UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS
DEPARTAMENTO DE ESTADÍSTICA

TESIS

COEFICIENTE DE DETERMINACIÓN Y ANÁLISIS DE VARIANZA, PARA REGRESIÓN POLINOMIAL LOCAL, EN MUESTREO DE POBLACIONES FINITAS

PRESENTA

LMA Luis Alejandro Escobar López

PARA OBTENER EL TÍTULO DE MAESTRÍA EN CIENCIAS EXACTAS, SISTEMAS Y DE LA INFORMACIÓN, EN ÁREA ESTADÍSTICA

TUTOR

Dr. José Elías Rodríguez Muñoz

COMITÉ TUTORAL

M. en C. José de Jesús Ruíz Gallegos
M. en C. María del Carmen Montoya Landeros

AGUASCALIENTES, AGS.,SEPTIEMBRE DEL 2011

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

Centro de Ciencias Básicas

L.M.A. LUIS ALEJANDRO ESCOBAR LÓPEZ
ALUMNO (A) DE LA MAESTRÍA EN CIENCIAS
EXACTAS, SISTEMAS Y DE LA INFORMACIÓN,
P R E S E N T E .

Estimado (a) alumno (a) Escobar:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o caso práctico titulado: **"COEFICIENTE DE DETERMINACIÓN Y ANÁLISIS DE VARIANZA PARA REGRESIÓN POLINOMIAL LOCAL EN MUESTREO DE POBLACIONES FINITAS",** hago de su conocimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., 9 de septiembre de 2011
"SE LUMEN PROFERRE"
LA DECANO

M. en C. MARTHA CRISTINA GONZÁLEZ DÍAZ

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES
CENTRO DE CIENCIAS BÁSICAS

c.c.p.- Archivo
MCGD,mjda

M. EN C. MARTHA CRISTINA GONZÁLEZ DÍAZ

DECANA DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE.

Por este medio le comunico que he revisado la versión final del escrito de tesis "**Coeficiente de determinación y análisis de varianza para regresión polinomial local en muestreo de poblaciones finitas**" presentada por el alumno **L.M.A. Luis Alejandro Escobar López** como requisito para obtener el grado de Maestro en Ciencias Exactas, Sistemas y de la Información en el área de Estadística. A mi juicio, este documento contiene ya todas las correcciones solicitadas al hacer la revisión del mismo. Por lo anterior, considero que se puede proceder ya a su impresión definitiva y a la programación del examen de grado.

ATENTAMENTE
AGUASCALIENTES, AGS., 9 DE SEPTIEMBRE DE 2011

DR. JOSÉ ELÍAS RODRÍGUEZ MUÑOZ
DIRECTOR DE TESIS Y MIEMBRO DEL COMITÉ TUTORAL
SINODAL DESIGNADO PARA EL EXAMEN DE GRADO

M. EN C. MARTHA CRISTINA GONZÁLEZ DÍAZ

DECANA DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por este medio le comunico que hemos revisado la versión final del escrito de tesis "**Coeficiente de determinación y análisis de varianza para regresión polinomial local en muestreo de poblaciones finitas**" presentada por el alumno **L.M.A. Luis Alejandro Escobar López** como requisito para obtener el grado de Maestro en Ciencias Exactas, Sistemas y de la Información en el área de Estadística. A nuestro juicio, este documento contiene ya todas las correcciones solicitadas al hacer la revisión del mismo. Por lo anterior, consideramos que se puede proceder ya a su impresión definitiva y a la programación del examen de grado.

ATENTAMENTE
AGUASCALIENTES, AGS., 9 DE SEPTIEMBRE DE 2011

M. EN C. JOSÉ DE JESÚS RUÍZ GALLEGOS

M. EN C. MARÍA DEL CARMEN MONTOYA LANDEROS

MIEMBROS DEL COMITÉ TUTORAL
SINODALES DESIGNADOS PARA EL EXAMEN DE GRADO

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES
08 SET. 2011
CENTRO DE CIENCIAS BÁSICAS

C.c.p. Archivo

Agradecimientos

A mi Director de Tesis, Dr. José Elías Rodríguez Muñoz por la oportunidad de realizar este trabajo bajo su dirección y apoyo.

A mi Tutor de Tesis, M. en C. José de Jesús Ruíz Gallegos por su apoyo y sugerencias para llevar a cabo este trabajo.

A M. en C. María del Carmen Montoya Landeros por su apoyo y sugerencias para llevar a cabo este trabajo.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca 329133/230696 proporcionada para la realización del programa de maestría.

A la Universidad Autónoma de Aguascalientes por el apoyo proporcionado durante la estancia de la maestría.

A todos los que de alguna manera me apoyaron en este proyecto.

Objetivo

Proponer una medida de bondad de ajuste y análisis de varianza, para los modelos estimados por regresión polinomial local, en muestreo de poblaciones finitas.

Índice general

Índice de cuadros

Índice de figuras

Resumen

En este trabajo se adapta una medida de bondad de ajuste, el análisis de varianza y se considera la estimación de parámetros, para modelos estimados mediante regresión polinomial local en muestreo de poblaciones finitas en presencia de información auxiliar.

Se define un coeficiente de determinación local para muestras seleccionadas de poblaciones finitas bajo un diseño de muestreo, que mide la proporción de variación explicada por el ajuste de regresión polinomial local.

Se obtiene una descomposición de la varianza global (ANOVA) a partir de la integración de las contrapartes locales y se define el estimador del coeficiente de determinación global.

Se realizan experimentos de simulación para dieciséis poblaciones hipotéticas, cuyos resultados numéricos ilustran los comportamientos del estimador del coeficiente de determinación y del estimador del estadístico de prueba propuestos.

Se utilizan datos reales de dos poblaciones para mostrar la utilidad de la metodología propuesta. Además se investigó la relación del coeficiente de determinación y la eficiencia del estimador del total por regresión polinomial local, en muestreo de poblaciones finitas.

Abstract

This document focuses on the adaptation of goodness of fit measurement, analysis of variance, and parameters estimation, for estimated models by local polynomial regression, in finite population survey sampling, in the presence of auxiliar information.

Local coefficient of determination is considered for finite population samples selected by a sampling design, which measures the proportion of local variation explained by local polynomial regression fitting.

A global analysis of variance (ANOVA) is obtained by integrating local counterparts, and a global coefficient of determination is defined.

Simulation experiments are performed for sixteen hypothetical populations, whose numerical results illustrate the behaviors of the coefficient of determination and hypothesis tests estimators proposed.

Real data is used for two population, to show the proposed methodology application. Also, relationship between the coefficient of determination and total estimator efficience is investigated for local polynomial regression in finite population sampling.

Capítulo 1

Introducción

Frecuentemente en el análisis de las principales variables de encuestas complejas por muestreo se tiene información auxiliar disponible para todas las unidades de la población. La cual puede utilizarse para incrementar la precisión de los estimadores de los parámetros de interés, que por lo regular son totales y/o promedios. Esta ganancia en la precisión de los estimadores está de manera implícita en la relación entre la información auxiliar y la característica de interés captada en la encuesta. Cuando dicha relación se puede describir a través de un modelo lineal se sugiere utilizar estimadores de regresión con la información auxiliar (Sarndal, Swensson, Wretman, 1992).

Sin embargo, cuando no se tiene evidencia de la forma de la relación entre la información auxiliar y la característica de interés, es necesario utilizar métodos de estimación alternativos en muestreo de poblaciones finitas que no presupongan conocida dicha relación.

Entre los métodos semiparamétricos propuestos están los estimadores por regresión polinomial local y los estimadores por splines. Estos métodos se utilizan para proponer estimadores de un total y derivar algunas de sus propiedades estadísticas. Por otro lado, en la estimación de un modelo lineal o cualquier otro de forma conocida, surge la pregunta ¿son adecuados los modelos estimados por métodos semiparamétricos?.

Una manera de responder a esta pregunta es proporcionar una medida de bondad de ajuste (coeficiente de determinación) para los modelos estimados y realizar un análisis de varianza que proporcione la suma de cuadrados de la curva ajustada, y que permita tomar la decisión sobre lo adecuado del modelo ajustado entre la información auxiliar y la característica de interés (Prueba global F).

Por ejemplo si se quiere estimar el total de impuestos pagados por las empresas del país en el ejercicio 2010 y se cuenta con información auxiliar sobre el empleo y los ingresos del ejercicio 2009, la metodología de regresión polinomial local, permite utilizar la relación desconocida entre las variables, para estimar el total.

En el trabajo de tesis, se propone una medida de bondad de ajuste y el análisis de varianza, para modelos estimados por regresión polinomial local en muestreo de poblaciones finitas, aplicado a datos reales de encuestas complejas.

Esto al adaptar el análisis de varianza, coeficiente de determinación y prueba F para regresión polinomial local en el contexto general de estadística (Huang y Chen, 2008) y los estimadores de regresión polinomial local en encuestas por muestreo (Breidt y Opsomer, 2000) al contexto de información auxiliar que se obtiene a través de muestreo en poblaciones finitas.

Para comprobar la efectividad de la propuesta, se compara la estimación del coeficiente de determinación global de una muestra bajo un diseño de muestreo dado de tamaño fijo, contra el respectivo de la población, la estimación del coeficiente de determinación local de una muestra contra el respectivo de la población para un conjunto determinado del rango de variación de la información auxiliar, el p-valor del análisis de varianza para una muestra contra el respectivo de la población, y además para los datos reales la estimación del total de la característica de interés con su respectivo de la población.

En el capítulo uno se da una introducción del trabajo de tesis.

En el capítulo dos se describe de forma analítica la metodología de regresión polinomial local en encuestas por muestreo de poblaciones finitas.

En el tercer capítulo se describe de forma analítica y apoyándose con gráficas el análisis de varianza, coeficiente de determinación y pueba F para regresión polinomial local.

En el cuarto capítulo se presenta la adaptación de las metodologías de regresión no paramétricas en el contexto de muestreo de poblaciones finitas.

El quinto capítulo trata del experimento por simulación para cuatro modelos hipotéticos y para dos poblaciones de datos sobre variables de los Censos General de Población y Vivienda y los Censos Económicos.

En el sexto capítulo se presentan los resultados para los experimentos por simulación de los cuatro modelos hipotéticos.

El séptimo capítulo presenta los resultados de los experimentos para las dos poblaciones con datos reales.

En el octavo capítulo se dan las conclusiones y comentarios finales sobre los resultados de la metodología propuesta y sobre el trabajo futuro.

En los apéndices se dan los aspectos técnicos, demostraciones, el código para generar los datos de las gráficas del ejemplo 1, la descripción del procedimiento para obtener la información de las poblaciones con datos reales y el programa en R del algoritmo para la simulación.

Capítulo 2

Regresión polinomial local en muestreo de poblaciones finitas

2.1. Introducción

Un problema central, en encuestas por muestreo, es el uso de información auxiliar en la estimación de parámetros de las variables de estudio en una población finita.

Esta información auxiliar, está disponible para todos los elementos de una población. Por ejemplo, los registros del último censo económico contienen información sobre el número de unidades económicas para cada municipio del país. Esta información puede ser de utilidad en estudios sobre la producción bruta total por persona ocupada.

Regresión polinomial local es un enfoque para el ajuste de curvas y superficies a los datos, mediante suavizadores en los que el ajuste en un punto x_0, del rango de la variable auxiliar X se realiza sólo con observaciones en una vecindad de x_0 y con un elemento de una familia paramétrica de funciones.
En regresión paramétrica, la forma de la función está dada por un modelo conocido, mientras que en regresión no paramétrica, la función está determinada por los datos disponibles.

En este capítulo, se describe la metodología de regresión polinomial local en muestreo de poblaciones finitas y su aplicación en la estimación de un total y un promedio poblacionales.

2.2. Definiciones de muestreo

Antes de conocer el estimador de un total y un promedio en muestreo de poblaciones finitas mediante regresión polinomial local, algunos conceptos de muestreo son:

Definición 1 *Una población es un conjunto de N elementos etiquetados con $k = \{1, \cdots, N\}$ denotado con $U = \{u_1, \cdots, u_N\}$, que sin perdida de generalidad se expresan como $U = \{1, \cdots, N\}$, sobre el cual se observan variables de interés $Y_1, \cdots, Y_q$.*

Definición 2 *Una muestra es un subconjunto de U con n elementos $(n \leq N)$ denotado con $S = \{u_1, \cdots, u_n\}$, y el conjunto de todas las muestras S de U se denota con L.*

Definición 3 *Un diseño de muestreo es una función $P : L \to [0, 1]$ que satisface las siguientes condiciones:*

1. $P(S) \geq 0 \ \forall S \in L$
2. $\sum_L P(S) = 1$

Definición 4 *Las probabilidades de inclusión de primer y segundo orden están dadas respectivamente por:*

1. $\pi_k = P(k \in S) = \sum_{S \ni k} P(S)$
2. $\pi_{k,l} = P(k, l \in S) = \sum_{S \ni k,l} P(S)$

donde $S \ni k$ indica que la suma se hace sobre todas las muestras que tienen al elemento k.

Definición 5 *Una muestra probabilística, es una muestra S seleccionada mediante un diseño de muestreo $P(\bullet)$, con $\pi_k > 0 \ \forall k \in U$.*

Definición 6 *Un diseño de muestreo $P(\bullet)$, se dice medible si cumple:*

1. $\pi_k > 0 \ \forall k \in U$
2. $\pi_{k,l} > 0 \ \forall k, l \in U$

2.3. Probabilidades de inclusión

Para una muestra S aleatoria simple sin reemplazo de tamaño n seleccionada de una población finita de tamaño N, el diseño de muestreo está dado por:

$$P(S) = \begin{cases} \dfrac{1}{\binom{N}{n}} & \text{si} \quad S \text{ tiene n elementos} \\ \\ 0 & \text{en otro caso.} \end{cases}$$

Las probabilidades de inclusión de primer orden se obtienen mediante:

$$\pi_k = \frac{n}{N},$$

con $k = 1, \cdots, N$
Mientras que las probabilidades de segundo orden están dadas por:

$$\pi_{k,l} = \frac{n(n-1)}{N(N-1)},$$

con $k \neq l$.

Sea U una población finita y S una muestra de tamaño n. La entropía de un diseño de muestreo $P(\bullet)$, sobre el conjunto de todas las muestras de tamaño n de U, es:

$$I(P) = -\sum p(S)log(p(s))$$

Generalmente, la maximización de la entropía, consiste en definir un diseño de muestreo aleatorio, y que la relación resultante entre la población y la muestra no siga patrón alguno, y se espera que los estimadores de la varianza se comporten bien.
Así, una muestra de máxima entropía es aquella que máximiza $I(P)$ bajo las restricciones dadas por las probabilidades de inclusión fijas. Esto para que se cumpla $\sum \pi_k = n$.

Para una muestra S de máxima entropía con probabilidades desiguales y tamaño de muestra fijo, seleccionada de una población finita U, el diseño de muestreo es:

$$P(s, S_n(U), \lambda) = \frac{exp(\lambda^T \mathbf{S})}{\sum_{z \in S_n(U)} exp(\lambda^T \mathbf{z})},$$

donde:

$S_n(U)$ es un conjunto de muestras de tamaño n de la población U

$\lambda \in \mathbb{R}^N$ es el vector de multiplicadores de Lagrange

$\mathbf{S} \in \mathbb{R}^{\mathbf{N}}$ es un vector tal que

$$s_k = \begin{cases} 1 & \text{si} \quad k \in S \\ 0 & \quad k \notin S. \end{cases}$$

Las probabilidades de inclusión de primer orden están dadas por:

$$\pi_k(\lambda, n) = \frac{exp\lambda_k \left[1 - \pi_k(\lambda, n-1)\right]}{\sum_{l \in U} exp\lambda_l \left[1 - \pi_l(\lambda, n-1)\right]}.$$

(Chen,1994)
Las probabilidades de inclusión conjuntas se obtienen mediante:

$$\pi_{k,l}(\lambda, n) = \frac{n(n-1)exp(\lambda_k)exp(\lambda_l \left[1 - \pi_k(\lambda, n-2) - \pi_l(\lambda, n-2) + \pi_{kl}(\lambda, n-2)\right])}{\sum_{i \in U} \sum_{j \in U j \neq i} exp(\lambda_i)exp(\lambda_j) \left[1 - \pi_i(\lambda, n-2) - \pi_j(\lambda, n-2) + \pi ij(\lambda, n-2)\right]}.$$

(Deville,2000)

2.4. Estimador de la función $m(x)$

Ahora se considera una variable respuesta Y, como una realización de una superpoblación, la cual se expresa $\forall k \in U$ como el modelo

$$y_k = m(x_k) + \epsilon_k, \tag{2.1}$$

$m(x_k)$ es una función suave de la variable auxiliar X, ϵ_k es la variable que absorbe lo no explicado de la variable respuesta por parte de la variable auxiliar.

Además, se considera que ϵ_k son variables aleatorias independientes con $E(\epsilon_k) = 0$ y varianza finita.

$m(x)$ alrededor de un punto x_0, en el rango de X, con x en la vecindad de x_0 de radio h se puede aproximar por una serie de Taylor de orden p. Es decir,

$$m(x) \approx m(x_0) + m^{(1)}(x_0)\frac{(x-x_0)}{1!} + \cdots + m^{(p)}(x_0)\frac{(x-x_0)^p}{p!},$$

que se expresa como

$$m(x) \approx \beta_0(x_0) + \beta_1(x_0)(x-x_0) +, \cdots + \beta_p(x_0)(x-x_0)^p,$$

con $\beta_j(x_0) = \frac{m^j(x_0)}{j!}$ para $j = 0, \cdots, p$.

Esta tendencia lineal está ajustada por mínimos cuadrados generalizados al resolver el problema de minimización:

$$Min_\beta N^{-1} \sum_{k \in U} (y_k - m(x_k))^2 K_h(x_k - x_0), \tag{2.2}$$

donde $K_h(\bullet)$ es una función de densidad unimodal simétrica alrededor de cero, denominada Kernel. Y asigna pesos grandes a observaciones cercanas a x_0 y pesos pequeños o nulos a observaciones lejanas de x_0.
El problema (2.2) con notación matricial se escribe como:

$$Min_\beta \, \epsilon^{\mathbf{T}} \, \mathbf{W}(\mathbf{x_0}; \mathbf{h}) \, \epsilon.$$

Esto es,

$$Min_\beta N^{-1} \left[Y^T W(x_0; h) Y - \beta^T(x_0) X^T(x_0) W(x_0; h) Y \right]$$

donde $Y = (y_1, \cdots, y_N)^T$, $W(x_0; h) = diag\left(K_h(x_1 - x_0), \cdots, K_h(x_N - x_0)\right)$,

$X(x_0) = [\mathbf{1} \ (\mathbf{x_k} - \mathbf{x_0}) \ \cdots \ (\mathbf{x_k} - \mathbf{x_0})^{\mathbf{p}}]$ y $\beta(x_0) = (\beta_0(x_0), \cdots, \beta_p(x_0))^T$.

El estimador de $\beta(x_0)$ por mínimos cuadrados generalizados es:

$$\widehat{\beta(x_0)} = \left[X^T(x_0) W(x_0; h) X(x_0) \right]^{-1} X^T(x_0) W(x_0; h) Y. \tag{2.3}$$

De esta manera, el estimador por regresión polinomial local de $m(x_k)$ está dado por:

$$\widehat{m(x_k)} \approx \widehat{m(x_0)} = \widehat{\beta_0(x_0)},$$

es decir,

$$\widehat{m(x_k)} = e_1^T \widehat{\beta(x_0)}, \tag{2.4}$$

con $e_1^T = (1, 0, \cdots, 0)^T$.

Cuando se asume la existencia de información auxiliar completa apriori de variables auxiliares, el valor de las variables se conoce para cada elemento de la población antes de

realizar el muestreo.

El objetivo es obtener un estimador de un parámetro con mejor precisión. Y la información auxiliar se utiliza en la etapa de estimación y por lo tanto estará explícitamente en la fórmula del estimador y no sólo mediante π_k.

Esto es, para un diseño de muestreo $P(\bullet)$ dado, se construyen estimadores que utilizan información auxiliar, con la finalidad de reducir la varianza en comparación con otros métodos.

El supuesto básico, detrás del uso de variables auxiliares, es que éstas se relacionan de alguna manera con la variable de estudio.

Sea una población U, con información auxiliar disponible $x_k \ \forall k \in U$, de tamaño N, de la cual se toma una muestra S de tamaño n $(n < N)$ bajo un diseño de muestreo $P(\bullet)$ medible. Y se quiere estimar un parámetro θ.
El problema es cómo estimar θ, cuando se observa (x_k, y_k) para $k \in S$ y cuando también se conoce x_k para $k \in U - S$. (Sarndal,Swensson y Wretman, 1992)

Este problema se transfiere a cómo estimar la función $m(\bullet)$ con la muestra y los valores de la variable auxiliar para los elementos de la población que no están en la muestra. Mediante la regresión polinomial local, el estimador de la función $m(\bullet)$ en muesteo de poblaciones finitas con una muestra S de tamaño fijo n seleccionada bajo un diseño de muestreo $P(\bullet)$ es:

$$\widehat{m_\pi(x_k)} = e_1^T \left[X_\pi^T(x_0) W_\pi(x_0; h) X_\pi(x_0)\right]^{-1} X_\pi^T(x_0) W_\pi(x_0; h) Y_\pi, \tag{2.5}$$

donde el subíndice π se refiere a que sólo se utiliza información de la muestra S

con $W_\pi(x_0; h) = diag\left(\frac{K_h(x_1 - x_0)}{\pi_1}, \cdots, \frac{K_h(x_n - x_0)}{\pi_n}\right)$, $Y_\pi = (y_1, \cdots, y_n)^T$,

π_k la probabilidad de inclusión de primer orden y

$X_\pi(x_0) = [\mathbf{1}\ (\mathbf{x_k} - \mathbf{x_0})\ \cdots\ (\mathbf{x_k} - \mathbf{x_0})^{\mathbf{p}}]$ para $k = 1, \cdots, n$.

2.5. Estimador de un total y un promedio poblacionales

Si el comportamiento de la variable respuesta Y está dado por el modelo (2.1), dado que se conoce $x_k \ \forall k \in U$, $m(x_k)$ se puede calcular para todos los elementos de la población. Esto es, se cree que la variable de estudio puede ser explicada como una extensión de la variable auxiliar a través de $m(x_k)$.

$m(x_k)$ sirve como vehículo para encontrar $\widehat{m(x_k)}$ y ponerlo en la expresión del estimador del parámetro de interés. Y la eficiencia del estimador del parámetro dependerá de la bondad de ajuste del modelo estimado, de la cual se habla en los siguientes capítulos.

El total de la variable Y de una población finita U, está dado por:

$$t_y = \textstyle\sum_{k \in U} y_k.$$

En encuestas por muestreo de poblaciones finitas existen varios estimadores para t_y, de los cuales se mencionan los siguientes dos:

1. Estimador de Horvitz-Thompson:

$$\widehat{t_{HT}} = \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{2.6}$$

2. Estimador por regresión:

$$\widehat{t_{Reg}} = \sum_{k \in U} \hat{y_k} + \sum_{k \in S} \frac{e_k}{\pi_k}, \tag{2.7}$$

con $\hat{y_k} = \hat{\beta}_0 + \hat{\beta}_1 x_k$, $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores usuales de los parámetros de un modelo de regresión lineal; y $e_k = y_k - \hat{y_k}$.

Y por regresión polinomial local mediante una muestra S, *Breidt y Opsomer (2000)* proponen el estimador de t_y por regresión polinomial local en muestreo de poblaciones finitas como:

$$\widehat{t_{RPL}} = \sum_{k \in S} \frac{y_k - \widehat{m(x_k)}}{\pi_k} + \sum_{k \in U} \widehat{m(x_k)}. \tag{2.8}$$

Esto con el objetivo de compararlos.

Ahora bien, cuando se quiere estimar un promedio de la variable Y de una población finita U, dado por:

$$\mu_y = \frac{\sum_{k \in U} y_k}{N},$$

los estimadores de Horvitz-Thompson, por regresión y por regresión polinomial local están dados respectivamente por:

1. Estimador de Horvitz-Thompson:

$$\widehat{\mu_{HT}} = \frac{\widehat{t_{HT}}}{\hat{N}}. \tag{2.9}$$

2. Estimador por regresión:

$$\widehat{\mu_{Reg}} = \frac{\widehat{t_{Reg}}}{\hat{N}}. \tag{2.10}$$

3. Estimador por regresión polinomial local:

$$\widehat{\mu_{RPL}} = \frac{\widehat{t_{RPL}}}{\hat{N}}, \tag{2.11}$$

con $\hat{N} = \sum_{k \in S} \frac{1}{\pi_k}$.

Capítulo 3

Coeficiente de determinación y análisis de varianza, para la regresión polinomial local

3.1. Introducción

La regresión polinomial local se utiliza para explorar la tendencia desconocida de un conjunto de datos. Para estimar esta tendencia se requiere determinar:

- El número equivalente de parámetros (g.l.).
- El ancho de banda (h).
- El grado del polinomio local (p).

Una vez que se establece la variable respuesta (Y) y la variable auxiliar (X), la tendencia desconocida se estima por regresión polinomial local mediante la expresión (2.4).

Y ¿qué tan bien se ajusta la estimación de dicha tendencia a la estructura real de los datos?

En este capítulo, se describe de forma analítica el coeficiente de determinación y el análisis de varianza para la regresión polinomial local en un contexto general de estadística.

3.2. Descomposición de la suma de cuadrados total

Si la población en estudio se puede describir a través del modelo

$$Y = X\left(x_0\right)\beta\left(x_0\right) + \epsilon\left(x_0\right), \tag{3.1}$$

con Y la variable de interés, $X(x_0)$ la información auxiliar, $\beta(x_0)$ el vector de parámetros de regresión polinomial local, $\epsilon(x_0)$ la variable aleatoria independiente de X y x_0 un valor dado en el rango de X.

El modelo (3.1) es otra forma de expresar el modelo (2.1). Luego la tendencia local se ajusta por mínimos cuadrados generalizados al resolver el problema de minimización

$$Min_{\beta} N^{-1} \sum_{k=1}^{N} \left(y_k - \sum_{j=0}^{p} \beta_j \left(x_k - x_0 \right)^j \right)^2 K_h(x_k - x_0), \tag{3.2}$$

el cual es equivalente al problema (2.2).

Se denota con $\widehat{\beta_{W(x_0;h)}}(x_0) = \left(\widehat{\beta}_0(x_0), \cdots, \widehat{\beta_p}(x_0) \right)^T$ la solución del problema (3.1) por mínimos cuadrados generalizados.

Luego el estimador (2.4) estima $m(x_k)$, esto es,

$$\widehat{m(x_k)} = \widehat{\beta_0(x_0)}, \tag{3.3}$$

donde $\widehat{\beta_0(x_0)}$ es una forma corta de expresar $\widehat{\beta}_0(x_k - x_0)$.

Y se necesita solo $\widehat{\beta_0(x_0)}$ debido a que interesa estimar el valor de Y en x_0 mediante $m(\bullet)$.

Teorema 1 *Una descomposición ANOVA local exacta para un conjunto finito de datos se obtiene para Regresión Polinomial Local en un punto x_0 en el rango de X como:*

$$SCT_p(x_0; h) = SCR_p(x_0; h) + SCE_p(x_0; h). \tag{3.4}$$

(Huang y Chen, 2008)
donde

$$\begin{aligned}
SCT_p(x_0; h) &= \frac{N^{-1} \sum_{k=1}^{N} (y_k - \bar{y})^2 K_h(x_k - x_0)}{\widehat{f(x_0; h)}}, \\
SCR_p(x_0; h) &= \frac{N^{-1} \sum_{k=1}^{N} \left[\sum_{j=0}^{p} \beta_j \left(x_k - x_0 \right)^j - \bar{y} \right]^2 K_h(x_k - x_0)}{\widehat{f(x_0; h)}}, y \\
SCE_p(x_0; h) &= \frac{N^{-1} \sum_{k=1}^{N} \left[y_k - \sum_{j=0}^{p} \beta_j \left(x_k - x_0 \right)^j \right]^2 K_h(x_k - x_0)}{\widehat{f(x_0; h)}},
\end{aligned}$$

con p el grado del polinomio local, N el número de observaciones (x_k, y_k) en el conjunto de datos, y $\widehat{f(x_0; h)}$ es el estimador de la densidad de kernel.

Se observa que, se utiliza $\bar{y} = N^{-1} \sum_{k=1}^{N} y_k$ y no un promedio ponderado, debido a que se quiere una medida de bondad de ajuste global para Regresión Polinomial Local. Y con esto, $\bar{y}$ no depende del x_0 que se elija.

Las sumas de cuadrados total, de regresión y del error se pueden expresar con notación matricial como:

$$
\begin{aligned}
SCT_p(x_0;h) &= \frac{N^{-1}\left(Y-\bar{y}\mathbf{1}\right)^T W(x_0;h)\left(Y-\bar{y}\mathbf{1}\right)}{\widehat{f(x_0;h)}},\\
SCR_p(x_0;h) &= \frac{N^{-1}\left(X(x_0)\widehat{\beta_{W(x_0;h)}}(x_0)-\bar{y}\mathbf{1}\right)^T W(x_0;h)\left(X(x_0)\widehat{\beta_{W(x_0;h)}}(x_0)-\bar{y}\mathbf{1}\right)}{\widehat{f(x_0;h)}}, y\\
SCE_p(x_0;h) &= \frac{N^{-1}\left(Y-X(x_0)\widehat{\beta_{W(x_0;h)}}(x_0)\right)^T W(x_0;h)\left(Y-X(x_0)\widehat{\beta_{W(x_0;h)}}(x_0)\right)}{\widehat{f(x_0;h)}}.
\end{aligned}
$$

Ahora se necesita la descomposición de ANOVA global, para esto se dan las siguientes condiciones.

Condiciones (A)

1. La densidad de X denotada con $f(x)$, es acotada fuera de 0 e ∞, $f(x)$ tiene segunda derivada continua sobre un soporte compacto.
2. La función Kernel denotada con $K(\bullet)$, es continua, acotada y es una función de densidad de probabilidad con un soporte sobre un intervalo compacto, digamos sobre $[-1,1]$.
3. El error ϵ proviene de una distribución simétrica con media 0 y varianza 1, y cuarto momento finito.
4. Existe la derivada de orden $(p+1)$ de $m(\bullet)$.
5. La varianza condicional $\sigma^2(\bullet)$ es acotada y continua.

Definición 7 *De la ecuación (3.4) del teorema* **1**, *se definen:*

$$
\begin{aligned}
SCT(h) &= \int SCT_p(x_0;h)\widehat{f(x_0;h)}dx_0,\\
SCR(h) &= \int SCR_p(x_0;h)\widehat{f(x_0;h)}dx_0, \ y\\
SCE(h) &= \int SCE_p(x_0;h)\widehat{f(x_0;h)}dx_0
\end{aligned}
$$

Teorema 2 *Bajo las condiciones (A) y el resultado del teorema* **1**, *la descomposición global ANOVA para un conjunto finito de datos para la Regresión Polinomial Local es*

$$SCT = SCR_p(h) + SCE_p(h). \tag{3.5}$$

Demostración

Sea un conjunto finito de datos de N observaciones (x_k, y_k) de las variables Y y X, cuya relación está dada por el modelo (3.1), y la estimación de $\beta(x_0)$ para la regresión polinomial local está dada por la ecuación (2.3).

Por el resultado (3.4) del teorema **1** y bajo las condiciones (A), se obtiene el valor esperado de la suma de cuadrados total local, es decir,

$$\int SCT_p(x_0;h)\widehat{f(x_0;h)}dx_0 = \int \left[SCR_p(x_0;h) + SCE_p(x_0;h)\right]\widehat{f(x_0;h)}dx_0,$$

por la linealidad de la integral

$$\int SCT_p(x_0;h)\widehat{f(x_0;h)}dx_0 = \int SCR_p(x_0;h)\widehat{f(x_0;h)}dx_0 + \int SCE_p(x_0;h)\widehat{f(x_0;h)}dx_0.$$

Y por la definición **7** se sigue

$$SCT(h) = SCR_p(h) + SCE_p(h).$$

Falta ver que $SCT(h) = SCT$.

En efecto,

$$\begin{aligned} SCT(h) = \int SCT_p(x_0;h)\widehat{f(x_0;h)}dx_0 &= \int \frac{N^{-1}\sum_{k=1}^{N}(y_k-\bar{y})^2 K_h(x_k-x_0)\widehat{f(x_0;h)}}{\widehat{f(x_0;h)}}dx_0 \\ &= N^{-1}\sum_{k=1}^{N}(y_k-\bar{y})^2\int K_h(x_k-x_0)dx_0 \\ &= N^{-1}\sum_{k=1}^{N}(y_k-\bar{y})^2 \\ &= SCT. \end{aligned}$$

Por lo tanto

$$SCT = SCR_p(h) + SCE_p(h).$$

Si bien, una medida de bondad de ajuste del modelo puede ser $SCE_p(x_0;h)$ para el caso local y $SCE_p(h)$ para el caso global. Esta suma de cuadrados debida al error está afectada por las unidades de las variables Y y X.

3.3. Coeficiente de determinación

Se necesita una medida relativa para la bondad de ajuste del modelo para la Regresión Polinomial Local, es decir, una medida sin unidades.

Definición 8 *De la ecuación (3.4) del teorema* **1** *se tiene* $SCT_p(x_0;h) = SCR_p(x_0;h) + SCE_p(x_0;h)$ *dividiendo entre* $SCT_p(x_0;h)$ *se obtiene*

$$1 = \frac{SCR_p(x_0;h)}{SCT_p(x_0;h)} + \frac{SCE_p(x_0;h)}{SCT_p(x_0;h)},$$

y se define el Coeficiente de Determinación Local en x_0*, denotado con* $R_p^2(x_0;h)$ *como:*

$$R_p^2(x_0;h) = \frac{SCR_p(x_0;h)}{SCT_p(x_0;h)} = 1 - \frac{SCE_p(x_0;h)}{SCT_p(x_0;h)}. \tag{3.6}$$

Se observa que, $R_p^2(x_0;h)$ está entre 0 y 1, y mide la proporción de la variación local de Y que se explica por el ajuste polinomial local alrededor de x_0. Además, $R_p^2(x_0;h)$ es invariante a transformaciones lineales de Y, es decir, es invariante respecto a traslaciones y cambios de escala.

Definición 9 *De la ecuación (3.5) del teorema* **2** *se tiene* $SCT = SCR_p(h) + SCE_p(h)$ *dividiendo entre* SCT *se obtiene*

$$1 = \frac{SCR_p(h)}{SCT} + \frac{SCE_p(h)}{SCT},$$

y se define el Coeficiente de Determinación Global, denotado con $R_p^2(h)$ *como:*

$$R_p^2(h) = \frac{SCR_p(h)}{SCT} = 1 - \frac{SCE_p(h)}{SCT}. \tag{3.7}$$

Por el teorema **2**, $R_p^2(h)$ está entre 0 y 1, y mide la proporción de la variación global de Y que se explica por el ajuste polinomial global.

Ejemplo 1 *Se tienen* $N = 1,000$ *observaciones respecto a las variables unidimensionales* Y *y* X *cuya relación se presenta en la figura 3.1.*

De la figura 3.1 se observa que existe una relación no lineal entre las variables, y alrededor de $x_0 = 0.5$ la naturaleza es cuadrática.

Se ajusta un modelo por regresión polinomial local a los datos, con un polinomio local de orden $p = 1$ y un ancho de banda $h = 0.202$. La determinación del valor de h se explicará más adelante.
En la figura 3.2 se muestra el ajuste global con el cálculo de $R_1^2(h)$. En la figura 3.3 se presentan los ajustes locales para $x_0 = 0.1, 0.3, 0.5, 0.7, 0.9$ con $R_1^2(x_0;h)$.

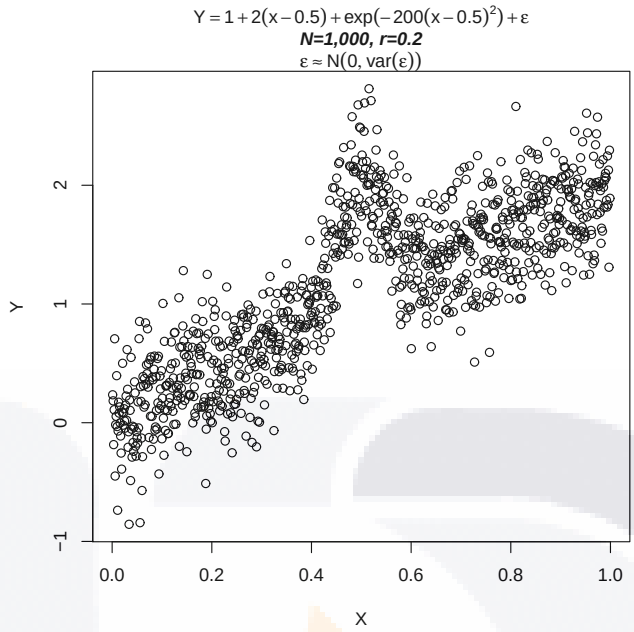

Figura 3.1: Diagrama de dispersión de las variables Y y X.

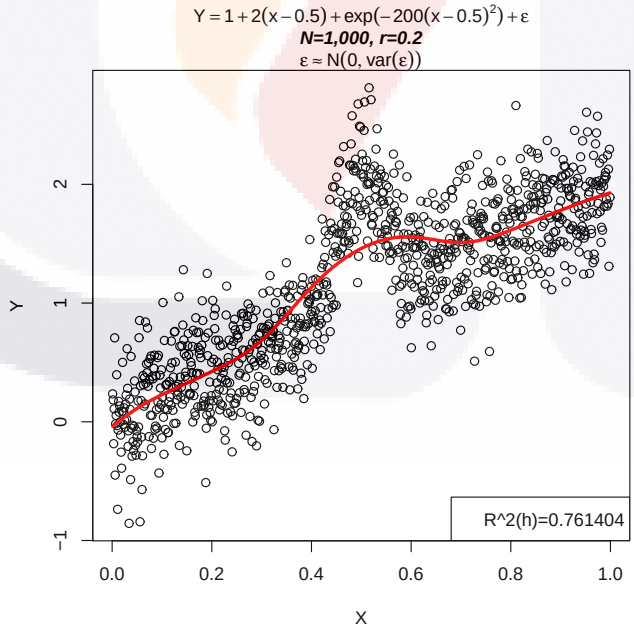

Figura 3.2: Ajuste con un modelo por R.P.L. a los datos.

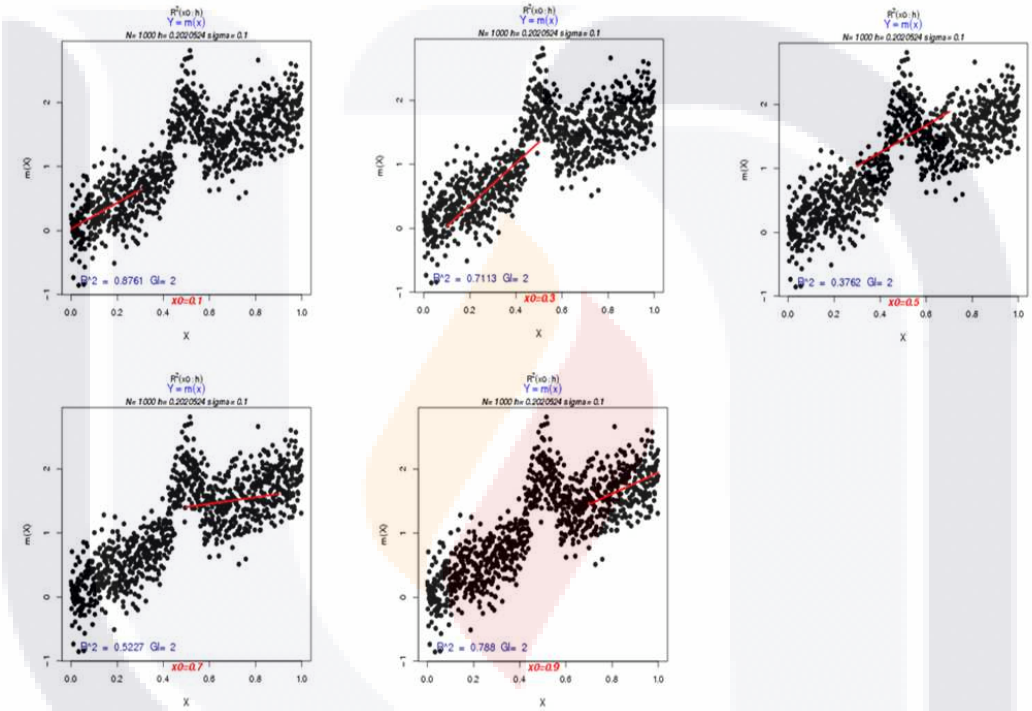

Figura 3.3: Ajuste local en diferentes valores x_0 del rango de la variable X.

Teorema 3 *Supóngase que conforme $N \to \infty$, $h = h(N) \to 0$. Al ajustar el modelo mediante Regresión Polinomial Local con un polinomio de orden p impar, bajo las condiciones (A) con $nh^{2p+2} \to 0$ y $nh^2 \to \infty$*

1. *El sesgo condicional asintótico de $R_p^2(h)$ es*

$$-h^2 \frac{\mu_2}{2\sigma_y^2} \int \sigma^2(x_0) f''(x_0) dx_0 (1 + O_P(1)).$$

2. *La varianza condicional asintótica de $R_p^2(h)$ es*

$$N^{-1} \left[\frac{Var(\epsilon^2)}{\sigma_y^4} E\left(\sigma^4(X)\right) \left(\int K_0^*(v) dv \right) + \frac{(m_4 - \sigma_y^4)\left(E(\sigma^2(X))\right)^2}{\sigma_y^8} \right],$$

   con σ_y^2 la varianza de Y, m_4 el cuarto momento central de Y y $K^*(v) = \int K(u)K(v-u)du$.

3. *Bajo el supuesto de homoscedasticidad y condicionado sobre $\{X_1, \cdots, X_N\}$, $R_p^2(h)$ converge en distribución a la distribución normal con sesgo y varianza condicionales anteriores.*

4. *Bajo los supuestos del inciso anterior, $SCE_p(h)$ es un estimador $\sqrt{N}$ consistente para σ^2. Su sesgo condicional asintótico es $O_P(N^{-1/2})$ si $\int f''(x_0)dx_0 = 0$ y su varianza condicional asintótica es*

$$N^{-1}\sigma^4 \left(\int K_0^*(v) dv \right) (1 + O_P(1)).$$

3.4. Matriz de proyección asintótica

De regresión lineal, se tiene que la estimación de Y está dado por

$$\widehat{Y} = X\widehat{\beta},$$

donde $\widehat{\beta} = \left[X^T W X\right]^{-1} X^T W Y$, es el estimador por mínimos cuadrados generalizados de β.

$\widehat{Y}$ se puede expresar a través de la denominada matriz de proyección H_W como:

$$\widehat{Y} = H_W Y,$$

con $H_W = X\left[X^T W X\right]^{-1} X^T W$.

De manera similar, Huang y Chen (2008) proponen la matriz de proyección asintótica para la regresión polinomial local.

Definición 10 *Bajo las condiciones (A1) y (A2) se define la matriz de proyección asintótica del ajuste global para la Regresión Polinomial Local, denotada con $H(h)$ como:*

$$H(h) = \int W(x_0;h)H_W(x_0;h)dx_0, \tag{3.8}$$

con

$$H_W(x_0;h) = X(x_0)\left[X^T(x_0)W(x_0;h)X(x_0)\right]^{-1}X^T(x_0)W(x_0;h),$$
$$W(x_0;h) = diag\left(K_h(x_1-x_0),\cdots,K_h(x_N-x_0)\right),\ y$$
$$X(x_0) = \left[\mathbf{1}\ (\mathbf{x_k}-\mathbf{x_0})\ \cdots\ (\mathbf{x_k}-\mathbf{x_0})^{\mathbf{p}}\right].$$

$H_W(x_0;h)$ es la matriz de proyección local y $H(h)$ sólo depende de X, $K(\bullet)$ y de h.

Se expresan las sumas de cuadrados globales debidas a la regresión y al error con matrices como:

$$SCR_p(h) = N^{-1}Y^T\left(H(h)-L\right)Y, \text{ y}$$
$$SCE_p(h) = N^{-1}Y^T\left(I-H(h)\right)Y,$$

con

$$L = \begin{pmatrix} \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \ddots & \vdots \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}, y$$

$$I = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

Luego, $\widehat{Y}_{RPL} = H(h)Y$, es el vector de valores de Y ajustado por Regresión Polinomial Local mediante la matriz de proyección asintótica $H(h)$.

3.5. Prueba de hipótesis de no efecto

En el análisis de regresión lineal o no lineal surge la pregunta ¿es adecuado el ajuste del modelo a los datos?, esta pregunta se responde mediante una prueba estadística de hipótesis denominada prueba global F para modelos de regresión.

Huang y Chen (2008), proponen una prueba global F de no efecto, para la regresión polinomial local que hereda las propiedades de la prueba global F para los modelos de regresión clásicos.

Las hipótesis que se plantean son:

H_0 : El ajuste del modelo a los datos no es adecuado.

H_a : El ajuste del modelo a los datos es adecuado.

Teorema 4 *Bajo las condiciones (A) y un ajuste por Regresión Polinomial Local de orden p a un conjunto finito de datos con* $h \to 0$ *conforme* $N \to \infty$ *condicionado sobre* $\{x_1, \cdots, x_N\}$

1. $(I - H(h))$ *y* $(H(h) - L)$ *son asintóticamente ortogonales sobre* $\{x_1, \cdots, x_N\}$ *en el sentido que* $E\left[(I - H(h))(H(h) - L)Y|x_1, \cdots, x_N\right] = E\left[\left(H(h) - H^2(h)\right)Y|x_1, \cdots, x_N\right]$, *el cual tiende al vector cero en probabilidad.*

2. *Bajo el supuesto de homoscedasticidad, un estadístico de prueba, denotado con* $F(h)$ *está dado por:*

$$F(h) = \frac{\frac{SCR_p(h)}{tr(H(h))-1}}{\frac{SCE_p(h)}{N-tr(H(h))}}, \tag{3.9}$$

   donde $tr\left(H(h)\right)$ *es la traza de la matriz* $H(h)$ *condicionada sobre* $\{x_1, \cdots, x_N\}$. *Con el supuesto de errores normales, el estadístico* $F(h)$ *dado en (3.9) está distribuido asintóticamente como* F *de Fisher con grados de libertad* $[tr\left(H(h)\right) - 1, N - tr\left(H(h)\right)]$ *respectivamente.*

3. *La traza condicional de* $H(h)$ *para la regresión lineal local* $(p = 1)$ *es asintóticamente:*

$$tr\left(H(h)\right) = h^{-1}|\Omega|\left(v_0 + v_2/\mu_2\right)\left(1 + O_P(1)\right), \tag{3.10}$$

   con $|\Omega|$ *el rango de* X *y* $v_j = \int u^j K^2(u)du$

Del teorema **4**, la tabla de análisis de varianza para la regresión polinomial local, con un nivel significancia α está dada en el cuadro 3.1. Asociado con la tabla (ANOVA) se define para la

Cuadro 3.1: Tabla ANOVA para Regresión Polinomial Local

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	$tr(H(h)) - 1$	$SCR_p(h)$	$\frac{SCR_p(h)}{tr(H(h))-1}$	$F(h)$	$Prob(F > F_{tr(H(h))-1, N-tr(H(h)), \alpha})$
Error	$N - tr(H(h))$	$SCE_p(h)$	$\frac{SCE_p(h)}{N-tr(H(h))}$		
Total	$N - 1$	SCT			

regresión polinomial local, el coeficiente de determinación ajustado, denotado con $R^2_{Ajus}(h)$ como:

$$R^2_{Ajus}(h) = 1 - \frac{\frac{SCE_p(h)}{N-tr(H(h))}}{\frac{SCT}{N-1}}. \tag{3.11}$$

De la ecuación (3.7) se tiene que $SCR_p(h) = (SCT)\,R^2(h)$ y $SCE_p(h) = (SCT)\left(1 - R^2(h)\right)$.

Sustituyendo en la ecuación (3.9) y simplificando, se tiene que el estadístico de prueba se puede expresar en términos de $R^2(h)$ como:

$$F(h) = \frac{\frac{R^2(h)}{tr(H(h))-1}}{\frac{(1-R^2(h))}{N-tr(H(h))}}. \tag{3.12}$$

Definición 11 *Hastie y Tibshirani (1990), definen el número equivalente de parámetros o número efectivo de parámetros, de un suavizador S como la traza del suavizador.*

Esto significa que la complejidad del modelo $m(\bullet)$ es similar que la complejidad de una regresión polinomial local de grado igual al número equivalente de parámetros (N.E.P.). Es decir:

$$N.E.P.=traza(S).$$

Así, para la regresión polinomial local, la matriz asintótica de proyección juega el papel de suavizador, pues $\widehat{Y} = H(h)Y$. Luego

$$N.E.P. = traza\left(H(h)\right). \tag{3.13}$$

$traza\left(H(h)\right)$ indica la cantidad de ajuste del modelo a los datos a través de la matriz $H(h)$.

Sin embargo, de la ecuación (3.8), $H(h)$ utiliza todos los coeficientes de $\widehat{\beta(x_0)}$, es decir, utiliza los p componentes del vector de coeficientes estimados para obtener $\widehat{Y}$.

Y para la propuesta, con el ajuste polinomial local en x_0 sólo se utiliza $\widehat{\beta_0(x_0)}$ y se descartan $\widehat{\beta_1(x_0)}, \cdots, \widehat{\beta_p(x_0)}$, para $p \geq 1$.

En el teorema **4**, tercer inciso, se da una forma para calcular $tr\left(H(h)\right)$, para regresión polinomial local que está relacionado con el estadístico de prueba dado en (3.12), cuyos grados de libertad para el denominador $N - tr\left(H(h)\right)$, deben calcularse adecuadamente para realizar el contraste de hipótesis.

Hastie y Tibshirani (1990) indican que, cuando las observaciones están ponderadas, los grados de libertad del error pueden calcularse mediante:

$$2tr\left(S\right) - tr\left[S^T W S W^{-1}\right],$$

con S la matriz del suavizador y W es la matriz diagonal de pesos. Y se da una aproximación:

$$2tr\left(S\right) - tr\left[S^T W S W^{-1}\right] \approx 1.25tr\left(S\right) - 0.5. \tag{3.14}$$

Luego, se propone adaptar la aproximación (3.14) al caso de regresión polinomial local en el calculo de $F(h)$.

Es decir:

$$N - tr\left(H(h)\right) \approx 1.25tr\left(H(h)\right) - 0.5. \tag{3.15}$$

Esto sólo para el cálculo de $F(h)$, con el objetivo de tener un criterio de rechazo de H_0 acorde con el teorema **4**.

Por lo anterior y con la ecuación (3.12) se propone el estadístico de prueba:

$$F_c(h) = \frac{\frac{R_p^2(h)}{tr(H(h))-1}}{\frac{\left(1-R_p^2(h)\right)}{1.25tr(H(h))-0.5}}. \tag{3.16}$$

Capítulo 4

Coeficiente de determinación y análisis de varianza, para la regresión polinomial local, en muestreo de poblaciones finitas

4.1. Introducción

Las técnicas de suavizamiento, son utilizadas en regresión no paramétrica, como métodos para encontrar la curva de regresión a partir de un conjunto de datos. Ajustar los datos a través de suavizadores polinomiales locales, tiene una serie de ventajas. Más aún, cuando se consideran los polinomios de orden $p = 1$, que se denominan, polinomios lineales locales.

Por ejemplo, al utilizar el kernel de Epanechnikov (Fan, 1993), entre los estimadores lineales, alcanza la eficiencia minimax completa, y se adapta automaticamente en las fronteras (Fan y Gijbels, 1992).

En el capítulo anterior se describió la forma de obtener una medida de bondad de ajuste de un modelo a un conjunto de datos. Así como la manera de realizar el análisis de varianza (ANOVA) para la regresión polinomial local.

En este capítulo, se adapta el coeficiente de determinación y el análisis de varianza para la Regresión Polinomial Local al Muestreo de Poblaciones Finitas.

4.2. Descomposición de la suma de cuadrados total

Sea un conjunto de datos bivariados $(x_1, y_1), \cdots, (x_n, y_n)$ observados al seleccionar una muestra aleatoria S de tamaño fijo n bajo un diseño de muestreo $P(S)$ de una población finita U de tamaño N.

La muestra S, puede considerarse como una realización de las variables Y y X, cuya relación en la población puede expresarse con el modelo (2.1), esto es, mediante

$$Y = X(x_0)\beta(x_0) + \epsilon(x_0).$$

Con estimador de regresión polinomial local dado por:

$$m(x_k) = e_1^T \left[X^T(x_0)W(x_0;h)X(x_0)\right]^{-1} X^T(x_0)W(x_0;h)Y.$$

Sin embargo, en el contexto de muestreo, $m(x_k)$ no puede calcularse debido a que se tiene sólo información de una muestra. Y se requiere un estimador de $m(x_k)$ para muestreo de poblaciones finitas, dado en la ecuación (2.5).
Luego, el estimador del modelo (3.1) con una muestra S mediante regresión polinomial local es:

$$\widehat{Y} = X_\pi(x_0)\widehat{\beta_\pi}(x_0), \tag{4.1}$$

con

$$\widehat{\beta_\pi}(x_0) = \left[X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\right]^{-1} X_\pi^T(x_0)W_\pi(x_0;h)Y_\pi. \tag{4.2}$$

Proposición 1 *Una descomposición ANOVA local exacta para la regresión polinomial local en muestreo de poblaciones finitas alrededor de un punto x_0 en el rango de X y ancho de banda h está dada por:*

$$\widehat{SCT_p}(x_0;h) = \widehat{SCR_p}(x_0;h) + \widehat{SCE_p}(x_0;h), \tag{4.3}$$

donde

$$\widehat{SCT_p}(x_0;h) = \frac{\hat{N}^{-1}}{\widehat{f(x_0;h)}} \sum_{k\in S} \left(y_k - \hat{\bar{y}}\right)^2 \frac{K_h(x_k-x_0)}{\pi_k},$$

$$\widehat{SCR_p}(x_0;h) = \frac{\hat{N}^{-1}}{\widehat{f(x_0;h)}} \sum_{k\in S} \left[\sum_{j=0}^p \hat{\beta}_j(x_k-x_0)^j - \hat{\bar{y}}\right]^2 \frac{K_h(x_k-x_0)}{\pi_k}, \text{ y}$$

$$\widehat{SCE_p}(x_0;h) = \frac{\hat{N}^{-1}}{\widehat{f(x_0;h)}} \sum_{k\in S} \left[y_k - \sum_{j=0}^p \hat{\beta}_j(x_k-x_0)^j\right]^2 \frac{K_h(x_k-x_0)}{\pi_k},$$

con $\hat{\bar{y}} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}}{\hat{N}}$, $\hat{N} = \sum_{k\in S}\frac{1}{\pi_k}$ *y* $\hat{\beta}_j$ *es el componente* j *de* $\widehat{\beta_\pi}(x_0)$.

Definición 12 *De la ecuación (4.3), los estimadores de las sumas de cuadrados total, de regresión y del error globales se definene como:*

$$\begin{aligned}
\widehat{SCT(h)} &= \int \widehat{SCT_p}(x_0;h)\widehat{f(x_0;h)}dx_0, \\
\widehat{SCR_p(h)} &= \int \widehat{SCR_p}(x_0;h)\widehat{f(x_0;h)}dx_0, \text{ y} \\
\widehat{SCE_p(h)} &= \int \widehat{SCE_p}(x_0;h)\widehat{f(x_0;h)}dx_0,
\end{aligned} \tag{4.4}$$

con $\widehat{f(x_0;h)} = \widehat{N^{-1}}\sum_{k\in S}\frac{K_h(x_k-x_0)}{\pi_k}$.

Proposición 2 *Bajo las condiciones (A) y el resultado de la proposición* **1**, *la descomposición global ANOVA para la regresión polinomial local en muestreo de poblaciones finitas es*

$$\widehat{SCT} = \widehat{SCR_p}(h) + \widehat{SCE_p}(h), \tag{4.5}$$

con

$$\widehat{SCT} = \sum_{k\in s}\frac{y_k^2}{\pi_k} - \widehat{\bar{y}}\widehat{N}.$$

4.3. Estimador del coeficiente de determinación local y global

Definición 13 *El estimador del coeficiente local para la regresión polinomial local, en muestreo de poblaciones finitas, alrededor de un punto x_0 en el rango de X y ancho de banda h, denotado con $\widehat{R^2_\pi}(x_0;h)$ está dado por:*

$$\widehat{R^2_\pi}(x_0;h) = \frac{\widehat{SCR_p}(x_0;h)}{\widehat{SCT}(x_0;h)} = 1 - \frac{\widehat{SCE_p}(x_0;h)}{\widehat{SCT}(x_0;h)}. \tag{4.6}$$

De la proposición **1**, $\widehat{R^2_\pi}(x_0;h)$ está entre 0 y 1, da una idea de la calidad de la estimación en diferentes regiones de los datos, y es una estimación de la proporción de la variación local de Y que se explica por el ajuste polinomial local alrededor de x_0.

Definición 14 *El estimador del coeficiente de determinación global para la regresión polinomial local en muestreo de poblaciones finitas con ancho de banda h, denotado con $\widehat{R^2_\pi(h)}$ está dado por:*

$$\widehat{R^2_\pi(h)} = \frac{\widehat{SCR_p(h)}}{\widehat{SCT}} = 1 - \frac{\widehat{SCE_p(h)}}{\widehat{SCT}}. \tag{4.7}$$

$\widehat{R^2_\pi(h)}$ siempre está entre 0 y 1 y es una estimación de la proporción de la variación global de Y que se explica por el ajuste global de regresión polinomial local.

Ahora, se requiere estimar la matriz de proyección asintótica $H(h)$ dada en la ecuación (3.8).

Proposición 3 *El estimador de la matriz de proyección asintótica del modelo de regresión polinomial local de orden p con ancho de banda h en muestreo de poblaciones finitas , denotado con $\widehat{H_\pi(h)}$ es:*

$$\widehat{H_\pi(h)} = \int W_\pi(x_0;h) H_{W_\pi}(x_0;h) dx_0, \tag{4.8}$$

con

$$H_{W_\pi}(x_0;h) = X_\pi(x_0)\left[X^T_\pi(x_0) W_\pi(x_0;h) X_\pi(x_0)\right]^{-1} X_\pi(x_0) W_\pi(x_0;h).$$

Obsérvese que $H_{W_\pi}(x_0;h)$ sirve como estimador de la matriz de proyección local y $\widehat{H_\pi(h)}$ depende únicamente de X_π, $K(\bullet)$ y de h.

4.4. Prueba de hipótesis de no efecto para la regresión polinomial local, en muestreo de poblaciones finitas

Con la finalidad de realizar una prueba de bondad de ajuste del modelo estimado con regresión polinomial local, en muestreo de poblaciones finitas, se tienen las hipótesis de la prueba de no efecto.

Las hipótesis son:

H_0 : El ajuste del modelo a los datos no es adecuado.

H_a : El ajuste del modelo a los datos es adecuado.

Las cuales deben verificarse para un nivel de significancia α a partir de una muestra S de tamaño fijo n seleccioanda bajo un diseño de muestreo $P(S)$.

Del teorema **4** y de la ecuación (3.12) se da la siguiente:

Proposición 4 *El estimador del estadístico de prueba de hipótesis de no efecto para la regresión polinomial local, en muestreo de poblaciones finitas, denotado con* $\widehat{F(h)}$ *es:*

$$\widehat{F(h)} = \frac{\frac{\widehat{R^2_\pi(h)}}{tr\left(\widehat{H_\pi(h)}\right)-1}}{\frac{\left(1-\widehat{R^2_\pi(h)}\right)}{\widehat{N}-tr\left(\widehat{H_\pi(h)}\right)}}, \tag{4.9}$$

con $tr\left(\widehat{H_\pi(h)}\right)$ *la traza del estimador dado en (4.8).*

Además, de la ecuación (3.16), se propone calcular el estimador (4.9) mediante la aproximación:

$$\widehat{F_c(h)} = \frac{\frac{\widehat{R^2_\pi(h)}}{tr\left(\widehat{H_\pi(h)}\right)-1}}{\frac{\left(1-\widehat{R^2_\pi(h)}\right)}{1.25*tr\left(\widehat{H_\pi(h)}\right)-0.5}}. \tag{4.10}$$

Para un nivel de significancia α, y con $\widehat{CMR} = \frac{\widehat{SCR_\pi(h)}}{tr(\widehat{H_\pi(h)})-1}$ y $\widehat{CME} = \frac{\widehat{SCE_\pi(h)}}{\widehat{N}-tr(\widehat{H_\pi(h)})}$.
De la proposición **4**, la tabla de análisis de varianza para la regresión polinomial local en muestreo de poblaciones finitas está dado en el cuadro 4.1

Definición 15 *Asociado con la tabla ANOVA para la regresión polinomial local, en muestreo de poblaciones finitas, se define el estimador del coeficiente de determinación ajustado, denotado con* $\widehat{R^2_{Ajus}(h)}$ *como:*

$$\widehat{R^2_{Ajus}(h)} = 1 - \frac{\frac{\widehat{SCE_p(h)}}{tr\left(\widehat{H_\pi(h)}\right)-1}}{\frac{\widehat{SCT}}{\widehat{N}-1}}. \tag{4.11}$$

Cuadro 4.1: Tabla ANOVA para Regresión polinomial local, en muestreo de poblaciones finitas.

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	$tr(\widehat{H_\pi(h)}) - 1$	$\widehat{SCR_p}(h)$	$\widehat{CMR}$	$\widehat{F_c(h)}$	$Prob\left(\widehat{F_c(h)} > F_{tr(\widehat{H_\pi(h)})-1,1.25*tr(\widehat{H_\pi(h)})-0.5,\alpha}\right)$
Error	$\widehat{N} - tr(\widehat{H_\pi(h)})$	$\widehat{SCE_p}(h)$	$\widehat{CME}$		
Total	$\widehat{N} - 1$	$\widehat{SCT}$			

Capítulo 5

Experimento por simulación

En este capítulo, se propone el experimento por simulación para comparar, con varias poblaciones,el desempeño del coeficiente de determinación local y global, la prueba F global, y además para los datos reales, la estimación de un total o un promedio mediante la regresión polinomial local en muestreo de poblaciones finitas.

5.1. Variables controladas.

1. **Tamaño de la población.** Se establece un tamaño de la población de $N = 5,000$ y no mayor por cuestiones de tiempo de cómputo.
2. **Variables de interés.** Se utilizan dos variables de interés:
   a) $m(x_k) = 2 + \left[Sen\left(2\pi x_k^3\right)\right]^3$,
   b) $m(x_k) = 1 + 2(x_k - 0.5) + exp\left[-200(x_k - 0.5)^2\right]$.
3. **Variable auxiliar.** Una variable auxiliar X, que es la misma para cada variable de interés.
4. **Tamaños de muestra.** Son dos tamaños de muestra los considerados.
   a) $n = 100$, 2 % de la población,
   b) $n = 250$, 5 % de la población.
5. **Diseño de muestreo.** Se contemplan dos diseños de muestreo.
   a) Diseño de muestreo de máxima entropía con probabilidades desiguales y tamaño de muestra fijo (Máx. Ent.),
   b) Diseño de muestreo aleatorio simple sin reemplazo (MASSR).
6. **Contribución de la parte residual a la variable de interés**. Dos valores para la contribución de la parte residual ϵ_k a la variable de interés, denotada con r y valores: (ver apéndice A)
   a) $r = 0.2$, 20 % de contribución,

b) $r = 0.8$, 80 % de contribución.

7. **Número de simulaciones**. Se realizan $M = 3,000$ simulaciones para cada experimento. (ver apéndice A)

5.2. Algoritmo de simulación.

Una vez que se determina la población o los datos de estudio, las funciones necesarias para realizar las simulaciones son:

Función 1 Cálculo de h(Población,NEP)

$N \leftarrow pobl$
for $i \leftarrow 1$, N **do**
  $tt \leftarrow \frac{i}{N+1}$
end for
$NEP \leftarrow g.l.$
$raiz \leftarrow biseccion(tt, NEP)$
$h \leftarrow raiz[1]$

Función 2 Generación de la variable auxiliar(N)

for $i \leftarrow 1$, N **do**
  $tt \leftarrow \frac{i}{N+1}$
end for
$X \leftarrow tt$

Función 3 Generación de los valores de la población Hardle(N,r,X)

$y \leftarrow 2 + \left[Sen(2\pi x^3)\right]^3$
$vare \leftarrow \frac{r}{1-r} Var(y)$
$y \leftarrow y + \epsilon$
Regresa
(x, y)

Función 4 Generación de los valores de la población Bump(N,r,X)

$y \leftarrow 1 + 2(x_k - 0.5) + exp^{[-200(x_k-0.5)^2]}$
$vare \leftarrow \frac{r}{1-r} Var(y)$
$y \leftarrow y + \epsilon$
Regresa
(x, y)

Función 5 Cálculo de parámetros(Población, h)

$R^2(h)$
$R^2(x_0; h)$
$F_c(h)$
$pvalor \leftarrow Prob(F_c(h) > F_{critica})$
if $pvalor < \alpha$ **then**
  $D \leftarrow 1$
else
  $D \leftarrow 0$
end if
μ_{RPL}
μ_{HT}
μ_{Reg}
$ANOVA$

Función 6 Probabilidades de inclusión de primer orden(Población, n)

if Diseño=MASSR **then**
  $\pi_k \leftarrow \frac{n}{N}$
else
  Establecer el valor de la correlación
  Generar tamaños de las unidades
  Obtener π_k de máxima entropía
end if

Función 7 Estimación(Población, M, n, h)

Para cada tamaño de muestra n
Calcular π_k según diseño de muestreo
for $i = 1$ a M **do**
  Seleccionar una muestra S
  Con la muestra S, calcular:
  $\widehat{R^2_\pi(h)}$
  $\widehat{R^2_\pi(x_0; h)}$
  $\widehat{F_c(h)}$
  $\widehat{pvalor}$
  if $\widehat{pvalor} < \alpha$ **then**
    $D_S \leftarrow 1$
  else
    $D_S \leftarrow 0$
  end if
  $\widehat{\mu_{RPL}}$
  $\widehat{\mu_{HT}}$
  $\widehat{\mu_{Reg}}$
end for

Función 8 Resultados(Muestras,$\theta = (\mu_y, R^2)$)

Con las M muestras calcular:
$R.E.C.M.R.(\widehat{\theta})$
$Promedio(\widehat{\theta})$
$Des.Est.(\widehat{\theta})$
$E.R.(\widehat{\theta})$
$\widehat{Prob}(D_S = D|D)$
Histogramas
Gráficas de lineas

Los cuales se describen acontinuación:

1. Determinar el número equivalente de parámetros, denotado con $N.E.P.$. Este número será el grado del polinomio que se ajustará globalmente a la población.

2. Calcular el ancho de banda h para el $N.E.P.$ correspondiente.
   Por la definición **11**, el ancho de banda se obtiene al resolver para h la igualdad

   $$N.E.P. = traza\left[H\left(h\right)\right],$$

   donde $H\left(h\right)$ es la matriz asintótica de proyección dada en la ecuación (3.8).
   Para calcular $H\left(h\right)$ se aplica regresión líneal local con la función kernel de Epanechnikov dado por

   $$K\left(u\right) = \frac{3}{4}\left(1-u^2\right) I_{|u|\leq 1},$$

   para 100 puntos equidistantes en el rango de la variable auxiliar X con distribución uniforme en $[0,1]$, y la integral se apróxima numéricamente mediante el método del trapecio.

   Luego para calcular h se resuleve la ecuación

   $$traza\left[H\left(h\right)\right] - N.E.P. = 0, \tag{5.1}$$

   esto es, se obtiene la aproximación de la raíz de la ecuación (5.1) mediante el método de bisección.

3. Se genera la variable auxiliar X con una distribución uniforme en $[0,1]$.

4. Se generan los valores de las funciones $m\left(x_k\right)$.

5. Se genera la parte residual de cada función $m\left(x_k\right)$ a través de una distribución normal con media cero y varianza $Var\left(\epsilon_k\right)$, con ϵ_k la parte residual. (ver apéndice A)

6. Generar los valores de las variables de interés con la parte residual, es decir, obtener

   a) $y_k = 2 + \left[Sen\left(2\pi x_k^3\right)\right]^3 + \epsilon_k$ (Hardle),

   b) $y_k = 1 + 2\left(x_k - 0.5\right) + exp\left[-200\left(x_k - 0.5\right)^2\right] + \epsilon_k$ (Bump).

7. Para el diseño Máx. Ent., se generan los tamaños de cada unidad poblacional, denotados con z_k, de forma tal que la correlación entre la variable auxiliar x_k y z_k sea igual que 0.5. Esto con la finalidad de asignar las probabilidades desiguales de primer orden a cada unidad de la población. (ver apéndice A)

8. Calcular los siguientes parámetros, denotados con θ, para cada población.

   a) El coeficiente de determinación global $R^2(h)$. Dado en la ecuación (3.7).

   b) El coeficiente de determinación local $R_p^2(x_0, h)$ en $x_0 = 0.1, \cdots, 0.9$. Dado en la ecuación (3.6).

   c) El estadístico de prueba $F_c(h)$. Dado en la ecuación (3.16).

   d) El valor p. Dado en el cuadro 7.1.

e) Con base en el estadístico de prueba $F_c(h)$ calculado para la población, tomar la decisión sobre el rechazo de H_0. Se denota la decisión con D que toma valores:

$$D = \begin{cases} 0 & \text{si no se rechaza } H_0, \\ \\ 1 & \text{si se rechaza } H_0. \end{cases}$$

9. Seleccionar $M = 3,000$ muestras aleatorias independientes para cada experimento.

10. Estimar los parámetros del paso 8 con cada una de las M muestras seleccionadas para cada experimento. Esto es, calcular

a) El coeficiente de determinación global estimado $\widehat{R^2_\pi(h)}$. Dado en la ecuación (4.7).

b) El coeficiente de derminación local estimado $\widehat{R^2_\pi}(x_0, h)$ en $x_0 = 0.1, \cdots, 0.9$. Dado en la ecuación (4.6).

c) El estadístico de prueba estimado $\widehat{F_c(h)}$. Dado en la ecuación (4.10).

d) El valor p estimado. Dado en el cuadro 4.1.

e) Con base en el estadístico de prueba estimado para cada muestra, tomar la decisión sobre el rechazo de H_0. Se denota la decisión con D_s que toma valores:

$$D_s = \begin{cases} 0 & \text{si no se rechaza } H_0, \\ \\ 1 & \text{si se rechaza } H_0. \end{cases}$$

f) Obtener la raíz cuadrada del error cuadrático medio relativo de cada estimación, para las M muestras, mediante

$$R.E.C.M.R. = \sqrt{\frac{1}{M}\sum_{s=1}^{M}\left(\frac{\hat{\theta}_s - \theta}{\theta}\right)^2}, \tag{5.2}$$

con $\hat{\theta}_s$ es el estimador del parámetro θ en estudio con la muestra S.

g) Obtener la estimación promedio del parámetro θ con las M muestras, denotada con $\bar{\hat{\theta}}$, mediante:

$$\bar{\hat{\theta}} = \frac{1}{M}\sum_{s=1}^{M}\hat{\theta}_s. \tag{5.3}$$

h) Calcular la desviación estándar de la estimación del parámetro θ con las M muestras, denotada con $Des.Est.(\hat{\theta})$, mediante:

$$Des.Est.(\hat{\theta}) = \sqrt{\frac{1}{M-1}\sum_{s=1}^{M}(\hat{\theta}_s - \bar{\hat{\theta}})^2}. \tag{5.4}$$

i) Aproximar la probabilidad de que $D_s = D$ dado que se conoce D, con las M muestras denotada con $\widehat{P(D_s = D|D)}$, mediante:

$$\widehat{P(D_s = D|D)} = \frac{\#(D_s = D|D)}{M}, \tag{5.5}$$

con # el número de muestras que cumplen con $(D_s = D|D)$.

j) Obtener el error relativo del estimador $\hat{\theta}$, denotado con $E.R.(\hat{\theta})$, mediante:

$$E.R.(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\theta}. \tag{5.6}$$

5.3. Descripción de los experimentos de simulación

Se consideran ocho experimentos de simulación para cada una de las dos poblaciones, bajo las siguientes condiciones:

1. Dos valores para la contribución de la parte residual a la variable respuesta.
   - a) $r = 0.2$, 20 % de contribución.
   - b) $r = 0.8$, 80 % de contribución.
2. Dos diseños de muestreo.
   - a) Máx. Ent.
   - b) MASSR
3. Dos tamaños de muestra.
   - a) $n = 100$, 2 % de la población.
   - b) $n = 250$, 5 % de la población.

En la figura 5.1 se da el diagrama de los 16 experimentos de simulación.
Los experimentos se plantean de esta forma con la finalidad de mostrar el desempeño de las

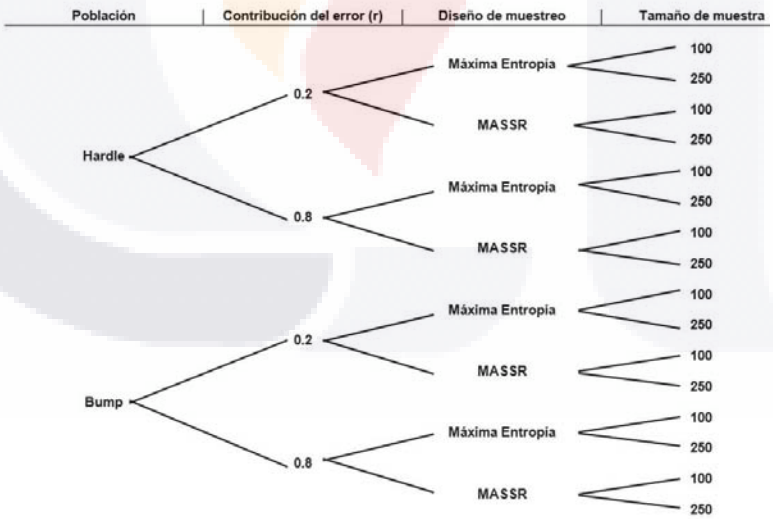

Figura 5.1: Experimentos de simulación.

estimaciones bajo los diferentes escenarios posibles.

Por ejemplo, con la primera población para $r = 0.8$, la decisión que seguramente se tomará en relación con la hipótesis nula será $D = 0$, esto significa que el ajuste del modelo a los datos no es adecuado. Lo cual se verá reflejado en el valor de $R^2(h)$, indicando así una pobre relación entre la variable respuesta y la variable auxiliar.

Luego, al seleccionar las M muestras bajo los diseños de muestreo para los diferentes tamaños de muestra, se esperaría que, se tome la misma decisión que con la población, con una alta probabilidad.

Capítulo 6

Resultados de las poblaciones hipotéticas

6.1. Introducción.

En este capítulo se dan los resultados obtenidos para las dos poblaciones descritas en el punto 6 del Algoritmo de Simulación. Primero se presentan los resultados para cada una de las poblaciones respecto del coeficiente de determinación, análisis de varianza y prueba global F para la regresión polinomial local. Además se muestra el desempeño local del coeficiente de determinación poblacional.
Después se realizan $3,000$ simulaciones para conocer el desempeño de las estimaciones de acuerdo con el diseño de muestreo y el tamaño de muestra dados en 5 y 4 de la sección de Variables Controladas respectivamente para las poblaciones ya mencionadas.

6.2. Población Hardle.

Cabe mencionar, que los cálculos de las tres primeras columnas del cuadro 6.1, se realizaron mediante las fórmulas del cuadro 3.1, y el cálculo del estadístico de prueba F se hace mediante la ecuación (3.16) y el $p-valor$ se obtiene con la F calculada con sus respectivos grados de libertad.

Cuadro 6.1: Tabla ANOVA para regresión polinomial local de la población Hardle con $r = 0.2$.

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	7	869.339900	124.191414	3.670000	0.0339897
Error	4,992	321.000000	0.064303		
Total	4,999	1,190.683300			

En el cuadro 6.1 se presenta la tabla del análisis de varianza para el ajuste del modelo

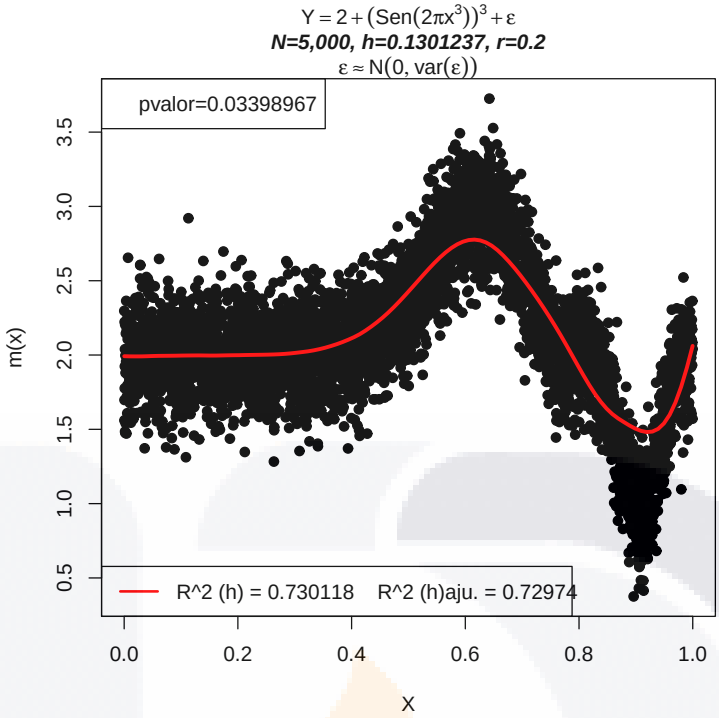

Figura 6.1: Se presenta la población Hardle, con una contribución de la parte residual de $r = 0.2$, y se muestra el ajuste del modelo por Regresión Polinomial Local a los datos con un coeficiente de determinación global de 0.730.

por regresión polinomial local de la población Hardle con $r = 0.2$. Se muestra un $p - valor$ de 0.033989, que para un nivel de significancia $\alpha = 0.05$ indica el rechazo de la hipótesis nula y por lo tanto el modelo ajustado a los datos es adecuado.

Cuadro 6.2: Tabla ANOVA para regresión polinomial local de la población Hardle con $r = 0.8$.

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	7	1,107.480300	158.211471	0.420000	0.868232
Error	4,992	3,580.000000	0.717147		
Total	4,999	4,687.837100			

En el cuadro 6.2, se presenta la tabla del análisis de varianza para el ajuste del modelo por regresión polinomial local de la población Hardle con $r = 0.8$. Se muestra un $p - valor$ de 0.8682, indica que la hipótesis nula no se rechaza y por lo tanto el modelo ajustado a los datos no es adecuado.

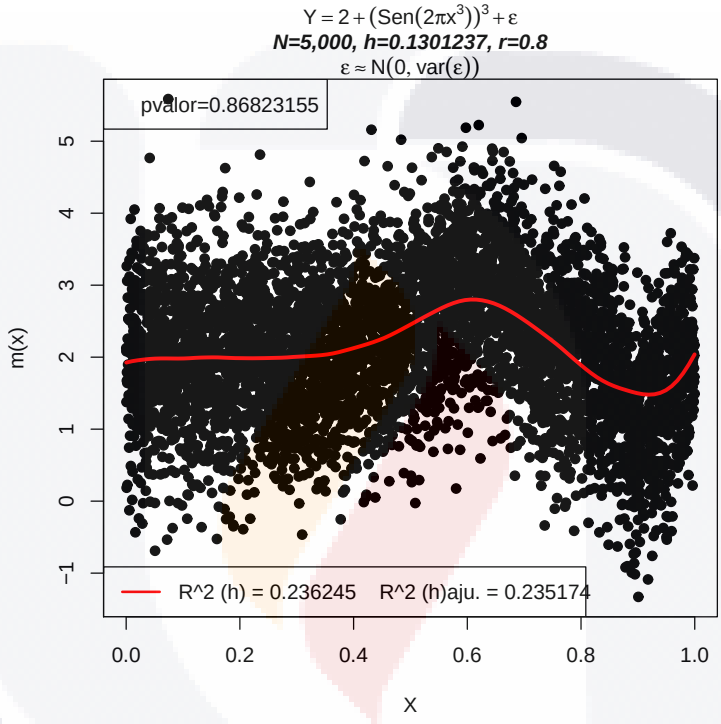

Figura 6.2: Se presenta la población Hardle, con una contribución de la parte residual de $r = 0.8$, y se muestra el ajuste del modelo por Regresión Polinomial Local a los datos con un coeficiente de determinación global de 0.2362.

6.3. Población Bump.

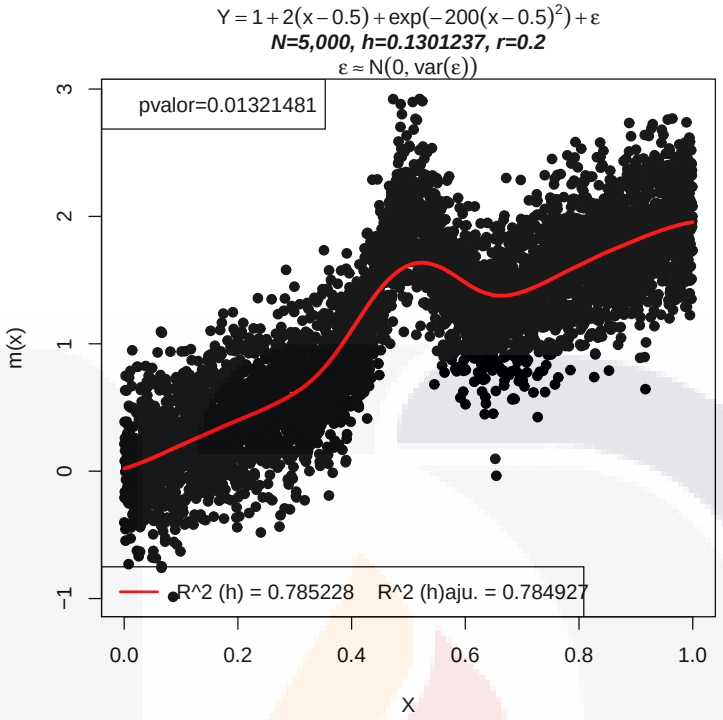

Figura 6.3: Se presenta la población Bump, con una contribución de la parte residual de $r = 0.2$, y se muestra el ajuste del modelo por regresión polinomial local a los datos con un coeficiente de determinación global de 0.7852.

Cuadro 6.3: Tabla ANOVA para regresión polinomial local de la población Bump con $r = 0.2$.

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	7	2,008.439800	286.919971	4.960000	0.013215
Error	4,992	549.000000	0.109976		
Total	4,999	2,557.778600			

En el cuadro 6.3, se presenta la tabla del análisis de varianza para el ajuste del modelo por regresión polinomial local de la población Bump con $r = 0.2$. Se muestra un $p - valor$ de 0.0132, que para un nivel de significancia $\alpha = 0.05$ indica el rechazo de la hipótesis nula y por lo tanto el modelo ajustado a los datos es adecuado.

En el cuadro 6.4, se presenta la tabla del análisis de varianza para el ajuste del modelo por regresión polinomial local de la población Bump con $r = 0.8$. Se muestra un $p - valor$

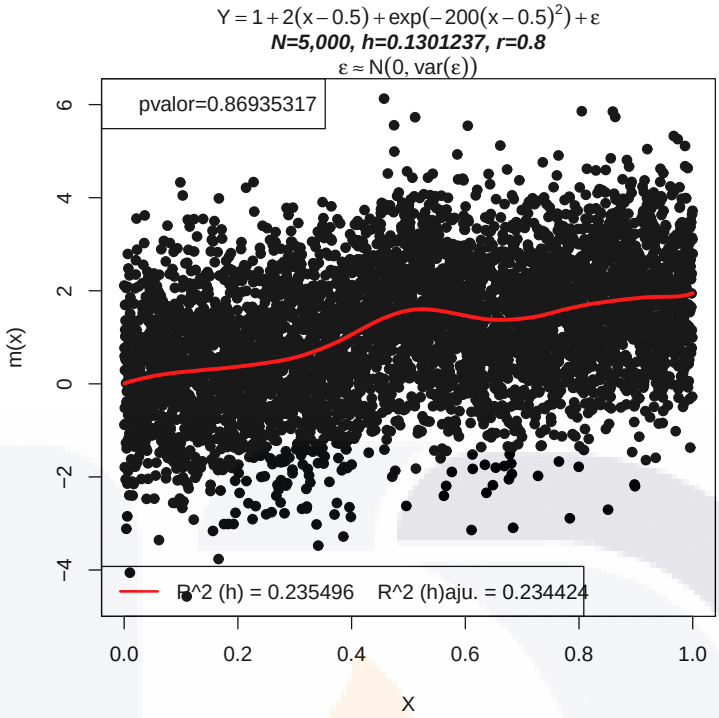

Figura 6.4: Se presenta la población Bump, con una contribución de la parte residual de $r = 0.8$, y se muestra el ajuste del modelo por regresión polinomial local a los datos con un coeficiente de determinación global de 0.2355.

Cuadro 6.4: Tabla ANOVA para regresión polinomial local de la población Bump con $r = 0.8$.

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p-valor
Regresión	7	2,411.149500	344.449929	0.420000	0.869353
Error	4,992	7,827.000000	1.567909		
Total	4,999	10,238.585300			

de 0.8694, indica que la hipótesis nula no se rechaza y por lo tanto el modelo ajustado a los datos no es adecuado.

6.4. Resultados de las 3,000 muestras para cada uno de los 16 experimentos.

Cuadro 6.5: Coeficiente de determinación estimado promedio, según diseño de muestreo.

Población	r	Diseño	n	$\widehat{\mathbf{R}^{\mathbf{2}}_{\pi}}(\mathbf{h})$	Des.Est.	$RECMR(\widehat{R^2_\pi}(h))$	$P(\widehat{D_s = D}\|D)$
Hardle	0.2	Máx.Ent.	100	0.751696	0.040418	0.062745	0.870000
			250	0.738758	0.025367	0.036698	0.910000
Hardle	0.2	MASSR	100	0.749861	0.040252	0.061397	0.870667
			250	0.738024	0.025365	0.036383	0.911333
Hardle	0.8	Máx.Ent.	100	0.314032	0.068398	0.438417	1.000000
			250	0.268009	0.043040	0.226404	1.000000
Hardle	0.8	MASSR	100	0.298585	0.065584	0.382979	1.000000
			250	0.258736	0.040269	0.195215	1.000000
Bump	0.2	Máx.Ent.	100	0.802817	0.036758	0.051888	0.990667
			250	0.792903	0.024043	0.032137	0.998000
Bump	0.2	MASSR	100	0.802330	0.030505	0.044532	0.998667
			250	0.792047	0.019414	0.026201	1.000000
Bump	0.8	Máx.Ent.	100	0.306195	0.078393	0.448222	1.000000
			250	0.265811	0.050916	0.251598	1.000000
Bump	0.8	MASSR	100	0.297057	0.065615	0.382023	0.998667
			250	0.259899	0.041380	0.203970	1.000000

En el cuadro 6.5, se presentan los resultados del coeficiente de determinación global estimado promedio, denotado con $\widehat{\widehat{R^2_\pi}}(h)$, calculado mediante la fórmula (5.3), la desviación estándar de $\widehat{R^2_\pi}(h)$ calculada con la fórmula (5.4), la raíz cuadrada del error cuadrático medio relativo $RECMR(\widehat{R^2_\pi}(h))$ calculado con la fórmula (5.2) y la probabilidad que la decisión tomada con la muestra sea igual a la decisión tomada con la población dado que se conoce la decisión tomada con la población, denotada con $P(\widehat{D_s = D}|D)$ y calculada mediante la fórmula (5.5), para las $3,000$ muestras y cada uno de los 16 experimentos.

Se observa que la desviación estándar de $\widehat{R^2_\pi}(h)$, disminuye alrededor del 37 % al incrementar el tamaño de muestra de $n = 100$ a $n = 250$, sin importar el diseño de muestreo y la población. Mientras que al comparar la desviación estándar de $\widehat{R^2_\pi}(h)$ entre los diseños de muestreo para la misma población y mismo tamaño de muestra, se observa que para la

población Hardle con $r = 0.2$, prácticamente no existe diferencia alguna. Para la población Hardle con $r = 0.8$, la desviación estándar de $\widehat{R^2_\pi}(h)$ con $n = 100$ es menor en 4 % con el diseño MASSR que con el diseño Máx. Ent., y menor en 6.4 % para $n = 250$ con el diseño MASSR para $n = 250$ que con el diseño Máx. Ent.
Similarmente, se observa para las poblaciones Bump con $r = 0.2$ y $r = 0.8$. Esto es, la desviación estándar de $\widehat{R^2_\pi}(h)$ es menor con el diseño MASSR que con el diseño Máx. Ent. para los dos tamaños de muestra.

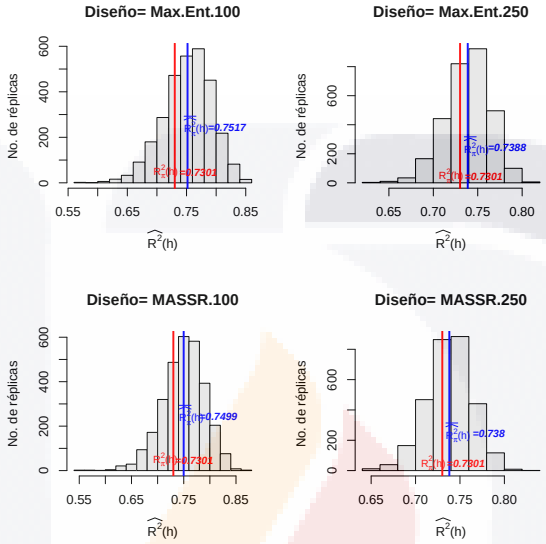

Figura 6.5: Histogramas de $\widehat{R^2_\pi(h)}$ con las 3,000 muestras de la población Hardle con $r = 0.2$, según diseño de muestreo.

La figura 6.5, muestra que para ambos diseños de muestreo, la distribución de $\widehat{R^2_\pi(h)}$ se aproxima a una distribución normal conforme se incrementa el tamaño de muestra.

La figura 6.6, muestra que para ambos diseños de muestreo, la distribución de $R^2_\pi(h)$ se aproxima a una distribución normal al incrementar el tamaño de muestra.

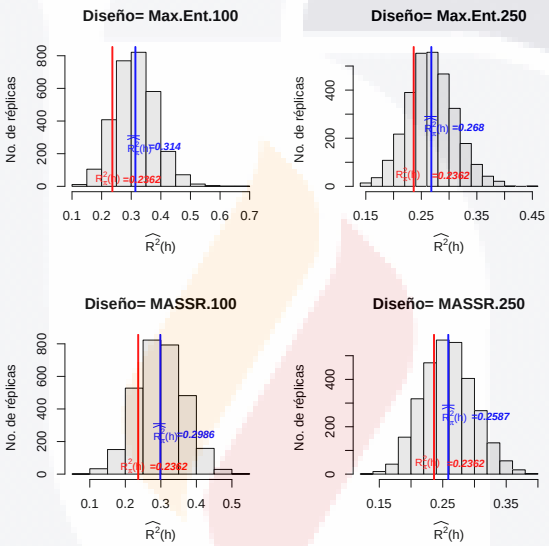

Figura 6.6: Histogramas de $\widehat{R^2_\pi(h)}$ con las 3,000 muestras de la población Hardle con $r = 0.8$, según diseño de muestreo.

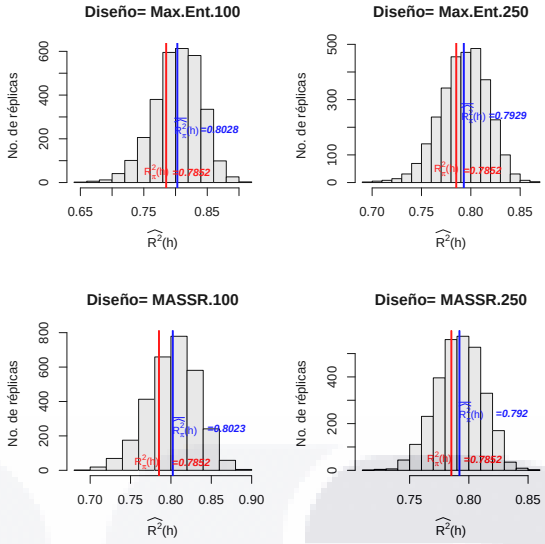

Figura 6.7: Histogramas de $\widehat{R^2_\pi(h)}$ con las 3,000 muestras de la población Bump con $r = 0.2$, según diseño de muestreo.

La figura 6.7, muestra que para el diseño de muestreo de máxima entropía con probabilidades desiguales, al incrementar el tamaño de muestra, la distribución de $\widehat{R^2_\pi(h)}$ se aproxima a una distribución normal con sesgo negativo. Mientras que para el diseño de muestreo aleatorio simple sin reemplazo, la distribución de $\widehat{R^2_\pi(h)}$ se aproxima a una distribución normal al aumentar el tamaño de muestra.

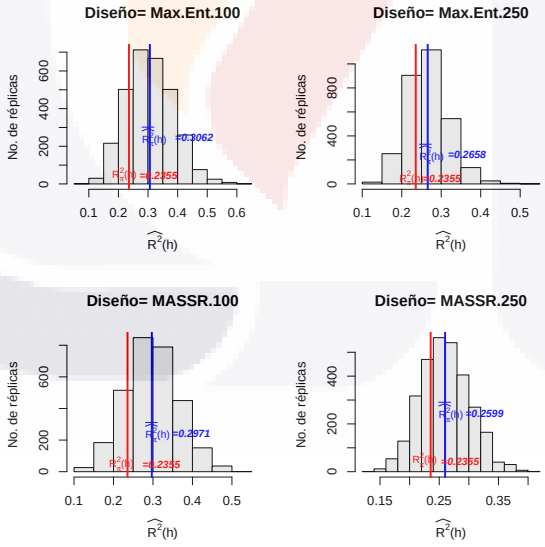

Figura 6.8: Histogramas de $\widehat{R^2_\pi(h)}$ con las 3,000 muestras de la población Bump con $r = 0.8$, según diseño de muestreo.

La figura 6.8, muestra que para el diseño de muestreo de máxima entropía, la distribución de $\widehat{R^2_\pi(h)}$ se aproxima a una distribución normal con sesgo positivo. En tanto que para el

diseño de muestreo aleatorio sin reemplazo, la distribución de $\widehat{R^2_\pi(h)}$ se aproxima a la normal al aumentar el tamaño de muestra.

Se observa que para un tamaño de muestra $n = 250$, los diagramas son más compactos y menos dispersos que con un tamaño de muestra $n = 100$, y los diagramas son similares entre diseños de muestreo para el mismo tamaño de muestra.

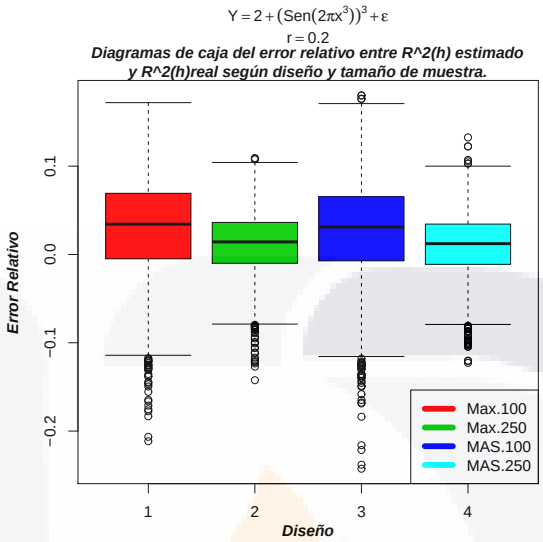

Figura 6.9: Diagrama de caja y brazos para las 3,000 muestras de la población Hardle con $r = 0.2$.

La figura 6.9, muestra que $\widehat{R^2_\pi(h)}$ sobreestima ligeramente a $R^2(h)$. Además el aumento del tamaño de muestra conlleva una reducción en el sesgo y varianza del estimador mencionado.

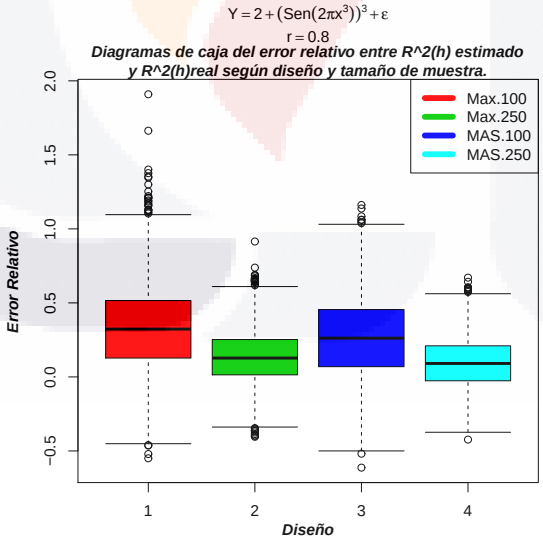

Figura 6.10: Diagrama de caja y brazos para las 3,000 muestras de la población Hardle con $r = 0.8$.

La figura 6.10, muestra que $\widehat{R^2_\pi(h)}$ sobreestima ligeramente a $R^2(h)$. Y el aumento del

tamaño de muestra conlleva una reducción en el sesgo y varianza del estimador mencionado.

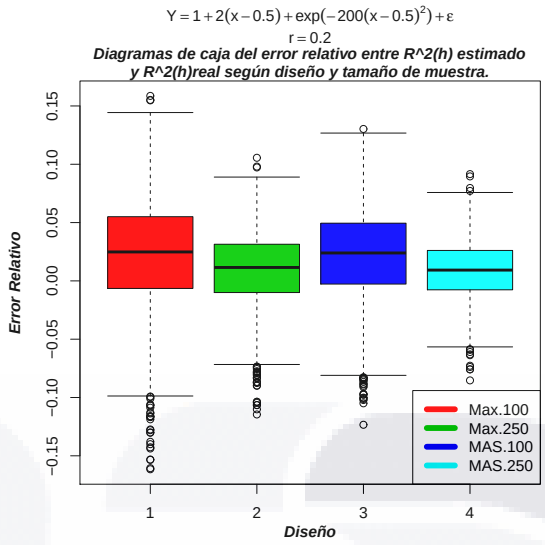

Figura 6.11: Diagrama de caja y brazos para las 3,000 muestras de la población Bump con $r = 0.2$.

La figura 6.11, muestra que $\widehat{R^2_\pi}(h)$ sobreestima ligeramente a $R^2(h)$. Además el aumento del tamaño de muestra conlleva una reducción en el sesgo y varianza del estimador mencionado.

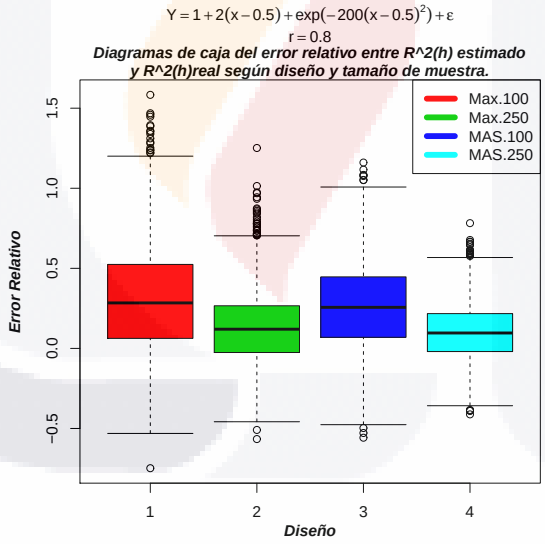

Figura 6.12: Diagrama de caja y brazos para las 3,000 muestras de la población Bump con $r = 0.8$.

La figura 6.12, muestra que $\widehat{R^2_\pi}(h)$ sobreestima ligeramente a $R^2(h)$. Además el aumento del tamaño de muestra conlleva una reducción en el sesgo y varianza del estimador mencionado.

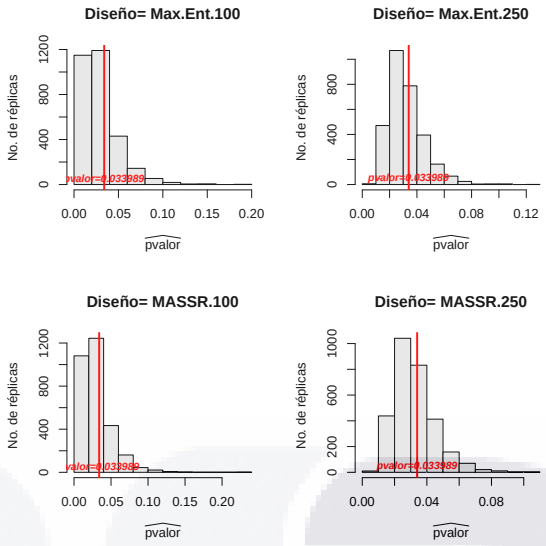

Figura 6.13: Histogramas de $\widehat{p-valor}$ con las 3,000 muestras de la población Hardle con $r = 0.2$, según diseño de muestreo.

La figura 6.13, muestra que la distribución de $\widehat{p-valor}$ es asimétrica positiva para ambos diseños de muestreo. Y el aumento del tamaño de muestra hace disminuir $\widehat{p-valor}$. Además se observa que la mayoría de las muestras dan un $p-valor$ estimado inferior que el nivel de significancia $\alpha = 0.05$, por lo tanto en la mayoría de las muestras, el modelo ajustado a los datos mediante regresión polinomial local es adecuado. Esto se puede ver en el cuadro 6.5 en la columna $\widehat{P(D_s = D|D)}$.

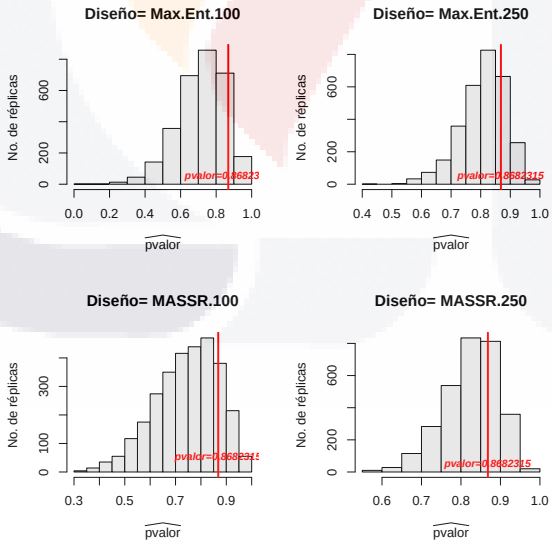

Figura 6.14: Histogramas de $\widehat{p-valor}$ con las 3,000 muestras de la población Hardle con $r = 0.8$, según diseño de muestreo.

La figura 6.14, muestra que la distribución de $\widehat{p-valor}$ es asimétrica negativa para los

dos diseños de muestreo. Se observa que la mayoría de las muestras dan un $p-valor$ estimado superior que el nivel de significancia $\alpha = 0.05$, lo que significa que en la mayoría de las muestras el modelo ajustado a los datos mediante regresión polinomial local no es adecuado. Esto se puede ver en el cuadro 6.5 en la columna $P(\widehat{D_s = D}|D)$.

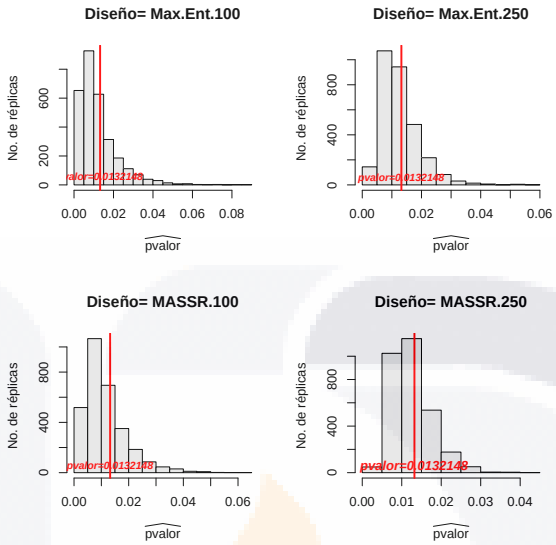

Figura 6.15: Histogramas de $\widehat{p-valor}$ con las 3,000 muestras de la población Bump con $r = 0.2$, según diseño de muestreo.

La figura 6.15, muestra que la distribución de $\widehat{p-valor}$ es asimétrica positiva para ambos diseños de muestreo. Y el aumento del tamaño de muestra hace que $\widehat{p-valor}$ disminuya. Además se observa que la mayoría de las muestras dan un $p-valor$ estimado inferior que el nivel de significancia $\alpha = 0.05$, y en la mayoría de las muestras el modelo ajustado a los datos mediante regresión polinomial local es adecuado. Esto se puede ver en el cuadro 6.5 en la columna $P(\widehat{D_s = D}|D)$.

La figura 6.16, muestra que la distribución de $\widehat{p-valor}$ es asimétrica negativa para ambos diseños de muestreo. Además se observa que la mayoría de las muestras dan un $p-valor$ estimado superior que el nivel de significancia $\alpha = 0.05$, y por lo tanto significa que en la mayoría de las muestras el modelo ajustado a los datos mediante regresión polinomial local no es adecuado. Esto se puede ver en el cuadro 6.5 en la columna $P(\widehat{D_s = D}|D)$.

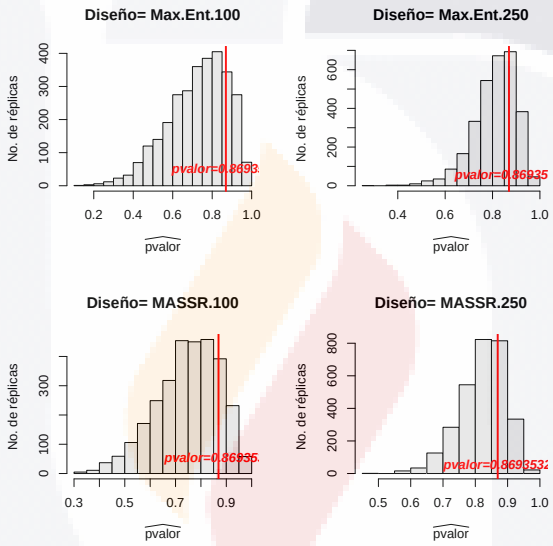

Figura 6.16: Histogramas de $p - \widehat{valor}$ con las 3,000 muestras de la población Bump con $r = 0.8$, según diseño de muestreo.

6.5. Desempeño local

En las siguientes gráficas se muestra el desempeño del coeficiente de determinación local para las 3,000 muestras seleccionadas de las poblaciones, según diseño de muestreo y tamaño de muestra.

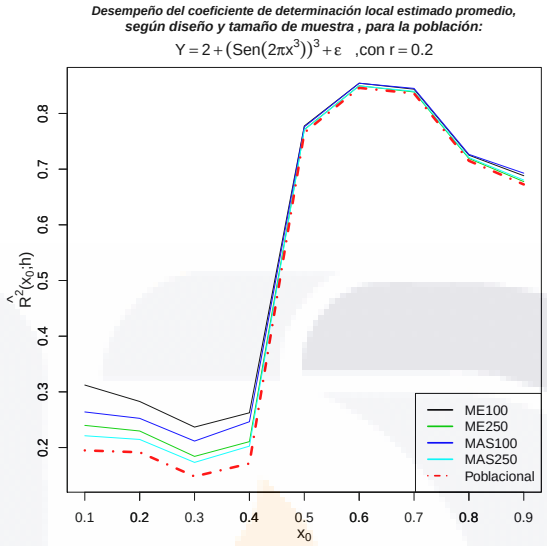

Figura 6.17: Desempeño de $\hat{R}^2(x_0, h)$ de las 3,000 muestras

En la figura 6.17, se compara el desempeño del coeficiente de determinación local estimado con cada diseño de muestreo y tamaño de muestra, y el desempeño del coeficiente de determinación local en la población. Se observa que las 3,000 muestras reproducen bien el desempeño del coeficiente de determinación local para ambos diseños de muestreo y tamaños de muestra. Además, para valores de la variable auxiliar entre 0.0 y 0.4, la estimación del coeficiente de determinación local fluctua alrededor de 0.25, que indica que el comportamiento de la variable respuesta es constante para cualquier valor de la variable auxiliar. Mientras que para valores superiores a 0.4, la estimación del coeficiente de determiandción local se incrementa.

En la figura 6.18, se compara el desempeño del coeficiente de determinación local estimado con cada diseño de muestreo y tamaño de muestra, y el desempeño del coeficiente de determinación local en la población. Se observa que las 3,000 muestras reproducen bien el desempeño del coeficiente de determinación local para ambos diseños de muestreo y tamaños de muestra. Además, para valores de la variable auxiliar entre 0.0 y 0.4, la estimación del coeficiente de determinación local fluctua alrededor de 0.1, que indica que el comportamiento de la variable respuesta es constante para cualquier valor de la variable auxiliar. Mientras que para valores superiores a 0.4, la estimación del coeficiente de determiandción local se incrementa.

En la figura 6.19, se compara el desempeño del coeficiente de determinación local estimado con el desempeño del coeficiente de determinación local en la población. Se observa que para valores de la variable auxiliar menores a 0.7, el coeficiente de determinación local estimado disminuye, y para valores superiores a 0.7, éste se incrementa.

En la figura 6.20, se compara el desempeño del coeficiente de determinación local estimado con el desempeño del coeficiente de determinación local en la población. Se observa que para valores de la variable auxiliar menores a 0.4, el coeficiente de determinación local estimado

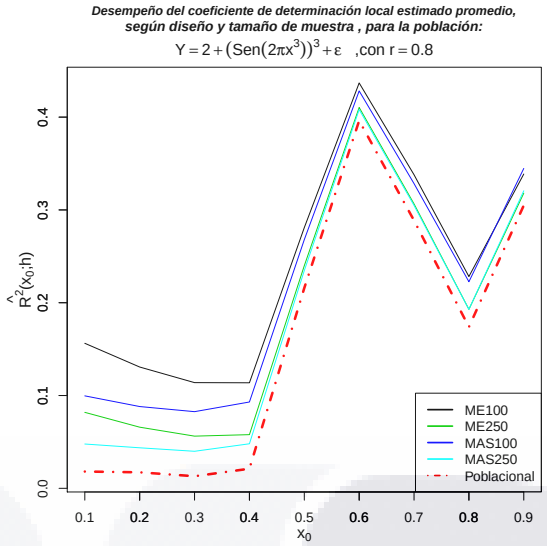

Figura 6.18: Desempeño de $\hat{R}^2(x_0, h)$ de las 3,000 muestras

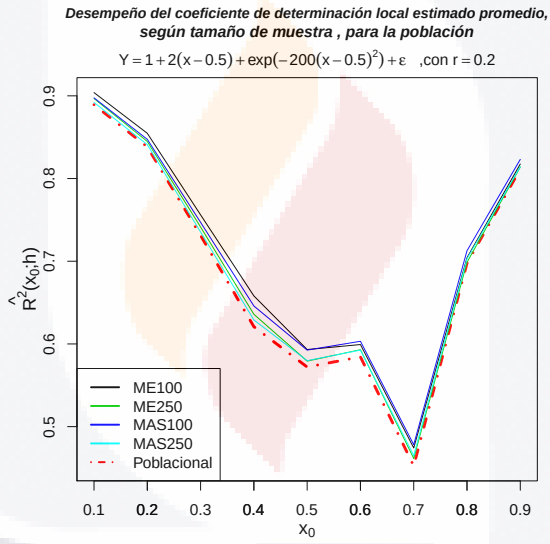

Figura 6.19: Desempeño de $\hat{R}^2(x_0, h)$ de las 3,000 muestras

disminuye, para valores entre 0.4 y 0.5, el coeficiente de determinación local estimado muestra un incremento leve, para valores entre 0.5 y 0.7, el coeficiente de determinación local estimado disminuye, y para valores superiores a 0.7, el coeficiente de determinación local estimado aumenta.

Por lo anterior, el desempeño de la estimación del coeficiente de determinación local, reproduce adecuadamente el comportamiento local de la variable respuesta en término de la variación que explica la variable auxiliar mediante un modelo lineal local.

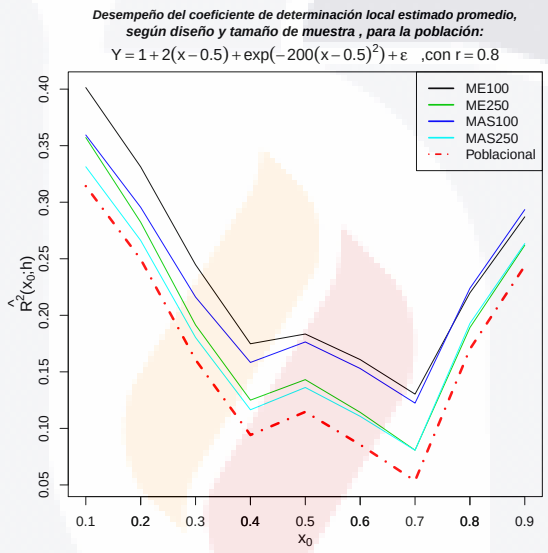

Figura 6.20: Desempeño de $\hat{R}^2(x_0, h)$ de las 3,000 muestras

Capítulo 7

Aplicación a datos reales

7.1. Primera población

La primera población está formada por 724 municipios de México para los años 2003 y 2008 referentes a unidades económicas en actividades comerciales. Se excluyerón municipios con información confidencial, los que no empataron en ambos años, con más de $20,000$ unidades económicas, con menos de $5,000$ viviendas y con más de $10,000,000$ en activos fijos por municipio.

Las variables que se consideran son las siguientes:

- Variable auxiliar (X).- Número de unidades económicas en actividades comerciales por municipio en 2003.
- Variable respuesta (Y).- Total de activos fijos (miles de pesos) por municipio en 2008.
- Tamaño de las unidades (Z).- Total de viviendas por municipio en 2005. [9]

Fuente de las variables X, Y [10]

En la figura 7.1, se muestra el diagrama de dispersión de la población y se observa una relación creciente entre las variables, esto es, a mayor número de unidades económicas por municipio, mayor es el valor total de los activos fijos.
Se puede describir está tendencia mediante un modelo de regresión lineal, sin embargo los datos están más dispersos conforme crece X.
Otra alternativa para describir dicha tendencia es aplicar regresión polinomial local.

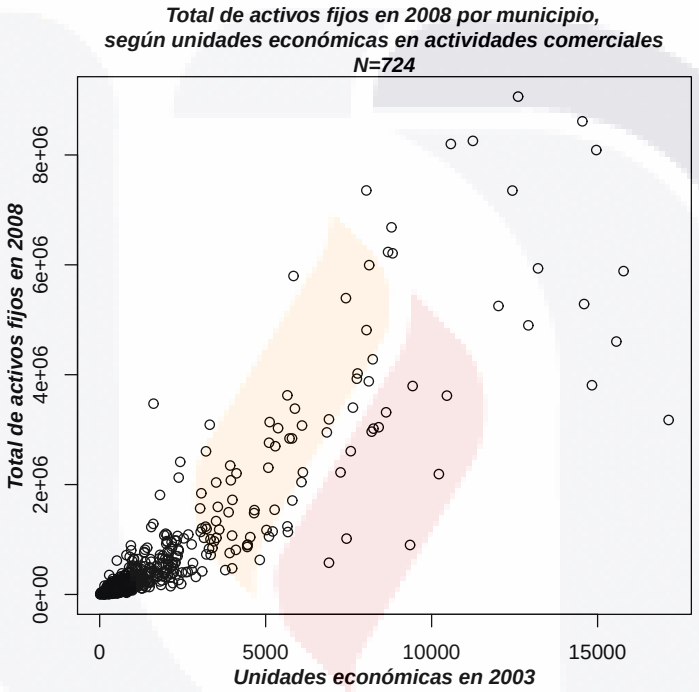

Figura 7.1: Población de municipios

7.2. Descripción de la simulación.

Con el objetivo de mostrar los resultados que se pueden obtener, mediante estimadores por regresión polinomial local en muestreo de poblaciones finitas, para el coeficiente de determinación global, el análisis de varianza y el promedio de la variable de estudio, se presentan los resultados obtenidos al relizar $3,000$ simulaciones para tres tamaños de muestra bajo un diseño de muestreo de máxima entropía con probabilidades desiguales.

Se consideran los siguientes elementos :

1. Tamaño de la población $N = 724$.
2. Número de parámetros equivalente 4.
3. Ancho de banda $h = 0.3$.
4. Tamaños de muestra $n = 35$, $n = 50$ y $n = 70$.
5. Diseño de muestreo de máxima entropía de probabilidades desiguales.
6. Parámetros de interés:
   - *a*) $\theta_1 = R^2(h)$ Coeficiente de determinación global.
   - *b*) $\theta_2 = R^2(x_0; h)$ Coeficiente de determinación local.
   - *c*) $\theta_3 = \mu_y$ Promedio del total de activos fijos por municipio.
   - *d*) $\theta_4 = F_c(h)$ Estadístico de prueba.
   - *e*) $\theta_5 = pvalor$ p-valor asociado con el estadístico de prueba calculado.
   - *f*) $\theta_6 = P\left(D_s = D | D = d\right)$ Probabilidad de tomar la misma decisión con la muestra (D_s) que la decisión tomada con la población $(D = d)$. Donde d puede tomar dos valores:
     - $d = 0$ El ajuste del modelo a los datos no es adecuado.
     - $d = 1$ El ajuste del modelo a los datos es adecuado.

7.3. Resultados

En esta sección se presentan los resultados de la simulación. Primero se dan los resultados para la población y después para las 3,000 muestras con cada tamaño de muestra.

En la figura 7.2, se muestra el diagrama de dispersión de la población con los ajustes por regresión lineal y regresión polinomial local. Se tiene que con regresión lineal el coeficiente de determinación es 0.7880, mientras que con regresión polinomial local es 0.8311.
Si bien la diferencia no es grande, se observa que mediante regresión polinomial local, el comportamiento de los datos, parece que se describe mejor.

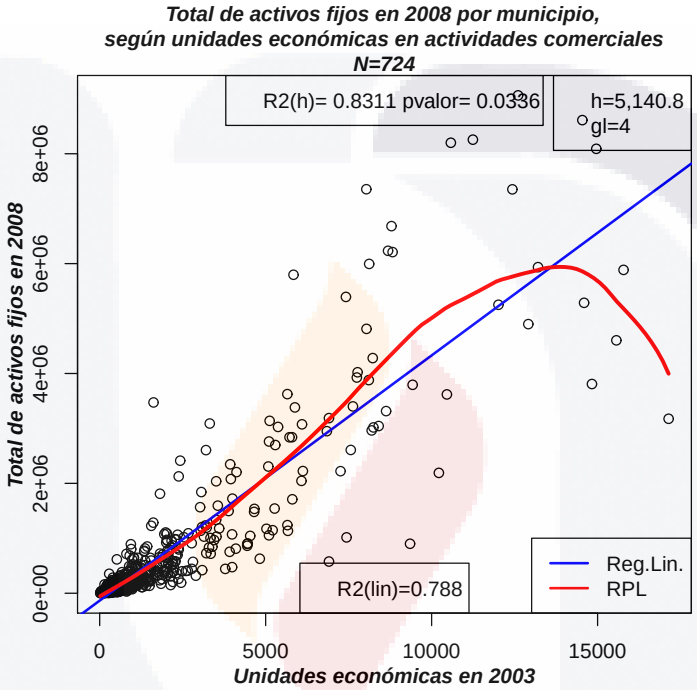

Figura 7.2: Población de municipios con ajustes

Cuadro 7.1: Tabla ANOVA de la regresión polinomial local para la población

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	$F(h)$	p-valor	Decisión
$m(x_k)$	3	$9.230135e^{14}$	$1.846027e^{14}$	7.38	0.033637	1
ϵ_k	720	$1.875777e^{14}$	$2.605246e^{11}$			
Total	723	$1.110591e^{15}$				

La tabla ANOVA se muestra en el cuadro 7.1. Se tiene que el modelo ajustado a los datos de la población mediante regresión polinomial local es adecuado. Esto se resume en la columna *Decisón* con el valor de uno.

$\theta_1 = R^2(h)$ **y** $\theta_3 = \mu_y$

La medida de bondad de ajuste, del modelo estimado por regresión polinomial local a los datos, es el coeficiente de determinación global, denotado con $R^2(h)$. En el cuadro 7.2 está el coeficiente de determinación global, el coeficiente de determinación ajustado de la población para regresión polinomial local, y el coeficiente de determinación para regresión lineal.

Cuadro 7.2: Coeficiente de determinación global de la población

$R^2(h)$	$R^2(h)_{ajustado}$	$R^2(lineal)$
0.831101	0.830397	0.788039

El promedio del total de activos fijos por municipio (miles de pesos) para la población es de $\mu_y = 575,910$.

Estimación de $\theta_1 = R^2(h)$ para $3,000$ muestras.

Se comparan los métodos de estimación del coeficiente de determinación y del promedio para los tres tamaños de muestra.

En el cuadro 7.3, se presentan las estadísticas para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según tamaño de muestra.

Cuadro 7.3: Estadísticas de $\widehat{R^2_{\pi}}(h)$,$\widehat{R^2_{\pi Aj}}(h)$, $\widehat{R^2_{lin}}$

	$\widehat{R^2_{\pi}}(h)$			$\widehat{R^2_{\pi Aj}}(h)$			$\widehat{R^2_{lin}}$		
n	35	50	70	35	50	70	35	50	70
Mín	0.5984	0.6727	0.7255	0.5966	0.6710	0.7243	0.4919	0.6013	0.6177
Q_1	0.8225	0.8206	0.8204	0.8220	0.8200	0.8197	0.7152	0.7233	0.7192
Q_2	0.8574	0.8477	0.8400	0.8570	0.8472	0.8394	0.7658	0.7604	0.7466
Media	0.8526	0.8458	0.8397	0.8521	0.8452	0.8392	0.7612	0.7582	0.7481
Q_3	0.8890	0.8737	0.8609	0.8886	0.8732	0.8605	0.8119	0.7957	0.7772
Máx.	0.9496	0.9337	0.9229	0.9495	0.9335	0.9227	0.9323	0.9118	0.8748
Des.Est.	0.0466	0.0371	0.0298	0.0468	0.0372	0.0299	0.0676	0.0516	0.0399

La figura 7.3, muestra que la estimación del coeficiente de determinación global por regresión polinomial local sobreestima levemente a $R^2(h)$ y al considerar un modelo no paramétrico para describir el comportamiento del total de activos fijos por municipio, el modelo de regresión lineal no logra captar dicho comportamiento.

La línea horizontal continua, representa el coeficiente de determinación ajustado para la población.

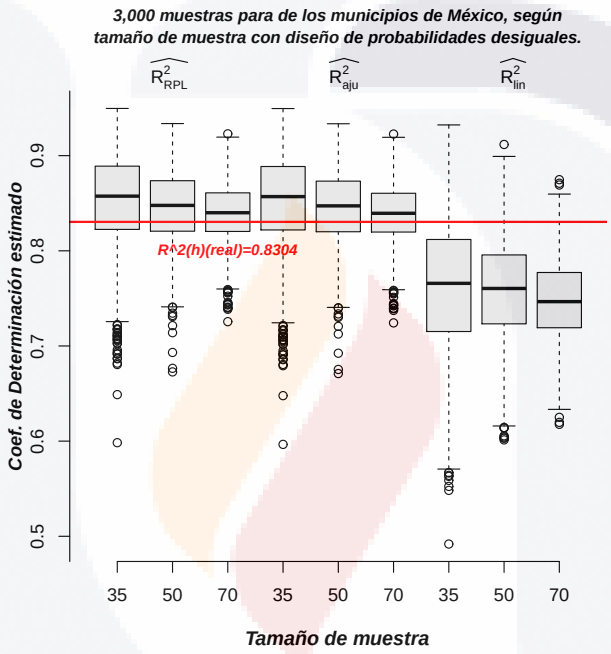

Figura 7.3: $\widehat{R^2_\pi}(h)$

Para analizar el comportamiento de las estimaciones del coeficiente de determinación (R^2), obtenemos el error relativo (E.R.) y la raíz cuadrada del error cuadrático medio relativo (R.E.C.M.R) para el estimador de R^2 mediante regresión polinomial local ($\widehat{R^2_\pi}(h)$) y regresión lineal ($\widehat{R^2_{lin}}$).

En el cuadro 7.4, se presentan las estadísticas del E.R. para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según tamaño de muestra.

Cuadro 7.4: Estadísticas de $E.R.\left(\widehat{R^2_\pi}(h)\right)$ y $E.R.\left(\widehat{R^2_{lin}}\right)$

	$\widehat{R^2_\pi}(h)$			$\widehat{R^2_{lin}}$		
n	35	50	70	35	50	70
Mín.	-0.28005	-0.19054	-0.12701	-0.40818	-0.27653	-0.25681
Q_1	-0.01034	-0.01261	-0.01291	-0.13948	-0.12976	-0.13469
Q_2	0.03169	0.02000	0.01073	-0.07855	-0.08504	-0.10168
Media	0.02590	0.01763	0.01038	-0.08411	-0.08770	-0.09987
Q_3	0.06962	0.05122	0.03587	-0.02309	-0.04264	-0.06487
Máx.	0.14261	0.12347	0.11050	0.12177	0.09704	0.05254
Des.Est.	0.05612	0.04465	0.03588	0.08129	0.06211	0.04799

La figura 7.4, muestra que la varianza y sesgo del estimador del coeficiente de determinación global para la regresión polinomial local es menor que para la regresión lineal. Además el estimador para la regresión polinomial local sobreestima levemente al coeficiente de determinación global, mientras que el estimador para la regresión lineal lo subestima.

En el cuadro 7.5, se presentan el R.E.C.M.R. para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según tamaño de muestra. Del cuadro 7.5,

Cuadro 7.5: $R.E.C.M.R.\left(\widehat{R^2_\pi}(h)\right)$ y $R.E.C.M.R.\left(\widehat{R^2_{lin}}\right)$

n	$\widehat{R^2_\pi}(h)$	$\widehat{R^2_{lin}}$	Razón
35	0.06179	0.11697	1.89302
50	0.04799	0.10746	2.23922
70	0.03734	0.11079	2.96706

se tiene que al incrementar el tamaño de muestra, $R.E.C.M.R.\left(\widehat{R^2_\pi}(h)\right)$ dismunuye, no así $R.E.C.M.R.\left(\widehat{R^2_{lin}}\right)$. Además, la razón entre $R.E.C.M.R.\left(\widehat{R^2_{lin}}\right)$ y $R.E.C.M.R.\left(\widehat{R^2_\pi}(h)\right)$ es mayor que uno para los tres tamaños de muestra. Es decir, para $n = 35$ $R.E.C.M.R.\left(\widehat{R^2_{lin}}\right)$ es 1.89 veces mayor que $R.E.C.M.R.\left(\widehat{R^2_\pi}(h)\right)$, para $n = 50$ es 2.24 veces y para $n = 70$ es 2.97 veces mayor.

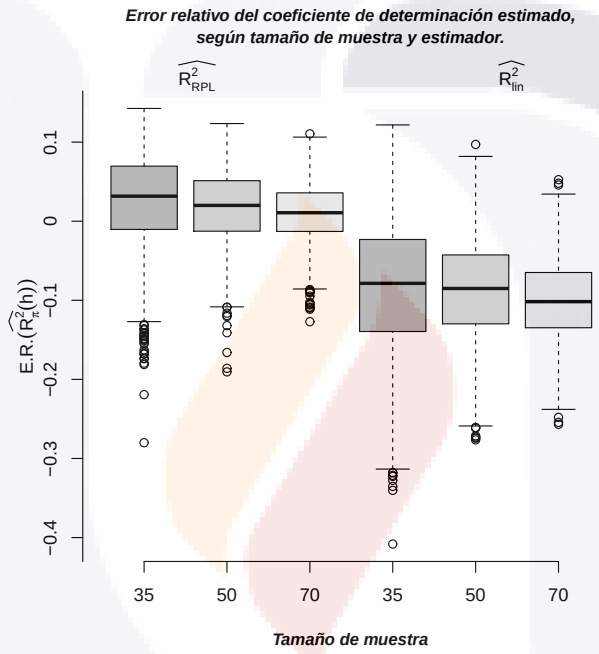

Figura 7.4: Error relativo de $\widehat{R^2_\pi}(h)$

Estimación de $\theta_2 = R^2(x_0; h)$ para $3,000$ muestras.

Se compara el desempeño del coeficiente de determinación local, de acuerdo con el tamaño de muestra, para la estimación por regresión polinomial local en muestreo de poblaciones finitas.

La figura 7.5, presenta el desempeño del coeficiente de determinación local para la población y los promedios del coeficiente de determinación local estimado con las $3,000$ muestras con tamaños de $n = 35$, $n = 50$ y $n = 70$ respectivamente.

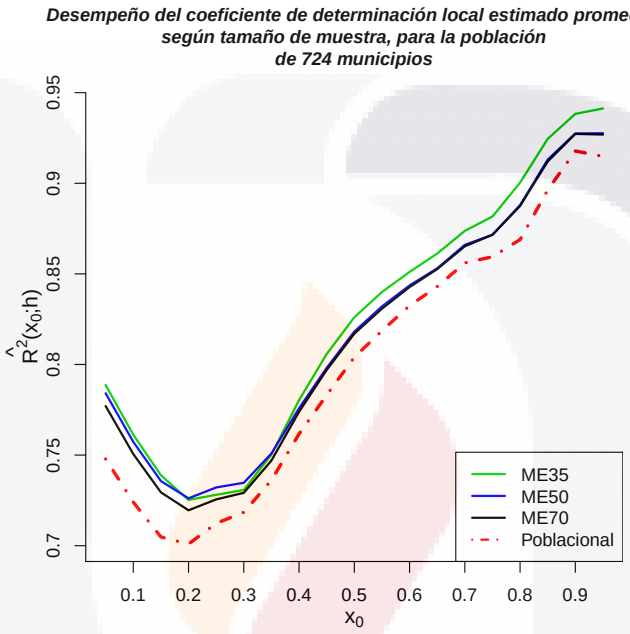

Figura 7.5: $\widehat{R^2_\pi}(x_0; h)$

Se observa, que el desempeño de $\widehat{R^2}(x_0; h)$ se aproxima a $R^2(x_0; h)$, al incrementar el tamaño de muestra. Además, para valores de x_0 alrededor de 0.2, $R^2(x_0; h)$ es aproximadamanete 0.7, cuyo valor es el mínimo que se observa en el rango de la variable auxiliar X. Esto, debido a que en esa región, se encuentra la mayoría de los municipios y el comportamiento local del total de activos fijos muestra una relación lineal moderada.

Estimación de $\theta_3 = \mu_y$ para $3,000$ muestras.

Se comparan los métodos de estimación del promedio del total de activos fijos por municipio para los tres tamaños de muestra. Los métodos que se utilizan para obtener $\widehat{\mu_y}$ son:

1. Regresión Polinomial local $\widehat{\mu_{RPL}}$, se calcula mediante la ecuación (2.11).
2. Estimador de Horvitz-Thompson $\widehat{\mu_{HT}}$, se calcula mediante la ecuación (2.9).
3. Regresión Lineal $\widehat{\mu_{Reg}}$, se calcula mediante la ecuación (2.10).

En el cuadro 7.6, se presentan las estadísticas para las estimaciones del promedio de los activos fijos por municipio con las $3,000$ muestras, según estimador y tamaño de muestra.

Cuadro 7.6: Estadísticas de $\widehat{\mu_{RPL}}$, $\widehat{\mu_{HT}}$, $\widehat{\mu_{Reg}}$

	$\widehat{\mu_{RPL}}$			$\widehat{\mu_{HT}}$			$\widehat{\mu_{Reg}}$		
n	35	50	70	35	50	70	35	50	70
Mín.	347,000	373,300	395,900	278,800	298,000	360,500	239,600	302,700	395,600
Q_1	516,200	526,600	531,500	499,900	514,600	522,000	509,800	520,800	526,300
Q_2	581,500	582,300	576,100	583,000	584,200	578,200	580,400	581,800	577,100
Media	597,500	592,400	582,100	602,900	597,100	585,200	603,100	595,900	584,900
Q_3	663,300	647,800	624,900	683,400	662,900	637,600	667,600	652,000	630,400
Máx.	1,189,000	1,031,000	859,900	1,471,000	1,199,000	936,500	1,479,000	1,228,000	1,009,000
Des.Est.	115,809.01	94,423.79	71,005.31	144,895.16	118,347.51	88,290.85	137,450.19	109,868.18	81,730.78

La figura 7.6, muestra que la estimación del promedio del total de activos fijos por municipio mediante regresión polinomial local, tiene más precisión al aumentar el tamaño de muestra que mediante el estimador de Horvitz-Thompson y regresión lineal.

La línea horizontal discontinua, representa el promedio de los activos fijos por municipio de la población.

Para analizar el comportamiento de las estimaciones del promedio del total de activos fijos por municipio (μ_y), obtenemos el error relativo (E.R.) y la raíz cuadrada del error cuadrático medio relativo (R.E.C.M.R) para el estimador de μ_y mediante regresión polinomial local ($\widehat{\mu_{RPL}}$), Estimador de Horvitz-Thompson ($\widehat{\mu_{HT}}$) y regresión lineal ($\widehat{\mu_{Reg}}$) .

En el cuadro 7.7, se presentan las estadísticas del E.R. para las estimaciones con las $3,000$ muestras, según estimador y tamaño de muestra.

La figura 7.7, muestra que las estimaciones del promedio del total de los activos fijos por municipio mediante las tres metodologías son insesgadas y consistentes.

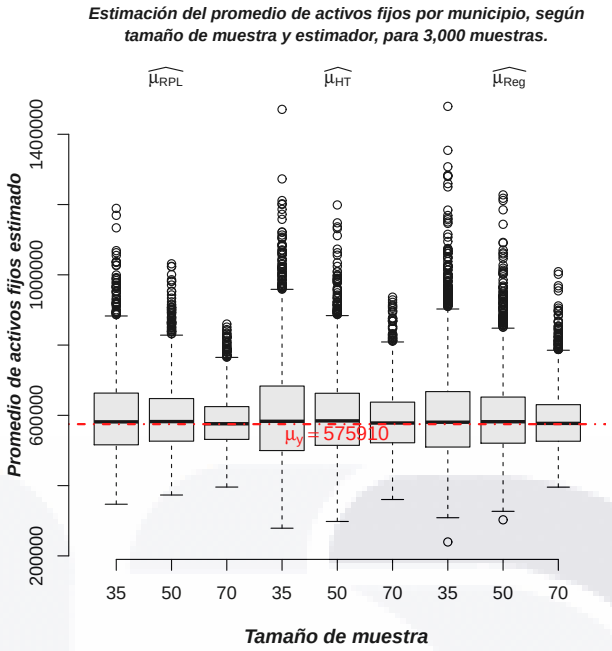

Figura 7.6: Promedio estimado de activos fijos

Cuadro 7.7: Estadísticas de $E.R.\left(\widehat{\mu_y}\right)$, según método y tamaño de muestra.

	$\widehat{\mu_{RPL}}$			$\widehat{\mu_{HT}}$			$\widehat{\mu_{Reg}}$		
n	35	50	70	35	50	70	35	50	70
Mín.	-0.39739	-0.35188	-0.31258	-0.51592	-0.48260	-0.37409	-0.58398	-0.47446	-0.31301
Q_1	-0.10374	-0.08568	-0.07708	-0.13194	-0.10647	-0.09363	-0.11481	-0.09564	-0.08611
Q_2	0.00969	0.01103	0.00032	0.01236	0.01444	0.00398	0.00787	0.01029	0.00211
Media	0.03755	0.02862	0.01075	0.04687	0.03677	0.01619	0.04714	0.03469	0.01565
Q_3	0.15181	0.12476	0.08506	0.18667	0.15096	0.10715	0.15924	0.13218	0.09465
Máx.	1.06474	0.79093	0.49311	1.55403	1.08168	0.62607	1.56885	1.13147	0.75193
Des.Est.	0.20109	0.16396	0.12329	0.25159	0.20550	0.15331	0.23867	0.19077	0.14192

En el cuadro 7.8, se presentan la R.E.C.M.R. para las estimaciones del promedio del total de los activos fijos por municipio, con las $3,000$ muestras, según estimador y tamaño de muestra. Del cuadro 7.8, se tiene que al incrementarse el tamaño de muestra, $R.E.C.M.R.\left(\widehat{\mu_{RPL}}\right)$, $R.E.C.M.R.\left(\widehat{\mu_{HT}}\right)$ y $R.E.C.M.R.\left(\widehat{\mu_{Reg}}\right)$ dismunuyen. Sin embargo, las razones entre $R.E.C.M.R.\left(\widehat{\mu_{HT}}\right)$, $R.E.C.M.R.\left(\widehat{\mu_{Reg}}\right)$ y $R.E.C.M.R.\left(\widehat{\mu_{RPL}}\right)$ son mayores que uno para los tres tamaños de muestra. Es decir, $R.E.C.M.R.\left(\widehat{R^2_{HT}}\right)$ es 1.25 veces mayor que $R.E.C.M.R.\left(\widehat{\mu_{RPL}}\right)$ para los tres tamaños de muestra. Mientras que $R.E.C.M.R.\left(\widehat{R^2_{Reg}}\right)$ es 1.19 veces mayor que $R.E.C.M.R.\left(\widehat{\mu_{RPL}}\right)$ para $n = 35$, para $n = 50$ es 1.17 veces mayor, y para $n = 70$ es 1.15 veces mayor.

Por tanto, al contar con información auxiliar, unidades económicas por municipio para el año 2004, para todos los municipios de la población, y al existir una relación con la variable de

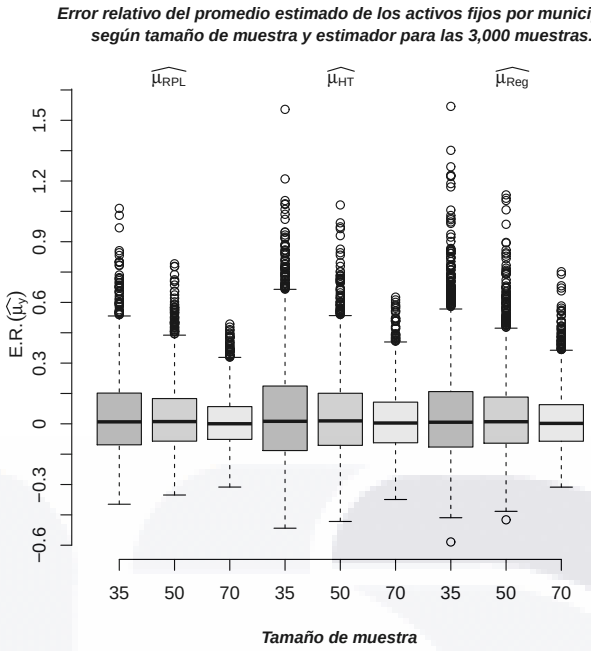

Figura 7.7: Error relativo de $\widehat{\mu_y}$

estudio (total de activos fijos por municipio), la cual puede captarse mediante algún modelo de regresión, o a través de un modelo no paramétrico como la regresión polinomial local, y puede utilizarse para mejorar la estimación del total y del promedio de la variable de estudio, en el sentido de disminuir la varianza y el sesgo de las estimaciones. Mientras que, si bien al utilizar sólo información del diseño de muestreo, como es con el estimador de Horvitz-Thompson, las estimaciones son insesgadas, éstas son menos precisas.

Cuadro 7.8: $R.E.C.M.R.(\widehat{\mu_y})$,según estimador.

n	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$	Razón(HT)	Razón (Reg)
35	0.20453	0.25588	0.24324	1.25106	1.18926
50	0.16640	0.20873	0.19387	1.25439	1.16508
70	0.12374	0.15413	0.14275	1.24559	1.15363

Prueba global F

Se realiza la prueba global del ajuste del modelo para la regresión polinomial local, para cada una de las $3,000$ muestras seleccionadas bajo un diseño de máxima entropía de probabilidades desiguales y tres tamaños de muestra.

En el cuadro 7.9, se presentan las estadísticas de $\widehat{F_c(h)}$ con las 3,000 muestras, según tamaño de muestra.

Cuadro 7.9: Estadísticas de $\widehat{F_c(h)}$.

n	35	50	70
Mín.	2.374	3.182	4.095
Q_1	7.376	7.179	7.138
Q_2	9.572	8.775	8.221
Media	10.336	9.275	8.573
Q_3	12.666	10.886	9.711
Máx.	29.913	22.326	19.003

En el cuadro 7.10, se presentan las estadísticas de $\widehat{pvalor}$ con las 3,000 muestras, según tamaño de muestra.

Cuadro 7.10: Estadísticas de $\widehat{pvalor}$.

n	35	50	70
Mín.	0.00205	0.00380	0.00531
Q_1	0.01205	:0.01621	0.02019
Q_2	0.02075	0.02445	0.02761
Media	0.02546	0.02712	0.02893
Q_3	0.03368	0.03537	0.03573
Máx.	0.19778	0.13324	0.09131

En la figura 7.8, se presenta el valor del estadístico de prueba en la población $F_c(h) = 7.38$ (línea horizontal continua) y el valor crítico de F para un nivel de significancia de $5\,\%$, con grados de libertad $[traza(H) - 1, 1.25 * traza(H) - 0.5]$ con $traza(H) = 4$, $F_{critica} = 5.9019$ (línea horizontal discontinua). Se grafican las estimaciones del estadístico de prueba según tamaño de muestra. Se observa que para $n = 35$, la proporción de muestras que llevan a la conclusión *el modelo es adecuado* es de alrededor de 0.9 (porción de la línea que corresponde a $n = 35$, que está por arriba de $F_{critica}$). Además al aumentar el tamaño de muestra la proporción mendionada se incrementa (ver cuadro 7.11). Así mismo, se muestra como las estimaciones del estadístico de prueba se aproximan a $F_c(h)$ al incrementar el tamaño de muestra.

En la figura 7.9, se presenta el $p - valor$ observado en la población $p - valor = 0.03364$ (línea horizontal continua) y el nivel de significancia $\alpha = 5\,\%$ (línea horizontal discontinua). Se grafican las estimaciones del $p-valor$ según tamaño de muestra. Se observa que para $n = 35$ la proporción de muestras que llevan a la conclusión *el modelo es adecuado* es de alrededor de 0.9

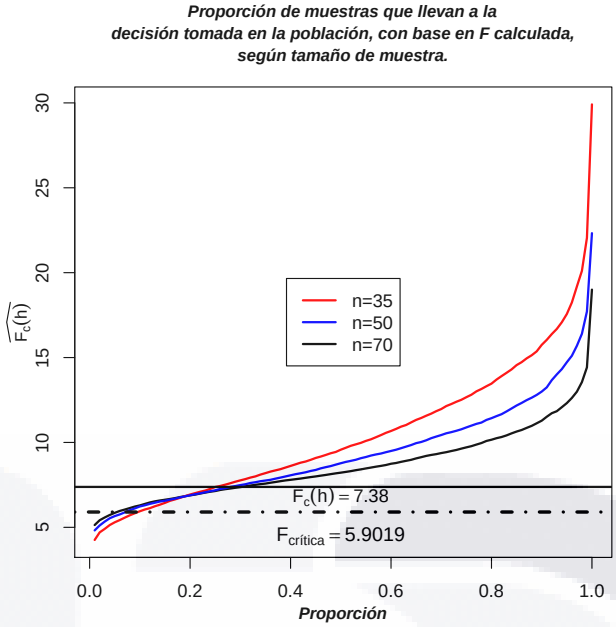

Figura 7.8: $\widehat{Prob}(D_s = D|D)$ con $\widehat{F_c}(h)$

(porción de la línea que corresponde a $n = 35$, que está por abajo de α). Además al aumentar el tamaño de muestra la proporción mendionada se incrementa (ver cuadro 7.11). Así mismo, se muestra como las estimaciones $p - \widehat{valor}$ se aproximan al $p-valor$ al incrementar el tamaño de muestra.

En la figura 7.10, se presenta el coeficiente de determinación global ajustado de la población $R^2(h) = 0.8304$ (línea horizontal discontinua central), las líneas horizontales discontinuas exteriores representan $R^2(h) \pm 0.05$.
Se comparan las estimaciones de $R^2(h)$, para regresión polinomial local y para regresión lineal, según tamaño de muestra. Se observa, para $n = 70$, que aproximadamente el $90\,\%$ de las estimaciones $\widehat{R^2_\pi}(h)$ están entre $R^2(h) - 0.05$ y $R^2(h) + 0.05$ (porción de la línea que corresponde a $n = 70$, que está entre las líneas horizontales exteriores). Mientras que para el mismo tamaño de muestra, sólo aproximadamente $23\,\%$ de las estimaciones de $R^2(h)$ para la regresión lineal están entre $R^2(h) - 0.05$ y $R^2(h) + 0.05$.

La figura 7.11, muestra el desempeño de $\widehat{\mu_{RPL}}$, $\widehat{\mu_{HT}}$ y $\widehat{\mu_{Reg}}$, en relación con la proporción de muestras de las $3,000$ muestras seleccionadas, con respecto al parámetro μ_y, según tamaño de muestra y estimador.

La figura 7.12, muestra que las estimaciones $\widehat{R^2_\pi}(h)$ mediante regresión polinomial local difieren de $R^2(h)$ menos que las estimaciones $\widehat{R^2}_{lin}$ en términos relativos. Se observa, para $n = 70$ que el $100\,\%$ de las estimaciones $\widehat{R^2_\pi}(h)$ difieren relativamente en no más de $10\,\%$ de $R^2(h)$. Mientras que para tal tamaño de muestra, cerca de $100\,\%$ de las estimaciones $\widehat{R^2}_{lin}$ difieren en no más de $20\,\%$ de $R^2(h)$. Esto es, se duplica el error relativo con las estimaciones mediante regresión lineal.

La figura 7.13, muestra que al aumentar el tamaño de muestra las estimaciones del promedio para el total de activos fijos por municipio mediante regresión polinomial local aventajan

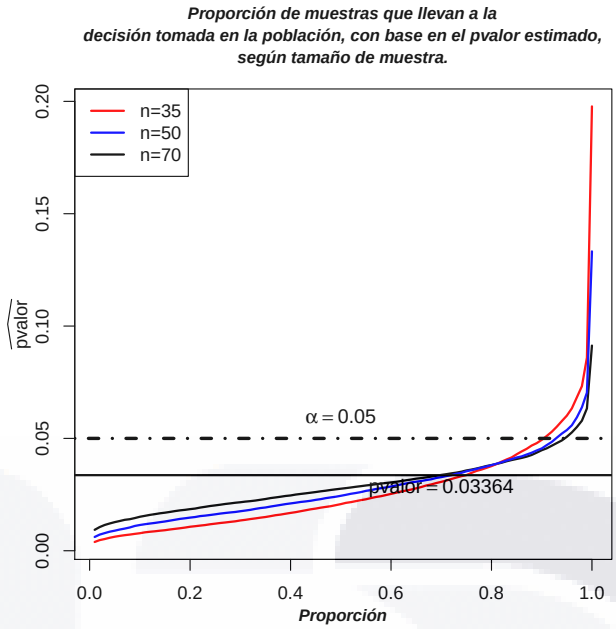

Figura 7.9: $\widehat{Prob}(D_s = D|D)$ con $\widehat{p-valor}$

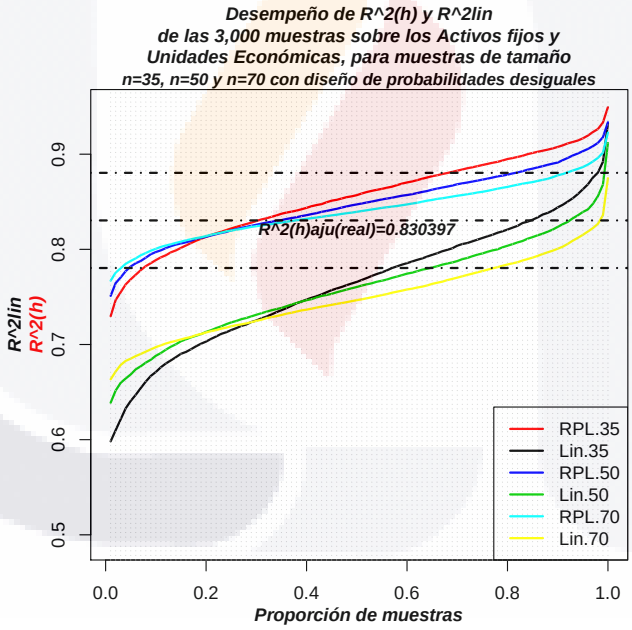

Figura 7.10: Desempeño de $\widehat{R^2_\pi}(h)$ contra $\widehat{R^2_{lin}}$

ligeramente a las estimaciones con los otros métodos. Se observa para $n = 70$ que el porcentaje de estimaciones$\widehat{\mu_{RPL}}$ que difieren en a los más $10\,\%$ de μ_y es $60\,\%$, para las estimaciones $\widehat{\mu_{Reg}}$ es $55\,\%$ y para las estimaciones $\widehat{\mu_{HT}}$ es $40\,\%$.

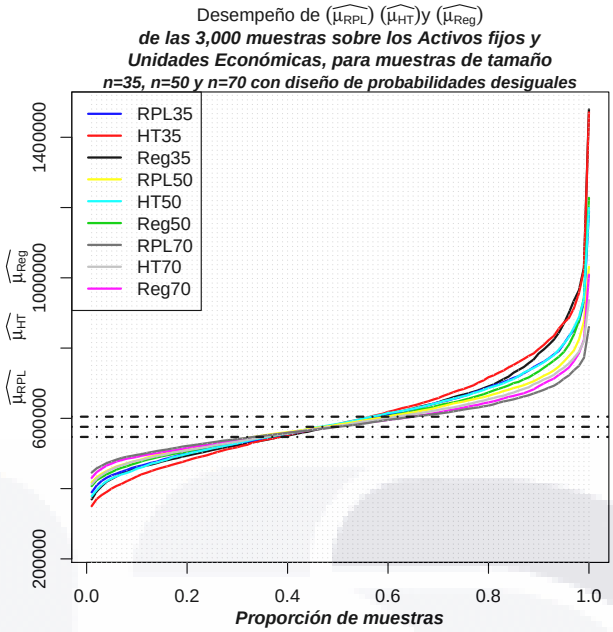

Figura 7.11: Desempeño de $\widehat{\mu_y}$

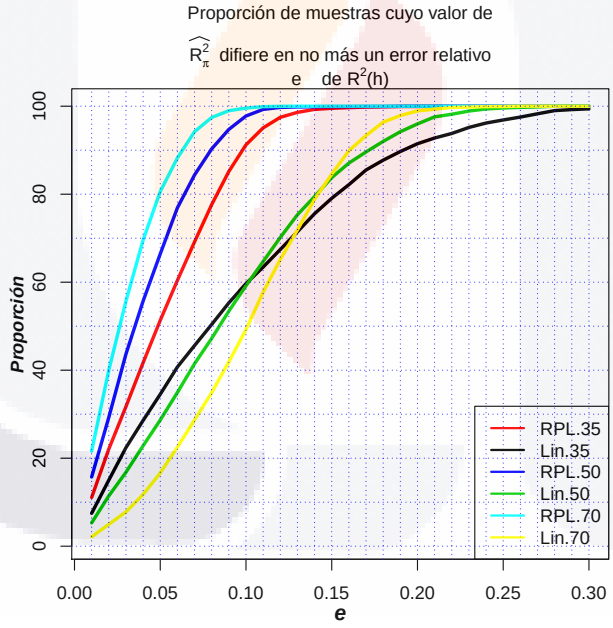

Figura 7.12: Diferencias entre $\widehat{R^2_\pi}(h)$ y $\widehat{R^2_{lin}}$

En el cuadro 7.11, se presenta la proporción de muestras que llevan a tomar la misma decisión tomada con la población, con base en el estadístico de prueba estimado $(\widehat{F_c}(h))$, según tamaño de muetsra y estimador.

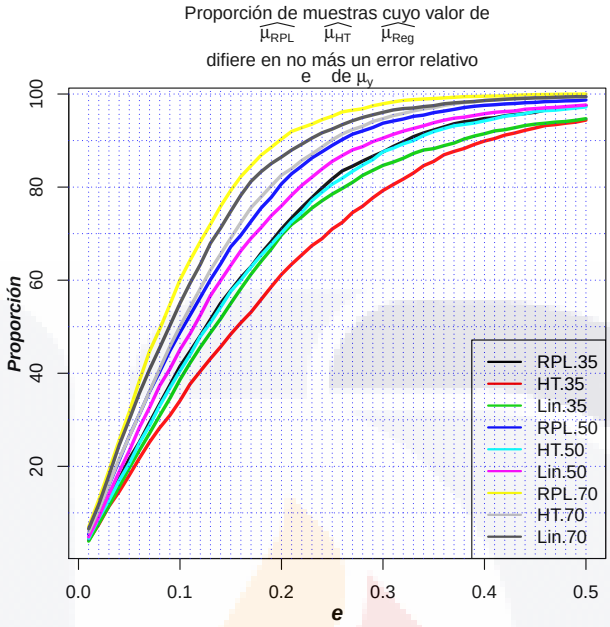

Figura 7.13: Diferencias entre $\widehat{\mu_{RPL}}$, $\widehat{\mu_{HT}}$ y $\widehat{\mu_{Reg}}$

Cuadro 7.11: Aproximación de $P(D_s = D|D = 0)$, según tamaño de muestra.

Tamaño de muestra	Proporción
35	0.90300
50	0.92867
70	0.94467

7.4. Segunda población

La segunda población está formada por 836 compañías del sector de alimentos, bebidas y medicina de Estados Unidos para los años 2009 y 2011, con un capital de mercado inferior a $10,001$ dólares y un precio promedio de almacen inferior a 81 dólares.

Las variables que se consideran son las siguientes:

- Variable auxiliar (X).- Capital de mercado de la compañía (dólares) en 2009.
- Variable respuesta (Y).- Precio promedio de almacén (dólares) en 2011.
- Tamaño de las unidades (Z).- Tamaño de la firma en 2009. Se consideran 10 clases.

Fuente de las variables X, Y, Z [8]

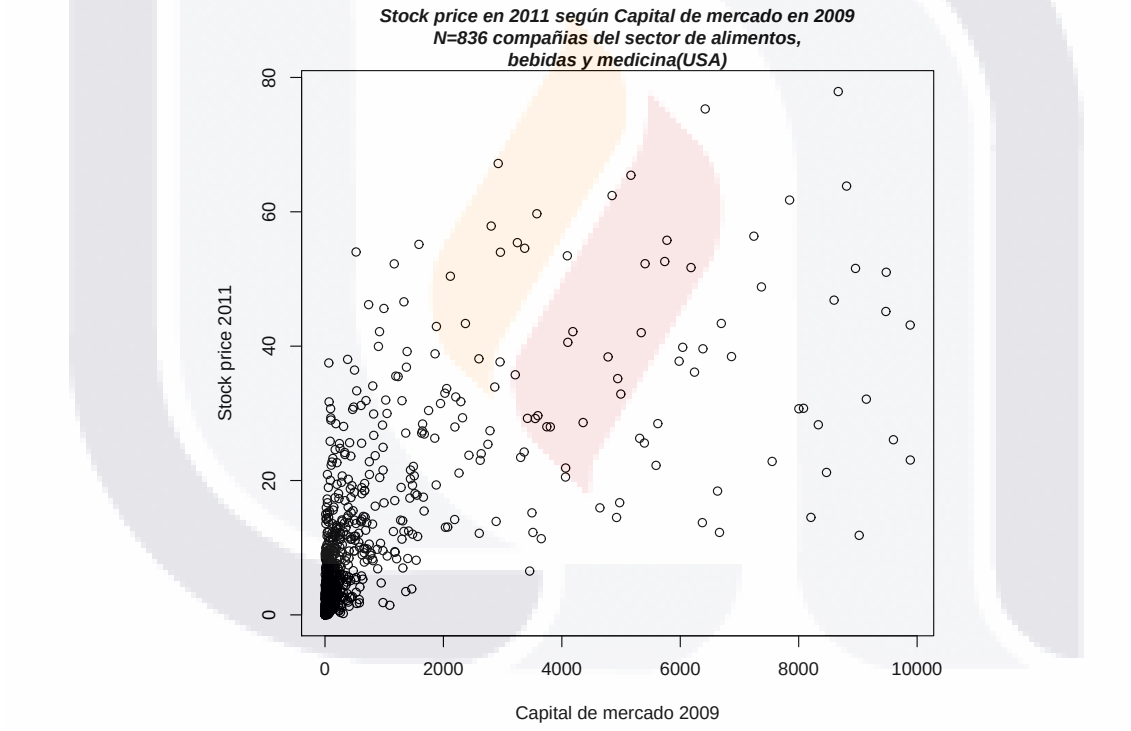

Figura 7.14: Diagrama de la población de compañías

En la figura 7.14, se muestra el diagrama de dispersión de la población sobre el precio promedio de almacen en 2011 como variable respuesta, y el capital de mercado de la compañía en 2009 como variable auxiliar. En el diagrama no se aprecia algún comportamiento bien definido. Sin embargo, se cree de alguna manera que el capital de mercado de las compañías afectan en el precio promedio de almacén. Tal relación se denota con $m(\bullet)$.

7.5. Descripción de la simulación.

Con el objetivo, de mostrar los resultados que se pueden obtener mediante estimadores por regresión polinomial local en muestreo de poblaciones finitas, para el coeficiente de determinación global, el análisis de varianza y el promedio de la variable de estudio, se presentan los resultados al realizar $3,000$ simulaciones con dos tamaños de muestra bajo un diseño de muestreo de máxima entropía con probabilidades desiguales.

Se consideran los siguientes elementos :

1. Tamaño de la población $N = 836$.
2. Número de parámetros equivalente 3.
3. Ancho de banda $h = 0.3333$.
4. Tamaños de muestra $n = 100$ y $n = 150$.
5. Diseño de muestreo de máxima entropía de probabilidades desiguales.
6. Parámetros de interés:
   - *a*) $\theta_1 = R^2(h)$ Coeficiente de determinación global.
   - *b*) $\theta_2 = R^2(x_0; h)$ Coeficiente de determinación local.
   - *c*) $\theta_3 = \mu_y$ Precio promedio de almacen.
   - *d*) $\theta_4 = F_c(h)$ Estadístico de prueba.
   - *e*) $\theta_5 = p - valor$ p-valor asociado con el estadístico de prueba calculado.
   - *f*) $\theta_6 = P\left(D_s = D | D = d\right)$ Probabilidad de tomar la misma decisión con la muestra (D_s) que la decisión tomada con la población $(D = d)$. Donde d puede tomar dos valores:
     - $d = 0$ El ajuste del modelo a los dato no es adecuado.
     - $d = 1$ El ajuste del modelo a los dato es adecuado.

7.6. Resultados.

Primero se dan los resultados para la población y después para las $3,000$ réplicas.

En la figura 7.15,se da el diagrama de dispersión de la población con los ajustes por regresión lineal y regresión polinomial local. Se observa una relación lineal débil con un coeficiente de determinación $R^2_{lin} = 0.464$, mientras que a través de regresión polinomial local se muestra una relación global no lineal con un coeficiente de determinación $R^2_{RPL} = 0.6798$.

La tabla ANOVA se muestra en el cuadro 7.12. Se tiene que la decisión es no rechazar la hipótesis nula, y por tanto el ajuste del modelo por regresión polinomial local a los datos no es adecuado. Sin embargo, se utiliza el método de regresión polinomial local para mostrar sus bondades y ventajas sobre otros métodos que son más rigidos en cuanto a los supuestos que hay que hacer.

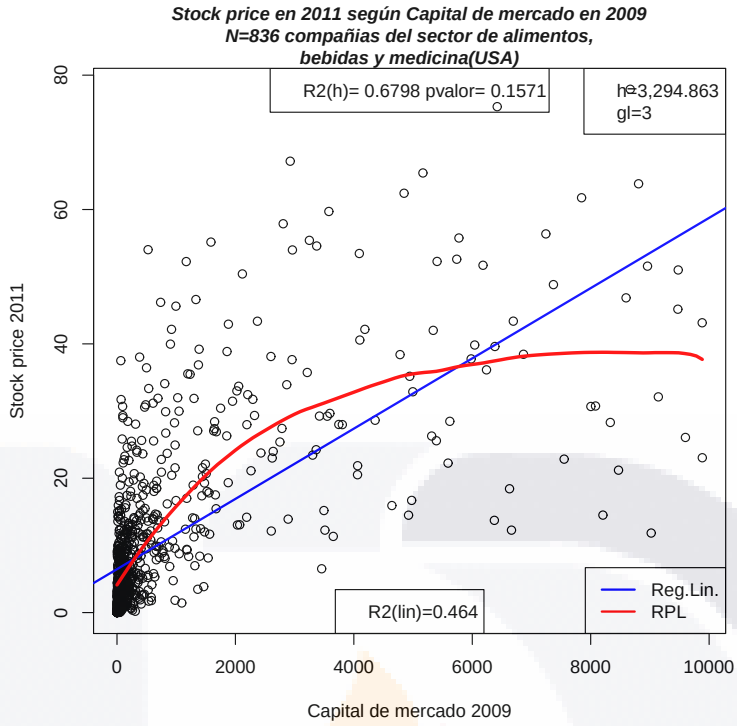

Figura 7.15: Diagrama de la población de compañias con ajustes

Cuadro 7.12: Tabla ANOVA de la regresión polinomial local para la población

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	$F(h)$	p-valor	Decisión
$m(x_k)$	2	99,676.5787	49,838.2894	3.45	0.151165	0
ϵ_k	833	46,941.0000	56.3517			
Total	835	146,617.7294				

La medida de bondad de ajuste del modelo estimado por regresión polinomial local a los datos, es el coeficiente de determinación global, denotado con $R^2(h)$. En el cuadro 7.13, está el coeficiente de determinación global, el coeficiente de determianción ajustado de la población para regresión polinomial local, y el coeficiente de determinación para regresión lineal.

Cuadro 7.13: Coeficiente de determinación global de la población

$R^2(h)$	$R^2(h)_{ajustado}$	$R^2(lineal)$
0.679840	0.679071	0.463951

El precio promedio de almacen para la población es $\mu_y = 10.5293$.

Ahora, se comparan los métodos de estimación del coeficiente de determinación y del promedio para los dos tamaños de muestra. Esto, con referencia a que el comportamiento de la población para las variables en estudio está descrita por la función $m(\bullet)$.

En el cuadro 7.14, se presentan las estadísticas para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según estimador y tamaño de muestra.

Cuadro 7.14: Estadísticas de $\widehat{R^2_{\pi}}(h)$,$\widehat{R^2_{\pi Aj}}(h)$, $\widehat{R^2_{lin}}$

	$n = 100$			$n = 150$		
Estadística	$\widehat{R^2_{\pi}}(h)$	$\widehat{R^2_{\pi Aj}}(h)$	$\widehat{R^2_{lin}}$	$\widehat{R^2_{\pi}}(h)$	$\widehat{R^2_{\pi Aj}}(h)$	$\widehat{R^2_{lin}}$
Mín	0.5556	0.5545	0.1670	0.5829	0.5817	0.2079
Q_1	0.6678	0.6669	0.3861	0.6673	0.6665	0.3997
Q_2	0.6959	0.6952	0.4456	0.6890	0.6882	0.4423
Media	0.6951	0.6943	0.4444	0.6886	0.6878	0.4418
Q_3	0.7222	0.7216	0.5014	0.7097	0.7090	0.4835
Máx.	0.8265	0.8262	0.7258	0.7879	0.7874	0.6629
Des.Est.	0.04097	0.04109	0.08264	0.03109	0.03119	0.06138

La figura 7.16, presenta los diagramas de caja y brazos del coeficiente de determinación estimado con las $3,000$ muestras seleccionadas, según tamaño de muestra. La línea horizontal continua, representa el coeficiente de determinación ajustado para la población. Se observa que el coeficiente de determinación global, estimado mediante regresión polinomial local refleja mejor el comportamiento de la población, y además la estimación es más precisa.

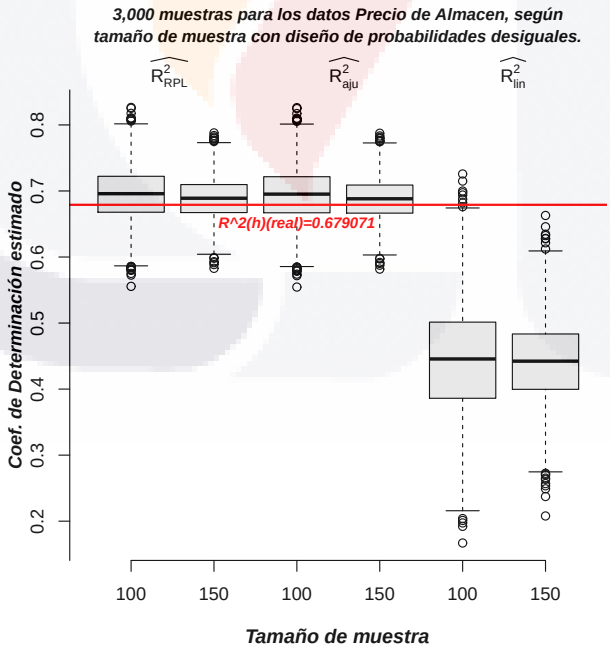

Figura 7.16: $\widehat{R^2_{\pi}(h)}$

Para analizar el comportamiento de las estimaciones del coeficiente de determinación (R^2), obtenemos el error relativo (E.R.) y la raíz cuadrada del error cuadrático medio relativo (R.E.C.M.R) para el estimador de R^2 mediante regresión polinomial local ($\widehat{R^2_\pi}(h)$) y regresión lineal ($\widehat{R^2_{lin}}$).

En el cuadro 7.15, se presentan las estadísticas del E.R. para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según tamaño de muestra.

Cuadro 7.15: Estadísticas de $E.R.\left(\widehat{R^2_\pi}(h)\right)$ y $E.R.\left(\widehat{R^2_{lin}}\right)$

Estadística	$n = 100$		$n = 150$	
	$\widehat{R^2_\pi}(h)$	$\widehat{R^2_{lin}}$	$\widehat{R^2_\pi}(h)$	$\widehat{R^2_{lin}}$
Mín.	-0.18279	-0.75441	-0.14261	-0.69420
Q_1	-0.01766	-0.43208	-0.01843	-0.41213
Q_2	0.02360	-0.34450	0.01347	-0.34943
Media	0.02242	-0.34633	0.01281	-0.35012
Q_3	0.06232	-0.26247	0.04393	-0.28885
Máx.	0.21578	0.06754	0.15896	-0.02489
Des.Est.	0.06027	0.12155	0.04574	0.09028

En el cuadro 7.16, se presentan el R.E.C.M.R. para las estimaciones del coeficiente de determinación con las $3,000$ muestras, según tamaño de muestra. En la que se observa que $R.E.C.M.R(\widehat{R^2_{lin}})$ es casi 6 veces mayor que $R.E.C.M.R.(\widehat{R^2_\pi}(h))$ para $n = 100$ y mayor que 7 veces para $n = 150$.
La figura 7.17, presenta el desempeño del coeficiente de determinación local para la población

Cuadro 7.16: $R.E.C.M.R.\left(\widehat{R^2_\pi}(h)\right)$ y $R.E.C.M.R.\left(\widehat{R^2_{lin}}\right)$

n	$\widehat{R^2_\pi}(h)$	$\widehat{R^2_{lin}}$	Razón
100	0.06429	0.36703	5.70897
150	0.04749	0.36157	7.61360

y los promedios del coeficiente de determinación local estimado con las $3,000$ muestras con tamaños de $n = 100$ y $n = 150$ respectivamente. Se observa que para valores de x_0 entre 0.1 y 0.4 las estimaciones $\widehat{R^2(x_0; h)}$ aproximan bien a $R^2(x_0; h)$ con ambos tamaños de muestra, mientras que para valores mayores que 0.4 al aumentar el tamaño de muestra la estimación mejora. Además los valores del coeficiente de determinación local entre de 0.1 y 0.3 son inferiores a 0.55 debido a la aglomeración de las compañías entorno a dichos valores de la variable auxiliar.

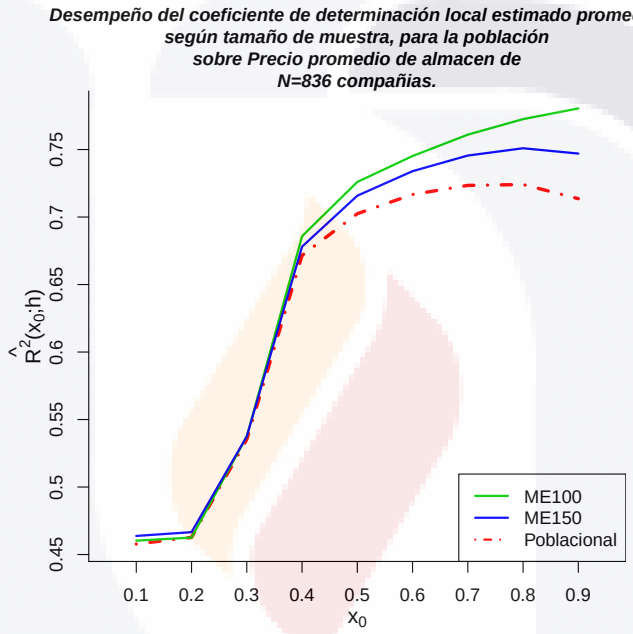

Figura 7.17: $\widehat{R^2(x_0;h)}$

Ahora, se comparan los métodos de estimación del promedio para los dos tamaños de muestra. Los métodos que se utilizan para obtener $\widehat{\mu_y}$ son:

1. Regresión Polinomial local $\widehat{\mu_{RPL}}$, calculado con la ecuación (2.11).
2. Estimador de Horvitz-Thompson $\widehat{\mu_{HT}}$, calculado con la ecuación (2.9).
3. Regresión Lineal $\widehat{\mu_{lin}}$, para el cual se tienen dos formas de obtenerlo:
   a) Mediante el estimador por regresión $\widehat{\mu_{r1}}$
   b) Mediante los residuos $\widehat{\mu_{r2}}$. Se utiliza el estimador por regresión lineal ($\widehat{\mu_{r2}}$), que se denotará con ($\widehat{\mu_{Reg}}$) , calculado con la ecuación (2.10).

En el cuadro 7.17, se presentan las estadísticas para las estimaciones del precio promedio de almacen con las 3,000 muestras, según estimador y tamaño de muestra.

Cuadro 7.17: Estadísticas de $\widehat{\mu_{RPL}}$, $\widehat{\mu_{HT}}$, $\widehat{\mu_{Reg}}$

	$n = 100$			$n = 150$		
Estadística	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$
Mín	-9.722	6.821	5.346	7.60	7.342	5.987
Q_1	9.692	9.754	9.508	9.90	9.888	9.711
Q_2	10.570	10.551	10.565	10.52	10.558	10.585
Media	10.566	10.610	10.691	10.56	10.582	10.635
Q_3	11.447	11.423	11.815	11.19	11.248	11.549
Máx.	19.776	15.099	16.690	14.53	14.606	15.496
Des.Est.	1.41715	1.26345	1.71427	0.98652	1.01439	1.37505

La figura 7.18, presenta los diagramas de caja y brazos del precio promedio de almacén estimado con las 3,000 muestras seleccionadas, según estimador y tamaño de muestra. La línea horizontal continua, representa el precio promedio de almacén de la población.

Para analizar el comportamiento de las estimaciones del precio promedio (μ_y), obtenemos el error relativo (E.R.) y la raíz cuadrada del error cuadrático medio relativo (R.E.C.M.R) para el estimador de μ_y mediante regresión polinomial local ($\widehat{\mu_{RPL}}$), Estimador de Horvitz-Thompson ($\widehat{\mu_{HT}}$) y regresión lineal ($\widehat{\mu_{Reg}}$) .

En el cuadro 7.18, se presentan las estadísticas del E.R. para las estimaciones con las 3,000 muestras, según estimador y tamaño de muestra.

En el cuadro 7.19, se presentan el R.E.C.M.R. para las estimaciones del precio promedio de almacen con las 3,000 muestras, según estimador y tamaño de muestra. Y se tiene que para $n = 100$ $R.E.C.M.R(\widehat{\mu_{HT}})$ es inferior a $R.E.C.M.R(\widehat{\mu_{RPL}})$, mientras que $R.E.C.M.R(\widehat{\mu_{lin}})$ es superior a $R.E.C.M.R(\widehat{\mu_{RPL}})$, en tanto que para $n = 150$ $R.E.C.M.R(\widehat{\mu_{HT}})$ y $R.E.C.M.R(\widehat{\mu_{lin}})$ están por arriba de $R.E.C.M.R(\widehat{\mu_{RPL}})$.

Se realiza la prueba global del ajuste del modelo para la regresión polinomial local,

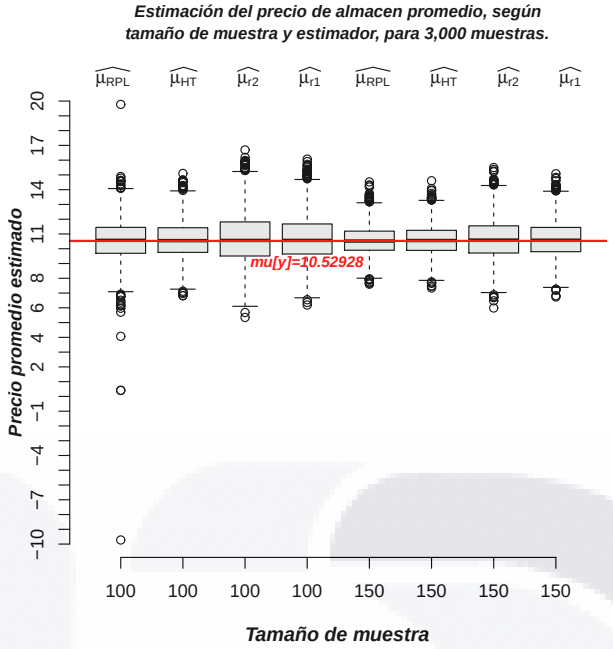

Figura 7.18: Estimación del promedio del precio de almacen

Cuadro 7.18: Estadísticas de $E.R.(\widehat{\mu_y})$, según método y tamaño de muestra.

Estadística	$n = 100$			$n = 150$		
	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$
Mín.	-1.92332	-0.35219	-0.49227	-0.27826	-0.30272	-0.43136
Q_1	-0.07951	-0.07359	-0.09699	-0.05973	-0.06086	-0.07770
Q_2	0.00384	0.00209	0.00337	-0.00075	0.00269	0.00533
Media	0.00348	0.00763	0.01532	0.00292	0.00502	0.01002
Q_3	0.08719	0.08486	0.12214	0.06302	0.06828	0.09689
Máx.	0.87822	0.43402	0.58511	0.37952	0.38721	0.47169
Des.Est.	0.13459	0.11999	0.16281	0.09369	0.09634	0.13059

Cuadro 7.19: $R.E.C.M.R.(\widehat{\mu_y})$,según estimador.

n	$\widehat{\mu_{RPL}}$	$\widehat{\mu_{HT}}$	$\widehat{\mu_{Reg}}$	Razón(HT)	Razón (Reg)
100	0.13461	0.12022	0.16350	0.89309	1.21462
150	0.09372	0.09646	0.13096	1.02924	1.39735

para cada una de las $3,000$ muestras seleccionadas bajo un diseño de máxima entropía de probabilidades desiguales y dos tamaños de muestra (100 y 150).

En el cuadro 7.20, se presentan las estadísticas de $\widehat{F_c(h)}$ y $\widehat{p-valor}$, según tamaño de

muestra.

Cuadro 7.20: Estadísticas de $\widehat{F_c(h)}$ y $\widehat{p-valor}$

Estadística	$n = 100$		$n = 150$	
	$\widehat{F_c(h)}$	$\widehat{p-valor}$	$\widehat{F_c(h)}$	$\widehat{p-valor}$
Mín.	2.052	0.05722	2.232	0.07766
Q_1	3.258	0.12270	3.226	0.13060
Q_2	3.706	0.14360	3.556	0.14710
Media	3.793	0.14540	3.605	0.14840
Q_3	4.219	0.16530	3.938	0.16510
Máx.	7.853	0.26770	6.011	0.24020

La figura 7.19, muestra la proporción de las 3,000 muestras seleccionadas, que lleva a la misma decisón tomada en la población (ver tabla ANOVA 7.12), con base en el estadístico de prueba estimado con un nivel de significancia de 5 %, según tamaño de muestra. Se aprecia que tanto para $n = 100$ como $n = 150$, el 100 % de las muestras arrojan un estadístico de prueba estimado inferior a $F_{critico}$. Además, al pasar de $n = 100$ a $n = 150$, dicho estadístico tiende al valor $F_c(h)$ de la población.

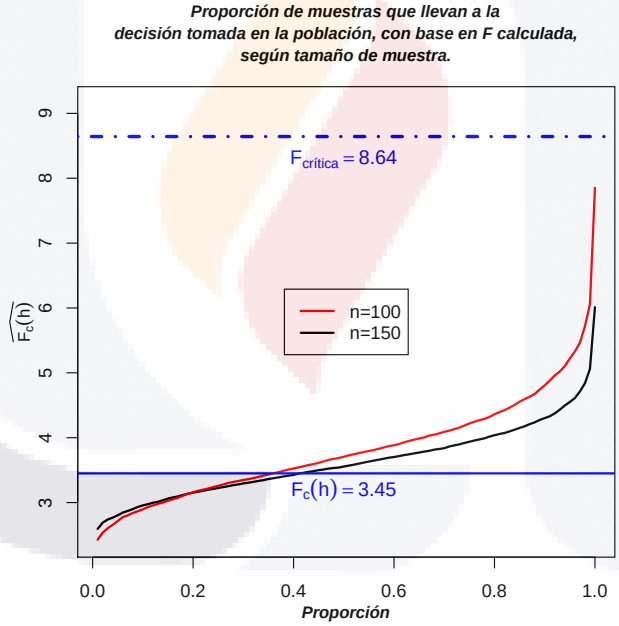

Figura 7.19: Decisión con $\widehat{F_c(h)}$

La figura 7.20, muestra la proporción de las 3,000 muestras seleccionadas, que lleva a la misma decisón tomada en la población (ver tabla ANOVA 7.12), con base en el $p - valor$ estimado con un nivel de significancia de 5 %, según tamaño de muestra. Se observa que el 60 % de las muestras dan un $p - valor$ estimado inferior que el $p - valor$ calculado con la población, pero no menor que α. Esto es, el 100 % de las muestras dan un $p - valor$ estimado

superior al 5 %. Además, al incrementar el tamaño de muestra de $n = 100$ a $n = 150$ el $p - valor$ estimado se acerca un poco más al $p - valor$ de la población.

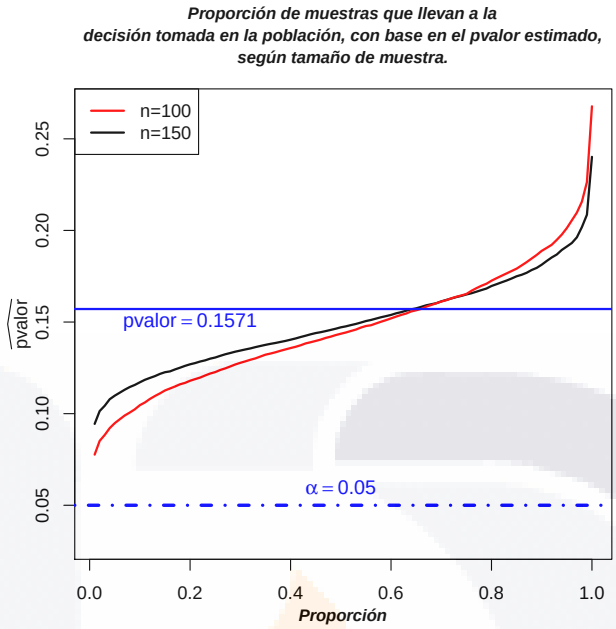

Figura 7.20: Decisión con $\widehat{p - valor}$

La figura 7.21, muestra el desempeño de $\widehat{R^2_\pi}(h)$ contra $\widehat{R^2_{lin}}$ en relación con la proporción de las $3,000$ muestras seleccionadas, con respecto al parámetro R^2, según tamaño de muestra. Y bajo el supuesto que el modelo $m(\bullet)$ es verdadero, se muestra que la regresión polinomial local refleja mejor el comportamiento de la población. Mientras que la regresión lineal sólo alcanza valores cercanos a R^2 del modelo $m(\bullet)$, en menos del 1 % de las $3,000$ muestras. Esto se observa para ambos tamaños de muestra.

La figura 7.22, presenta la proporción de muestras de las $3,000$ muestras seleccionadas cuyo valor de $\widehat{R^2_\pi}(h)$ y $\widehat{R^2_{lin}}$, difieren en no más de un error relativo e del parámetro R^2, según tamaño de muestra. Y bajo el supuesto que el modelo $m(\bullet)$ es verdadero, se muestra que la regresión polinomial local aventaja a la regresión lineal. Por ejemplo, mediante regresión polinomial local para $n = 100$, el 65 % de las muestras dan un error relativo inferior al 6 %, y para $n = 150$ es casi 80 % para el mismo error. Mientras que la regresión lineal para $n = 100$ y $n = 150$, menos del 5 % de las muestras dan un error relativo inferior al 15 %.

El comportamiento de la variable precio de almacén se describe mejor mediante regresión polinomial local que con regresión lineal. Esto se observa con las estimaciones del coeficiente de determinación global.

Esto es, el comportamiento local de la variable de estudio no se puede caracterizar a través de un modelo de regresión lineal con la misma pendiente para cada vecindad de la variable auxiliar. Ya que el modelo de regresión lineal sobreestima el comportamiento para valores altos de la variable auxiliar y lo subestima para valores menores. Esto impacta en la estimación del precio de almacén promedio al sobreestimarlo.

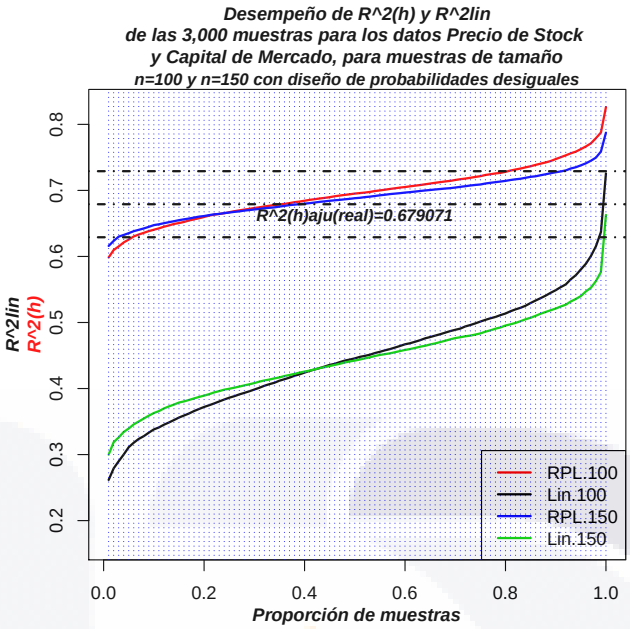

Figura 7.21: Desempeño de $\widehat{R^2_\pi}(h)$ contra $\widehat{R^2_{lin}}$

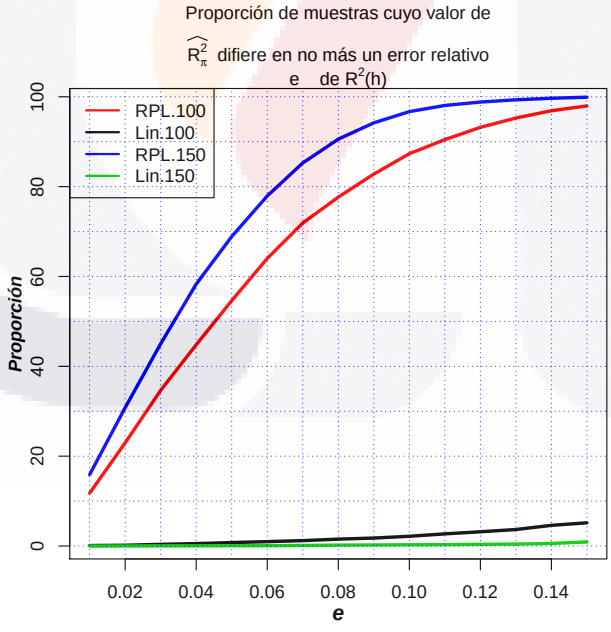

Figura 7.22: Diferencias entre $\widehat{R^2_\pi}(h)$ y $\widehat{R^2_{lin}}$

Capítulo 8

Conclusiones

La propuesta que se presenta en este trabajo, da una opción no paramétrica para la estimación del coeficiente de determinación, el análisis de varianza y la estimación de parámetros en muestreo de poblaciones finitas.

Se espera que la metodología sea útil en el análisis de datos y en la aplicación en encuestas complejas por muestreo de poblaciones finitas.

Los resultados de las simulaciones muestran que los estimadores del coeficiente de determinación y del estadístico de prueba mediante regresión polinomial local en muestreo de poblaciones finitas, tienen un comportamiento asintótico consistente, y convergen en distribución a una distribución normal y distribución F de Fisher, lo que se justifica en los teoremas 3 y 4 respectivamente.

Bajo le evidencia de una relación entre las variables, no necesariamente una relación de tipo paramétrica, el uso de información auxiliar disponible para todas las unidades de una población, en la modelación no paramétrica de estimadores de parámetros puede competir con los estimadores usuales.

Esto es, el estimador de regresión polinomial local para la media, es más eficiente que el estimador de Horvitz-Thompson y que el estimador de regresión lineal. Se pueden consultar los teoremas correspondientes en [2].

Se observó además, que en un alto porcentaje de las muestras seleccionadas en cada uno de los experimentos de simulación y con los datos reales, la decisión tomada con base en el estadístico de prueba, es la decisión tomada con toda la población, independientemente de ésta.

Por lo anterior, permite decir que para un tamaño de muestra acorde con la estructura de la población, el diseño de muestreo y las especificaciones de confianza y error, la metodología puede considerarse como una alternativa viable en la inferencia del análisis de varianza para los métodos de estimación mediante regresión no paramétrica.

La implemetación del procedimiento se realizó en el paquete R [14], en cual tiene funciones que facilitan los cálculos con matrices, ciclos y manejo de tablas de datos. Sin embargo,

los tiempos de ejecución dependen entre otros aspectos de, la memoria disponible del equipo, la velocidad del procesador, el número de simulaciones, el tamaño de la población, etc., ya que para el manejo de gran cantidad de datos R no cuenta con las funciones por defecto, existen funciones en R que permiten su manejo, como la paralelización o el uso de apuntadores para trabajar directamente en disco duro, siempre que el procedimiento lo permita.

Finalmente, existen aspectos que se pueden considerar para trabajos posteriores. En particular, estudiar y adaptar métodos de elección de diferentes valores del ancho de banda, de acuerdo con la naturaleza de los datos. Y adaptarlo en muestreo de poblaciones finitas.

Bibliografía

[1] Huang, L. S. y J. Chen (2008). Analysis of variance, coefficient of determination and F-test for local polynomial regression. The Annals of Statistics. 36:5, 2085-2109.

[2] Breidt, F. J. y J. D. Opsomer (2000). Local Polynomial Regression Estimators in surveys sampling. The Annals of statistics, 28(4), 1026-1053.

[3] Breidt, F. J., Claeskens y J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalized splines. Biometrika, 92(4), 831-846.

[4] Sarndal, C. E., B. Swensson y Jan Wretman (1992). Model Assisted Survey Sampling. Springer.

[5] Seber, G. A. F. (2003). Linear Regression Analysis. John Wiley and Sons.

[6] Draper, N. y Smith H. (1998). Applied Regression Analysis. Wiley Series in Probability and Statistics.

[7] Hastie, T.J. y Tibshirani R. J. (1990). Generalized Additive Models. Chapman and Hall.

[8] $http://pages.stern.nyu.edu/\ adamodar/New_{H}ome_{P}age/data.html$

[9] INEGI. Conteo de Población y Vivienda 2005.

[10] INEGI Sistema Estatal y Municipal de Bases de Datos (SIMBAD). Información de los Censos Económicos 2004 y 2009.

[11] Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local Plynomial Regression: Optimal Kernels and asymptotic minimax efficiency. Annals Inst. statistics Math. Vol. 49, No. 1, 79-99.

[12] Matei Alina and Tillé Yves (2005). Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size. Journal of Official Statistics, Vol. 21, No. 4, pp. 543-570.

[13] Matei Alina (2005). Computational aspects of sample surveys. Université de Neuchâtel. Faculte des Sciences.

[14] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Apéndice A. Aspectos Técnicos

Contribución de la parte residual a la variable de interés.

Para obtener las proporciones de la variabilidad de la variable respuesta que es explicada por la parte residual mediante el modelo $m(x)$, denominada contribución de la parte residual a la variable de interés y denotada con r, se considerarón las poblaciones $Hardle$ y $Bump$ ambas de tamaño $N = 5,000$.

Se buscaron valores de r para los cuales:

- Se rechace H_0.
- No se rechace H_0.

con un nivel de significancia $\alpha = 0.05$.

Esto es, el objetivo fue encontrar valores de la contribución de la parte residual a la variable de interés tales, que uno indique que el ajuste del modelo a los datos es adecuado y otro que indique lo contrario.

El estadístico de prueba está dado por la ecuación (3.16). Y se rechaza H_0 con un nivel de significancia $\alpha = 0.05$ si $F_c(h) > F_{\alpha,tr(H(h))-1,1.25*tr(H(h))-0.5}$. Para las poblaciones $F_{\alpha,tr(H(h))-1,1.25*tr(H(h))-0.5} = F_{0.05,7,9.25}$.

El cuadro 8.1, muestra los resultados para ambas poblaciones, según valor de la contribución de la parte residual a la variable de interés. Se observa que para valores de $r = 0.1$ y $r = 0.2$ la decisión es rechazar H_0 en ambas poblaciones, y para $r = 0.3, \cdots, 0.9$ la decisión es no rechazar H_0 en ambas poblaciones.
De acuerdo con el objetivo de encontrar dos valores de r, uno con el cual se rechace H_0 y otro no se rechzace H_0, se establecen $r = 0.2$ y $r = 0.8$. Esto, debido a que al seleccionar muestras aleatorias, un alto porcentaje lleven a tomar la misma decisión tomada en la población.

Número de simulaciones.

Para determinar el número de muestras aleatorias que se seleccionarán de cada población, se tomó un escenario díficil de modelar y de estimar. La población considerada fue $y_k = 2 + Sen(2\pi x_k) + \epsilon_k$ con $N = 5,000$, $r = 0.5$ y número equivalente de parámetros

Cuadro 8.1: Contribución de la parte residual.

			Hardle				*Bump*	
r	$R^2(h)$	$F_c(h)$	$p-valor$	D	$R^2(h)$	$F_c(h)$	$p-valor$	D
0.1	0.8205	6.20	0.00663	1	0.8809	10.05	0.00113	1
0.2	0.7301	3.67	0.03547	1	0.7852	4.96	0.01405	1
0.3	0.6545	2.57	0.09180	0	0.6961	3.11	0.05633	0
0.4	0.5626	1.75	0.21069	0	0.6073	2.10	0.14582	0
0.5	0.4884	1.30	0.34682	0	0.5175	1.46	0.28976	0
0.6	0.4053	0.92	0.53215	0	0.4325	1.03	0.47063	0
0.7	0.3038	0.59	0.75065	0	0.3393	0.70	0.67357	0
0.8	0.2363	0.42	0.86763	0	0.2355	0.42	0.86763	0
0.9	0.1426	0.23	0.96747	0	0.1466	0.23	0.96747	0

$N.E.P. = 5$.
Se calculó el ancho de banda $h = 0.2020524$ y se obtuvó un coeficiente de determinación global $R^2(h) = 0.526414$.
Después se consideró un diseño de muestreo de máxima entropía con probabilidades desiguales de tamaño fijo $n = 100$.
Se obtuvieron $6,000$ muestras independientes y se calculó $R.E.C.M.R$ dado en la ecuación (5.2) para $\theta = R^2(h)$ y $\widehat{\theta} = \widehat{R^2_\pi}(h)$.

Los resultados para 100, 200,$\cdots$,$6,000$ muestras se presenta en la siguiente figura.

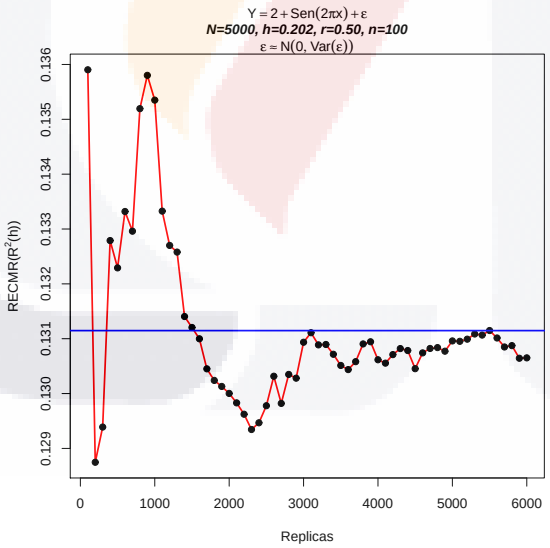

Figura 8.1: Número de muestras

De la figura 8.1, se observa que apartir de $3,000$ muestras $R.E.C.M.R.(R^2(h))$ se estabiliza alrededor del promedio. Luego con esto se determinó que el número de simulaciones para cada población fuera de $M = 3,000$.

Generación de la parte residual de cada función $m(x)$.

Para cada población $y = m(x) + \epsilon$, la parte residual ϵ se genera mediante una distribución normal con media cero y varianza $Var(\epsilon)$.
$Var(\epsilon)$ se determina en función de su contribución a $Var(y)$.
Dado que $Var(y) = Var(m(x) + \epsilon) = Var(m(x)) + Var(\epsilon)$, y que se consideran dos valores para la contribución de la parte residual a la variable de interés, $r = 0.2, 0.8$, $Var(\epsilon)$ se obtiene de la siguiente manera:

$$\begin{aligned} Var(\epsilon) &= rVar(y) \\ &= r\left[Var(m(x) + Var(\epsilon))\right] \\ &= r\left(Var(m(x))\right) + r\left(Var(\epsilon)\right). \end{aligned}$$

Luego

$$Var(\epsilon)\left(1 - r\right) = rVar(m(x)),$$

y por tanto

$$Var(\epsilon) = \frac{r}{(1-r)} Var(m(x)).$$

Así, la parte residual de cada función $m(x)$ se genera mediante una distribución $N\left(0, Var(\epsilon)\right)$.

Tamaño de las unidades poblacionales.

Con cierta probabilidad se usa la misma fuente de aleatoriedad en la construcción de dos variables, y con dos fuentes independientes en otro caso. La probabilidad considerada para determinar cuál fuente se utilizará, está muy relacionada con el coeficiente de correlación.
Consideremos distribuciones que pueden generarse al utilizar una variable aleatoria uniforme y mediante un algoritmo determinístico.
Supongáse que A y B son dos distribuciones con varianzas finitas. El problema puede determinarse como:
Construir X y Y tales que X tenga distribución A y Y tenga distribución B, y además que la correlación entre X y Y, denotada con $\rho = corr(X, Y)$, este en $[-1, 1]$.
Sean μ_A, σ_A, μ_B y σ_B los primeros momentos y las desviaciones estándar de A y B respectivamente.
Sean f y g algoritmos tales que $f(U)$ tenga distribución A y $g(U)$ tenga distribución B, donde U es una variable aleatoria uniforme sobre $[0, 1]$.
Se sabe que una transformación lineal simple da respuesta al problema antes planteado. Si X y X_1 son dos variables aleatorias independientes, entonces $\rho X + \sqrt{1 - \rho^2} X_1$ es una varaible

aleatoria con distribución normal estándar, y su correlación con X es exactamente ρ. (Dukić, Marić, 2010).

Luego, para generar los tamaños de cada unidad de la población, denotados con Z y que servirán para asignar las probabilidades desiguales de inclusión, en el caso de las poblaciones hipotéticas se considera un coeficiente de correlación $\rho = 0.5$, esto con la finalidad de que entre mayor sea el valor de la variable auxiliar la probabilidad de inclusión de la unidad esté dada en una proporción de 50 % respecto a su tamaño.
Por ejemplo, si el diseño de muestreo fuera de probabilidades proporcionales al tamaño de la empresa, el cual, por ejemplo está dado por los ingresos de la empresa, entre mayores ingresos mayor será la probabilidad de incluir dicha empresa en la muestra, mientras que tal vez la variable auxiliar sea el personal ocupado el cual se supone tiene una correlación de alrededor de 0.5 con los ingresos.
Así, el tamaño de cada unidad de la población se obtiene mediante:

$$Z = \rho X + \sqrt{1 - \rho^2} X_1,$$

con X la variable auxiliar, X_1 es una varaible aleatoria con distribución uniforme sobre $[0, 1]$ que se genera de manera independiente de X.

Para el caso, de la población real sobre el precio promedio de almacén, el tamaño de las compañias se elige como los rangos preestablecidos con base en el capital de la firma, los cuales se muestran en el cuadro 8.2. El coeficiente entre la variable auxiliar (Capital del

Cuadro 8.2: Tamaño de las compañias.

Tamaño	Capital de la firma (miles de dólares)
1	$[0, 10)$
2	$[10, 20)$
3	$[20, 40)$
4	$[40, 100)$
5	$[100, 250)$
6	$[250, 500)$
7	$[500, 1,000)$
8	$[1,000, 2,500)$
9	$[2,500, 10,000)$
10	$[10,000, +)$

mercado de la compañía en 2009) y el tamaño de la compañía, es de $\rho = 0.6587$.

Apéndice B. Demostraciones

Proposión 1

Demostración

Sean $W_\pi(x_0;h) = diag\left(\frac{K_h(x_1-x_0)}{\pi_1},\cdots,\frac{K_h(x_n-x_0)}{\pi_n}\right)$, $Y_\pi = (y_1,\cdots,y_n)^T$, $\widehat{\bar{y}} = \frac{\sum_{k\in S}\frac{y_k}{\pi_k}}{\hat{N}}$, $\hat{N} = \sum_{k\in S}\frac{1}{\pi_k}$ y $1 = (1,\cdots,1)^T$.

$$\widehat{SCT_p}(x_0;h) = \frac{\hat{N}^{-1}\sum_{k\in S}\left(y_k-\widehat{\bar{y}}\right)^2\frac{K_h(x_k-x_0)}{\pi_k}}{\hat{f}(x_0;h)},$$

con $\hat{f}(x_0;h) = \hat{N}^{-1}\sum_{k\in S}\frac{K_h(x_k-x_0)}{\pi_k}$.
La expresión con notación matricial, es

$$\widehat{SCT_p}(x_0;h) = \left(Y_\pi - \widehat{\bar{y}}1\right)^T W_\pi(x_0;h)\left(Y_\pi - \widehat{\bar{y}}1\right),$$

desarrollando,

$$\begin{aligned}\widehat{SCT_p}(x_0;h) &= \left[\left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right) + \left(X_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}1\right)\right]^T \\ &= \left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right)^T W_\pi(x_0;h)\left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right) \\ &+2\left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right)W_\pi(x_0;h)\left(X_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}\right) \\ &+\left(x_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}\right)^T W_\pi(x_0;h)\left(x_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}\right),\end{aligned}$$

desarrollando el segundo término de la expresión anterior,

$$Y_\pi W_\pi(x_0;h)X_\pi(x_0)\hat{\beta}_\pi(x_0) - Y_\pi^T W_\pi(x_0;h)\widehat{\bar{y}}1 - \hat{\beta}_\pi{}^T(x_0)X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\hat{\beta}_\pi(x_0) + \hat{\beta}_\pi{}^T(x_0)X_\pi^T(x_0)W_\pi(x_0;h)\widehat{\bar{y}}1.$$

Sabemos que,

$$\hat{\beta}_\pi(x_0) = \left[X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\right]^{-1}X_\pi^T(x_0)W_\pi(x_0;h)Y_\pi,$$

luego se tiene que,

$$\hat{\beta}_\pi{}^T(x_0)X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\hat{\beta}_\pi(x_0) =$$
$$\left[\left[X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\right]^{-1}X_\pi^T(x_0)W_\pi(x_0;h)Y_\pi\right]^T X_\pi^T(x_0)W_\pi(x_0;h)X_\pi(x_0)\hat{\beta}_\pi(x_0) =$$
$$Y_\pi^T W_\pi(x_0;h)X_\pi(x_0)I_{n\times n}\hat{\beta}_\pi(x_0).$$

Además,

$$Y_\pi = X_\pi(x_0)\hat{\beta}_\pi(x_0),$$

luego,

$$Y_\pi^T = \hat{\beta}_\pi{}^T(x_0)X_\pi^T(x_0),$$

por lo que,

$$Y_\pi^T W_\pi(x_0;h)\widehat{\bar{y}}1 = \hat{\beta}_\pi{}^T(x_0)X_\pi^T(x_0)W_\pi(x_0;h)\widehat{\bar{y}}1,$$

por lo tanto,

$$\left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right)W_\pi(x_0;h)\left(X_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}\right) = 0.$$

Y así se tiene que,

$$\begin{aligned}\widehat{SCT_p}(x_0;h) &= \left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right)^T W_\pi(x_0;h)\left(Y_\pi - X_\pi(x_0)\hat{\beta}_\pi(x_0)\right)\\ &+ \left(X_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}1\right)^T W_\pi(x_0;h)\left(X_\pi(x_0)\hat{\beta}_\pi(x_0) - \widehat{\bar{y}}1\right).\end{aligned}$$

Por otro lado,

$$\begin{aligned}\widehat{SCR_p}(x_0;h) &= \tfrac{\hat{N}^{-1}}{\hat{f}(x_0;h)}\sum_{k\in s}\left\{\left(\sum_{j=0}^{p}\hat{\beta}_j\left(x_k-x_0\right)^j - \widehat{\bar{y}}\right)^2\frac{K_h(x_k-x_0)}{\pi_k}\right\}\\ &= \tfrac{\hat{N}^{-1}}{\hat{f}(x_0;h)}\sum_{k\in s}\left\{\left(\sum_{j=0}^{p}\hat{\beta}_j\left(x_k-x_0\right)^j - \widehat{\bar{y}}\right)\frac{K_h(x_k-x_0)}{\pi_k}\left(\sum_{j=0}^{p}\hat{\beta}_j\left(x_k-x_0\right)^j - \widehat{\bar{y}}\right)\right\}\\ &= \tfrac{\hat{N}^{-1}}{\hat{f}(x_0;h)} * \sum_{k\in s}\left\{\left(\sum_{j=0}^{p}\hat{\beta}_j(x_k-x_0)^j\right)\frac{K_h(x_k-x_0)}{\pi_k}\left(\sum_{j=0}^{p}\hat{\beta}_j(x_k-x_0)^j\right)\right\}\\ &-\sum_{k\in s}\left\{\left(\sum_{j=0}^{p}\hat{\beta}_j(x_k-x_0)^j\right)\frac{K_h(x_k-x_0)}{\pi_k}\widehat{\bar{y}}\right\}\\ &-\sum_{k\in s}\left\{\widehat{\bar{y}}\frac{K_h(x_k-x_0)}{\pi_k}\left(\sum_{j=0}^{p}\hat{\beta}_j(x_k-x_0)^j\right)\right\}\\ &+\sum_{k\in s}\left\{\widehat{\bar{y}}\frac{K_h(x_k-x_0)}{\pi_k}\widehat{\bar{y}}\right\} * .\end{aligned}$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} * \hat{\beta_\pi}^T(x_0) X_\pi^T(x_0) W_\pi(x_0;h) X_\pi(x_0) \hat{\beta_\pi}(x_0)$$
$$-\hat{\beta_\pi}^T(x_0) X_\pi^T(x_0) W_\pi(x_0;h) \widehat{\bar{y}} 1 - \widehat{\bar{y}} 1^T W_\pi(x_0;h) X_\pi(x_0) \hat{\beta_\pi}(x_0) + \widehat{\bar{y}} 1^T W_\pi(x_0,h) \widehat{\bar{y}} 1 *$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} \left[\left(X_\pi(x_0)\hat{\beta_\pi}(x_0) - \widehat{\bar{y}} 1 \right)^T W_\pi(x_0;h) \left(X_\pi(x_0)\hat{\beta_\pi}(x_0) - \widehat{\bar{y}} 1 \right) \right].$$

Ahora, para la suma de cuadrados de los errores,

$$\widehat{SCE_p}(x_0;h) = \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} \sum_{k\in s} \left\{ \left(y_k - \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \right)^2 \frac{K_h(x_k - x_0)}{\pi_k} \right\}$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} \sum_{k\in s} \left\{ \left(y_k - \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \right) \frac{K_h(x_k - x_0)}{\pi_k} \left(y_k - \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \right) \right\}$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} * \sum_{k\in s} \left\{ y_k \frac{K_h(x_k - x_0)}{\pi_k} y_k \right\} - \sum_{k\in s} \left\{ y_k \frac{K_h(x_k - x_0)}{\pi_k} \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \right\}$$
$$- \sum_{k\in s} \left\{ \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \frac{K_h(x_k - x_0)}{\pi_k} y_k \right\}$$
$$+ \sum_{k\in s} \left\{ \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \frac{K_h(x_k - x_0)}{\pi_k} \sum_{j=0}^{p} \hat{\beta}_j (x_k - x_0)^j \right\} *$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} * Y_\pi^T W_\pi(x_0;h) Y_\pi - Y_\pi^T W_\pi(x_0;h) X_\pi(x_0) \hat{\beta_\pi}(x_0)$$
$$-\hat{\beta_\pi}^T X_\pi^T(x_0) W_\pi(x_0;h) Y_\pi + \hat{\beta_\pi}^T X_\pi^T(x_0) W_\pi(x_0;h) x_\pi(x_0) \hat{\beta_\pi}(x_0) *$$

$$= \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} \left[\left(Y_\pi - X_\pi(x_0)\hat{\beta_\pi}(x_0) \right)^T W_\pi(x_0;h) \left(Y_\pi - X_\pi(x_0)\hat{\beta_\pi}(x_0) \right) \right].$$

Por lo tanto,

$$\widehat{SCT_p}(x_0;h) = \widehat{SCR_p}(x_0;h) + \widehat{SCE_p}(x_0;h).$$

Proposión 2

Demostración

Sean $(x_1, y_1), \cdots, (x_n, y_n)$ los valores de una muestra S de tamaño n, seleccionada bajo un diseño de muestreo medible $P(S)$ de una población U finita de tamaño N.
La estimación de la variable respuesta Y por regresión polinomial local mediante la muestra S está dada por la ecuación (28).

Por el resultado de la proposición 1 y bajo las condiciones (A), se tiene que el valor

esperado de la suma de cuadrados total estimada con la muestra es:

$$\int \widehat{SCT_p}(x_0;h)\hat{f}(x_0;h)dx_0 = \int \left[\widehat{SCR_p}(x_0;h) + \widehat{SCE_p}(x_0;h)\right] \hat{f}(x_0;h)dx_0.$$

Por la linealidad de la integral,

$$\int \widehat{SCT_p}(x_0;h) = \int \widehat{SCR_p}(x_0;h)\hat{f}(x_0;h)dx_0 + \int \widehat{SCE_p}(x_0;h)\hat{f}(x_0;h)dx_0,$$

por la definición 12 se tiene,

$$\begin{aligned}\widehat{SCT}(h) &= \int \widehat{SCT_p}(x_0;h)\hat{f}(x_0;h)dx_0 \\ &= \widehat{SCR_p}(h) + \widehat{SCE_p}(h).\end{aligned}$$

Además.

$$\begin{aligned}\widehat{SCT}(h) &= \int \frac{\hat{N}^{-1}}{\hat{f}(x_0;h)} \sum_{k\in s} \left(y_k - \widehat{\bar{y}}\right)^2 \frac{K_h(x_k - x_0)}{\pi_k} \hat{f}(x_0;h)dx_0 \\ &= \hat{N}^{-1} \int \sum_{k\in s} \left(y_k - \widehat{\bar{y}}\right)^2 \frac{K_h(x_k - x_0)}{\pi_k} dx_0 \\ &= \hat{N}^{-1} \sum_{k\in s} \int \left(y_k - \widehat{\bar{y}}\right)^2 \frac{K_h(x_k - x_0)}{\pi_k} dx_0 \\ &= \hat{N}^{-1} \sum_{k\in s} \frac{\left(y_k - \widehat{\bar{y}}\right)^2}{\pi_k} \int K_h(x_k - x_0)dx_0 \\ &= \hat{N}^{-1} \sum_{k\in s} \frac{\left(y_k - \widehat{\bar{y}}\right)^2}{\pi_k} \\ &= \widehat{SCT}.\end{aligned}$$

Apéndice C. Código para la creación de datos

Los datos para las figuras 1,2 y 3 se crearon mediante el siguiente código utilizando el paquete R:

```
GeneraM1< −function(N,r)
{
x< −as.numeric(lapply(1:N, function(a) a/(N+1)))
y< −2+sin(2*pi*x)
vare< −(r/(1-r))*var(y)
y< −y+rnorm(N,0,sqrt(vare))
variables< −matrix(c(x,y),nrow=N)
variables
}
```

Con N el número de datos y r la contribución de la parte residual a la variable respuesta. Para el ejemplo de las figuras se crearon $N = 1000$ datos con una contribución del $20\,\%$ ($r = 0.2$), esto mediante la siguiente instrucción:

```
datos< −GeneraM1(1000,0.2)
```

Apéndice D. Poblaciones con datos reales

Primera población

Para obtener los datos sobre las unidades económicas en actividades comerciales por municipio para los años 2003 y 2008, accesar al portal de internet del INEGI http://inegi.org.mx, seleccionar la liga "Estadística" y elegir dentro de la columna de "Banco de datos" la liga "Sistema Estatal y Municipal de Bases de Datos (SIMBAD)". Se abrirá otra ventana, dentro de la cual en el recuadro de "Filtro de contenidos:" escribir "Unidades Económicas" y seleccionar la opción "Principales características de las unidades económicas" y pulsar el botón filtrar. Esto llevará a una nueva ventana, y en la pestaña "1.Variables" seleccionar "Actividades comerciales"; en la pestaña "2.Años a consultar" elegir "2003" y "2008"; en la pestaña "3.Área geográfica" seleccionar todo; pulse el botón Ver municipios y pulse el botón Ver consulta. Esto mostrará los datos sobre las principales características de las unidades económicas en actividades comerciales. Para obtener la información por municipio, pulsar cada uno de los signos "+" en la columna "Clave", después en la parte inferior de la ventana elegir el formato del archivo y pulsar el botón Exportar.
Por otro lado, para obtener los datos sobre población y vivienda por municipio para el año 2005, regresar al portal del INEGI y seleccionar "Estadísticas" y elegir dentro de la columna "Bases de datos" la liga "Sistema Estatal y municipal de Bases de Datos (SIMBAD)". Se abre otra ventana y en el recuadro de "Filtro de contenidos" escribir "Población 2005", elegir "II Conteo de Población y Vivienda 2005" y pulsar el botón Filtrar, para la población en el 2005, pulsar "II conteo de Población y vivienda 2005" y seleccionar "Población Total"; en la pestaña "1.Variables" elegir "Sexo"; en la pestaña "2.Años a consultar" está por defecto "2005"; en la pestaña "3.Área geográfica" seleccionar todo; pulse el botón Ver municipios y puslse el botón Ver consulta. Para obtener los datos pulsar cada uno de los signos "+" en la columna "Clave", después en la parte inferior de la ventana elegir el formato del archivo y pulsar el botón Exportar.
De manera similar, para el "Total de viviendas particulares en 2005", regresar a la ventana de consulta al pulsar el botón Modificar Consulta y en la parte superior pulsar en el texto marcado en color amarillo "II Conteo de Población y Vivienda 2005", seleccionar "Vivienda" (última opción de la lista), en la pestaña "1.Variables" elegir "Tipo y clase de vivienda"; en la pestaña "2.Años a consultar" está por defecto "2005"; en la pestaña "3.Área geográfica" seleccionar todo; pulse el botón Ver municipios y puslse el botón Ver consulta. Para obtener los datos pulsar cada uno de los signos "+" en la columna "Clave", después en la parte inferior de la ventana elegir el formato del archivo y pulsar el botón Exportar.

De la base con la información sobre unidades económicas, se eligen las variables:

- Clave de la entidad y del municipio.
- Unidades económicas en 2003.
- Total de activos fijos en 2008,

y se le agrega la información de viviendas particulares y la población total en 2005.

Las variables que se utilizan para este trabajo son:

Variable	Descripción	Valores
Clave	Identificador del municipio	01001 a 32058
UE2003	Número de unidades económicas por municipio en 2003	Mín.=18, Máx.=17,140
Activos2008	Total de activos fijos por municipio en 2008	Mín.=520 Máx.=9,062,000
Viviendas	Total de viviendas particulares por municipio en 2005	Mín.=5,014, Máx.=357,100
Total	Población total por municipio en 2005	Mín.=102, Máx.=1,821,000

Segunda población

Para obtener los datos sobre las compañías del sector de alimentos, bebidas y medicina de los Estados Unidos de América en los años 2009 y 2011, accesar al portal de la escuela de negocios "Leonard N. Stern" de la universidad de Nueva York http://www.stern.nyu.edu/~ adamodar/New_Home_Page/data.html.
Del segundo cuadro sobre "Individual company information" se descargan directamente las bases de datos. Para este trabajo se descargan los archivos correspondientes a la columna US para "current(January 2011)" y "Jan 09".
Se empataron las bases por nombre de la compañia y se eligieron aquellas compañías cuyo código de clasificación industrial estándar corresponde a: bebidas, química básica, carbón, medicamentos, procesamiento de alimentos, suministros médicos, empaques y contenedores, productos de papel, farmacias y tabaco.

Las variables que se utilizan para este trabajo son:

Variable	Descripción
CompanyName	Nombre de la compañía
IndustryName	Nombre de la industria a la que pertenece la compañía
SICCode	Código de clasificación industrial estándar
Sizeclass	Tamaño de la compañía en 2009
MarketCap	Capital de mercado de la compañía en 2009
StockPrice	Precio promedio de almacén en 2011

Apéndice E. Programa en R

```
    GeneraM1< −function(N,r)
{
x< −as.numeric(lapply(1:N, function(a) a/(N+1)))
y< −2+(sin(2*pi*x^3))^3
vare< −(r/(1-r))*var(y)
y< −y+rnorm(N,0,sqrt(vare))
variables< −matrix(c(x,y),nrow=N)
variables
}

# La población "Hardle" con una contribución r = 0.2 se genera mediante la instrucción:
datosM12< −GeneraM1(5000,0.2)
# se toma la variable auxiliar X
# la cual será la misma para las demás poblaciones hipotéticas.
x< −datosM12[, 1]
Hardle< −function(N,r,x)
{
y< −2+(sin(2*pi*x^3))^3
vare< −(r/(1-r))*var(y)
y< −y+rnorm(N,0,sqrt(vare))
variables< −matrix(c(x,y),nrow=N)
variables
}
}
# La población "Hardle" con una contribución r = 0.8 se genera mediante la instrucción:
datosM18< −Hardle(5000,0.8,x)

GeneraM2< −function(N,r,x)
{
y< −1+2*(x-0.5)+exp(-200*(x-0.5)^2)
vare< −(r/(1-r))*var(y)
y< −y+rnorm(N,0,sqrt(vare))
variables< −matrix(c(x,y),nrow=N)
variables
}
# Las poblaciones "Bump" con una contribución r = {0.2, 0.8} se generan mediante la
```

instrucción:
datosM2r< −GeneraM1(5000,r,x)

```
# Kernel de Epanechnikov
kern< −function(x)
{
k< −rep(0, length(x))
k[abs(x) <= 1] < − 0.75 * (1 - x[abs(x) <= 1]^2)
return(k)
}

# Integración por el método del trapecio
Ha< −function(x,h)
{ N< −length(x)
a< −min(x)
b< −max(x)
v< −(b-a)/100
xp< −seq(a,b,length=100)
X< −cbind(rep(1,N),x-a)
W< −diag(kern((x-a)/h))/h
aprox< −v*W %* %X %* %solve(t(X) %* %W %* %X) %* %t(X) %* %W/2
for(j in 2:99)
{ X=cbind(rep(1,N),x-xp[j])
W=diag(kern((x-xp[j])/h))/h
H=W %* %X %* %solve(t(X) %* %W %* %X) %* %t(X) %* %W
aprox< −aprox+v*H
}
X=cbind(rep(1,N),x-b)
W=diag(kern((x-b)/h))/h
aprox< −aprox+v*W %* %X %* %solve(t(X) %* %W %* %X) %* %t(X) %* %W/2
return(aprox)
}

# Calculo de la suma de cuadrados del error con la población
Ha.SCE< −function(datos,h)
{ N< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
a< −min(x)
b< −max(x)
v< −(b-a)/100
xp< −seq(a,b,length=100)
Xa=cbind(rep(1,N),(x-a))
Wa=diag(kern((x-a)/h))/h
aprox< −v*(y %* %Wa %* %(diag(1,N)-Xa %* %solve(t(Xa) %* %Wa %* %Xa) %* %
t(Xa) %* %Wa) %* %y)/2
for(j in 2:99)
```

```
{
Xj=cbind(rep(1,N),(x-xp[j]))
Wj=diag(kern((x-xp[j])/h))/h
H=y %* %Wj %* %(diag(1,N)-Xj %* %solve(t(Xj) %* %Wj %* %Xj) %* %
t(Xj) %* %Wj) %* %y
aprox< −aprox+v*H
}
Xb=cbind(rep(1,N),(x-b))
Wb=diag(kern((x-b)/h))/h
aprox< −aprox+((v/2)*y %* %Wb %* %(diag(1,N)-Xb %* %solve(t(Xb) %* %Wb %* %Xb)
%* %t(Xb) %* %Wb) %* %y)
return(aprox)
}

# R2 global poblacional
R2g< −function(datos,h,tr.H.h)
{
N< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
SCEg< −Ha.SCE(datos,h)
ybar< −mean(y)
SCTg< −sum((y-ybar)^2)
L< −matrix(1/N,N,N)
SCRg< −SCTg-SCEg
R2g< −1-(SCEg/SCTg)
R2g.adj< −1-((SCEg/(N-tr.H.h))/ (SCTg/(N-1)))
R2.lin< −summary.lm(lm(y x))$r.squared
glnum< −tr.H.h-1
glden1< −(N-tr.H.h)
gldenF< −(1.25*tr.H.h)-0.5
F.h< −(R2g/glnum)/((1-R2g)/gldenF)
pvalor< −pf(as.numeric(F.h),as.numeric(glnum),as.numeric(gldenF),lower.tail=F)
return(c(round(SCTg,4),round(SCRg,4),round(SCEg),round(R2g,6),round(R2g.adj,6),
round(R2.lin,6),round(F.h,2),round(glnum,4),round(glden1,4),round(pvalor,8)))
}

#Matriz de proyección local, sólo para graficar
m.x0< −function(datos,x0,h)
{
N< −dim(datos)[1]
d< −length(x0)
x< −datos[,1]
y< −datos[,2]
W< −diag(kern((x-x0)/h))/h
X< −cbind(rep(1,N),x)
mx0< −solve(t(X) %* %W %* %X) %* %t(X) %* %W %* %y
```

```
mx0< −c(1,0) %* %mx0+((c(0,1) %* %mx0)*x0)
return(mx0)
}

# Calculo del coeficiente de determinación local
R2< −function(datos,x0,h)
{
N< −dim(datos)[1]
d< −length(x0)
x< −datos[,1]
y< −datos[,2]
W< −diag(kern((x-x0)/h))/h
X< −cbind(rep(1,N),x)
H< −X %* %solve(t(X) %* %W %* %X) %* %t(X) %* %W
SCE< −t(y) %* %W %* %(diag(1,N)-H) %* %y
ybar< −mean(y)
SCT< −t(y-rep(ybar,N)) %* %W %* %(y-rep(ybar,N))
R2< −1-(SCE/SCT)
return(R2)
}
HR2< −function(datos,x0,h)
{
N< −dim(datos)[1]
d< −length(x0)
x< −datos[,1]
y< −datos[,2]
W< −diag(kern((x-x0)/h))/h
X< −cbind(rep(1,N),x)
H< −X %* %solve(t(X) %* %W %* %X) %* %t(X) %* %W
return(H)
}

# Cálculo de probabilidades de inclusión para las poblaciones hipotéticas
# x datos de var aux. en población, n tamaño de muestra,
# c correlación entre x y tamaño de las unidades
library(sampling)
ProbInclu< −function(x,n,c)
{
N< −length(x)
x1< −runif(N)
z< −(c*x)+(sqrt(1-c^2)*x1)
pik< −inclusionprobabilities(z,n)
pikt=UPMEpiktildefrompik(pik)
w=pikt/(1-pikt)
q=UPMEqfromw(w,n)
resulta< −list(pik=pik,pikt=pikt,w=w,q=q)
resulta
```

```
}
Mpik< −function(x,n,c)
{
N< −dim(x)[1]
res< −ProbInclu(x[,1],n,c)
q< −res$q
pik< −res$pik
return(list(q=q,pik=pik))
}
#Selección de la muestra
Msel< −function(x,q,pik)
{
s=UPMEsfromq(q)
muestra< −list(x[s==1,],pik=pik[s==1])
muestra
}

#Selección de la muestra bajo MASSR
MASSR< −function(datos,n)
{
N< −dim(datos)[1]
s< −sample(1:N,n)
muestra< −datos[s,]
pik< −n/N
muestra< −cbind(muestra,pik)
return(muestra)
}

# Calculo de probabilidades de inclusión para las poblaciones reales
# x datos de var aux. en población, n tamaño de muestra,
# z es la variable que define el tamaño de las unidades.
ProbInclu< −function(x,n,z)
{
N< −length(x)
pik< −inclusionprobabilities(z,n)
pikt=UPMEpiktildefrompik(pik)
w=pikt/(1-pikt)
q=UPMEqfromw(w,n)
resulta< −list(pik=pik,pikt=pikt,w=w,q=q)
resulta
}
Mpik< −function(x,n,z)
{
N< −dim(x)[1]
res< −ProbInclu(x[,1],n,z)
q< −res$q
pik< −res$pik
```

```
return(list(q=q,pik=pik))
}

# Estimación del coeficiente de determinación global
Ha.SCE< −function(datos,h)
{
n< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
pik< −datos[,3]
a< −min(x)
b< −max(x)
v< −(b-a)/100
xp< −seq(a,b,length=100)
Xa=cbind(rep(1,n),(x-a))
Wa=diag(kern((x-a)/h)/pik)/h
aprox< −v*(t(y) %* %Wa %* %(diag(1,n)-Xa %* %solve(t(Xa) %* %Wa %* %Xa)
%* %t(Xa) %* %Wa) %* %y)/2
for(j in 2:99)
{
Xj=cbind(rep(1,n),(x-xp[j]))
Wj=diag(kern((x-xp[j])/h)/pik)/h
H=t(y) %* %Wj %* %(diag(1,n)-Xj %* %solve(t(Xj) %* %Wj %* %Xj)
%* %t(Xj) %* %Wj) %* %y
aprox< −aprox+v*H
}
Xb=cbind(rep(1,n),(x-b))
Wb=diag(kern((x-b)/h)/pik)/h
aprox< − aprox+((v/2)*t(y) %* %Wb %* %(diag(1,n)-Xb %* %
solve(t(Xb) %* %Wb %* %Xb) %* %t(Xb) %* %Wb) %* %y)
return(aprox)
}
R2g.Est< −function(datos,h,H.h.e)
{
n< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
pik< −datos[,3]
N.e< −sum(1/pik)
tr.H.h.e< −sum(diag(H.h.e))
SCEg< −Ha.SCE(datos,h)
ybar< −sum(y/pik)/N.e
SCTg< −sum(y^2/pik)-(ybar^2*N.e)
L< −matrix(1/n,n,n)
SCRg−SCTg-SCEg
R2g< −1-(SCEg/SCTg)
R2g.adj< −1-((SCEg/(N.e-tr.H.h.e))/ (SCTg/(N.e-1)))
```

```
R2.lin< −summary.lm(lm(y x))$r.squared
glnum< −tr.H.h.e-1
glden1< −N.e-tr.H.h.e
gldenF< −(1.25*tr.H.h.e)-0.5
F.h< −(R2g/glnum)/((1-R2g)/gldenF)
pvalor< −pf(F.h,glnum,gldenF,lower.tail=F)
return(c(round(SCTg),round(SCRg,4),round(SCEg,4),round(R2g,6),round(R2g.adj,6),
round(R2.lin,6), round(F.h,4),round(glnum,4),round(glden1,4),round(pvalor,8)))
}

# Gráfica del ajuste por RPL con datos de la muestra
m.x0< −function(datos,x0,h,pik)
{
N< −dim(datos)[1]
d< −length(x0)
x< −datos[,1]
y< −datos[,2]
W< −diag(kern((x-x0)/h)/pik)/h
X< −cbind(rep(1,N),x)
mx0< −solve(t(X) %* %W %* %X) %* %t(X) %* %W %* %y
mx0< −c(1,0) %* %mx0+((c(0,1) %* %mx0)*x0)
return(mx0)
}

# Estimación del coeficiente de determinación local
R2e.loc< −function(datos,x0,h)
{
n< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
pik< −datos[,3]
Wpi=diag(kern((x-x0)/h)/pik)/h
X< −cbind(rep(1,n),x)
Hpi< −X %* %solve(t(X) %* %Wpi %* %X) %* %t(X) %* %Wpi
SCE.e< −t(y) %* %Wpi %* %(diag(1,n)-Hpi) %* %y
N.est< −sum(1/pik)
ybar< −sum(y/pik)/N.est
SCT.e< −t(y-rep(ybar,n)) %* %Wpi %* %(y-rep(ybar,n))
R2loc< −1-(SCE.e/SCT.e)
return(R2loc)
}
HR2.e< −function(datos,x0,h)
{
n< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
pik< −datos[,3]
```

```
Wpi=diag(kern((x-x0)/h)/pik)/h
X< −cbind(rep(1,n),x)
Hpi< −X %* %solve(t(X) %* %Wpi %* %X) %* %t(X) %* %Wpi
return(Hpi)
}

# Estimación del promedio de la variable respuesta
# Sirve para la estimación de m(x0) y se pasa la muestra con 3 columnas X,Y,pik
m.x0.e< −function(muestra,x0,h)
{
n< −dim(muestra)[1]
x< −muestra[,1]
y< −muestra[,2]
pik< −muestra[,3]
W< −diag(kern((x-x0)/h)/pik)/h
X< −cbind(rep(1,n),(x-x0))
mx0e< −solve(t(X) %* %W %* %X) %* %t(X) %* %W %* %y
mx0e< −c(1,0) %* %mx0e
return(mx0e)
}
# Aqui los datos son la poblacion (X,Y), x0 es la variable auxiliar y pik son las piks en la
población
m.x0.aux< −function(datos,x0,h,pik)
{
N< −dim(datos)[1]
x< −datos[,1]
y< −datos[,2]
W< −diag(kern((x-x0)/h))/h
X< −cbind(rep(1,N),(x-x0))
mx0a< −solve(t(X) %* %W %* %X) %* %t(X) %* %W %* %y
mx0a< −c(1,0) %* %mx0a
return(mx0a)
}
# Se estiman los valores de m(x0) con la muestra
est.mx0< −function(muestra,h)
{
n< −dim(muestra)[1]
yest.x0< −rep(NA,n)
for(k in 1:n)
{
yest.x0[k]< −m.x0.e(muestra,muestra[k,1],h)
} return(yest.x0)
}
# Se guarda la variable auxiliar de la población (Xk) con sus piks
mx0.pob< −function(Pob,piks,h)
{
auxiliar< −cbind(Pob[,1],piks)
```

```
return(auxiliar)
}
# Calculo de m(xk) para k en la población
est.mx0N< −function(pob,auxiliar,h)
{ N< −dim(pob)[1]
mx0< −rep(NA,N)
for(k in 1:N)
{
mx0[k]< −m.x0.aux(pob,auxiliar[k,1],h,auxiliar[,2])
}
suma2< −sum(mx0)
return(list(mx0,suma2=suma2))
}
# Estimación del total de la variable respuesta por RPL
TotRPL< −function(pob,muestra,h,auxiliar,suma2)
{
N< −dim(pob)[1]
n< −dim(muestra)[1]
yest.x0< −rep(NA,n)
for(k in 1:n)
{
yest.x0[k]< −m.x0.e(muestra,muestra[k,1],h)
}
suma1< −sum(muestra[,2]/muestra[,3])-sum(yest.x0/muestra[,3])
tot.RPL< −suma1+suma2
return(tot.RPL)
}
# Estimación de Horvitz-Thompson para el promedio de la variable respuesta.
TotHT< −function(muestra)
{ ys< −muestra[,2]
piks< −muestra[,3]
tot.HT< −sum(ys/piks)
med.HT< −tot.HT/sum(1/piks)
return(med.HT)
}
# Estimación por regresión para el promedio de la variable respuesta.
Totrls< −function(poblacion,muestra)
{
x< −poblacion[,1]
xs< −muestra[,1]
ys< −muestra[,2]
piks< −muestra[,3]
modelo2< −lm(ys˜ xs)
residuos< −(modelo2$residuals)/piks
suma< −sum(modelo2$coefficients[1]+modelo2$coefficients[2]*x)
tot.rls< −suma+sum(residuos)
med.rls< −tot.rls/sum(1/piks)
```

```
return(med.rls)
}

# Algoritmo para estimar el coeficiente de determinación global
# Se cargan los datos de la población hipotética
PobM280< −read.table(file.choose(),header=T)
#Se calculan la probabilidades de inclusión con máx. ent.
Mm< −Mpik(PobM280,100,0.5)
#Se determina el valor del ancho de banda
h< −0.1301237
# Se da el número de simulaciones en R
R< −3000
R2gEstimada< −matrix(0,nrow=R,ncol=10)
system.time(
for(i in 1:R)
{
Muestra< −Msel(PobM280,Mmq, Mmpik)
H.h.e< −Ha.e(Muestra[[1]][,1],h)
datos.m.M280< −cbind(Muestra[[1]],Muestra$pik)
# Muestra< −MASSR(PobM280,100)
# H.h.e< −Ha.e(Muestra[,1],h)
# datos.m.M280< −cbind(Muestra[,1:2],Muestra[,3])
R2gEstimada[i,]< −R2g.Est(datos.m.M280,h,H.h.e)
}
)
# Se guardan los resultados, según experimento
write.table(R2gEstimada,"R2EM280ME1.txt")

# Algoritmo para estimar el coeficiente de determinación global
# Se cargan los datos de la población con datos reales

Pob< −read.table(file.choose(),header=T)
# Determinar el valor de h
h< −0.3*(max(Pob[,1])-min(Pob[,1]))
# Tamaño de la muestra
n< −60
# Tamaño de las unidades
z< −Pob[,3]
xU< −mean(Pob[,1])
system.time(Mm< −Mpik(Pob[,1:2],n,z))
auxiliar< −mx0.pob(Pob[,1:2],Mm$pik,h)
x0< −seq(0.05,0.95,0.05)
nx0< −length(x0)
system.time(mx0N< −est.mx0N(Pob[,1:2],auxiliar,h))
M< −3000
Resulta< −matrix(NA,nrow=M,ncol=3)
R2estx0< −matrix(NA,nrow=M,ncol=nx0)
```

```
totalRPL< −rep(NA,M)
media.est< −rep(NA,M)
EstimaHT< −rep(NA,M)
Estimarls< −rep(NA,M)
mediarls< −rep(NA,M)
R2gEstimada< −matrix(0,nrow=M,ncol=10)
system.time(
for(i in 1:M)
{
Muestra< −Msel(Pob[,1:2],Mm$q,Mm$pik)
muestra< −cbind(Muestra[[1]],Muestra$pik)
# Muestra< −MASSR(Pob[,1:2],n)
# muestra< −cbind(Muestra[,1:2],Muestra[,3])
for(j in 1:nx0)
{
R2estx0[i,j]< −R2e.loc(muestra,x0[j],h) }
totalRPL[i]< −TotRPL(Pob[,1:2],muestra,h,auxiliar,mx0N[1,3])
media.est[i]< −totalRPL[i]/sum(1/muestra[,3])
EstimaHT[i]< −TotHT(muestra)
Estimarls[i]< −Totrls(Pob,muestra)
Resulta[i,]< −cbind(media.est[i],EstimaHT[i],Estimarls[i])
H.h.e< −Ha.e(Muestra[[1]][,1],h)
# H.h.e< −Ha.e(Muestra[,1],h)
R2gEstimada[i,]< −R2g.Est(muestra,h,H.h.e)
}
) Resultas< −cbind(R2gEstimada,Resulta)
write.table(Resultas,R2ActivoME503.txt")
write.table(R2estx0,R2locME503.txt")
```