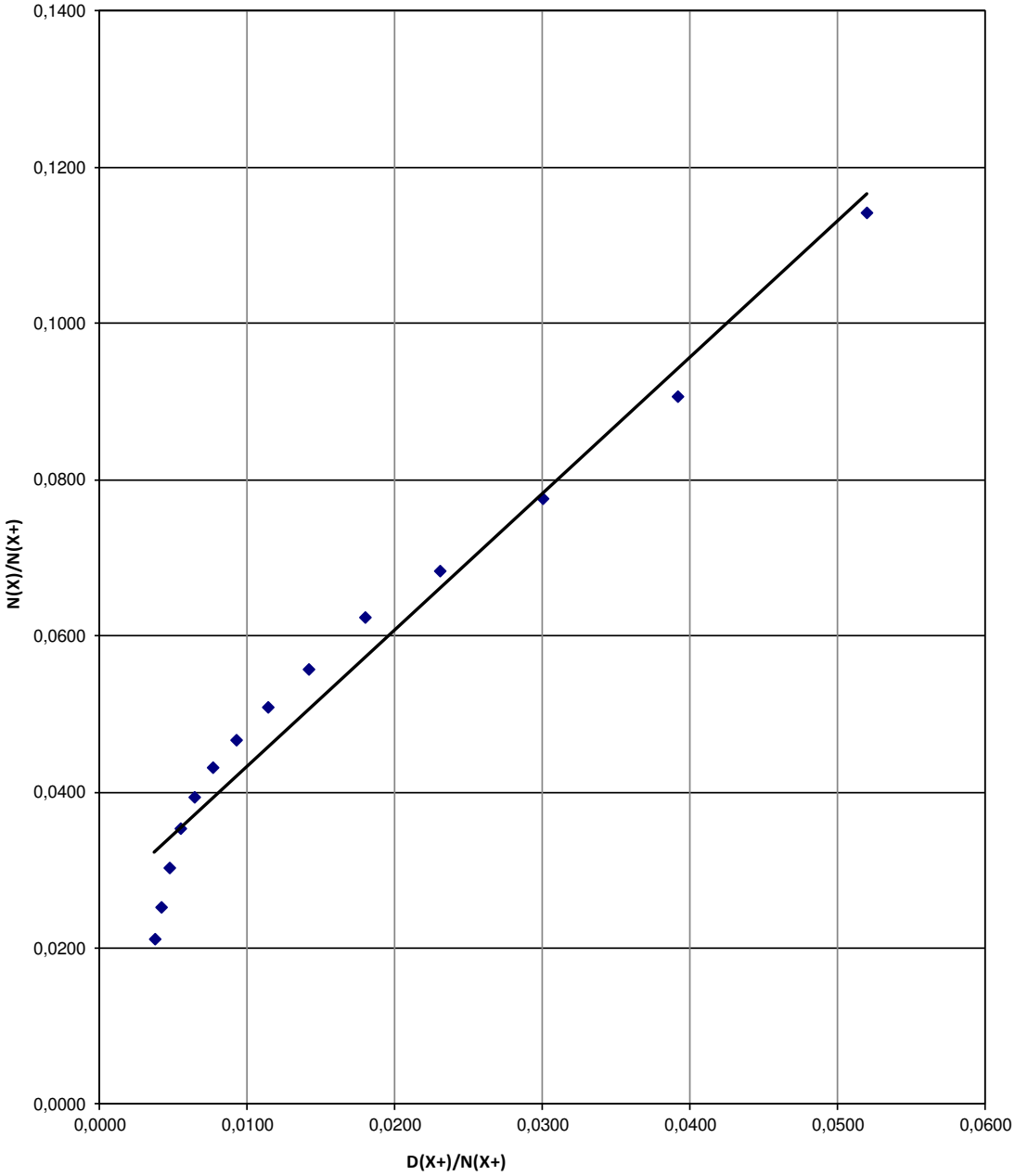
CUADRO Nº D.2

PERÚ: TASAS PARCIALES DE MORTALIDAD Y NATALIDAD, HOMBRES, 2004.

Edad	D(x+)/N(x+)	N(x)/N(x+)
Total		
00-04		
05-09		
10-14	0,0041	0,0264
15-19	0,0048	0,0307
20-24	0,0055	0,0329
25-29	0,0063	0,0354
30-34	0,0073	0,0382
35-39	0,0085	0,0419
40-44	0,0102	0,0465
45-49	0,0124	0,0514
50-54	0,0153	0,0559
55-59	0,0192	0,0608
60-64	0,0241	0,0663
65-69	0,0308	0,0763
70-74	0,0398	0,0901
75-79	0,0517	0,1141
80 y +	0,0713	0,1971

Una vez calculadas las tasas parciales de mortalidad y natalidad, se grafican y se observa su comportamiento lineal, de no escapar demasiado a la tendencia lineal, se procede con los cálculos finales para obtener el porcentaje de omisión de las defunciones.

GRÁFICO Nº A.1.1
DEPARTAMENTO DE LIMA: GRÁFICO DE LAS TASAS PARCIALES DE NATALIDAD, N(X)/N(X+), CONTRA LAS TASAS PARCIALES DE MORTALIDAD, D(X+)/N(X+), PARA HOMBRES, 2007



Paso 5, determinación del porcentaje de omisión de defunciones.

CUADRO Nº D.3
PERÚ: AJUSTE A UNA LÍNEA RECTA,
HOMBRES, 2004.

Edad		D(x+)/N(x+)		N(x)/N(x+)
Grupo 1				
10-14		0,0041		0,0264
15-19		0,0048		0,0307
20-24		0,0055		0,0329
25-29		0,0063		0,0354
30-34		0,0073		0,0382
35-39		0,0085		0,0419
40-44		0,0102		0,0465
Total		0,0466		0,2520
Promedio	**x1 =**	**0,0067**	**y1 =**	**0,0360**
Grupo 2				
45-49		0,0124		0,0514
50-54		0,0153		0,0559
55-59		0,0192		0,0608
60-64		0,0241		0,0663
65-69		0,0308		0,0763
70-74		0,0398		0,0901
75-79		0,0517		0,1141
Total		0,1934		0,5150
Promedio	**x2 =**	**0,0276**	**y2 =**	**0,0736**

Con los datos del Cuadro Nº D.3 se procede a calcular los valores siguientes:

K = $(y_2 - y_1) / (x_2 - x_1)$ = (0.0736 - 0.0360) / (0.027646 - 0.0067) = 1.79128

C = 1 / K = 1 / 1.79128 = 0.5581

r = $y_1 - K\, x_1$ = 0.0360 – 1.79128 * 0.0067 = 0.0241 = 2.41 por ciento

Porcentaje de omisión = ((K – 1) / K) * 100 = ((1.79128 – 1) / 1.7929)) * 100 = 44.19 %.

donde: **K,** factor de ajuste

C, cobertura de registro

r, tasa de crecimiento

Porcentaje de omisión, omisión de las defunciones por ciento.

ANEXOS

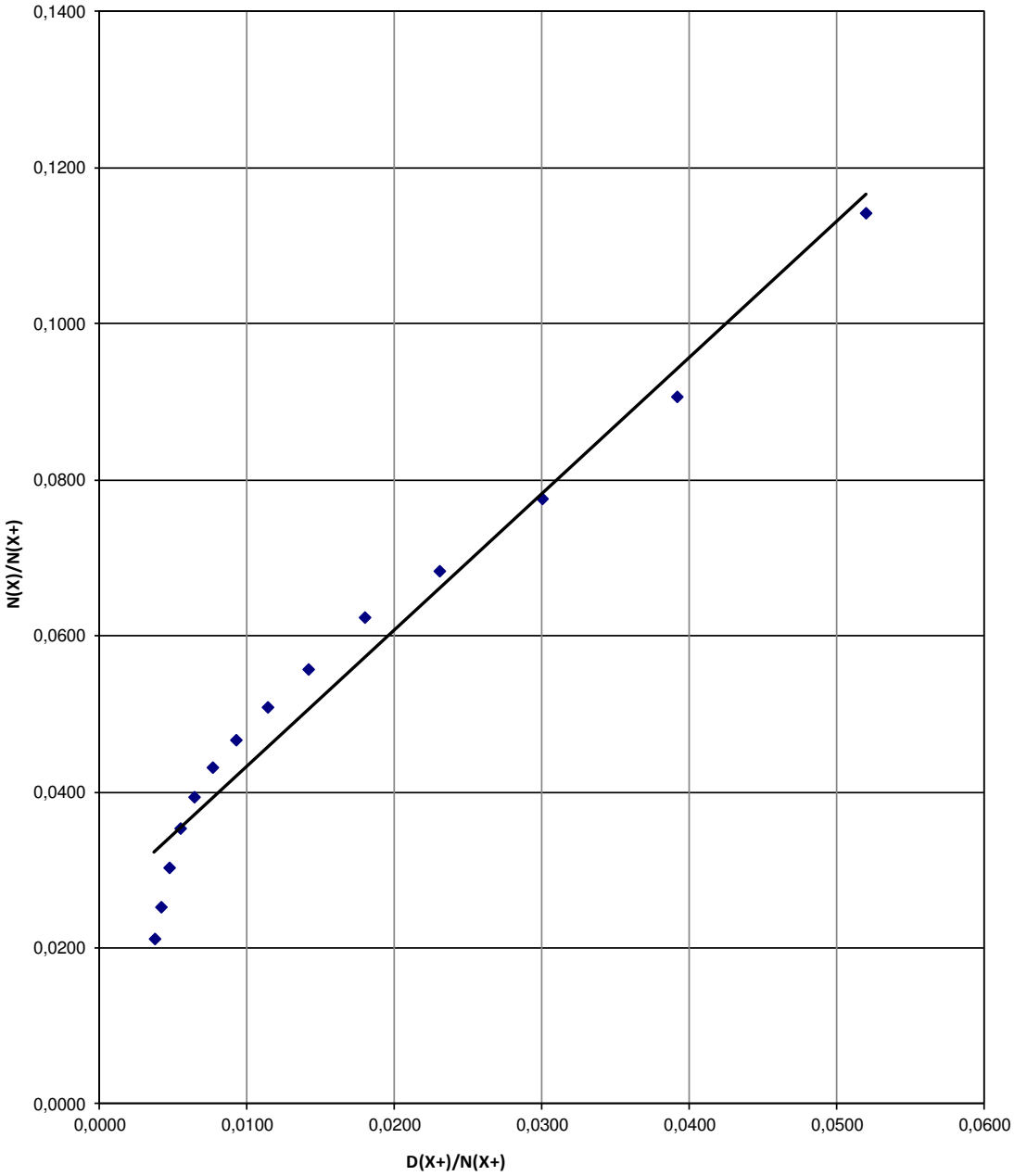
Anexo 1

CUADRO Nº A.1.1
DEPARTAMENTO DE LIMA: APLICACIÓN DEL MÉTODO DE LA ECUACIÓN DE EQUILIBRIO DE BRASS, HOMBRES 2007
(CENSO DE POBLACIÓN 2007 Y ESTADÍSTICAS VITALES 2007)

EDAD	POB. MASCULINA	DEF.EST.VIT. HOMBRES	N(x+)	N(x)	D(x+)	N(x)/N(x+)	D(x+)/N(x+)
TOTAL	4114683	13254					
00-04	362292	665			13254		
05-09	348293	71	3752391	71059	12589		
10-14	375542	76	3404098	72384	12518	0,0213	0,0037
15-19	391795	140	3028556	76734	12442	0,0253	0,0041
20-24	409293	243	2636761	80109	12302	0,0304	0,0047
25-29	379680	317	2227468	78897	12059	0,0354	0,0054
30-34	348958	351	1847788	72864	11742	0,0394	0,0064
35-39	298990	365	1498830	64795	11391	0,0432	0,0076
40-44	261924	394	1199840	56091	11026	0,0467	0,0092
45-49	215973	456	937916	47790	10632	0,0510	0,0113
50-54	186949	591	721943	40292	10176	0,0558	0,0141
55-59	147223	672	534994	33417	9585	0,0625	0,0179
60-64	118032	834	387771	26526	8913	0,0684	0,0230
65-69	91487	1111	269739	20952	8079	0,0777	0,0300
70-74	70279	1365	178252	16177	6968	0,0908	0,0391
75-79	53120	1616	107973	12340	5603	0,1143	0,0519
80 y +	54853	3987	54853	10797	3987	0,1968	0,0727

x1 =	**0,0059**	**Promedio de los 7 primeros D(x+)/N(x+)**
y1 =	**0,0345**	**Promedio de los 7 primeros N(x)/N(x+)**
x2 =	**0,0268**	**Promedio de los 7 siguientes D(x+)/N(x+)**
y2 =	**0,0743**	**Promedio de los 7 siguientes N(x)/N(x+)**
f =	**1,9044**	**(y2-y1)/(x2-x1)**
r =	**0,0234**	**y1-x1*f**
omisión =	**47,49 %**	**(f-1)/f**

GRÁFICO Nº A.1.1
DEPARTAMENTO DE LIMA: GRÁFICO DE LAS TASAS PARCIALES DE NATALIDAD, N(X)/N(X+), CONTRA LAS TASAS PARCIALES DE MORTALIDAD, D(X+)/N(X+), PARA HOMBRES, 2007



CUADRO Nº A.1.2

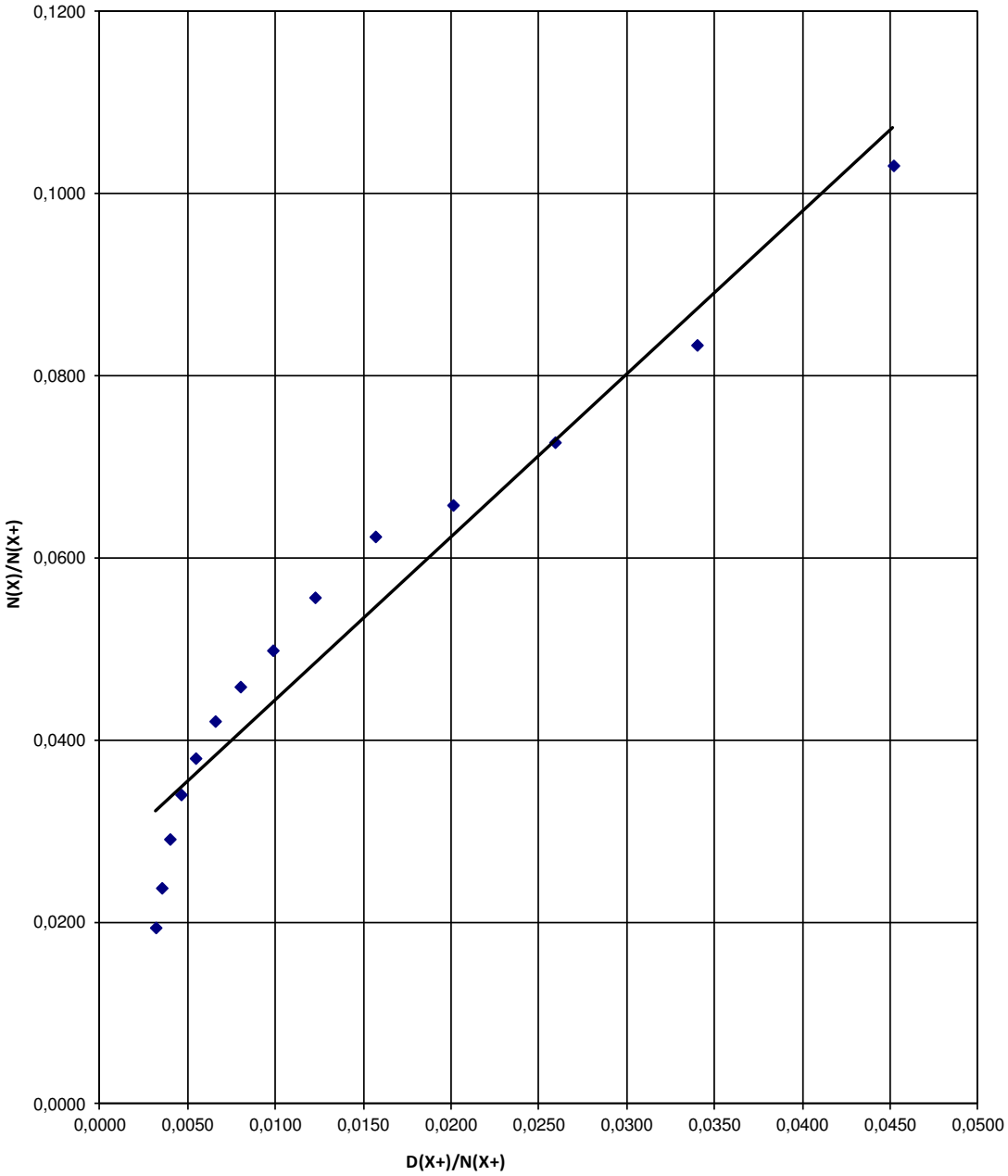
DEPARTAMENTO DE LIMA: APLICACIÓN DEL MÉTODO DE LA ECUACIÓN DE EQUILIBRIO DE BRASS, MUJERES 2007

(CENSO DE POBLACIÓN 2007 Y ESTADÍSTICAS VITALES 2007)

EDAD	POB. FEMENINA	DEF.EST.VIT. MUJERES	N(x+)	N(x)	D(x+)	N(x)/N(x+)	D(x+)/N(x+)
TOTAL	4279517	11929					
00-04	347202	515			11929		
05-09	332911	51	3932315	68011	11414		
10-14	366619	49	3599404	69953	11363	0,0194	0,0032
15-19	402543	100	3232785	76916	11314	0,0238	0,0035
20-24	422692	144	2830242	82524	11214	0,0292	0,0040
25-29	397073	160	2407550	81977	11070	0,0340	0,0046
30-34	367795	166	2010477	76487	10910	0,0380	0,0054
35-39	324028	229	1642682	69182	10744	0,0421	0,0065
40-44	281225	322	1318654	60525	10515	0,0459	0,0080
45-49	236389	400	1037429	51761	10193	0,0499	0,0098
50-54	209866	534	801040	44626	9793	0,0557	0,0122
55-59	159019	580	591174	36889	9259	0,0624	0,0157
60-64	125584	742	432155	28460	8679	0,0659	0,0201
65-69	97453	832	306571	22304	7937	0,0728	0,0259
70-74	77035	1140	209118	17449	7105	0,0834	0,0340
75-79	59227	1391	132083	13626	5965	0,1032	0,0452
80 y +	72856	4574	72856	13208	4574	0,1813	0,0628

x1 =	**0,0050**	**Promedio de los 7 primeros D(x+)/N(x+)**
y1 =	**0,0332**	**Promedio de los 7 primeros N(x)/N(x+)**
x2 =	**0,0233**	**Promedio de los 7 siguientes D(x+)/N(x+)**
y2 =	**0,0705**	**Promedio de los 7 siguientes N(x)/N(x+)**
f =	**2,0422**	**(y2-y1)/(x2-x1)**
r =	**0,0230**	**y1-x1*f**
omisión =	**51,03 %**	**(f-1)/f**

GRÁFICO Nº A.1.2
DEPARTAMENTO DE LIMA: GRÁFICO DE LAS TASAS PARCIALES DE NATALIDAD, N(X)/N(X+), CONTRA LAS TASAS PARCIALES DE MORTALIDAD, D(X+)/N(X+), PARA MUJERES, 2007



Anexo 2

CUADRO Nº A.2.1

DEPARTAMENTO DE LIMA: DEFUNCIONES SIN CORREGIR Y CORREGIDAS CON EL MÉTODO DE BRASS, POR SEXO Y EDAD, 2007

Edad	Defunciones sin corregir			Defunciones corregidas con el método de Brass			Edad	Defunciones sin corregir			Defunciones corregidas con el método de Brass		
	Total	Hombre	Mujer	Total	Hombre	Mujer		Total	Hombre	Mujer	Total	Hombre	Mujer
Total	**25183**	**13254**	**11929**	**49601**	**25241**	**24360**	**46**	160	81	79	315	154	161
							47	161	86	75	317	164	153
0	978	554	424	1921	1055	866	**48**	171	92	79	336	175	161
1	93	48	45	183	91	92	**49**	193	103	90	380	196	184
2	49	24	25	97	46	51	**50**	234	124	110	461	236	225
3	32	20	12	63	38	25	**51**	213	107	106	420	204	216
4	28	19	9	54	36	18	**52**	234	123	111	461	234	227
5	23	13	10	45	25	20	**53**	228	129	99	448	246	202
6	23	14	9	45	27	18	**54**	216	108	108	427	206	221
7	20	11	9	39	21	18	**55**	217	126	91	426	240	186
8	32	16	16	63	30	33	**56**	245	126	119	483	240	243
9	24	17	7	46	32	14	**57**	256	128	128	505	244	261
10	23	15	8	45	29	16	**58**	253	132	121	498	251	247
11	20	11	9	39	21	18	**59**	281	160	121	552	305	247
12	23	14	9	45	27	18	**60**	339	186	153	666	354	312
13	26	16	10	50	30	20	**61**	294	148	146	580	282	298
14	33	20	13	65	38	27	**62**	310	155	155	612	295	317
15	32	19	13	63	36	27	**63**	301	158	143	593	301	292
16	44	28	16	86	53	33	**64**	332	187	145	652	356	296
17	49	27	22	96	51	45	**65**	382	203	179	753	387	366
18	57	30	27	112	57	55	**66**	335	198	137	657	377	280
19	58	36	22	114	69	45	**67**	371	226	145	726	430	296
20	80	46	34	157	88	69	**68**	413	232	181	812	442	370
21	73	46	27	143	88	55	**69**	442	252	190	868	480	388
22	73	49	24	142	93	49	**70**	514	286	228	1011	545	466
23	71	49	22	138	93	45	**71**	464	250	214	913	476	437
24	90	53	37	177	101	76	**72**	500	276	224	983	526	457
25	101	72	29	196	137	59	**73**	496	262	234	977	499	478
26	85	53	32	166	101	65	**74**	531	291	240	1044	554	490
27	97	66	31	189	126	63	**75**	569	312	257	1119	594	525
28	93	64	29	181	122	59	**76**	563	300	263	1108	571	537
29	101	62	39	198	118	80	**77**	580	316	264	1141	602	539
30	105	76	29	204	145	59	**78**	668	361	307	1315	688	627
31	100	67	33	195	128	67	**79**	627	327	300	1236	623	613
32	93	64	29	181	122	59	**80**	661	340	321	1303	647	656
33	105	68	37	205	129	76	**81**	649	344	305	1278	655	623
34	114	76	38	223	145	78	**82**	658	354	304	1296	675	621
35	120	77	43	235	147	88	**83**	608	316	292	1198	602	596
36	112	71	41	219	135	84	**84**	616	303	313	1215	577	638
37	105	67	38	206	128	78	**85**	644	322	322	1270	613	657
38	123	70	53	241	133	108	**86**	583	291	292	1150	554	596
39	134	80	54	262	152	110	**87**	549	270	279	1084	514	570
40	144	84	60	283	160	123	**88**	503	227	276	996	432	564
41	128	68	60	252	129	123	**89**	445	208	237	880	396	484
42	150	80	70	295	152	143	**90**	423	189	234	838	360	478
43	143	75	68	282	143	139	**91**	391	166	225	775	316	459
44	151	87	64	297	166	131	**92**	341	138	203	678	263	415
45	171	94	77	336	179	157	**93**	330	118	212	658	225	433
							94	245	89	156	488	169	319
							95y+	915	312	603	1825	594	1231

Anexo 3

CUADRO Nº A.3.1

DEPARTAMENTO DE LIMA: POBLACIÓN CENSADA SIN CORREGIR Y CORREGIDA CON PORCENTAJE DE OMISIÓN AL 30 DE JUNIO DEL 2007, POR SEXO Y EDAD

Edad	Población censada sin corregir			Población censada corregida			Edad	Población censada sin corregir			Población censada corregida		
	Total	Hombre	Mujer	Total	Hombre	Mujer		Total	Hombre	Mujer	Total	Hombre	Mujer
Total	**8394200**	**4114683**	**4279517**	**8513387**	**4173114**	**4340273**	**46**	87503	41953	45550	88746	42549	46197
							47	100555	48567	51988	101983	49257	52726
0	131850	67254	64596	133722	68209	65513	**48**	89176	42351	46825	90442	42952	47490
1	136255	69847	66408	138190	70839	67351	**49**	80108	38158	41950	81246	38700	42546
2	150353	76314	74039	152488	77398	75090	**50**	91349	42450	48899	92646	43053	49593
3	149289	76172	73117	151409	77254	74155	**51**	67189	31698	35491	68143	32148	35995
4	141747	72705	69042	143759	73737	70022	**52**	84461	40350	44111	85660	40923	44737
5	134917	69035	65882	136832	70015	66817	**53**	79028	37296	41732	80151	37826	42325
6	132580	67611	64969	134462	68571	65891	**54**	74788	35155	39633	75850	35654	40196
7	138979	71191	67788	140953	72202	68751	**55**	66781	31823	34958	67729	32275	35454
8	140254	71644	68610	142245	72661	69584	**56**	62959	30437	32522	63853	30869	32984
9	134474	68812	65662	136383	69789	66594	**57**	66713	31900	34813	67660	32353	35307
10	147207	74878	72329	149297	75941	73356	**58**	57557	27600	29957	58374	27992	30382
11	145820	74101	71719	147890	75153	72737	**59**	52232	25463	26769	52974	25825	27149
12	151235	76726	74509	153382	77815	75567	**60**	63158	29919	33239	64055	30344	33711
13	146325	73869	72456	148403	74918	73485	**61**	40250	19962	20288	40821	20245	20576
14	151574	75968	75606	153727	77047	76680	**62**	50270	24480	25790	50984	24828	26156
15	155500	76595	78905	157708	77683	80025	**63**	46907	22654	24253	47573	22976	24597
16	144880	70859	74021	146937	71865	75072	**64**	43031	21017	22014	43642	21315	22327
17	153320	75975	77345	155497	77054	78443	**65**	48393	23363	25030	49080	23695	25385
18	167998	83738	84260	170383	84927	85456	**66**	34676	16904	17772	35168	17144	18024
19	172640	84628	88012	175091	85830	89261	**67**	41412	19899	21513	42000	20182	21818
20	176950	86919	90031	179461	88153	91308	**68**	34384	16185	18199	34872	16415	18457
21	154390	75683	78707	156583	76758	79825	**69**	30075	15136	14939	30502	15351	15151
22	166918	82941	83977	169287	84119	85168	**70**	37865	17828	20037	38403	18081	20322
23	165511	80958	84553	167861	82108	85753	**71**	24270	12122	12148	24614	12294	12320
24	168216	82792	85424	170604	83968	86636	**72**	31141	14791	16350	31583	15001	16582
25	166669	81553	85116	169036	82711	86325	**73**	27388	13180	14208	27777	13367	14410
26	152951	74280	78671	155123	75335	79788	**74**	26650	12358	14292	27028	12533	14495
27	162798	80526	82272	165109	81669	83440	**75**	28497	13255	15242	28901	13443	15458
28	151093	74103	76990	153238	75155	78083	**76**	22000	10591	11409	22312	10741	11571
29	143242	69218	74024	145276	70201	75075	**77**	23129	10978	12151	23458	11134	12324
30	165582	80612	84970	167933	81757	86176	**78**	21680	10116	11564	21988	10260	11728
31	132476	64010	68466	134357	64919	69438	**79**	17041	8180	8861	17283	8296	8987
32	146661	71091	75570	148743	72100	76643	**80**	20215	8911	11304	20503	9038	11465
33	142588	70220	72368	144613	71217	73396	**81**	12242	5766	6476	12416	5848	6568
34	129446	63025	66421	131284	63920	67364	**82**	14019	6335	7684	14218	6425	7793
35	131168	63200	67968	133030	64097	68933	**83**	11569	5186	6383	11734	5260	6474
36	121274	57972	63302	122996	58795	64201	**84**	10995	4864	6131	11151	4933	6218
37	133172	63988	69184	135063	64897	70166	**85**	10709	4486	6223	10861	4550	6311
38	122493	58506	63987	124233	59337	64896	**86**	8678	3801	4877	8801	3855	4946
39	114911	55324	59587	116543	56110	60433	**87**	8420	3492	4928	8540	3542	4998
40	128103	61259	66844	129922	62129	67793	**88**	5609	2423	3186	5688	2457	3231
41	95682	46219	49463	97040	46875	50165	**89**	4912	2050	2862	4982	2079	2903
42	120341	58440	61901	122050	59270	62780	**90**	4592	1724	2868	4657	1748	2909
43	103231	49412	53819	104697	50114	54583	**91**	2461	1027	1434	2496	1042	1454
44	95792	46594	49198	97153	47256	49897	**92**	2795	1055	1740	2835	1070	1765
45	95020	44944	50076	96369	45582	50787	**93**	2330	844	1486	2363	856	1507
							94	1876	704	1172	1903	714	1189
							95y+	6287	2185	4102	6376	2216	4160

Anexo 4

CUADRO Nº A.4.1
DEPARTAMENTO DE LIMA: TASA DE MORTALIDAD SIN SUAVIZAR POR SEXO Y EDAD, 2007

Edad	Hombre	Mujer	Edad	Hombre	Mujer
0	0,0155	0,0132	**46**	0,0036	0,0035
1	0,0013	0,0014	**47**	0,0033	0,0029
2	0,0006	0,0007	**48**	0,0041	0,0034
3	0,0005	0,0003	**49**	0,0051	0,0043
4	0,0005	0,0003	**50**	0,0055	0,0045
5	0,0004	0,0003	**51**	0,0063	0,0060
6	0,0004	0,0003	**52**	0,0057	0,0051
7	0,0003	0,0003	**53**	0,0065	0,0048
8	0,0004	0,0005	**54**	0,0058	0,0055
9	0,0005	0,0002	**55**	0,0074	0,0052
10	0,0004	0,0002	**56**	0,0078	0,0074
11	0,0003	0,0002	**57**	0,0075	0,0074
12	0,0003	0,0002	**58**	0,0090	0,0081
13	0,0004	0,0003	**59**	0,0118	0,0091
14	0,0005	0,0004	**60**	0,0117	0,0093
15	0,0005	0,0003	**61**	0,0139	0,0145
16	0,0007	0,0004	**62**	0,0119	0,0121
17	0,0007	0,0006	**63**	0,0131	0,0119
18	0,0007	0,0006	**64**	0,0167	0,0133
19	0,0008	0,0005	**65**	0,0163	0,0144
20	0,0010	0,0008	**66**	0,0220	0,0155
21	0,0011	0,0007	**67**	0,0213	0,0136
22	0,0011	0,0006	**68**	0,0269	0,0200
23	0,0011	0,0005	**69**	0,0313	0,0256
24	0,0012	0,0009	**70**	0,0301	0,0229
25	0,0017	0,0007	**71**	0,0387	0,0355
26	0,0013	0,0008	**72**	0,0351	0,0276
27	0,0015	0,0008	**73**	0,0373	0,0332
28	0,0016	0,0008	**74**	0,0442	0,0338
29	0,0017	0,0011	**75**	0,0442	0,0340
30	0,0018	0,0007	**76**	0,0532	0,0464
31	0,0020	0,0010	**77**	0,0541	0,0437
32	0,0017	0,0008	**78**	0,0671	0,0535
33	0,0018	0,0010	**79**	0,0751	0,0682
34	0,0023	0,0012	**80**	0,0716	0,0572
35	0,0023	0,0013	**81**	0,1120	0,0949
36	0,0023	0,0013	**82**	0,1051	0,0797
37	0,0020	0,0011	**83**	0,1144	0,0921
38	0,0022	0,0017	**84**	0,1170	0,1026
39	0,0027	0,0018	**85**	0,1347	0,1041
40	0,0026	0,0018	**86**	0,1437	0,1205
41	0,0028	0,0025	**87**	0,1451	0,1140
42	0,0026	0,0023	**88**	0,1758	0,1746
43	0,0029	0,0025	**89**	0,1905	0,1667
44	0,0035	0,0026	**90**	0,2059	0,1643
45	0,0039	0,0031	**91**	0,3033	0,3157
			92	0,2458	0,2351
			93	0,2629	0,2873
			94	0,2367	0,2683
			95y+	0,2681	0,2959

Anexo 5

CUADRO Nº A.5.1
DEPARTAMENTO DE LIMA: LN DE LAS TASAS DE MORTALIDAD SIN SUAVIZAR POR SEXO Y EDAD, 2007

Edad	Hombre	Mujer	Edad	Hombre	Mujer
0	-4,1690	-4,3261	**46**	-5,6215	-5,6593
1	-6,6573	-6,5959	**47**	-5,7049	-5,8424
2	-7,4281	-7,2946	**48**	-5,5031	-5,6869
3	-7,6173	-7,9950	**49**	-5,2855	-5,4434
4	-7,6247	-8,2662	**50**	-5,2064	-5,3955
5	-7,9376	-8,1140	**51**	-5,0600	-5,1159
6	-7,8398	-8,2054	**52**	-5,1641	-5,2836
7	-8,1427	-8,2479	**53**	-5,0354	-5,3449
8	-7,7924	-7,6538	**54**	-5,1537	-5,2034
9	-7,6875	-8,4673	**55**	-4,9014	-5,2502
10	-7,8704	-8,4305	**56**	-4,8569	-4,9107
11	-8,1828	-8,3042	**57**	-4,8873	-4,9073
12	-7,9663	-8,3424	**58**	-4,7142	-4,8122
13	-7,8230	-8,2091	**59**	-4,4388	-4,6997
14	-7,6146	-7,9516	**60**	-4,4511	-4,6826
15	-7,6769	-7,9943	**61**	-4,2738	-4,2348
16	-7,2123	-7,7297	**62**	-4,4328	-4,4129
17	-7,3204	-7,4635	**63**	-4,3351	-4,4336
18	-7,3065	-7,3484	**64**	-4,0922	-4,3232
19	-7,1260	-7,5927	**65**	-4,1146	-4,2393
20	-6,9095	-7,1879	**66**	-3,8172	-4,1647
21	-6,7711	-7,2803	**67**	-3,8488	-4,3001
22	-6,8074	-7,4606	**68**	-3,6146	-3,9097
23	-6,7832	-7,5526	**69**	-3,4651	-3,6648
24	-6,7231	-7,0387	**70**	-3,5018	-3,7753
25	-6,4031	-7,2883	**71**	-3,2514	-3,3390
26	-6,6146	-7,1127	**72**	-3,3506	-3,5914
27	-6,4741	-7,1887	**73**	-3,2879	-3,4061
28	-6,4233	-7,1880	**74**	-3,1190	-3,3872
29	-6,3884	-6,8442	**75**	-3,1193	-3,3825
30	-6,3348	-7,2866	**76**	-2,9344	-3,0703
31	-6,2289	-6,9435	**77**	-2,9175	-3,1296
32	-6,3818	-7,1694	**78**	-2,7022	-2,9288
33	-6,3137	-6,8729	**79**	-2,5890	-2,6852
34	-6,0887	-6,7612	**80**	-2,6368	-2,8609
35	-6,0777	-6,6636	**81**	-2,1892	-2,3554
36	-6,0765	-6,6390	**82**	-2,2532	-2,5297
37	-6,2285	-6,8019	**83**	-2,1676	-2,3853
38	-6,1006	-6,3984	**84**	-2,1459	-2,2769
39	-5,9112	-6,3088	**85**	-2,0045	-2,2624
40	-5,9618	-6,3120	**86**	-1,9400	-2,1161
41	-5,8954	-6,0109	**87**	-1,9302	-2,1712
42	-5,9660	-6,0845	**88**	-1,7383	-1,7455
43	-5,8592	-5,9730	**89**	-1,6582	-1,7914
44	-5,6513	-5,9425	**90**	-1,5801	-1,8060
45	-5,5399	-5,7791	**91**	-1,1932	-1,1530
			92	-1,4033	-1,4476
			93	-1,3362	-1,2471
			94	-1,4410	-1,3157
			95y+	-1,3166	-1,2177

Anexo 6

CUADRO Nº A.6.1

DEPARTAMENTO DE LIMA: COEFICIENTES DE LOS POLINOMIOS POR SECCIONES SPLINE, POR EDAD, HOMBRES, 2007

Edad	c1	c2	c3	c4	Edad	c1	c2	c3	c4
0	-5,8233	-0,4725	0,0000	0,0055	**46**	-5,5943	0,0708	0,0017	0,0002
1	-6,2903	-0,4560	0,0165	0,0043	**47**	-5,5216	0,0749	0,0023	-0,0004
2	-6,7255	-0,4100	0,0294	0,0019	**48**	-5,4448	0,0783	0,0011	-0,0006
3	-7,1041	-0,3453	0,0353	0,0002	**49**	-5,3660	0,0787	-0,0007	-0,0003
4	-7,4140	-0,2741	0,0360	-0,0005	**50**	-5,2884	0,0762	-0,0017	-0,0001
5	-7,6525	-0,2035	0,0346	-0,0014	**51**	-5,2139	0,0726	-0,0019	0,0004
6	-7,8229	-0,1386	0,0303	-0,0015	**52**	-5,1428	0,0701	-0,0006	0,0004
7	-7,9326	-0,0823	0,0259	-0,0022	**53**	-5,0729	0,0701	0,0006	0,0005
8	-7,9911	-0,0369	0,0194	-0,0015	**54**	-5,0018	0,0727	0,0021	0,0000
9	-8,0102	-0,0026	0,0149	-0,0004	**55**	-4,9270	0,0769	0,0021	0,0001
10	-7,9983	0,0259	0,0136	0,0000	**56**	-4,8480	0,0812	0,0023	0,0001
11	-7,9589	0,0530	0,0136	-0,0008	**57**	-4,7644	0,0860	0,0025	-0,0004
12	-7,8930	0,0779	0,0113	-0,0010	**58**	-4,6762	0,0899	0,0014	-0,0005
13	-7,8047	0,0976	0,0083	-0,0011	**59**	-4,5854	0,0913	0,0000	0,0000
14	-7,6999	0,1111	0,0052	-0,0008	**60**	-4,4942	0,0912	0,0000	0,0001
15	-7,5844	0,1191	0,0028	-0,0011	**61**	-4,4029	0,0916	0,0004	0,0006
16	-7,4635	0,1215	-0,0004	-0,0002	**62**	-4,3103	0,0942	0,0022	0,0002
17	-7,3427	0,1200	-0,0011	-0,0002	**63**	-4,2138	0,0990	0,0027	-0,0002
18	-7,2240	0,1172	-0,0017	-0,0004	**64**	-4,1123	0,1037	0,0020	-0,0002
19	-7,1090	0,1125	-0,0030	-0,0005	**65**	-4,0069	0,1071	0,0015	-0,0005
20	-6,9999	0,1051	-0,0045	-0,0002	**66**	-3,8988	0,1085	-0,0001	-0,0003
21	-6,8996	0,0955	-0,0051	0,0002	**67**	-3,7907	0,1075	-0,0009	-0,0004
22	-6,8090	0,0860	-0,0044	0,0002	**68**	-3,6844	0,1045	-0,0022	-0,0002
23	-6,7271	0,0779	-0,0037	0,0000	**69**	-3,5823	0,0995	-0,0028	0,0002
24	-6,6529	0,0706	-0,0036	-0,0002	**70**	-3,4855	0,0943	-0,0023	0,0001
25	-6,5861	0,0629	-0,0041	0,0004	**71**	-3,3934	0,0901	-0,0019	0,0006
26	-6,5269	0,0559	-0,0029	0,0001	**72**	-3,3046	0,0880	-0,0002	0,0004
27	-6,4737	0,0506	-0,0025	0,0001	**73**	-3,2163	0,0890	0,0012	0,0002
28	-6,4254	0,0460	-0,0021	0,0001	**74**	-3,1259	0,0920	0,0018	0,0002
29	-6,3813	0,0423	-0,0017	0,0001	**75**	-3,0319	0,0962	0,0025	-0,0001
30	-6,3406	0,0393	-0,0013	0,0001	**76**	-2,9333	0,1010	0,0023	-0,0001
31	-6,3025	0,0370	-0,0009	0,0004	**77**	-2,8301	0,1053	0,0021	-0,0004
32	-6,2661	0,0362	0,0002	0,0000	**78**	-2,7231	0,1084	0,0010	-0,0003
33	-6,2297	0,0366	0,0002	-0,0003	**79**	-2,6140	0,1095	0,0001	-0,0002
34	-6,1932	0,0361	-0,0007	0,0001	**80**	-2,5047	0,1091	-0,0005	-0,0006
35	-6,1578	0,0348	-0,0005	0,0003	**81**	-2,3967	0,1062	-0,0024	0,0000
36	-6,1231	0,0348	0,0005	0,0005	**82**	-2,2930	0,1014	-0,0023	0,0002
37	-6,0874	0,0371	0,0019	0,0000	**83**	-2,1937	0,0973	-0,0018	0,0003
38	-6,0484	0,0410	0,0019	-0,0002	**84**	-2,0979	0,0946	-0,0010	0,0001
39	-6,0056	0,0444	0,0014	0,0002	**85**	-2,0042	0,0929	-0,0007	0,0001
40	-5,9597	0,0477	0,0019	0,0001	**86**	-1,9119	0,0919	-0,0004	0,0000
41	-5,9099	0,0519	0,0023	0,0002	**87**	-1,8204	0,0912	-0,0003	-0,0004
42	-5,8554	0,0572	0,0029	-0,0002	**88**	-1,7299	0,0894	-0,0014	-0,0004
43	-5,7955	0,0625	0,0024	-0,0004	**89**	-1,6423	0,0854	-0,0026	-0,0004
44	-5,7310	0,0661	0,0012	-0,0001	**90**	-1,5599	0,0790	-0,0039	-0,0005
45	-5,6637	0,0682	0,0009	0,0003	**91**	-1,4853	0,0697	-0,0054	0,0005
					92	-1,4205	0,0603	-0,0040	0,0005
					93	-1,3636	0,0539	-0,0024	0,0006
					94	-1,3115	0,0510	-0,0006	0,0002
					95y+	-1,2609	0,0000	0,0000	0,0000

CUADRO Nº A.6.2

DEPARTAMENTO DE LIMA: COEFICIENTES DE LOS POLINOMIOS POR SECCIONES SPLINE, POR EDAD, MUJERES, 2007

Edad	c1	c2	c3	c4	Edad	c1	c2	c3	c4
0	-5,8906	-0,5152	0,0000	0,0052	**46**	-5,7449	0,0784	-0,0008	0,0005
1	-6,4005	-0,4996	0,0156	0,0046	**47**	-5,6669	0,0782	0,0006	-0,0001
2	-6,8799	-0,4546	0,0293	0,0032	**48**	-5,5882	0,0791	0,0002	-0,0005
3	-7,3019	-0,3864	0,0389	0,0009	**49**	-5,5094	0,0781	-0,0011	-0,0002
4	-7,6485	-0,3060	0,0415	-0,0012	**50**	-5,4326	0,0752	-0,0018	-0,0001
5	-7,9142	-0,2266	0,0379	-0,0019	**51**	-5,3594	0,0711	-0,0022	0,0007
6	-8,1047	-0,1563	0,0324	-0,0022	**52**	-5,2898	0,0688	-0,0001	0,0007
7	-8,2308	-0,0981	0,0258	-0,0022	**53**	-5,2203	0,0708	0,0021	0,0003
8	-8,3054	-0,0533	0,0191	-0,0001	**54**	-5,1471	0,0759	0,0030	0,0001
9	-8,3397	-0,0154	0,0188	-0,0005	**55**	-5,0681	0,0822	0,0033	-0,0005
10	-8,3367	0,0208	0,0173	-0,0008	**56**	-4,9831	0,0874	0,0019	-0,0002
11	-8,2993	0,0531	0,0149	-0,0008	**57**	-4,8940	0,0904	0,0011	-0,0003
12	-8,2322	0,0804	0,0124	-0,0012	**58**	-4,8028	0,0918	0,0002	-0,0003
13	-8,1406	0,1016	0,0088	-0,0014	**59**	-4,7111	0,0913	-0,0007	-0,0003
14	-8,0316	0,1150	0,0046	-0,0012	**60**	-4,6209	0,0889	-0,0016	-0,0005
15	-7,9132	0,1206	0,0011	-0,0014	**61**	-4,5341	0,0843	-0,0031	0,0005
16	-7,7930	0,1185	-0,0032	-0,0012	**62**	-4,4524	0,0797	-0,0016	0,0006
17	-7,6789	0,1084	-0,0069	-0,0005	**63**	-4,3736	0,0785	0,0004	0,0004
18	-7,5778	0,0933	-0,0083	0,0003	**64**	-4,2944	0,0805	0,0017	0,0003
19	-7,4926	0,0774	-0,0075	-0,0001	**65**	-4,2119	0,0849	0,0027	0,0002
20	-7,4229	0,0621	-0,0077	0,0007	**66**	-4,1241	0,0910	0,0034	0,0001
21	-7,3678	0,0487	-0,0056	0,0010	**67**	-4,0295	0,0982	0,0038	-0,0008
22	-7,3237	0,0405	-0,0026	0,0006	**68**	-3,9283	0,1034	0,0014	-0,0007
23	-7,2852	0,0370	-0,0009	-0,0003	**69**	-3,8242	0,1041	-0,0008	-0,0002
24	-7,2494	0,0342	-0,0019	0,0004	**70**	-3,7210	0,1020	-0,0013	-0,0004
25	-7,2168	0,0314	-0,0008	0,0001	**71**	-3,6208	0,0982	-0,0025	0,0006
26	-7,1861	0,0301	-0,0005	0,0004	**72**	-3,5245	0,0949	-0,0008	0,0003
27	-7,1561	0,0302	0,0006	0,0003	**73**	-3,4300	0,0944	0,0002	0,0004
28	-7,1250	0,0323	0,0014	0,0001	**74**	-3,3350	0,0961	0,0015	0,0002
29	-7,0912	0,0353	0,0016	0,0009	**75**	-3,2371	0,0999	0,0022	-0,0002
30	-7,0535	0,0410	0,0042	0,0001	**76**	-3,1352	0,1037	0,0015	0,0000
31	-7,0082	0,0497	0,0045	0,0003	**77**	-3,0300	0,1067	0,0015	-0,0004
32	-6,9538	0,0596	0,0054	-0,0004	**78**	-2,9223	0,1085	0,0004	-0,0004
33	-6,8892	0,0692	0,0042	-0,0004	**79**	-2,8137	0,1082	-0,0007	0,0001
34	-6,8161	0,0765	0,0031	-0,0002	**80**	-2,7062	0,1069	-0,0006	-0,0005
35	-6,7366	0,0823	0,0026	0,0001	**81**	-2,6003	0,1044	-0,0019	0,0004
36	-6,6516	0,0878	0,0029	0,0001	**82**	-2,4975	0,1016	-0,0009	0,0002
37	-6,5609	0,0939	0,0032	-0,0007	**83**	-2,3966	0,1005	-0,0001	0,0003
38	-6,4645	0,0982	0,0012	-0,0005	**84**	-2,2959	0,1011	0,0007	0,0003
39	-6,3656	0,0991	-0,0002	-0,0003	**85**	-2,1937	0,1036	0,0017	0,0001
40	-6,2670	0,0978	-0,0011	-0,0004	**86**	-2,0883	0,1074	0,0021	0,0000
41	-6,1707	0,0944	-0,0024	0,0001	**87**	-1,9787	0,1117	0,0022	-0,0006
42	-6,0785	0,0900	-0,0020	0,0001	**88**	-1,8655	0,1142	0,0003	-0,0002
43	-5,9905	0,0862	-0,0018	0,0001	**89**	-1,7511	0,1143	-0,0003	-0,0003
44	-5,9060	0,0830	-0,0013	0,0000	**90**	-1,6375	0,1126	-0,0014	-0,0009
45	-5,8242	0,0804	-0,0013	0,0002	**91**	-1,5272	0,1071	-0,0041	0,0003
					92	-1,4238	0,1000	-0,0031	0,0003
					93	-1,3267	0,0946	-0,0023	0,0005
					94	-1,2339	0,0915	-0,0007	0,0002
					95y+	-1,1429	0,0000	0,0000	0,0000

REFERENCIAS BIBLIOGRÁFICAS

Asís López, E.H. (2010). *Métodos numéricos con MatLab*. Lima: Fondo Editorial Universidad de Ciencias y Humanidades.

Burden, R.L., Faires, J.D. (1985). *Análisis numérico*. México: Grupo Editorial Iberoamérica.

Daniel, W.W. (1974). *Applied nonparametric statistics*. Boston: Houghton Mifflin Company.

Eubank, R.L. (1999). *Nonparametric regression and Spline smoothing* (2th ed). New York: Marcel Dekker, Inc.

Harrison, D. & Rubimfeld, D.L. (1978). *Boston Housing Data*. Hedonic prices and the demand forclean air. Boston: Orange. Retrieved from http://orange.biolab.si/doc/datasets/housing.htm

INEI. Perú (2010). *Situación y perspectivas de la mortalidad por sexo y grupos de edad, nacional y por departamentos, 1990-2025. (metodología y tablas de mortalidad).* Lima: Imprenta INEI.

Ipanaqué, R., Urbina, R.T. y Correa, S.B. (1998). *B-Splines con Mathematica 5.1*. Piura: Universidad de Piura.

Montgomery, D.C., Peck, E.A. y Vining, G.G (2004). *Introducción al análisis de regresión lineal* (primera reimpresión). México: Compañía Editorial Continental.

Ortega, A. (1987). *Tablas de mortalidad*. San José, Costa Rica: Centro Latinoamericano de Demografía (CELADE).

Schumaker, L.L. (2007). *Spline functions: basic theory* (3th ed.). Cambridge: Cambridge University Press, Third edition, 2007.

Siegel, J.S. & Swanson, D.A. (2008). *The methods and materials of demography* (2th ed.). Bingley, United Kingdom: Emerald Group Publishing Ltd.

Simonoff, J.S.(1996). *Smoothing methods in statistics*. New York: Springer.

Wang, Y. (2011). *Smoothing Splines, methods and applications.* Boca Raton: Taylor & Francis Group, LLC.