UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2013, Vol. 6, No. 2, pp. 175–195

Comparación entre árboles de regresión CART y regresión lineal

Comparison between CART regression trees and linear regression

Juan Felipe Díaz[a]
jfdiazs0@unal.edu.co

Juan Carlos Correa[b]
jccorrea@unal.edu.co

Resumen

La regresión lineal es el método más usado en estadística para predecir valores de variables continuas debido a su fácil interpretación, pero en muchas situaciones los supuestos para aplicar el modelo no se cumplen y algunos usuarios tienden a forzarlos llevando a conclusiones erróneas. Los árboles de regresión CART son una alternativa de regresión que no requiere supuestos sobre los datos por analizar y es un método de fácil interpretación de los resultados. En este trabajo se comparan a nivel predictivo la regresión lineal con CART mediante simulación. En general, se encontró que cuando se ajusta el modelo de regresión lineal correcto a los datos, el error de predicción de regresión lineal siempre es menor que el de CART. También se encontró que cuando se ajusta erróneamente un modelo de regresión lineal a los datos, el error de predicción de CART es menor que el de regresión lineal solo cuando se tiene una cantidad de datos suficientemente grande.

***Palabras clave*:** simulación, error de predicción, regresión lineal, árboles de clasificación y regresión CART.

Abstract

Linear regression is the most widely used method in statistics to predict values of continuous variables due to its easy interpretation, but in many situations the suppositions to apply the model are not met and some users tend to force them leading them to erroneous conclusions. CART regression trees is a regression alternative that does not require suppositions on the data to be analyzed and is a method of easy interpretation of results. This work compares predictive levels of linear regression with CART through simulation. In general, it was found that when the correct linear regression model is adjusted to the data, the prediction

[a]Maestría en Ciencias - Estadística. Universidad Nacional de Colombia, sede Medellín, Colombia.
[b]Profesor Asociado. Universidad Nacional de Colombia, sede Medellín, Colombia.

error of linear regression is always lower than that of CART. It was also found that when linear regression model is erroneously adjusted to the data, the prediction error of CART is lower than that of linear regression only when it has a sufficiently large amount of data.

***Keywords*:** simulation, prediction error, linear regression, CART classification and regression trees.

1. Introducción

El modelo lineal clásico ha sido utilizado extensivamente y con mucho éxito en múltiples situaciones. Tiene ventajas que lo hacen muy útil para el usuario, debido a que es fácil de interpretar, fácil de estimar y poco costoso. La facilidad de interpretación de este modelo lo ha popularizado bastante y no es raro ver su ajuste en situaciones inapropiadas, por ejemplo, en respuestas que son discretas o sesgadas; y el desespero por parte de los usuarios por aproximarse a él, por ejemplo, mediante transformaciones de los datos, sin considerar los cambios en la estructura del error. Por lo anterior, es necesario un modelo que tenga similares ventajas y que no sea tan rígido con los supuestos, para que el usuario final lo pueda aplicar tranquilamente.

Los árboles de clasificación y regresión (CART) es un método que utiliza datos históricos para construir árboles de clasificación o de regresión, los cuales son usados para clasificar o predecir nuevos datos. Estos árboles CART pueden manipular fácilmente variables numéricas y categóricas. Entre otras ventajas está su robustez a *outliers*, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

Desde el planteamiento de los árboles de clasificación y regresión CART por Leo Breiman y otros en 1984 (Breiman et al. 1984), se presentó gran interés en la utilización de esta metodología por parte de la comunidad científica debido a su fácil implementación en todo tipo de problemas y su clara interpretación de los resultados.

Muchos investigadores después de la publicación del libro de Breiman (Breiman et al. 1984) han planteado variaciones del método en sus distintas etapas, pero en muchos casos la idea inicial del particionamiento recursivo es la misma, otros han aplicado CART y sus variaciones en distintos campos como la medicina, la biología y el aprendizaje de máquinas. Algunos investigadores han comparado esta metodología con otras técnicas de modelamiento como Tamminen, Laurinen y Roning (Tamminen et al. 1999) quienes en 1999, debido a que el sistema físico de los humanos es altamente no lineal y la regresión lineal tradicional no puede ser usada como modelo de aproximación, compararon los árboles de regresión con las redes neuronales en un conjunto de datos obtenidos por un método de medición de aptitud aeróbica, concluyendo que las redes neuronales son una potente herra-

mienta de aproximación, pero se dificulta la interpretación del modelo, mientras que los árboles de regresión son fáciles de visualizar y su estructura es más comprensible. Ankarali, Canan, Akkus, Bugdayci y Ali Sungur (Ankarali et al. 2007) en 2007 compararon los métodos de árboles de clasificación y regresión logística en la determinación de factores de riesgo sociodemográficos que influyen en el estado de depresión de 1447 mujeres en periodos separados de posparto, y concluyeron que los árboles de clasificación dan información más detallada sobre el diagnóstico mediante la evaluación conjunta de una gran cantidad de factores de riesgo que el modelo de regresión logística.

El problema central es comparar por medio de un estudio de simulación, a nivel predictivo, el método no paramétrico CART con el método paramétrico Regresión lineal, dos técnicas que tienen similares ventajas en cuanto a la simplicidad de sus modelos y su fácil interpretación de los resultados. En la sección 2 se presenta el método CART: particionamiento recursivo, árboles de clasificación y árboles de regresión. En la sección 3 se describe el estudio de simulación: errores de predicción y pasos. En las secciones 4 y 5 se simulan conjuntos de datos cuyo verdadero modelo es un modelo de regresión lineal y se ajusta a estos datos tanto los modelos de regresión correctos como modelos de regresión incorrectos, para comparar luego sus errores de predicción con los errores de predicción de árboles de regresión ajustados a los mismos datos. En las secciones 6 y 7 se dan las conclusiones y agradecimientos.

2. CART

2.1. Particionamiento recursivo

El algoritmo conocido como particionamiento recursivo es el proceso paso a paso para construir un árbol de decisión y es la clave para el método estadístico no paramétrico CART (Izenman 2008).

Sea Y una variable respuesta y sean p variables predictoras $x_1, x_2, \ldots, x_p$, donde las x´s son tomadas fijas y Y es una variable aleatoria. El problema estadístico es establecer una relación entre Y y las x´s de tal forma que sea posible predecir Y basado en los valores de las x´s. Matemáticamente, se quiere estimar la probabilidad condicional de la variable aleatoria Y,

$$P[Y = y|x_1, x_2, \ldots, x_p]$$

cuando la variable Y es discreta, o un funcional de su probabilidad tal como la esperanza condicional

$$E[Y|x_1, x_2, \ldots, x_p].$$

cuando la variable Y es continua.

2.1.1. Elementos de la construcción del árbol

Según Zhang & Singer (2010) para ilustrar las ideas básicas considere el diagrama de la Figura 1.

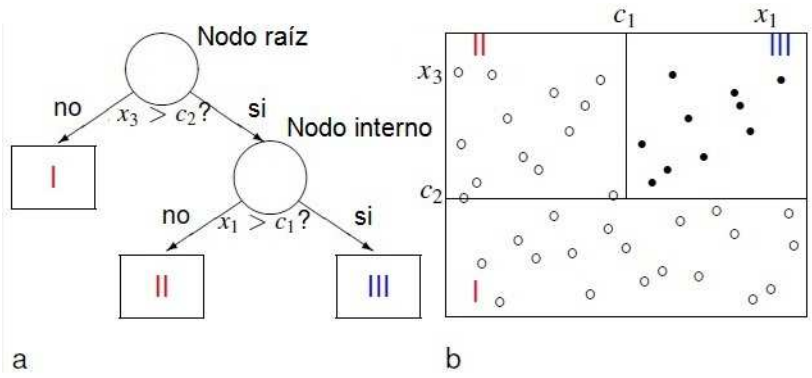


Figura 1: *Ejemplo árbol. Fuente: modificado de Zhang & Singer 2010.*

El árbol tiene tres niveles de nodos. El primer nivel tiene un único nodo en la cima (el círculo) llamado nodo raíz. Un nodo interno (el círculo) en el segundo nivel, y tres nodos terminales (las cajas) que están respectivamente en el segundo y tercer nivel. El nodo raíz y el nodo interno son particionados cada uno en dos nodos en el siguiente nivel, los cuales son llamados nodos hijos izquierdo y derecho.

El nodo raíz contiene una muestra de sujetos desde la cual se aumenta el árbol, es decir, desde donde se desprenden los demás nodos. Estos sujetos constituyen lo que se llama una muestra de aprendizaje, la cual puede ser la muestra total en estudio o una parte de esta.

El objetivo del particionamiento recursivo es acabar en nodos terminales que sean homogéneos en el sentido de que ellos contengan solo puntos o círculos Figura 1b.

Una medida cuantitativa de la homogeneidad de un nodo es la noción de impureza, para la cual se define el siguiente indicador:

$$\text{Impureza del nodo} = \frac{\#\text{ sujetos que cumplen la característica en el nodo}}{\#\text{ total de sujetos en el nodo}}. \quad (1)$$

En la Figura 1, si la característica es ser círculo, el nodo hijo izquierdo del nodo raíz tiene impureza igual a 1, debido a que en este nodo solo hay círculos, pero, si la característica es ser punto, la impureza es igual a 0, debido a que no hay ningún punto en este nodo. Nótese que en el nodo hijo derecho del nodo raíz el número de círculos es aproximadamente igual al número de puntos, teniendo este nodo una medida de la impureza de aproximadamente 0.5 independientemente de si la característica sea ser círculo o punto. Mientras más homogéneo sea el nodo el límite del cociente en la ecuación (1) es 0 o 1.

2.1.2. División de un nodo

Para dividir el nodo raíz en dos nodos homogéneos, se debe seleccionar entre los rangos de todas las variables predictoras el valor de la división que más lleve al límite de 0 o 1 el cociente en la ecuación (1) para cada nodo hijo. En la Figura 1 a) se seleccionó como división el valor c_2 entre el rango de la variable x_3. El proceso continua para los dos nodos hijos, teniendo en cuenta para cada nodo el rango resultante de la variable con la que se dividió el nodo padre y el rango de las demás variables involucradas.

Antes de seleccionar la mejor división, se debe definir la bondad de una división. Se busca una división que resulte en dos nodos hijos puros (u homogéneos). Sin embargo, en la realidad los nodos hijos son usual y parcialmente puros. Además, la bondad de una división debe poner en una balanza la homogeneidad (o la impureza) de los dos nodos hijos simultáneamente.

2.1.3. Nodos terminales

El proceso de particionamiento recursivo continúa hasta que el árbol sea saturado en el sentido de que los sujetos en los nodos descendientes no se pueden partir en una división adicional. Esto sucede, por ejemplo, cuando en un nodo queda solo un sujeto. El número total de divisiones permitidas para un nodo disminuye cuando aumentan los niveles del árbol. Cualquier nodo que no pueda o no sea dividido es un nodo terminal. El árbol saturado generalmente es bastante grande para utilizarse, porque los nodos terminales son tan pequeños que no se puede hacer inferencia estadística razonable, debido a que los datos quedan "sobre-ajustados", es decir, el árbol alcanza un ajuste tan fiel a la muestra de aprendizaje que cuando en la práctica se aplique el modelo obtenido a nuevos datos los resultados pueden ser muy malos, y por tanto, no es necesario esperar hasta que el árbol sea saturado. En lugar de esto, se escoge un tamaño mínimo de nodo apriori. Se detiene la división cuando el tamaño del nodo es menor que el mínimo. La escogencia del tamaño mínimo depende del tamaño de muestra (uno por ciento) o se puede tomar simplemente como cinco sujetos (los resultados generalmente no son significativos con menos de cinco sujetos).

Breiman et al. (1984) argumentan que dependiendo del límite de parada, el particionamiento tiende a terminar muy pronto o muy tarde. En consecuencia, ellos hacen un cambio fundamental introduciendo un segundo paso llamado "poda".

La poda consiste en encontrar un subárbol del árbol saturado que sea el más "predictivo" de los resultados y menos vulnerable al ruido en los datos. Los subárboles se obtienen podando el árbol saturado desde el último nivel hacia arriba. Por ejemplo, el árbol de la Figura 2a es un subárbol del árbol de la Figura 2b.

Los pasos de particionamiento y poda se pueden ver como variantes de los procesos paso a paso *forward* y *backward* en regresión lineal.

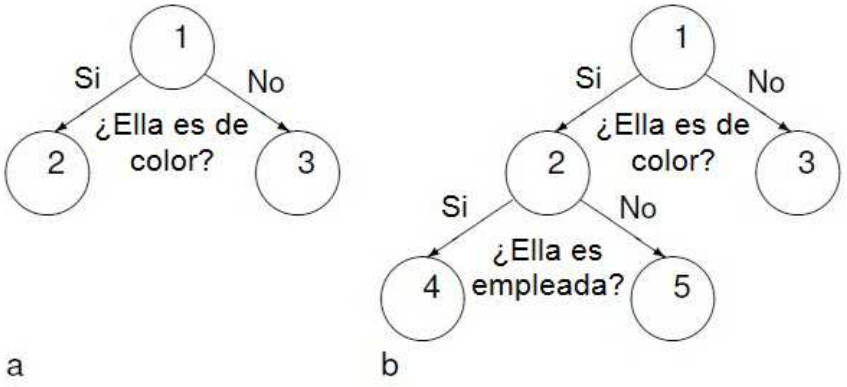


Figura 2: *El nodo 1 se divide en los nodos 2 y 3, luego, el nodo 2 se divide en los nodos 3 y 4. Fuente: modificado de Zhang & Singer 2010.*

2.2. Árboles de clasificación

Los árboles de clasificación y regresión (CART) fueron desarrollados en los años ochenta por Breiman, Freidman, Olshen y Stone en el libro *Classification and regression trees* (Breiman et al. 1984).

La metodología CART utiliza datos históricos para construir árboles de clasificación o de regresión, los cuales son usados para clasificar o predecir nuevos datos. Estos árboles CART pueden manipular fácilmente como variable respuesta variables numéricas y categóricas. Entre otras ventajas está su robustez a *outliers*, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

Esta metodología consiste de tres pasos:

- Construcción del árbol saturado.
- Escogencia del tamaño correcto del árbol.
- Clasificación de nuevos datos usando el árbol construido.

La construcción del árbol saturado se hace con particionamiento recursivo. La diferencia en la construcción de los árboles de clasificación y los árboles de regresión es el criterio de división de los nodos, es decir, la medida de impureza y la bondad de una división es diferente para los árboles de clasificación y de regresión. En esta sección se considera primero la construcción de árboles de clasificación.

2.2.1. Determinación de la división de un nodo

Sea Y una variable dicotómica con valores 0 y 1, y sea τ un nodo. Para construir el árbol saturado, en el proceso de particionamiento recursivo se tiene que, si τ es el nodo menos impuro la impureza es 0 y debe tener como resultado $P[Y = 1|\tau] = 0$ o $P[Y = 1|\tau] = 1$. El nodo τ es más impuro cuando su impureza es 1 con $P[Y = 1|\tau] = \frac{1}{2}$. Por tanto, la función impureza tiene una forma cóncava y se puede definir formalmente como

$$i(\tau) = \phi(P[Y = 1|\tau]), \tag{2}$$

donde ϕ tiene las siguientes propiedades,

(i) $\phi \geq 0$ y

(ii) para cualquier $p \in (0, 1)$, $\phi(p) = \phi(1 - p)$ y $\phi(0) = \phi(1) < \phi(p)$.

Las escogencias más comunes de funciones de impureza para la construcción de árboles de clasificación son:

- $\phi(p) = min(p, 1 - p)$, (mínimo error o error de Bayes)
- $\phi(p) = -p\log(p) - (1 - p)\log(1 - p)$, (entropía)
- $\phi(p) = p(1 - p)$, (índice Gini)

donde, se define $0\log(0) := 0$.

Además, se define la bondad de una división s como

$$\Delta I(\tau) = i(\tau) - P[\tau_L]i(\tau_L) - P[\tau_R]i(\tau_R), \tag{3}$$

donde τ es el nodo padre del nodo izquierdo τ_L y del nodo derecho τ_R, y $P[\tau_L]$ y $P[\tau_R]$ son respectivamente las probabilidades de que un sujeto caiga dentro de los nodos τ_L y τ_R.

La ecuación (3) mide el grado de reducción de la impureza cuando se pasa del nodo padre a los nodos hijos. Se selecciona s tal que $\Delta I(\tau)$ sea máxima.

2.2.2. Determinación de los nodos terminales

Una vez se tiene construido el árbol saturado se inicia la etapa de poda. La poda consiste en encontrar el subárbol del árbol saturado con la mejor calidad en cuanto a que sea lo más predictivo posible y lo menos sensible al ruido de los datos. Es decir, se debe definir una medida de calidad de un árbol. Para esto se debe recordar que el objetivo de los árboles de clasificación es el mismo que el

del particionamiento recursivo: extraer subgrupos homogéneos de la población o muestra en estudio. Para alcanzar este objetivo se debe tener certeza de que los nodos terminales son homogéneos, es decir, la calidad de un árbol es simplemente la calidad de sus nodos terminales. Por tanto, para un árbol $\mathcal{T}$ se define

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} \mathrm{P}[\tau]\mathrm{r}(\tau), \tag{4}$$

donde $\tilde{\mathcal{T}}$ es el conjunto de nodos terminales de $\mathcal{T}$, $\mathrm{P}[\tau]$ es la probabilidad de que un sujeto pertenezca al nodo τ y $\mathrm{r}(\tau)$, es una medida de calidad del nodo τ la cual es similar a la suma de cuadrados de los residuales en regresión lineal.

El propósito de la poda es seleccionar el mejor subárbol, $\mathcal{T}^*$, de un árbol saturado inicialmente, $\mathcal{T}_0$, tal que $R(\mathcal{T}^*)$ sea mínimo.

Una escogencia obvia para $\mathrm{r}(\tau)$ es la medida de impureza del nodo τ, aunque en general se toma como el costo de mala clasificación, es decir, $\mathrm{r}(\tau) = \sum_{i=1}^{l}\{c(j|i)\mathrm{P}[Y = i|\tau]\}$, donde $c(i|j)$ es el costo de mala clasificación de que un sujeto de la clase j sea clasificado en la clase i, con $i, j = 1, \ldots, l$. Cuando $i = j$, se tiene la clasificación correcta y el costo debería ser cero, es decir, $c(i|i) = 0$.

Generalmente, es difícil en la práctica medir el costo relativo $c(j|i)$ para $i \neq j$, y por tanto, no se puede asignar el costo de mala clasificación de cada nodo antes de aumentar cualquier árbol, incluso cuando se conoce el perfil del árbol. Por otra parte, existe suficiente evidencia empírica en la literatura que demuestra que el uso de una función de impureza como la entropía usualmente lleva a árboles útiles con tamaños de muestra razonables.

Estimación del costo de mala clasificación

Sea $R^s(\tau)$ la proporción de elementos mal clasificados del nodo τ, también conocida como *estimación por resustitución del costo de mala clasificación para el nodo* τ. Se define la *estimación por resustitución del costo de mala clasificación para el árbol* $\mathcal{T}$ como,

$$R^s(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} R^s(\tau). \tag{5}$$

Zhang & Singer (2010) afirman que esta estimación por resustitución generalmente subestima el costo. Además, Breiman et al. (1984) prueban que a medida que aumentan los nodos en el árbol disminuye la estimación por resustitución (5), y como consecuencia, este estimador tiene el problema de seleccionar árboles sobreajustados.

2.2.3. Costo-complejidad

El tamaño del árbol es importante a la hora de dar conclusiones sobre la muestra o población en estudio, debido a que un árbol con una gran cantidad de nodos puede tener problemas de sobreajuste. Una medida de la calidad de un árbol debe tener en cuenta tanto la calidad de los nodos terminales como el tamaño del árbol (número de nodos del árbol), y tener en cuenta solo el costo de mala clasificación puede llevar a árboles muy grandes.

Se define el *costo-complejidad* del árbol $\mathcal{T}$ como

$$R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha|\tilde{\mathcal{T}}|, \tag{6}$$

donde α (≥ 0) es el parámetro de complejidad y $|\tilde{\mathcal{T}}|$ es el número de nodos terminales en $\mathcal{T}$ llamado *complejidad* del árbol $\mathcal{T}$. La diferencia entre $R(\mathcal{T})$ y $R_\alpha(\mathcal{T})$ como una medida de la calidad del árbol reside en que $R_\alpha(\mathcal{T})$ penaliza un gran árbol.

Aunque se dijo anteriormente que la aproximación por resustitución tiene sus problemas al estimar el costo de mala clasificación para un nodo, es muy útil al estimar el costo-complejidad.

El uso del costo-complejidad permite construir una secuencia de *subárboles óptimos anidados* (ver Zhang & Singer 2010) desde cualquier árbol $\mathcal{T}$ dado. La idea es construir una secuencia de subárboles anidados para un árbol saturado $\mathcal{T}$, minimizando el costo-complejidad $R_\alpha(\mathcal{T})$, y seleccionar como subárbol final el que tenga el más pequeño costo de mala clasificación de estos subárboles.

Cuando se tiene una muestra de prueba, estimar $R(\mathcal{T})$ es sencillo para cualquier subárbol $\mathcal{T}$, porque solo se necesita aplicar los subárboles a la muestra de prueba y luego se escoge el mejor valor de α, pero, si no se tiene una muestra de prueba, se pueden crear muestras artificiales utilizando el proceso de *validación cruzada* (ver Zhang & Singer 2010) para estimar $R(\mathcal{T})$ y así escoger el mejor valor de α.

2.3. Árboles de regresión

En la construcción de árboles de clasificación se indicó que es necesario una medida de impureza dentro de un nodo, es decir, un criterio de división de nodo para construir un gran árbol y luego un criterio de costo-complejidad para podarlo. Estas directrices generales se aplican cada vez que se intenta desarrollar métodos basados en árboles. Para la construcción de árboles de clasificación la variable respuesta debe ser categórica, mientras que para la construcción de árboles de regresión la variable respuesta debe ser continua. En general, la metodología para construir árboles de clasificación y árboles de regresión es la misma, por tanto, los pasos vistos anteriormente para construir árboles de clasificación son aplicables en la construcción de árboles de regresión. La diferencia radica en la escogencia de la función impureza para dividir un nodo y en la estimación del costo-complejidad

para podar el árbol. Para una respuesta continua, una escogencia natural de la impureza para un nodo τ es la varianza de la respuesta dentro del nodo:

$$i(\tau) = \sum_{\text{sujeto } i \in \tau} (Y_i - \bar{Y}(\tau))^2, \tag{7}$$

donde $\bar{Y}(\tau)$ es el promedio de Y_i´s dentro del nodo τ. Para dividir un nodo τ en dos nodos hijos, τ_L y τ_R, se define la bondad de una división s como

$$\Delta I(\tau) = i(\tau) - i(\tau_L) - i(\tau_R). \tag{8}$$

A diferencia de la ecuación (3), la ecuación (8) no necesita pesos. Además, se puede hacer uso de $i(\tau)$ para definir el costo del árbol como

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} i(\tau), \tag{9}$$

y luego sustituirlo en la ecuación (6) para formar el costo-complejidad.

3. Descripción del estudio de simulación

3.1. Medidas del error de predicción

Suponga que se tiene un conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ que sigue un modelo de regresión lineal:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \text{ donde } \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n. \tag{10}$$

De lo anterior se sabe que

$$y_{verd_i} = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},\ i = 1, \ldots, n.$$

Se ajusta un modelo de regresión lineal a los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, luego, los valores predichos son de la forma:

$$y_{reg_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip},\ i = 1, \ldots, n,$$

donde, $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ son las estimaciones por mínimos cuadrados de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$.

Por tanto, el error de predicción por regresión lineal se calcula como

$$EPRL = \frac{\sum_{i=1}^{n}(y_{reg_i} - y_{verd_i})^2}{n}. \tag{11}$$

Además, se ajusta un árbol de regresión a los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, obteniendo un árbol de l nodos terminales. Sean $C_1, C_2, \ldots, C_l$ las clases correspondientes a los l nodos terminales, luego, los valores predichos por el árbol de regresión son de la forma:

$$y_{cart_i} = \begin{cases} r_k & \text{si} \quad x_i \in C_k \ ; \ \ k = 1, \ldots, l \\ 0 & \text{si} \quad \text{en otro caso} \end{cases}$$

donde,

$$r_k = \frac{\sum\{y_i | x_i \in C_k, \ i = 1, \ldots, n\}}{\#(\{y_i | x_i \in C_k, \ i = 1, \ldots, n\})} \ ; \quad k = 1, \ldots, l.$$

Por tanto, el error de predicción por CART se calcula como

$$EPCART = \frac{\sum_{i=1}^{n}(y_{cart_i} - y_{verd_i})^2}{n}. \tag{12}$$

3.2. Pasos del estudio de simulación

Los conjuntos de datos simulados en este trabajo se generan de modelos de regresión lineal de la forma:

$$Y_i = F(x_{i1}, x_{i2}, \ldots, x_{ip}) + \varepsilon_i, \text{ donde } \varepsilon_i \sim N(0, \sigma^2), \ i = 1, \ldots, n \tag{13}$$

donde

$$E[Y_i] = F(x_{i1}, x_{i2}, \ldots, x_{ip}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x_i) = f(x_i), \ i = 1, \ldots, n \tag{14}$$

mediante los siguientes pasos:

1. Se especifican las funciones $g_1(x), \ldots, g_p(x)$ y valores de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ en la ecuación (14).
2. Se genera una secuencia de n números $x_1, x_2, \ldots, x_n$ igualmente espaciados del conjunto (soporte) $X = [1, 100]$.
3. Se generan aleatoriamente n números $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ de la distribución $N(0, \sigma^2)$.

4. Se calculan los valores $y_i = f(x_i) + \varepsilon_i$ para todo $i = 1, \ldots, n$.

5. Se estandarizan los datos $y_1, y_2, \ldots, y_n$ obteniendo $y_1^*, y_2^*, \ldots, y_n^*$, donde,

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \tag{15}$$

6. Se toma como muestra de aprendizaje $\mathcal{L} = \{(x_1, y_1^*), (x_2, y_2^*), \ldots, (x_n, y_n^*)\}$ la cual sigue el modelo de regresión lineal descrito por la ecuación (13).

7. Para la muestra de aprendizaje $\mathcal{L}$ se ajusta un modelo de regresión lineal utilizando la librería *MASS* y se ajusta un árbol de regresión utilizando la librería *rpart* del paquete estadístico R.

8. Se estiman los errores de predicción para el modelo de regresión lineal ajustado y para el árbol de regresión ajustado, los cuales se definen respectivamente en las ecuaciones (11) y (12).

9. Se repiten los pasos 3 a 8 para obtener 1000 errores de predicción por regresión lineal $EPRL_1$, $EPRL_2$,..., $EPRL_{1000}$ y 1000 errores de predicción por árboles de clasificación $EPCART_1$, $EPCART_2$,..., $EPCART_{1000}$.

10. Se calcula el promedio de los 1000 errores de predicción para regresión lineal y el promedio de los 1000 errores de predicción para árboles de regresión, los cuales son respectivamente $EPRL = \frac{\sum_{k=1}^{1000} EPRL_k}{1000}$ y $EPCART = \frac{\sum_{k=1}^{1000} EPCART_k}{1000}$.

11. Se calcula la diferencia de logaritmos de los errores de predicción, $DIFLOG = Log(EPCART) - Log(EPRL)$, la cual es una medida de proximidad de los dos errores. A medida que $DIFLOG \to 0$, los dos errores de predicción se van acercando entre ellos. Si $DIFLOG > 0$ entonces $EPCART > EPRL$ y la regresión lineal predice mejor los datos que los árboles de regresión, pero, si $DIFLOG < 0$ entonces $EPCART < EPRL$ y los árboles de regresión predicen mejor los datos que la regresión lineal. Si $DIFLOG = 0$ entonces $EPCART = EPRL$ y ambos modelos predicen igual.

4. Comparación de las predicciones cuando el modelo lineal ajustado es el correcto

En esta sección se supone que los datos siguen un modelo de regresión lineal específico. Se ajusta un árbol de regresión CART y el modelo correcto a los datos para predecir la respuesta. El objetivo es comparar las magnitudes de los errores de predicción de CART y de regresión lineal, cambiando el tamaño y la varianza de los errores de los datos. A continuación, se simularán conjuntos de datos para cinco modelos de regresión lineal, dos modelos cuadráticos y tres trigonométricos, variando el número de datos y la desviación estándar de los errores.

4.1. Predicción de modelos de regresión lineal cuadráticos

Suponga que se tiene un conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ que sigue un modelo de regresión cuadrático de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \text{ donde } \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n. \tag{16}$$

De lo anterior, se sabe que

$$y_{verd_i} = E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,\ i = 1, \ldots, n. \tag{17}$$

Para simular los conjuntos de datos se siguen los pasos descritos en la sección 1.5. En el paso 1, se toma $p = 2$ y se especifican las funciones

$$g_1(x) = x, \quad g_2(x) = x^2. \tag{18}$$

El primer modelo por analizar se obtiene al sustituir $\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 3$ en la ecuación (16) y se llamará modelo cuadrático 1. El segundo modelo por analizar se obtiene al sustituir $\beta_0 = 680,\ \beta_1 = -22,\ \beta_2 = 0.25$ en la ecuación (16) y se llamará modelo cuadrático 2.

En la Tabla 1 se puede observar para los modelos cuadráticos 1 y 2, que para cualquier valor de n fijo, al aumentar la desviación estándar σ, los errores de predicción de la regresión lineal y de CART se aproximan entre sí, siendo en todos los casos menor el error de predicción de la regresión lineal.

En los gráficos de los modelos cuadráticos de la Tabla 2, se puede ver cómo las predicciones de CART describen la forma del verdadero modelo de los datos simulados para $n = 100$ y $n = 1000$.

4.2. Predicción de modelos de regresión lineal trigonométricos

Suponga que se tiene un conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ que sigue un modelo trigonométrico de la forma:

$$y_i = a \sin(bx_i + c) + d + \varepsilon_i, \text{ donde } \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n \tag{19}$$

donde el valor de b es conocido. De lo anterior se tiene que

$$y_{verd_i} = E(y_i) = a \sin(bx_i + c) + d,\ i = 1, \ldots, n. \tag{20}$$

El modelo (19) se puede reescribir como

$$a \sin(bx_i + c) + d + \varepsilon_i = a \sin(c) \cos(bx_i) + a \cos(c) \sin(bx_i) + d + \varepsilon_i,\ i = 1, \ldots, n. \tag{21}$$

Tabla 1: *Comparación de los errores de predicción para los modelos cuadráticos. Fuente: elaboración propia*

Cuadrático 1			Cuadrático 2		
n	σ	$DIFLOG$	n	σ	$DIFLOG$
	1	8.8029		1	5.4355
	10	6.7680		5	4.0401
50	100	4.7744	50	10	3.4510
	500	3.3271		25	2.5974
	1000	2.6916		50	1.8948
	2000	1.9102		100	1.1250
	1	8.6704		1	5.7342
	10	6.6368		5	4.2644
100	100	4.5679	100	10	3.5766
	500	3.0888		25	2.6834
	1000	2.3725		50	1.8463
	2000	1.5971		100	0.9506
	1	9.1452		1	6.0454
	10	7.0944		5	4.5622
500	100	5.0816	500	10	3.9003
	500	3.5544		25	2.7633
	1000	2.6974		50	1.7254
	2000	1.6999		100	0.7671
	1	9.4044		1	6.3086
	10	7.4077		5	4.8416
1000	100	5.3838	1000	10	4.0910
	500	3.7601		25	2.8315
	1000	2.7961		50	1.7460
	2000	1.7307		100	0.7736
	1	10.1050		1	7.0107
	10	8.1027		5	5.3742
5000	100	6.0591	5000	10	4.3912
	500	3.9929		25	2.9208
	1000	2.8959		50	1.7731
	2000	1.7763		100	0.7815

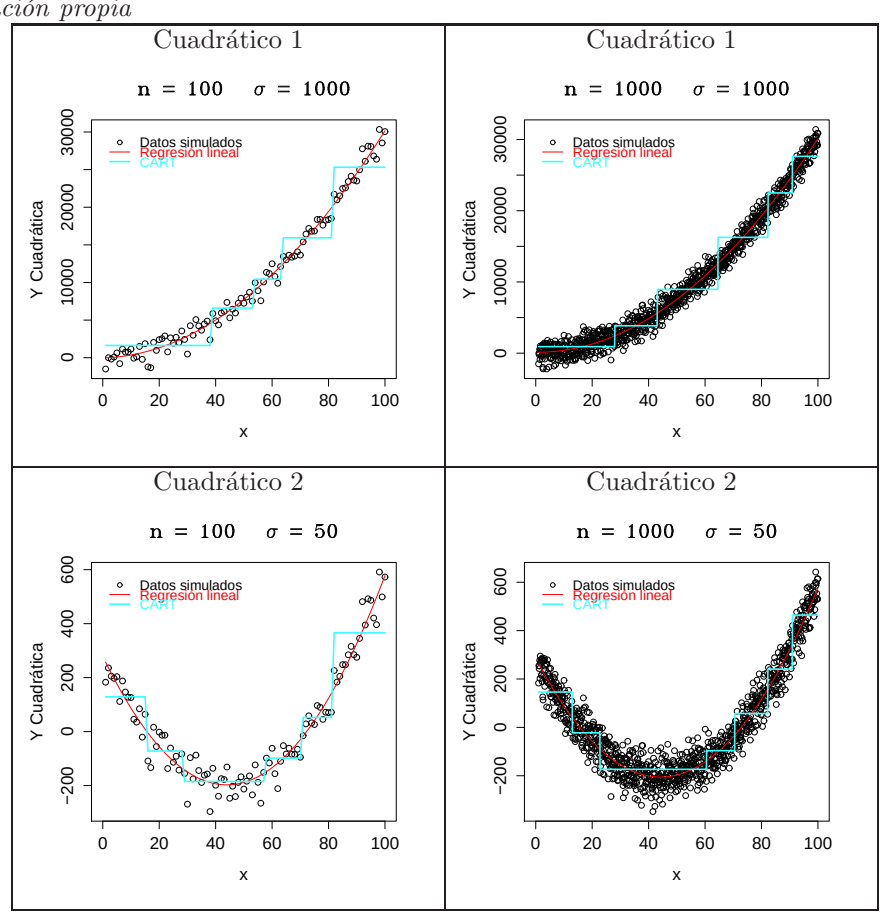
Para simular los conjuntos de datos se siguen los pasos descritos en la sección 1.5. En el paso 1, se toma $p = 2$, se especifican las funciones

$$g_1(x) = cos(bx), \;\; g_2(x) = sin(bx), \tag{22}$$

y se especifican los valores de los parámetros $\beta_0 = d$, $\beta_1 = a\sin(c)$ y $\beta_2 = a\cos(c)$. Para encontrar a, c y d en términos de β_0, β_1 y β_2, se resuelven las ecuaciones

$$a = \pm\sqrt{(}\beta_1^2 + \beta_2^2), \;\; c = \arctan(\beta_1/\beta_2), \;\; d = \beta_0. \tag{23}$$

Tabla 2: *Gráficos de las predicciones para los modelos cuadráticos. Fuente: elaboración propia*



El tercer modelo por analizar se obtiene al sustituir $a = 10,\ b = 0.1,\ c = 1,\ d = 12$ en la ecuación (19) y se llamará modelo trigonométrico 1. El cuarto modelo por analizar se obtiene al sustituir $a = 10,\ b = 0.5,\ c = 1,\ d = 12$ en la ecuación (19) y se llamará modelo trigonométrico 2. El quinto y último modelo por analizar se obtiene de sustituir $a = 10,\ b = 1,\ c = 1,\ d = 12$ en la ecuación (19) y se llamará modelo trigonométrico 3.

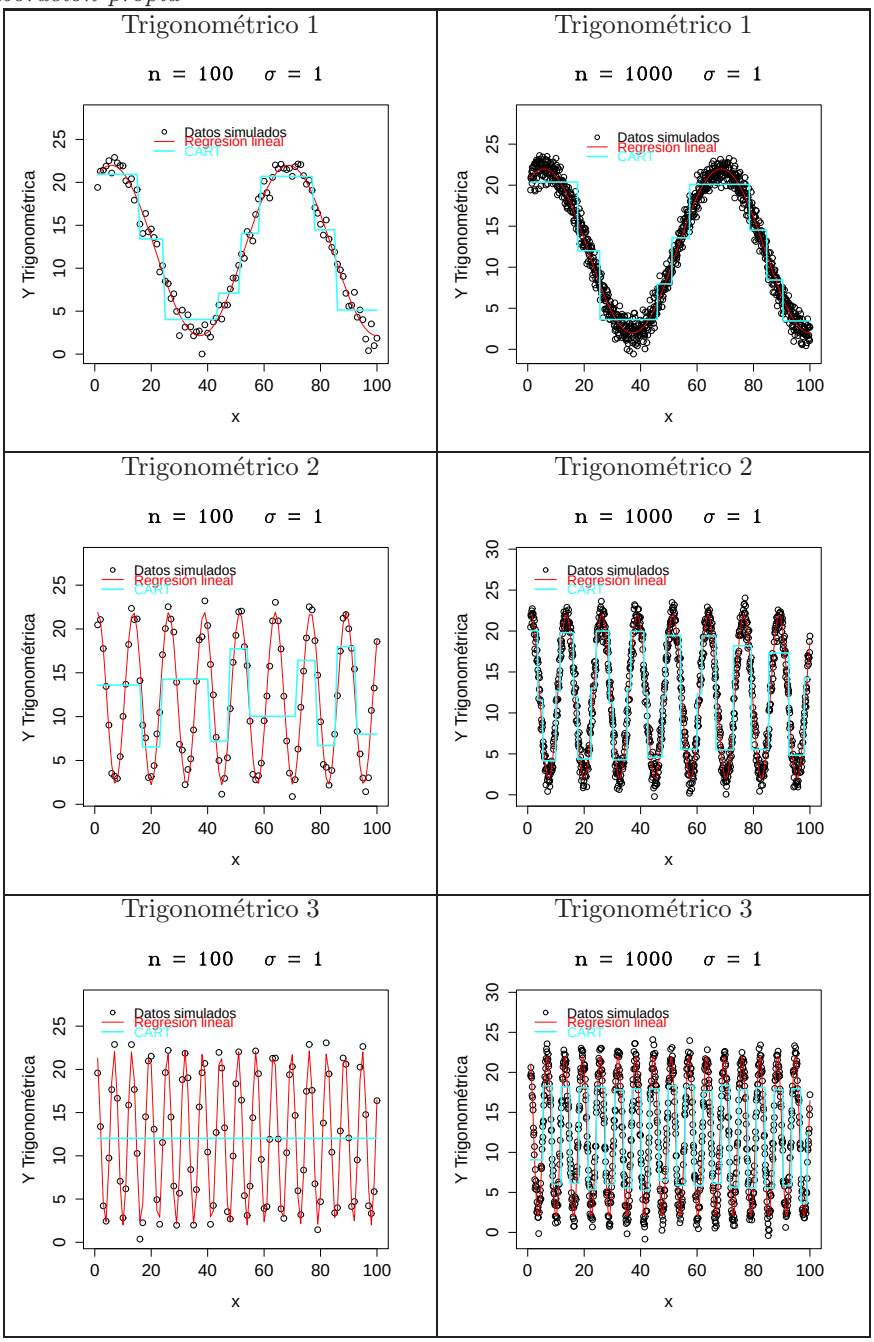
De igual manera que para los modelos cuadráticos, en la Tabla 3 se puede observar para los modelos trigonométricos 1, 2 y 3, que para cualquier valor de n fijo, al aumentar la desviación estándar σ, los errores de predicción de la regresión lineal y de CART se aproximan entre sí, siendo en todos los casos menor el error de predicción de la regresión lineal.

Tabla 3: *Comparación de los errores de predicción para los modelos trigonométricos. Fuente: elaboración propia*

Trigonométrico 1			Trigonométrico 2			Trigonométrico 3		
n	σ	$DIFLOG$	n	σ	$DIFLOG$	n	σ	$DIFLOG$
	0.1	4.6871		0.1	5.3747		0.1	5.4009
	0.3	3.7664		0.3	4.3779		0.3	4.4733
50	0.5	3.3019	50	0.5	3.9477	50	0.5	3.9775
	0.8	2.9084		0.8	3.5096		0.8	3.5627
	1	2.6753		1	3.3064		1	3.3305
	2	1.9181		2	2.5195		2	2.5511
	0.1	4.6585		0.1	5.5911		0.1	5.6893
	0.3	3.7042		0.3	4.5872		0.3	4.7109
100	0.5	3.1851	100	0.5	4.1200	100	0.5	4.2248
	0.8	2.7160		0.8	3.6437		0.8	3.8107
	1	2.4882		1	3.4060		1	3.5492
	2	1.6586		2	2.5312		2	2.6679
	0.1	5.1325		0.1	5.4799		0.1	5.8220
	0.3	4.1113		0.3	4.4435		0.3	4.8207
500	0.5	3.5220	500	0.5	3.9073	500	0.5	4.2554
	0.8	2.9345		0.8	3.3196		0.8	3.6504
	1	2.6243		1	3.0206		1	3.3276
	2	1.5876		2	2.0048		2	2.2863
	0.1	5.4128		0.1	5.7053		0.1	6.1348
	0.3	4.3312		0.3	4.6271		0.3	5.0262
1000	0.5	3.7003	1000	0.5	4.0429	1000	0.5	4.3948
	0.8	3.0333		0.8	3.3894		0.8	3.7299
	1	2.6955		1	3.0631		1	3.3895
	2	1.6081		2	2.0030		2	2.2809
	0.1	6.0474		0.1	6.3035		0.1	6.7301
	0.3	4.7021		0.3	4.9564		0.3	5.3738
5000	0.5	3.9106	5000	0.5	4.1899	5000	0.5	4.5899
	0.8	3.1304		0.8	3.4461		0.8	3.8144
	1	2.7607		1	3.0925		1	3.4445
	2	1.6242		2	2.0165		2	2.2933

En los gráficos de los modelos trigonométricos 1 y 2 de la Tabla 4 se puede ver cómo las predicciones de CART describen la forma del verdadero modelo de los datos simulados para $n = 100$ y $n = 1000$. En los gráficos del modelo trigonométrico 3 de la Tabla 4 se ve que las predicciones de CART no describen la forma verdadera de los datos con $n = 100$, pero, si la describen con $n = 1000$. Nótese que este modelo de regresión tiene una forma más compleja que los modelos anteriores en cuanto al número de máximos y mínimos locales que tiene su gráfica.

Tabla 4: *Gráficos de las predicciones para los modelos trigonométricos. Fuente: elaboración propia*



5. Comparación de las predicciones cuando el modelo lineal ajustado es incorrecto

A continuación se tomarán tres modelos de regresión lineal de los descritos en la sección 3 para generar conjuntos de datos, a los cuales se ajustan rectas de regresión lineal como modelo equivocado para comparar estas predicciones con las de CART. Se escogieron estos modelos porque hay casos en el estudio de simulación en que la recta de regresión predice mejor los datos que los árboles de regresión cuando el tamaño muestral es pequeño. El objetivo es ver como CART toma ventaja del aumento del tamaño muestral para predecir mejor los datos que la recta de regresión en estos modelos.

En la Tabla 5 se puede observar para el modelo cuadrático 1, que en general CART predice mejor la respuesta que la recta de regresión, exceptuando para $n = 50$, donde los errores de predicción de la recta de regresión son más pequeños que los de CART. En los gráficos del modelo cuadrático 1 de la Tabla 6, se puede ver cómo las predicciones de CART se adaptan a la forma del verdadero modelo de los datos simulados.

En la Tabla 5 se observa para el modelo trigonométrico 2, que CART es más preciso que la recta de regresión, es decir, el error de predicción de CART es menor que el error de la recta de regresión para cualquier valor de n y cualquier valor de σ. En los gráficos de la Tabla 6 para el modelo trigonométrico 2, se puede observar cómo las predicciones de CART con $n = 50$ descubren patrones en los datos que pueden no notarse a simple vista. Aunque se puede decir para $n = 50$ y $n = 100$ que las predicciones de CART se adaptan a la forma del verdadero modelo de los datos simulados, es claro que con $n = 50$ es más difícil describir la verdadera forma del modelo por su cantidad de máximos y mínimos relativos. Para $n = 100$ es más clara la verdadera forma del modelo debido a que se tienen más cantidad de datos para describirlo.

En la Tabla 5 se observa para el modelo trigonométrico 3, que el error de predicción de CART es mayor que el de la recta de regresión para $n = 50$ cuando $\sigma = 0.1, 0.3, 0.5, 0.8$, y para $n = 100$ cuando $\sigma = 0.1, 0.3, 0.5$, pero, en los otros casos, el error de predicción de CART es menor. En los gráficos de la Tabla 6 para el modelo trigonométrico 3, se observa que las predicciones de CART aparentemente forman una recta, es decir, CART carece de capacidad de captar la verdadera forma del modelo con $n = 100$ datos, al igual que con $n = 50$. Se puede decir, para este modelo, que con $n = 50$ y $n = 100$ es más difícil describir la verdadera forma del modelo por su cantidad de máximos y mínimos relativos. Se observa que con $n = 500$ las predicciones de CART se adaptan a la verdadera forma del modelo debido a que se tiene más cantidad de datos para describirlo. Si bien no existe evidencia para todos los modelos que el aumento de n implica un aumento en la precisión de las predicciones de CART con respecto a la recta de regresión (disminución de la diferencia de logaritmos de los errores en la tabla), se puede observar globalmente que esta precisión para $n = 50$ y $n = 100$ es notablemente menor que para $n = 500$, $n = 1000$ y $n = 5000$.

En general se puede concluir que a medida que aumenta el número de máximos y mínimos relativos en el modelo trigonométrico los árboles de regresión tienen más problemas en describir la forma del verdadero modelo de los datos cuando el número de datos no es suficiente.

Tabla 5: *Comparación de los errores de predicción con modelos lineales incorrectos. Fuente: elaboración propia*

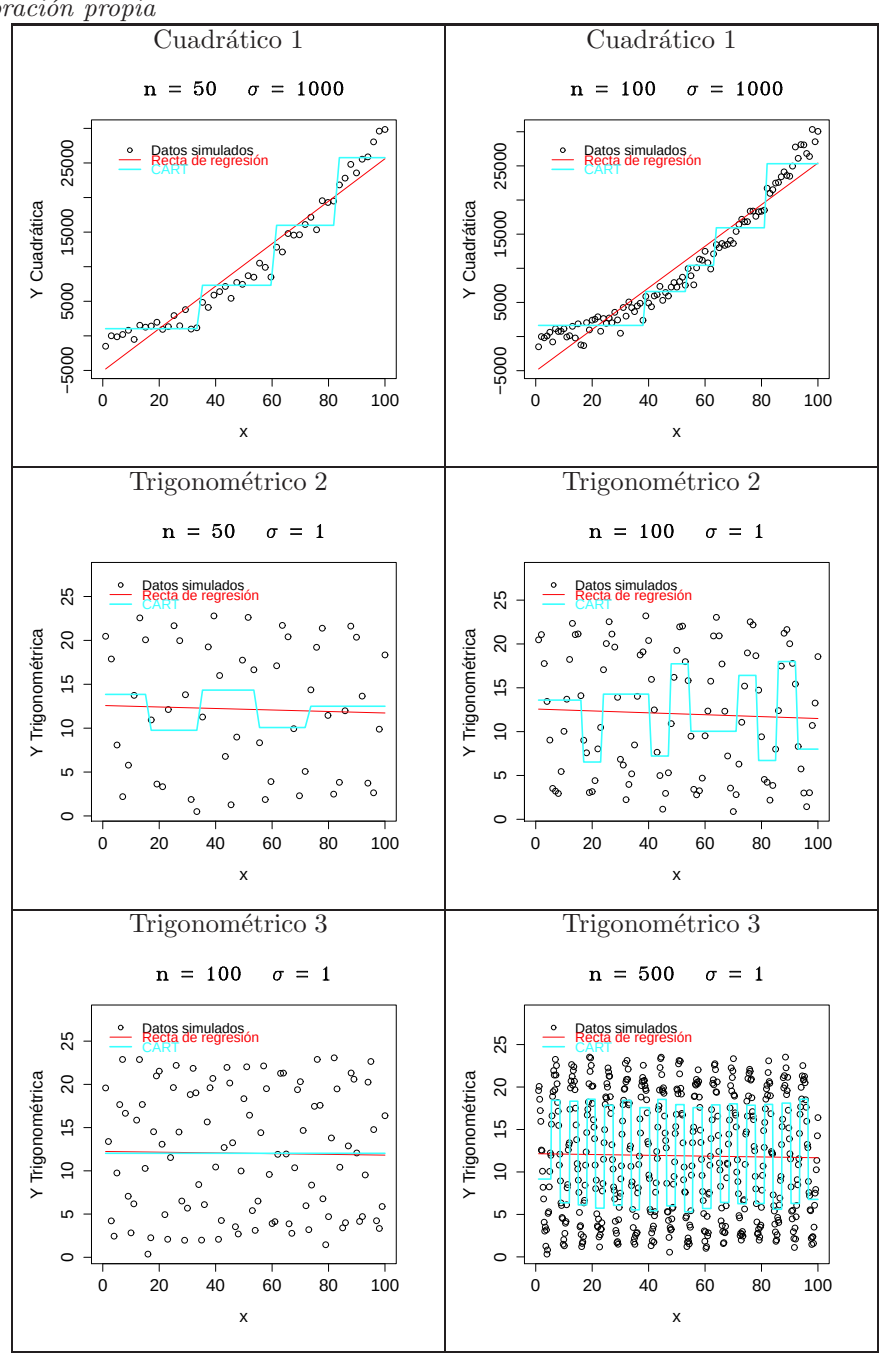
Cuadrático 1			Trigonométrico 2			Trigonométrico 3		
n	σ	$DIFLOG$	n	σ	$DIFLOG$	n	σ	$DIFLOG$
	1	0.3610		0.1	-0.0278		0.1	0.0001
	10	0.3574		0.3	-0.0277		0.3	0.0001
50	100	0.3499	50	0.5	-0.0273	50	0.5	0.0001
	500	0.3344		0.8	-0.0256		0.8	0.0000
	1000	0.2969		1	-0.0239		1	-0.0001
	2000	0.2600		2	-0.0175		2	0.0000
	1	-0.0528		0.1	-0.1268		0.1	0.0001
	10	-0.0528		0.3	-0.1305		0.3	0.0001
100	100	-0.1328	100	0.5	-0.1338	100	0.5	0.0001
	500	-0.1891		0.8	-0.1363		0.8	-0.0003
	1000	-0.2038		1	-0.1379		1	-0.0009
	2000	-0.1915		2	-0.1397		2	-0.0086
	1	-0.2721		0.1	-0.9144		0.1	-0.5408
	10	-0.2722		0.3	-0.9007		0.3	-0.5430
500	100	-0.2993	500	0.5	-0.8847	500	0.5	-0.5414
	500	-0.2933		0.8	-0.8636		0.8	-0.5369
	1000	-0.2883		1	-0.8492		1	-0.5336
	2000	-0.2611		2	-0.7981		2	-0.5127
	1	-0.2709		0.1	-0.9292		0.1	-0.5459
	10	-0.2713		0.3	-0.9292		0.3	-0.5459
1000	100	-0.2929	1000	0.5	-0.9090	1000	0.5	-0.5455
	500	-0.2972		0.8	-0.8765		0.8	-0.5436
	1000	-0.2830		1	-0.8642		1	-0.5426
	2000	-0.2588		2	-0.8119		2	-0.5364
	1	-0.2705		0.1	-0.9817		0.1	-0.5494
	10	-0.2705		0.3	-0.9646		0.3	-0.5490
5000	100	-0.2761	5000	0.5	-0.9465	5000	0.5	-0.5485
	500	-0.2927		0.8	-0.9136		0.8	-0.5478
	1000	-0.2804		1	-0.8941		1	-0.5470
	2000	-0.2414		2	-0.8179		2	-0.5415

6. Conclusiones

Del estudio de simulación se concluye que, cuando se comparan las predicciones de los árboles de regresión y las de regresión lineal al predecir la respuesta de cualquier modelo de regresión analizado, sea cuadrático o trigonométrico, el error de predicción de la regresión lineal siempre es menor que el de CART. Aunque el aumento de la varianza de los errores de los datos hace que el error de predicción de la regresión lineal se aproxime al de CART, el estudio de simulación no muestra ningún caso en que este error supere al de CART.

Al comparar las predicciones de los árboles de regresión y las de la recta de regresión al predecir la respuesta del modelo cuadrático 1 y de los modelos trigonométricos 2 y 3, se observa que siempre que se tenga la cantidad de datos suficiente para

Tabla 6: *Gráficos de los modelos lineales ajustados incorrectamente. Fuente: elaboración propia*



describir la forma funcional de la media de los datos, el error de predicción de CART es menor que el de la recta de regresión.

De lo anterior se puede concluir que, el modelo CART es una alternativa que prueba ser una buena opción cuando el usuario desconoce la forma funcional verdadera del modelo, lo cual es común en investigaciones reales, y puede utilizarse como una primera etapa en la parte exploratoria en modelación. Si el usuario está seguro de cuál es la forma funcional de su modelo, entonces CART no es una opción viable.

7. Agradecimientos

A los profesores Víctor Ignacio López Ríos y René Iral Palomino por sus invaluables comentarios y sugerencias. A Diana Guzmán Aguilar, Jorge Iván Vélez y en general a la Escuela de Estadística y la Facultad de Ciencias de la Universidad Nacional de Colombia, sede Medellín por haber propiciado un ambiente idóneo para la realización de este trabajo.

Recibido: 27 de mayo de 2013
Aceptado: 20 de septiembre de 2013

Referencias

Ankarali, H., Canan, A., Akkus, Z., Bugdayci, R. & Ali Sungur, M. (2007), 'Comparison of logistic regression model and classification tree: An application to postpartum depression data', *Expert Systems with Applications* **32**, 987–994.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification And Regression Trees*, CHAPMAN & HALL/CRC, Boca Raton.

Izenman, A. (2008), *Modern Multivariate Statistical Techniques*, Springer, New York.

Tamminen, S., Laurinen, P. & Roning, J. (1999), 'Comparing regression trees with neural networks in aerobic fitness approximation'.

Zhang, H. & Singer, B. (2010), *Recursive Partitioning and Applications*, Springer, New York.