Comparación entre Árboles de Regresión CART y Regresión Lineal

Juan Felipe Díaz Sepúlveda

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2012

Comparación entre Árboles de Regresión CART y Regresión Lineal

Juan Felipe Díaz Sepúlveda

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Ciencias - Estadística

Director:
Ph.D. Juan Carlos Correa Morales

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2012

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 a 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 sólo 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 statistical method most used to predict values of continuous variables because of its easy interpretation, but in many situations to apply the model assumptions are not met and some users tend to force leading to erroneous conclusions. CART regression trees are an alternative regression requires no assumptions about the data to be analyzed and a method of easy interpretation of the results. In this paper we compare the predictive level from both CART and linear regression through simulation. In general, it was found that when adjusting the correct linear regression model to the data, the linear regression prediction error is always less than the CART prediction error. We also found that when adjusted erroneously linear regression model to the data, CART prediction error is smaller than the linear regression prediction error only when it has a sufficiently large amount of data.

Keywords: Simulation, Prediction error, Linear Regression, CART: Classification and Regression Trees.

Contenido

1. Introducción

1.1. Planteamiento del problema

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, entre ellas se tienen:

- Interpretabilidad
- Teóricamente atractivo
- Fácil de estimar
- Poco costoso

Tal vez la interpretabilidad del modelo lineal clásico ha popularizado tanto este modelo, que no es raro ver su ajuste en situaciones inapropiadas, por ejemplo, respuestas que son discretas o sesgadas; y el desespero por parte de los usuarios por aproximarse a él, por ejemplo mediante transformaciones, sin considerar los cambios en la estructura del error. De aquí la necesidad de tener un modelo que tenga similares ventajas, pero 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/o 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.

El problema central es comparar, a nivel predictivo, los árboles de regresión CART con el método de regresión lineal por medio de un estudio de simulación, simulando conjuntos de datos cuyo verdadero modelo es un modelo de regresión lineal y ajustando 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 CART ajustados a los mismos datos.

1.2. Antecedentes

Desde el planteamiento de los árboles de clasificación y regresión CART por Leo Breiman y otros en 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 [4] 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 de estos autores son:

En 1995 Chaudhuri, Lo, Loh y Yang [9] estudiaron un método de regresión generalizado que mezcla regresión no paramétrica de árboles estructurados y particionamiento recursivo adaptativo con estimación de máxima verosimilitud. La función estimada es polinómica por tramos determinados por los nodos terminales de un árbol de decisión binario. El árbol de decisión es construido particionando recursivamente los datos de acuerdo a los signos de los residuales de un modelo ajustado por máxima verosimilitud en cada nodo.

En 1999 Tamminen, Laurinen y Roning [28] compararon los árboles de regresión con las redes neuronales en un conjunto de datos obtenidos por un método de medición de aptitud aeróbica, los cuales consisten de mediciones del consumo máximo de oxígeno como valores de referencia y características físicas, incluyendo intervalos R-R de latidos medios del corazón en reposo. Debido a que el sistema físico de los humanos es altamente no lineal la regresión lineal tradicional no puede ser usada como modelo de aproximación de los datos, por tanto, los árboles de regresión y las redes neuronales son considerados como candidatos en este estudio para modelar los datos.

En 2000 Li, Lue y Chen [22] introducen una aproximación iterativa a la regresión con estructura de árbol, centrándose en la exploración de la información geométrica en los datos. El procedimiento comienza con la búsqueda de una dirección a lo largo de la cual la superficie de regresión es más curva. Esta dirección es usada para dividir los datos en dos regiones. En cada región se encuentra una dirección y luego se divide de la misma manera. El proceso continua hasta que la variable regresora es descompuesta en regiones donde se pueda aproximar una regresión lineal. Para implementar la búsqueda de la dirección se aplica el método "Principal Hessian Directions" (PHD) (Li, 1992). Por último hacen una comparación con los métodos CART, SUPPORT y MARS.
Balac, Gaines y Fisher [2] en 2000 presentan una aplicación de los árboles de regresión que permiten a un robot aprender modelos de acción a través de experiencias de modo que puedan hacer predicciones similares.
Lewis [21] en 2000 da una visión general de la metodología CART, enfatizando más en su uso práctico que en la teoría estadística subyacente.

En 2001 Izrailev y Agrafiotis [18] introducen un método novedoso de particionamiento basado en hormigas artificiales. Este método muestra un mejor desempeño que el particionamiento recursivo sobre tres conjuntos de datos bien estudiados.
Kramer, Widmer, Pfahringer y DeGroeve [19] en 2001 se dedican al problema de aprender a predecir clases ordinales usando árboles de clasificación y regresión. Los autores utilizan un algoritmo de

árbol inductivo llamado S-CART y estudian varios caminos de transformación dentro de un aprendiz de tareas de clasificación ordinal. Estas variantes de algoritmos son comparados en conjuntos de datos que son referencia para verificar las fortalezas y debilidades de las estrategias y estudiar el intercambio entre la precisión de la clasificación categórica óptima y el mínimo error basado en la distancia.

En 2002 Loh [23] propone un algoritmo para la construcción de árboles de regresión llamado GUIDE. Es diseñado específicamente para eliminar el sesgo de selección de variables. GUIDE controla el sesgo empleando análisis chi-cuadrado de residuales y calibración bootstrap de probabilidades de significancia. En un experimento con datos reales compara las predicciones por medio del error cuadrático medio con CART.

Chaudhuri y Loh [10] en 2002 estudian un método de regresión no paramétrica que mezcla características claves de la regresión cuantil polinomial por tramos y la regresión estructural de árbol basada en particionamiento recursivo adaptativo del espacio de covariables. A diferencia de la regresión por mínimos cuadrados, la cual se concentra en modelar la relación entre la respuesta y las covariables en el centro de los datos, estos árboles de regresión cuantil proporcionan una visión de la naturaleza de esa relación en el centro tan bien como en las colas de la distribución de la respuesta.

Carmack, Sain y Schucany [8] en 2002 presentan un procedimiento utilizando pruebas de permutación aplicadas a estadísticos de orden para determinar cuales divisiones en un árbol de regresión son significativas. Generalmente no se disponen de procedimientos formales para este tipo de prueba. La tradicional validación cruzada y el procedimiento de pruebas de permuatación son comparados en un ejemplo específico.

Torgo [30] en 2002 describe un método para obtener árboles de regresión usando modelos de regresión lineal en los nodos terminales en una forma computacionalmente eficiente que permite el uso de este método en grandes conjuntos de datos.

Cappelli, Mola y Siciliano [6] en 2002 sugieren la introducción de una tercera etapa en la construcción del árbol saturado. El objetivo es encontrar un árbol *honesto*, es decir, un árbol que no sólo sea comprensible y preciso, sino también estadísticamente confiable. Los autores introducen procedimientos de prueba tanto para árboles clasificación como de regresión los cuales orientan la búsqueda hacia aquellas partes en la estructura del árbol que son estadísticamente significativas.

En 2003 Scott, Willett y Nowak [26] plantean un procedimiento para podar inicialmente el árbol máximo en la construcción de árboles de clasificación y regresión. Proponen un enfoque al modelamiento del árbol iniciando con una estructura de árbol diádico y una partición fija. Ellos muestran que los árboles diádicos son flexibles, fáciles de construir y producen resultados óptimos cuando están debidamente podados. También defienden el uso de la log-verosimilitud negativa como medida del riesgo empírico en problemas de regresión no gaussianos, en contraste al criterio de sumas de cuadrados del error usados en CART.

Engle-Warnick [13] en 2003 introduce un enfoque a un árbol de clasificación binario no paramétrico para inferir estrategias no observadas desde acciones observadas, y son interpretables con afirmaciones de la forma if-then. Define los árboles de clasificación binaria y sus medidas de desempeño,

y un resumen del algoritmo de regresión.

Dudoit, Gentleman y Van der Laan [12] en 2003 tienen como propósito una estrategia unificada para la construcción, selección y evaluación del desempeño de estimadores en presencia de censura y proponen una metodología para estimación basada en árboles con datos censurados. El enfoque abarca predicción univariada, predicción multivariada y estimación de densidad, definiendo una función de pérdida adecuada para cada uno de estos problemas. El método propuesto es evaluado usando estudios de simulación y datos de supervivencia de pacientes con cáncer de seno.

En 2004 Larsen y Speckman [20] desarrollan una metodología de árboles de regresión multivariada la cual es ilustrada en un estudio de predicción de la abundancia de varias especies de plantas que se producen en los bosques de Missouri Ozark. La técnica es una variación de la aproximación de Segal (1992) para datos longitudinales. Tiene el potencial de ser aplicada en gran variedad de problemas en los cuales el analista busca predecir la ocurrencia simultánea de muchas variables dependientes.

Cappelli y Reale [7] en 2004 proponen un enfoque no paramétrico que explota en la estructura de árboles de regresión por mínimos cuadrados la propiedad de contiguidad del método de agrupamiento de Fisher (1958) propuesto para agrupar una sola variable real. Este enfoque es aplicado en el estudio de los cambios en los niveles medios de agua del lago Michigan-Huron.

Cappelli y Mola [5] en 2004 muestran como el algoritmo STP planteado por Capelli y otros en 2002 [6] es una herramienta útil entre los métodos de árboles de clasificación para evitar sobreajuste. El problema del sobreajuste es la presencia de subdivisión falsa, la cual, si bien reduce el error total no corresponde a la verdadera relación entre predictores y variable respuesta. Los autores muestran como el proceso STP estudia la dependencia entre la variable respuesta y las variables a dividir, y aplicado a simulaciones y ejemplos reales puede evaluar la presencia de sobreajuste preservando solo subdivisiones significantes.

De Carvalho, De Souza y Verde [11] en 2004 presentan un algoritmo para clasificación simbólica de datos. Los datos de entrada para la etapa de aprendizaje son conjuntos de objetos simbólicos, descritos por variables en intervalos simbólicos (o conjuntos de valores). Al final de la etapa de aprendizaje cada grupo es representado por un objeto simbólico (modal) el cual es descrito por variables de un histograma simbólico (o diagrama de barras). La asignación de nuevas observaciones a un grupo es basada en una función de disimilaridad la cual mide la diferencia en contenido y posición entre ellos. Los autores muestran la utilidad de este clasificador de patrón simbólico modal en un conjunto de imágenes simuladas.

Miglio y Soffritti [24] en 2004 comparan dos metodologías para la comparación de dos árboles de clasificación. La primera es una distancia que mide la cantidad de reasignaciones necesitada para cambiar uno de los árboles de tal manera que resulte en una estructura idéntica a la del otro, y la segunda es una medida de similaridad que compara las particiones asociadas a los árboles tomando en cuenta su poder predictivo. Los autores analizan características y limitaciones de estas medidas de proximidad y proponen una nueva medida de disimilaridad que tiene en cuenta aspectos explorados separadamente por las dos medidas analizadas.

Piccarreta [25] en 2004 proponen un nuevo criterio para generar árboles de clasificación en el caso de que la variable respuesta sea categórica ordenada. Este criterio es obtenido midiendo la impureza

dentro de un nodo haciendo referencia a una medida general de dispersión mutua (el índice Gini), el cual puede ser aplicado a cualquier tipo de variable.

En 2005 Struyf y Dzeroski [27] proponen un sistema basado en restricciones para construir árboles de regresión multiobjetivo. Un árbol de regresión multiobjetivo es un árbol de decisión capaz de predecir muchas variables numéricas de una vez. Su enfoque es primero construir un gran árbol basado en los datos de entrenamiento y luego podarlo para satisfacer las restricciones de usuario. Esto tiene la ventaja que el árbol puede ser almacenado en la base de datos inductiva y usado para responder consultas inductivas con diferentes restricciones. Evalúan su sistema en varios conjuntos de datos de palabras reales y miden el equilibrio entre tamaño y precisión.

Huang [16] en 2005 propone un método (REH)y una variación de éste para resolver el problema de encontrar pocos eventos raros (una proporción de 0.05 o menos de la muestra de estudio) de un conjunto de observaciones. Estos son aplicados a tres conjuntos de datos reales los cuales son caracterizados por una larga cola derecha en la variable de respuesta. se compara el desempeño para encontrar eventos raros de la variación REH con la metodología *Random Forest*.

En 2006 Vens y Blockeel [31] proponen una heurística alternativa que da igual precisión que los modelos de árboles pero que arroja árboles simples con mejor poder explicativo. Los modelos de árboles, generalmente, son árboles de regresión que contienen algun modelo no trivial en sus nodos terminales. Las implementaciones más populares de los modelos de árboles construyen árboles con modelos de regresión lineal en sus nodos terminales. Estos usan la reducción de la varianza como heurística para seleccionar las pruebas durante el proceso de construcción del árbol. Los autores muestran que sistemas que emplean esta heurística pueden exhibir un comportamiento débil en algunos casos bastante simples, ya que no es visible en la precisión predictiva del árbol, pero reduce su poder interpretativo.

Hothorn, Hornik y Zeileis [15] en 2006 proponen un marco unificado para particionamiento recursivo el cual incorpora modelos de regresión de estructura de árbol dentro de una teoría bien definida de procedimientos de inferencia condicional. El criterio de parada basado en procedimientos de prueba múltiple son implementados y muestran que el desempeño predictivo de los árboles resultantes es tan bueno como el desempeño del procedimiento de búsqueda exhaustiva establecido. También muestran que la precisión de la predicción de árboles con parada anticipada es equivalente a la precisión de la predicción de árboles podados con selección de variables insesgadas. Se analizan datos de estudios sobre clasificación de glaucoma, supervivencia de cáncer de seno y experiencias de mamografía.

He [14] en 2006 implementó el método boostrap no paramétrico para imputar valores faltantes retirando datos en el árbol construido (CART o *Random Forest*), y la clasificación resultante fue comparada entre los datos completos y la clasificación resultante utilizando variables sustitutas. Los autores encontraron significativas mejoras en la capacidad de predecir para los modelos CART y *Random Forest*.

En 2007 Ankarali, Canan, Akkus, Bugdayci y Ali Sungur [1] comparan 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 postparto. De acuerdo al árbol de clasificación óptimo, se determinaron un total de seis factores de riesgo, pero, en el modelo de regresión logística tres de estos efectos fueron significativos. Los autores concluyen que los árboles de clasificación frente al modelo de regresión logística proporcionan información más detallada en el diagnóstico mediante la evaluación de una gran cantidad de factores de riesgo.

1.3. 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. (Izeman, [17])

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,

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

o un funcional de su probabilidad tal como la esperanza condicional

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

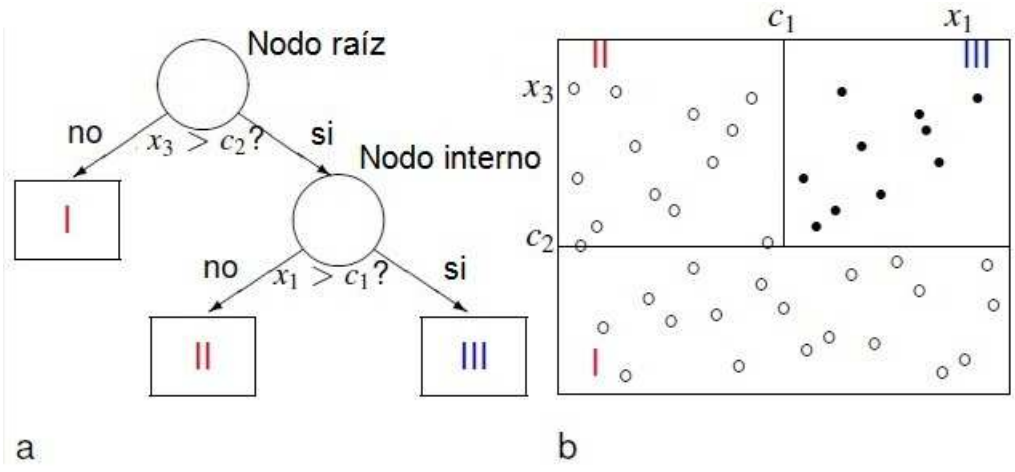

Figura 1-1.: Ejemplo árbol. Fuente (Zhang [32]).

1.3.1. Elementos de la construcción del árbol

Según Zhang [32] para ilustrar las ideas básicas, considere el diagrama de la figura **1-1**. 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.

Para entender la construcción de la figura **1-1**, se necesita responder tres preguntas básicas:

- ¿Qué contienen los nodos?
- ¿Por qué y cómo se divide un nodo padre en dos nodos hijos?
- ¿Cuándo se declara un nodo terminal?

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 ésta.

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 **1-1** b).

La completa homogeneidad de los nodos terminales es un ideal raramente alcanzado en el análisis de datos real. De esta manera, el objetivo del particionamiento recursivo es hacer las variables resultantes en los nodos terminales tan homegéneas como sea posible.

Una medida cuantitativa de la homogeneidad es la noción de impureza. La idea es la siguiente:

$$\text{Impureza de un nodo} = \frac{\text{Número de sujetos que cumplen la característica en el nodo}}{\text{Número total de sujetos en el nodo}}. \qquad (1\text{-}1)$$

En la figura **1-1**, si la característica es ser círculo, el nodo hijo terminal (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, el nodo hijo terminal del nodo raíz tiene impureza igual a 0 debido a que no hay ningún punto en este nodo. Nótese que para el nodo hijo interno (nodo hijo derecho) del nodo raíz hay aproximadamente igual número de círculos y número de puntos teniendo este nodo una medida de la impureza de aproximadamente 0,5 independientemente de si la característica es ser círculo o punto. Mientras más homogéneo sea el nodo el límite del cociente en la ecuación 1-1 es 0 o 1.

1.3.2. 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 1-1 para cada nodo hijo. En la figura **1-1** 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, tomando 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 (o homogéneos). Sin embargo, en la realidad los nodos hijos son usualmente parcialmente puros. Además, la bondad de una división debe poner en una balanza la homogeneidad (o la impureza) de los dos nodos hijos simultáneamente.

Si se toma la covariable x_1 con el valor de corte c como alternativa para dividir un nodo, como resultado de la pregunta "¿es $x_1 > c$?" se tiene la siguiente tabla:

	$Y=0$	$Y=1$	
Nodo Izquierdo (τ_L) $x_1 \leq c$	n_{11}	n_{12}	$n_{1\cdot}$
Nodo Derecho (τ_R) $x_1 > c$	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	

Sea Y una variable dicotómica con valores 0 y 1. Se estima $P[Y=1|\tau_L]$ y $P[Y=1|\tau_R]$ por $\frac{n_{12}}{n_{1\cdot}}$ y $\frac{n_{22}}{n_{2\cdot}}$, respectivamente. Se introduce la noción de impureza "entropía" en el nodo hijo izquierdo definida como

$$i(\tau_L) = -\frac{n_{11}}{n_{1\cdot}} \log\left(\frac{n_{11}}{n_{1\cdot}}\right) - \frac{n_{12}}{n_{1\cdot}} \log\left(\frac{n_{12}}{n_{1\cdot}}\right) \tag{1-2}$$

De la misma manera, se define la impureza en el nodo hijo derecho como

$$i(\tau_R) = -\frac{n_{21}}{n_{2\cdot}} \log\left(\frac{n_{21}}{n_{2\cdot}}\right) - \frac{n_{22}}{n_{2\cdot}} \log\left(\frac{n_{22}}{n_{2\cdot}}\right). \tag{1-3}$$

Entonces, la bondad de una división, s, es medida por

$$\Delta I(s,\tau) = i(\tau) - P[\tau_L] i(\tau_L) - P[\tau_R] i(\tau_R), \tag{1-4}$$

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

Aquí, $P[\tau_L]$ se puede tomar como $\frac{n_{1\cdot}}{n_{1\cdot}+n_{2\cdot}}$ y $P[\tau_R]$ como $\frac{n_{2\cdot}}{n_{1\cdot}+n_{2\cdot}}$.

La ecuación 1-4 mide el grado de reducción de la impureza cuando se pasa del nodo padre a los nodos hijos.

1.3.3. Nodos terminales

El proceso de particionamiento recursivo continua 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 queda solo un sujeto en un nodo. 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 [4] argumenta 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.

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

1.4. Árboles de clasificación

Los árboles de clasificación y regresión (CART) fueron desarrollados en los años 80 por Breiman, Freidman, Olshen y Stone en el libro *Classification and Regression Trees* publicado en 1980 [4].

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 variables numéricas y/o 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 es diferente para los árboles de clasificación y de regresión. En esta sección se considera primero la construcción de árboles de clasificación.

1.4.1. Impureza del nodo

Sea Y una variable dicotómica con valores 0 y 1. Para construir el árbol saturado, en el proceso de particionamiento recursivo se tiene que para 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

$$i(\tau) = \phi(\{Y = 1|\tau\}), \tag{1-5}$$

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 $0\log(0) := 0$.

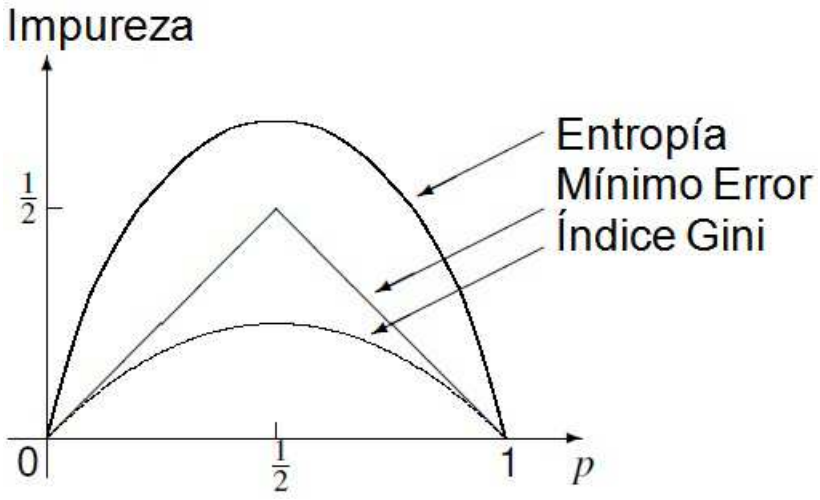

Figura 1-2.: funciones de impureza. Fuente (Zhang [32]).

1.4.2. 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 el más predictivo de los resultados y menos vulnerable al ruido en 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

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} \mathrm{P}[\tau]\mathrm{r}(\tau), \tag{1-6}$$

donde $\tilde{\mathcal{T}}$ es el conjunto de nodos terminales de $\mathcal{T}$ 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 debido a que los árboles de clasificación trabajan sobre respuestas binarias.

Costo de mala clasificación

Sea Y una variable dicotómica con valores 0 y 1 y sea $c(i|j)$ el costo de mala clasificación de que un sujeto de la clase j sea clasificado en la clase i. Cuando $i = j$, se tiene la clasificación correcta y el costo debería ser cero, es decir, $c(i|i) = 0$. Sin pérdida de generalidad se puede tomar $c(1|0) = 1$ y suponer que $c(0|1) \geq c(1|0)$, pero, medir el costo relativo $c(0|1)$ es difícil debido a que es una decisión subjetiva que requiere un amplio conocimiento del problema aplicado.
El nodo τ es asignado a la clase j si

$$\sum_i \{c(j|i)\mathrm{P}[Y = i|\tau]\} \leq \sum_i \{c(1 - j \mid i)\mathrm{P}[Y = i|\tau]\}. \tag{1-7}$$

Sea $\mathrm{r}(\tau)$ el lado izquierdo de 1-7, es decir,

$$\mathrm{r}(\tau) = \sum_i \{c(j|i)\mathrm{P}[Y = i|\tau]\} \tag{1-8}$$

el cual es el costo esperado de cualquier sujeto dentro del nodo, y usualmente se conoce como el costo de mala clasificación dentro del nodo τ, o también como el costo de mala clasificación condicional del nodo τ. Para encontrar el costo de mala clasificación incondicional del nodo τ se multiplica $\mathrm{r}(\tau)$ por $\mathrm{P}[\tau]$ obteniendo,

$$R(\tau) = \mathrm{P}[\tau]\mathrm{r}(\tau), \tag{1-9}$$

el cual se conoce simplemente como el *costo de mala clasificación del nodo* τ. Si se reemplaza la ecuación 1-9 en la ecuación 1-6 se obtiene,

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} R(\tau), \tag{1-10}$$

el cual se conoce como el *costo de mala clasificación del árbol* $\mathcal{T}$.

Generalmente es difícil en la práctica asignar la función de costo 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,

$$R^s(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} R^s(\tau). \tag{1-11}$$

La estimación por resustitución generalmente subestima el costo. Si se tiene un conjunto de datos independiente, se pueden asignar los nuevos sujetos a varios nodos del árbol y calcular el costo basado en estos nuevos sujetos. Este costo tiende a ser más grande que la estimación del costo por resustitución porque el criterio de división está de alguna manera relacionado al costo, y como resultado, la estimación por resustitución del costo de mala clasificación usualmente es muy optimista. Adicionalmente, Breiman [4] prueba que a medida que aumentan los nodos en el árbol disminuye la estimación por resustitución 1-11, y como consecuencia, este estimador tiene el problema de seleccionar árboles sobre-ajustados.

Como ejemplo, suponga que se tiene una muestra de 3861 mujeres quienes después de estar embarazadas, tuvieron un bebé con vida. Se quiere determinar si el hecho de una mujer ser de color o estar trabajando influye en un parto prematuro. La figura **1-3** ilustra la situación. De las 3861 mujeres, 205 tuvieron partos prematuros (clase 1), mientras que, 3656 tuvieron partos en el tiempo normal (clase 0). El objetivo del árbol construido es clasificar un bebé al nacer como prematuro o no para poder brindarle el cuidado especial de ser necesario, por tanto, $c(1|0)$ es el costo de clasificar un bebé no prematuro como prematuro (el costo de cometer un falso-positivo), y $c(0|1)$ es el costo de clasificar un bebé prematuro como no prematuro (el costo de cometer un falso-negativo). Al cometer un falso-positivo se le brinda cuidado especial a un bebé que no lo necesita, mientras que, al cometer un falso-negativo se le niega cuidado especial a un bebe prematuro lo cual puede ser fatal. Por esta razón se asume que el costo que se paga al cometer un falso-negativo es mayor o igual que el de cometer un falso-positivo, es decir, $c(0|1) \geq c(1|0)$.
Para este ejemplo, se toma un rango de valores entre 1 y 18 para $c(0|1)$. El límite superior de 18 se basa en el hecho de que 3656:205=17.8:1, donde 205 y 3656 son respectivamente las cantidades de partos prematuros y no prematuros en el nodo raíz. La tabla **1-1** reporta los costos de mala clasificación para los cinco nodos de la figura **1-3**b). Cuando $c(0|1) = 10$, significa que cada error falso-negativo cuenta como 10 falsos-positivos. Si al nodo raíz se le asigna la clase 1 el costo es 3656, pero, si al nodo raíz se le asigna la clase 0 el costo es $205 * 10$=2050. En otras palabras, la pertenencia de un nodo a la clase 0 o 1 utilizando la ecuación 1-7, depende de si el costo de los errores falsos-positivos es más pequeño o no que el de los errores falsos-negativos. La tabla **1-2** muestra las estimaciones por resustitución del costo de mala clasificación para los cinco nodos en

el árbol de la figura **1-3**b) con $c(0|1) = 10$.

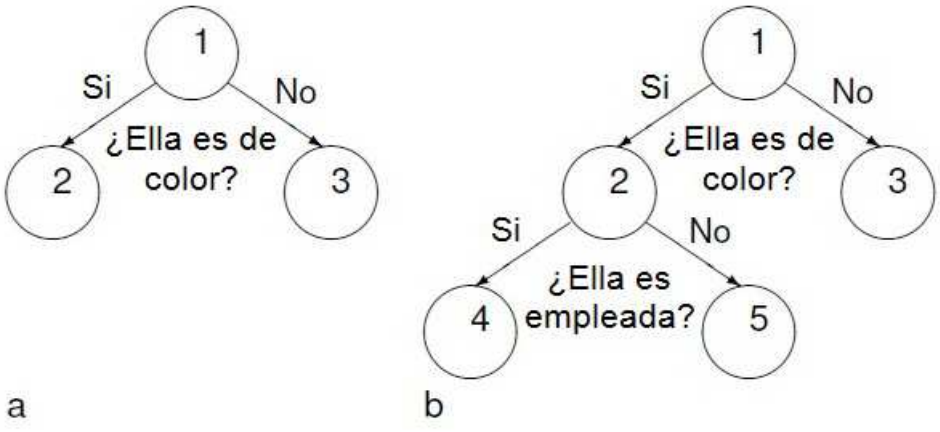

Figura 1-3.: El nodo 1 se divide en los nodos 2 y 3, luego, el nodo 2 se divide en los nodos 3 y 4. Fuente (Zhang [32]).

Tabla 1-1.: Costo de mala clasificación. Fuente (Zhang [32]).

		Nodo				
	Clase	1	2	3	4	5
$c(0\|1)$	1	3656	640	3016	187	453
1	0	205	70	135	11	59
10	0	2050	700	1350	110	590
18	0	3690	1260	2430	198	1062

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 sobre-ajuste. 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 $\tilde{\mathcal{T}}$ como

$$R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha|\tilde{\mathcal{T}}|, \tag{1-12}$$

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. Como

Tabla 1-2.: Estimaciones por resustitución del costo de mala clasificación con $c(0|1) = 10$. Fuente (Zhang [32]).

Nodo	Clase	P[τ]	r[τ]	$R^s(\mathcal{T})$
1	0	$\frac{3861}{3861}$	$\frac{10*205}{3861}$	$\frac{2050}{3861} = 0{,}531$
2	1	$\frac{710}{3861}$	$\frac{1*640}{710}$	$\frac{640}{3861} = 0{,}166$
3	0	$\frac{3151}{3861}$	$\frac{10*135}{3151}$	$\frac{1350}{3861} = 0{,}35$
4	0	$\frac{198}{3861}$	$\frac{10*11}{198}$	$\frac{110}{3861} = 0{,}028$
5	1	$\frac{506}{3861}$	$\frac{1*453}{506}$	$\frac{453}{3861} = 0{,}117$

ejemplo, denote por $\mathcal{T}_1$ y $\mathcal{T}_0$ los árboles de las figuras **1-3**a) y **1-3**b), respectivamente, y sea $\mathcal{T}_2$ el árbol que contiene solamente el nodo raíz. Nótese que $\mathcal{T}_1$ y $\mathcal{T}_2$ son los únicos subárboles de $\mathcal{T}_0$ distintos de él mismo. Usando las estimaciones por resustitución en la tabla **1-2**, el costo de $\mathcal{T}_0$ es 0.350+0.028+0.117=0.495 y su complejidad es 3, por tanto, su costo-complejidad es $0.495+3\alpha$ para un parámetro de complejidad α dado. La pregunta es: ¿existe un subárbol de $\mathcal{T}_0$ más pequeño que tenga el mismo costo-complejidad? El siguiente teorema es fundamental para responder esta pregunta.

Teorema
(Breiman [4]) Sea $\mathcal{T}_0$ un árbol dado. Para cualquier valor del parámetro de complejidad α, existe un único subárbol más pequeño de $\mathcal{T}_0$ que minimiza el costo-complejidad.

El teorema anterior afirma que no se pueden tener dos subárboles de $\mathcal{T}_0$ de tamaño más pequeño y el mismo costo-complejidad. Este subárbol más pequeño se conoce como subárbol óptimo con respecto al parámetro de complejidad.
Cuando $\alpha = 0$, el subárbol óptimo es el mismo $\mathcal{T}_0$. Para el ejemplo, el costo complejidad de $\mathcal{T}_1$ es $0.166+0.350+ 0*2=0.516$ y el de $\mathcal{T}_2$ es $0.531+0*1=0.531$, los cuales son más grandes que 0.495 el cual es el costo complejidad de $\mathcal{T}_0$.

Se puede escoger un α lo suficientemente grande para que el subárbol óptimo correspondiente sea de un solo nodo. Si se toma $\alpha \geq 0.018$, se tiene que

$$R_{0,018}(\mathcal{T}_2) = 0{,}531 + 0{,}018 * 1 = 0{,}495 + 0{,}018 * 3 = R_{0,018}(\mathcal{T}_0)$$

y

$$R_{0,018}(\mathcal{T}_2) = 0{,}531 + 0{,}018 * 1 < 0{,}516 + 0{,}018 * 2 = R_{0,018}(\mathcal{T}_1),$$

por tanto, $\mathcal{T}_2$ es el subárbol óptimo ya que tiene menor tamaño que $\mathcal{T}_0$.
Se debe tener en cuenta que no todos los subárboles son óptimos con respecto a un parámetro de complejidad, por ejemplo, $\mathcal{T}_1$ no es óptimo con respecto a ningún parámetro, ya que para $\alpha \in [0; 0{,}018)$ el subárbol óptimo es $\mathcal{T}_0$, mientras que si $\alpha \in [0{,}018; +\infty)$ el subárbol óptimo es

$\mathcal{T}_2$. Lo anterior muestra que si bien el parámetro de complejidad tiene un rango continuo de valores, solo se dispone de un número finito de subárboles, además, un subárbol óptimo es óptimo para un intervalo en el rango del parámetro de complejidad, y el número de tales intervalos es finito.

El uso del costo-complejidad permite construir una secuencia de *subárboles óptimos anidados* (ver Zhang [32]) desde cualquier árbol $\mathcal{T}$ dado. La idea es construir la secuencia de subárboles anidados para el 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 dispone de una muestra de prueba, estimar $R(\mathcal{T})$ es sencillo para cualquier subárbol $\mathcal{T}$, porque sólo se necesita aplicar los subárboles a la muestra de prueba, pero, si no se tiene una muestra de prueba, se puede utilizar el proceso de *validación cruzada* (ver Zhang [32]) para crear muestras artificiales y así estimar $R(\mathcal{T})$.

1.5. 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:

$$i(\tau) = \sum_{\text{sujeto } i \in \tau} (Y_i - \bar{Y}(\tau))^2, \tag{1-13}$$

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 función de división

$$\phi(s, \tau) = i(\tau) - i(\tau_L) - i(\tau_R), \tag{1-14}$$

donde s es la división permitida para el nodo τ. A diferencia de la bondad de una división en 1-4, la función de división 1-14 no necesita pesos. Además, se puede hacer uso de $i(\tau)$ para definir el costo del árbol como

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} i(\tau), \tag{1-15}$$

y luego sustituirlo en la ecuación 1-12 para formar el costo-complejidad.

1.6. La librería rpart del paquete estadístico R

En este trabajo se utiliza la librería *rpart* del paquete estadístico R para ajustar los árboles de regresión en el estudio de simulación.

Los programas de rpart construyen modelos de clasificación o de regresión de una estructura muy general usando el proceso de construcción de árboles visto anteriormente con algunas variaciones.

En la parte de particionamiento recursivo, esta librería tiene la opción de asignar el número mínimo de observaciones, n_{min}, que debe tener un nodo para dividirlo. De manera predeterminada es 20.

Tiene la opción de asignar el número mínimo de observaciones que debe tener un nodo terminal. De manera predeterminada es $\frac{n_{min}}{3}$.

Tiene dos opciones de medida de impureza para el particionamiento recursivo: el índice Gini y la entropía. El programa de manera predeterminada trabaja con el índice Gini.

Tiene la opción de asignar la matriz de costo $[c(i|j)]_{i\times j}$, $i, j = 1, \ldots, C$ donde C es el número de clases de la variable Y. De manera predeterminada se toma $c(i|j) = 1$ para todo $i \neq j$.

Trabaja con el método de *la apriori alterada* (ver Therneau [29]), el cual sirve para calcular las probabilidades apriori de cada clase utilizando la matriz de costo. La apriori alterada simplemente ayuda a la función de impureza a escoger para cada nodo la división que sea probablemente la mejor en términos del costo.

Tiene la opción de asignar el parámetro de complejidad α. Computacionalmente, este parámetro significa que cualquier división que no disminuya la falta general de ajuste en un factor de α no se intenta. La principal función de este parámetro es ahorrar tiempo de cálculo mediante la poda de divisiones que, obviamente, no valen la pena. Esencialmente, el usuario informa al programa que cualquier división que no mejore el ajuste con α es probable que se pode por *validación cruzada* (ver Zhang [32], Therneau [29]), y que por tanto el programa no necesita calcularla.

Para construir árboles de regresión emplea el método *anova* (ver Therneau [29]), el cual utiliza como criterio de división de un nodo la fórmula $SST - (SSL + SSR)$, donde $SST = \sum (y_i - \overline{y})^2$ es la suma de cuadrados para el nodo, y SSR, SSL son las sumas de cuadrados para el nodo hijo derecho e izquierdo, respectivamente. Esto es equivalente a elegir la división que maximice la suma de cuadrados entre grupos en un simple análisis de varianza. Este es el método que `rpart` tiene predeterminado cuando la variable dependiente es continua.

Para ajustar árboles CART con los valores predeterminados de los parámetros en la librería *rpart* se utiliza la instrucción `rpart(y` $\sim$ `x`$_1$ `+ x`$_2$ `+ ... + x`$_p$`)`, donde `y` es la variable respuesta y `x`$_1$`, x`$_2$`,...,x`$_p$ son la variables predictoras. Si `y` es discreta la función ajusta un árbol de clasificación y si es continua un árbol de regresión. En el estudio de simulación realizado en este trabajo se tiene solo una variable predictora, `x`, por tanto, la instrucción utilizada para ajustar los árboles de regresión es `rpart(y` $\sim$ `x)`.

Para el uso de las rutinas de *rpart* en R remítase a Therneau [29].

1.7. Regresión por mínimos cuadrados

Según Breiman [4], la regresión consiste de datos (x, y) donde $\mathtt{x}$ es un vector que cae en un espacio de medida X e y es un número real. La variable y es usualmente llamada variable respuesta o dependiente. Las variables en $\mathtt{x}$ son conocidas como variables predictoras o independientes.

Una regla de predicción o predictor es una función $d(\mathtt{x})$, definida en X que toma valores reales. El análisis de regresión es el término genérico involucrado alrededor de la construcción de un predictor $d(\mathtt{x})$ comenzando con una muestra de aprendizaje $\mathcal{L}$. La construcción de un predictor puede tener dos propósitos:

1. predecir la variable respuesta correspondiente a medidas futuras de las variables predictoras tan preciso como sea posible;
2. entender la relación estructural entre la variable respuesta y las variables independientes.

Suponga que una muestra de aprendizaje $\mathcal{L} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ de tamaño N, fue usada para construir un predictor $d(\mathtt{x})$. Entonces la pregunta es cómo medir la precisión de este predictor. Si se toma una muestra de prueba muy grande $(x_1, y'_1), (x_2, y'_2), \ldots, (x_{N_2}, y'_{N_2})$ de tamaño N_2, la precisión de $d(\mathtt{x})$ podría ser la *medida del error cuadrático*,

$$\frac{\sum_{i=1}^{N_2}(y'_i - d(\mathtt{x}))^2}{N_2}, \tag{1-16}$$

la cual es la medida de precisión clásicamente usada en regresión. La metodología que hay alrededor de esta medida es la regresión por mínimos cuadrados. Asuma que el vector aleatorio (X, y) y la muestra de aprendizaje $\mathcal{L}$ son independientemente extraídas de la misma distribución subyacente (fundamental).

Definición
Se define el *error cuadrático medio* $R^*(d)$ del predictor d como

$$R^*(d) = E(Y - d(\mathtt{x}))^2 \tag{1-17}$$

Esto es, $R^*(d)$ es el error cuadrático esperado usando $d(\mathtt{x})$ como un predictor de Y cuando la esperanza es tomada con el soporte $\mathcal{L}$ fijo. Usando la anterior definición, el predictor óptimo tiene una forma simple.

Proposición
El predictor d_B que minimiza $R^*(d)$, llamado *predictor óptimo de Bayes*, es

$$d_B(\mathtt{x}) = E(Y|X = \mathtt{x}) \tag{1-18}$$

En otras palabras, $d_B(\mathtt{x})$ es la esperanza condicional de la respuesta, dado que las variables predictoras toman el valor $\mathtt{x}$.

Importante: El valor del error cuadrático medio, $R^*(d)$, depende del rango de la variable respuesta.

En el siguiente capítulo se definen las medidas del error de predicción para el modelo de regresión lineal y los árboles de regresión CART, las cuales se basan en la medida del error cuadrático 1-16.

1.8. Descripción del estudio de simulación

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

$$Y = F(\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_p) + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2) \tag{1-19}$$

donde

$$Y = F(\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_p) = \beta_0 + \sum_{j=1}^{p} \beta_j \mathtt{x}_j = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x) = f(x) \tag{1-20}$$

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 1-20.

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,

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \tag{1-21}$$

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 1-19.

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 ecuaciónes 2-3 y 2-4.

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 el cociente $COCEP = \frac{EPCART}{EPRL}$ para comparar los dos errores de predicción. Si $COCEP > 1$, la regresión lineal predice mejor los datos que los árboles de regresión, pero, si $COCEP < 1$ los árboles de regresión predicen mejor los datos que la regresión lineal. Cuando $COCEP = 1$ ambos modelos predicen igual. Este cociente se toma para comparar cuántas veces es más grande el error de predicción de los árboles de regresión que el error de predicción de la regresión lineal cuando $COCEP > 1$. Adicionalmente, en las tablas se reporta 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 y es equivalente a $COCEP$. A medida que $DIFLOG \to 0$, los dos errores de predicción se van acercando entre ellos. Si $DIFLOG > 0$ entonces $EPCART > EPRL$, pero, si $DIFLOG < 0$ entonces $EPCART < EPRL$. Si $DIFLOG = 0$ entonces $EPCART = EPRL$.

2. Predicción de un modelo de regresión lineal utilizando CART

2.1. Medida 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:

$$y = \beta_0 + \beta_1 \mathtt{x}_1 + \ldots + \beta_p \mathtt{x}_p + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2). \tag{2-1}$$

De lo anterior se sabe que

$$y_{verd} = E[y|\mathtt{x}_1, \ldots, \mathtt{x}_p] = \beta_0 + \beta_1 \mathtt{x}_1 + \ldots + \beta_p \mathtt{x}_p.$$

Por tanto, el predictor óptimo de Bayes 1-18 que minimiza el error cuadrático medio es,

$$d_B(\mathtt{x}) = y_{verd}.$$

Suponga que se construye un predictor $d(\mathtt{x})$ con la muestra de aprendizaje $\mathcal{L} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ y sean $y'_1, y'_2, \ldots, y'_n$ los valores predichos por el predictor $d(\mathtt{x})$ para $y_1, y_2, \ldots, y_n$. Si en la ecuación 1-16 se toma como muestra de prueba $(x_1, y'_1), (x_2, y'_2), \ldots, (x_n, y'_n)$ y se sustituye $d(\mathtt{x})$ por $d_B(\mathtt{x})$, se obtiene la medida 2-2 que ya no es la medida de precisión del predictor $d(\mathtt{x})$, sino más bien, una medida de la precisión del predictor $d(\mathtt{x})$ con respecto al predictor óptimo $d_B(\mathtt{x})$, el cual es la verdadera media de los datos.

$$\frac{\sum_{i=1}^{n}(y'_i - d_B(\mathtt{x}))^2}{n} = \frac{\sum_{i=1}^{n}(y'_i - y_{verd})^2}{n} \tag{2-2}$$

A continuación, se definirán los errores de predicción para el modelo de regresión lineal y los árboles de regresión CART, los cuales se basan en la medida 2-2.

2.1.1. Medida del error para la predicción por regresión lineal

Suponga que para el conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, donde n es grande, se ajusta un modelo de regresión lineal, por tanto, los valores predichos son de la forma:

$$y_{reg} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p,$$

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

Reemplazando y_{reg} en 2-2, el error de predicción se calcula como

$$EPRL = \frac{\sum_{i=1}^{n}(y_{reg} - y_{verd})^2}{n}. \tag{2-3}$$

2.1.2. Medida del error para la predicción por CART

Suponga, además, que para el conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ se estima un árbol de regresión, obteniendo un árbol de k nodos terminales. Sean $C_1, C_2, \ldots, C_k$ las clases correspondientes a los k nodos terminales, por tanto, los valores predichos por el árbol de regresión son de la forma:

$$y_{cart} = f(x) = \begin{cases} r_i & \text{si} \quad x \in C_i \; ; \;\; i = 1, \ldots, k \\ \\ 0 & \text{si} \quad \text{en otro caso} \end{cases}$$

donde,

$$r_i = \frac{\sum\{y_j | x_j \in C_i, \; j = 1, \ldots, n\}}{|\{y_j | x_j \in C_i, \; j = 1, \ldots, n\}|} \; ; \quad i = 1, \ldots, k.$$

Reemplazando y_{cart} en 2-2, el error de predicción se calcula como

$$EPCART = \frac{\sum_{i=1}^{n}(y_{cart} - y_{verd})^2}{n}. \tag{2-4}$$

2.2. Sensibilidad del error de predicción de CART a cambios en el rango de la variable respuesta

Breiman [4], afirma que el error cuadrático medio de CART depende del rango de la variable respuesta. Como la medida del error cuadrático 1-16 de un predictor $d(\mathtt{x})$ es una estimación del error cuadrático medio 1-17, y a su vez, $EPCART$ se definió en términos del error cuadrático 1-16, es de esperarse que $EPCART$ también dependa del rango de la variable respuesta. A continuación, se muestra que $EPCART$ depende del rango de la variable respuesta.

Si en la ecuación 2-1 se toma $p = 1$, se obtiene el modelo de regresión lineal

$$y = \beta_0 + \beta_1 \mathtt{x} + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2). \tag{2-5}$$

Para generar modelos de regresión lineal con variables respuesta de diferente rango, se generan modelos de regresión de la forma 2-5, donde, los errores y el intercepto son los mismos en ambos modelos, pero, la pendiente es distinta. Es decir, sean $e_1, e_2, \ldots, e_n$, extraídos de una distribución $N(0, \sigma^2)$, β_0, β_1 y β_1^* constantes, $\beta_1 \neq \beta_1^*$.
Sean

$$y_i = \beta_0 + \beta_1 \mathtt{x}_i + e_i \quad \text{y} \quad y_i^* = \beta_0 + \beta_1^* \mathtt{x}_i + e_i, \quad i = 1, \ldots, k, \tag{2-6}$$

entonces, $\{y_i\}_{i=1}^n$ y $\{y_i^*\}_{i=1}^n$ son dos conjuntos de datos extraídos respectivamente de las variables y y $y*$ con rangos diferentes.
En efecto, de las ecuaciones en 2-6 se obtiene que

$$y_i = y_i^* + (\beta_1 - \beta_1^*)\mathtt{x}_i, \tag{2-7}$$

lo cual implica que para cualquier i, el valor de y_i es el valor de y_i^* más un término distinto de cero que depende únicamente de $\mathtt{x}_i$. Por tanto, los rangos de y y $y*$ son diferentes.

En la tabla **2-1** se muestran los errores de predicción para conjuntos de $n = 1000$ datos generados del modelo descrito por la ecuación 2-5 con $\beta_0 = 10$, $e_1, e_2, \ldots, e_n$ fijos y distintos valores de la pendiente β_1. Se puede observar que para un valor fijo de σ el error de predicción de CART aumenta cuando la pendiente de la recta aumenta ($EPCART \longrightarrow \infty$ cuando $\beta_1 \longrightarrow \infty$) y disminuye cuando la pendiente de la recta de regresión disminuye ($EPCART \longrightarrow 0$ cuando $\beta_1 \longrightarrow 0$). Nótese que cuando $\beta_1 = 0{,}001$ el error de predicción de CART es más pequeño que el error de predicción de la regresión para cualquier valor de σ. También se puede observar, que para un valor fijo de σ el error de predicción de la regresión lineal permanece constante para cualquier valor de la pendiente β_1, mostrando así, que $EPRL$ es invariante a cambios en el rango de la variable respuesta.

Tabla 2-1.: Sensibilidad de $EPCART$ a cambios en la pendiente β_1 para $n = 1000$ observaciones.

σ	β_1	EPRL	EPCART	EPCART/EPRL
1	3	0.0036	115.4785	31717.3345
	2	0.0036	51.4280	14125.2206
	1	0.0036	12.9760	3563.9940
	1/2	0.0036	5.2036	1429.2163
	1/10	0.0036	0.3253	89.3392
	1/100	0.0036	0.0182	4.9918
	1/1000	0.0036	0.0011	0.3084
$\sqrt{2}$	3	0.0073	115.7094	15890.3719
	2	0.0073	51.6817	7097.4434
	1	0.0073	12.9477	1778.1062
	1/2	0.0073	5.2139	716.0202
	1/10	0.0073	0.3200	43.9410
	1/100	0.0073	0.0235	3.2324
	1/1000	0.0073	0.0014	0.1960
$\sqrt{3}$	3	0.0109	115.7240	10594.9179
	2	0.0109	51.6955	4732.8950
	1	0.0109	12.9531	1185.9024
	1/2	0.0109	5.2690	482.3903
	1/10	0.0109	0.3708	33.9521
	1/100	0.0109	0.0248	2.2721
	1/1000	0.0109	0.0017	0.1585
2	3	0.0146	116.2768	7984.1490
	2	0.0146	51.9041	3563.9940
	1	0.0146	20.8143	1429.2163
	1/2	0.0146	5.2797	362.5315
	1/10	0.0146	0.3749	25.7450
	1/100	0.0146	0.0831	5.7031
	1/1000	0.0146	0.0020	0.1398
$\sqrt{5}$	3	0.0182	116.2907	6388.0789
	2	0.0182	51.7531	2842.9011
	1	0.0182	20.8246	1143.9379
	1/2	0.0182	5.2905	290.6162
	1/10	0.0182	0.5339	29.3307
	1/100	0.0182	0.0834	4.5792
	1/1000	0.0182	0.0023	0.1286
3	3	0.0328	116.7842	3563.9940
	2	0.0328	51.8033	1580.9220
	1	0.0328	20.9317	638.7878
	1/2	0.0328	7.9486	242.5746
	1/10	0.0328	0.5473	16.7014
	1/100	0.0328	0.0846	2.5812
	1/1000	0.0328	0.0036	0.1086

2.3. Estandarización de los datos

Teóricamente, para la recta de regresión descrita por la ecuación 2-5, como $y \sim N(\beta_0 + \beta_1 x,\ \sigma^2)$, la estandarización está dada por:

$$z = \frac{y - y_{verd}}{\sigma}$$

donde la variable z se puede ver como un modelo de regresión lineal de la forma

$$z = \beta_0^* + \beta_1^* \mathtt{x} + \varepsilon^*,$$

con

$$\beta_0^* = 0,\ \beta_1^* = 0,\ \varepsilon^* = \frac{\varepsilon}{\sigma} \sim N(0,1),$$

lo que implica que si n es suficientemente grande, el modelo de regresión lineal estandarizado estimará un $\widehat{\beta_1^*} \approx 0$ (β_1^* será no significativo) y por tanto $EPCART$ sería tan bueno o quizás mejor que $EPRL$ como se observa en la tabla **2-1**. Gráficamente se puede ver en la figura **2-1**, que cuando la pendiente de la recta disminuye, también disminuye el rango de la variable respuesta Y, es decir, para $\beta_1 = 2$ se tiene un rango aproximado de 0 a 200 para la variable Y, para $\beta_1 = 1$ se tiene un rango aproximado de 0 a 100, para $\beta_1 = 0{,}5$ se tiene un rango aproximado de 0 a 55, ... , hasta llegar a $\beta_1 = 0{,}001$ que tiene un rango aproximado de $\{10\}$ y es donde las predicciones de CART y regresión lineal coinciden.

Debido a que la medida el error cuadrático medio de CART (ecuación 1-17) es afectado seriamente por el rango de la variable respuesta, Breiman [4] sugiere la estandarización de los datos para que este error sea comparable. Como ya se mostró que $EPCART$ depende del rango de la variable respuesta, se deben estandarizar los datos para comparar dicho error.

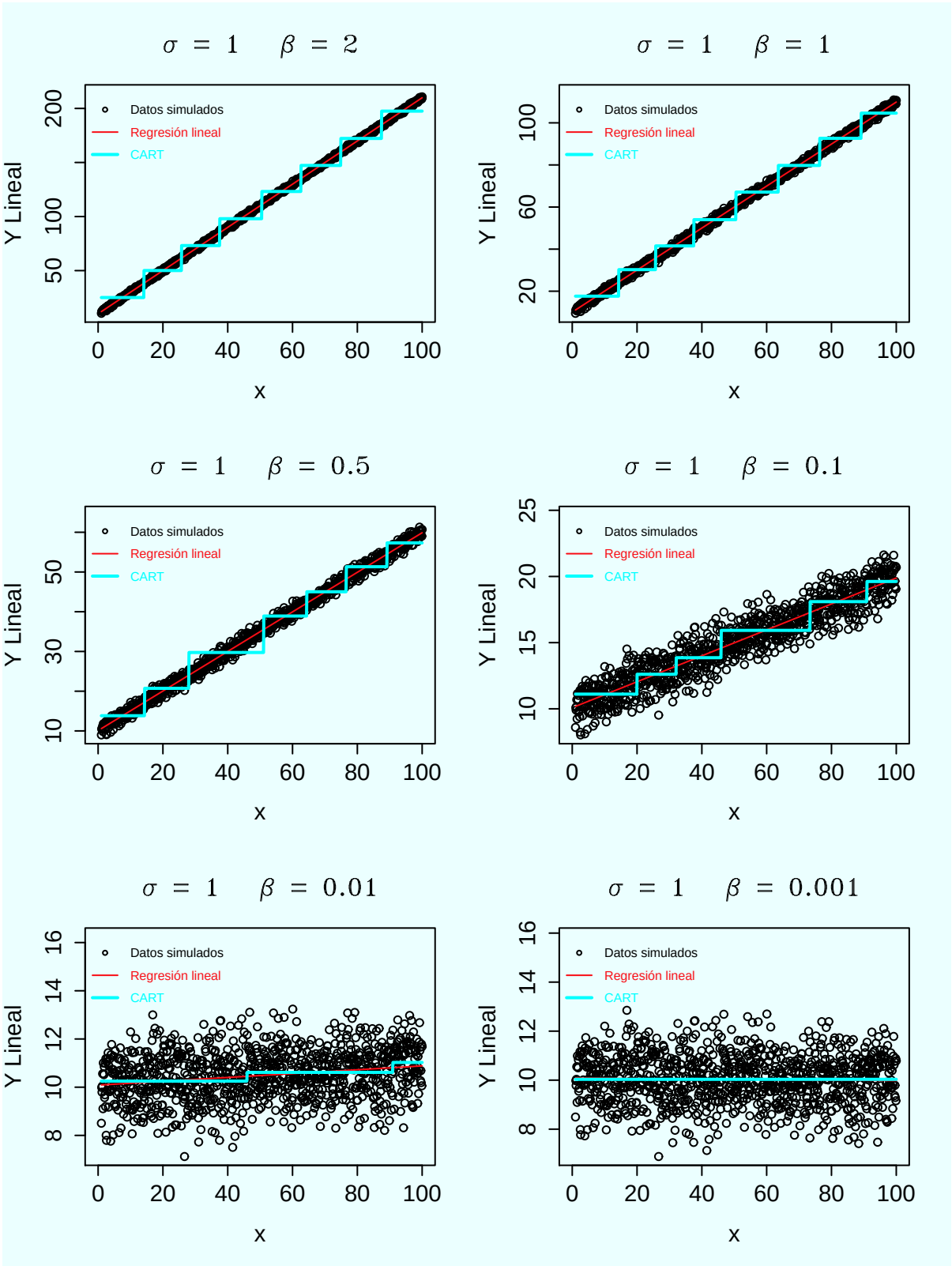

Figura 2-1.: Predicciones de CART y regresión lineal cuando cambia la pendiente $\beta_1 = \beta$ para $n = 1000$ observaciones.

3. Comparación de las predicciones de CART y modelos de regresión lineal ajustados correctamente

En este capítulo 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.

3.1. Predicción de modelos de regresión lineal cuadráticos

En esta sección se comparan los errores de predicción de CART y de regresión lineal para datos que siguen modelos de regresión 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:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2). \tag{3-1}$$

De lo anterior, se sabe que

$$y_{verd} = E(y) = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{3-2}$$

Para simular los conjuntos de datos se siguen los pasos descritos en la sección 1.8. En el paso 1, se toma $p = 2$ y se especifican las funciones

$$g_1(x) = x, \;\; g_2(x) = x^2. \tag{3-3}$$

Los valores de β_0, β_1 y β_2 se especifican a continuación.

3.1.1. Errores de predicción de CART vs Regresión Lineal para el modelo cuadrático 1

El primer modelo a analizar se obtiene al sustituir $\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 3$ en la ecuación 3-1 y se llamará modelo cuadrático 1.

En la tabla **3-1** se puede observar que para cualquier valor de n fijo, al aumentar la desviación estándar σ de los errores de los datos el error de predicción de la regresión lineal se aproxima al error de predicción de CART, siendo en todos los casos menor el error de predicción de la regresión lineal.
En los gráficos **3-1**, **3-2** y **3-3** se puede ver como las predicciones de CART describen la forma del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ cuando $n = 100$ o $n = 1000$, pero, el modelo de regresión lineal describe mejor los datos que CART. Nótese como el aumento de la desviación estándar no influye en la forma de las predicciones de CART para un n en particular en este modelo.

Tabla 3-1.: Comparación de los errores de predicción para el modelo cuadrático 1.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	1	0.0000	0.1388	635211003.6813	8.8029	0.1388
	10	0.0000	0.1376	5861152.6880	6.7680	0.1376
50	100	0.0000	0.1353	59490.3550	4.7744	0.1353
	500	0.0001	0.1311	2123.5917	3.3271	0.1311
	1000	0.0002	0.1208	491.5798	2.6916	0.1206
	2000	0.0014	0.1110	81.3178	1.9102	0.1096
	1	0.0000	0.0531	468159989.4510	8.6704	0.0531
	10	0.0000	0.0531	4333545.2818	6.6368	0.0531
100	100	0.0000	0.0440	36973.8326	4.5679	0.0440
	500	0.0000	0.0391	1226.7490	3.0888	0.0390
	1000	0.0002	0.0374	235.7695	2.3725	0.0372
	2000	0.0010	0.0386	39.5417	1.5971	0.0376
	1	0.0000	0.0318	1396968620.7102	9.1452	0.0318
	10	0.0000	0.0318	12427210.3594	7.0944	0.0318
500	100	0.0000	0.0297	120668.9173	5.0816	0.0297
	500	0.0000	0.0299	3584.0208	3.5544	0.0299
	1000	0.0001	0.0307	498.2235	2.6974	0.0306
	2000	0.0007	0.0327	50.1098	1.6999	0.0321
	1	0.0000	0.0319	2537541889.5718	9.4044	0.0319
	10	0.0000	0.0319	25568133.5129	7.4077	0.0319
1000	100	0.0000	0.0304	241990.3523	5.3838	0.0304
	500	0.0000	0.0300	5755.6709	3.7601	0.0300
	1000	0.0000	0.0306	625.2991	2.7961	0.0306
	2000	0.0006	0.0330	53.7893	1.7307	0.0324
	1	0.0000	0.0319	12734206954.8149	10.1050	0.0319
	10	0.0000	0.0319	126673715.2247	8.1027	0.0319
5000	100	0.0000	0.0315	1145813.9717	6.0591	0.0315
	500	0.0000	0.0298	9837.1635	3.9929	0.0298
	1000	0.0000	0.0314	786.8097	2.8959	0.0314
	2000	0.0006	0.0347	59.7462	1.7763	0.0341

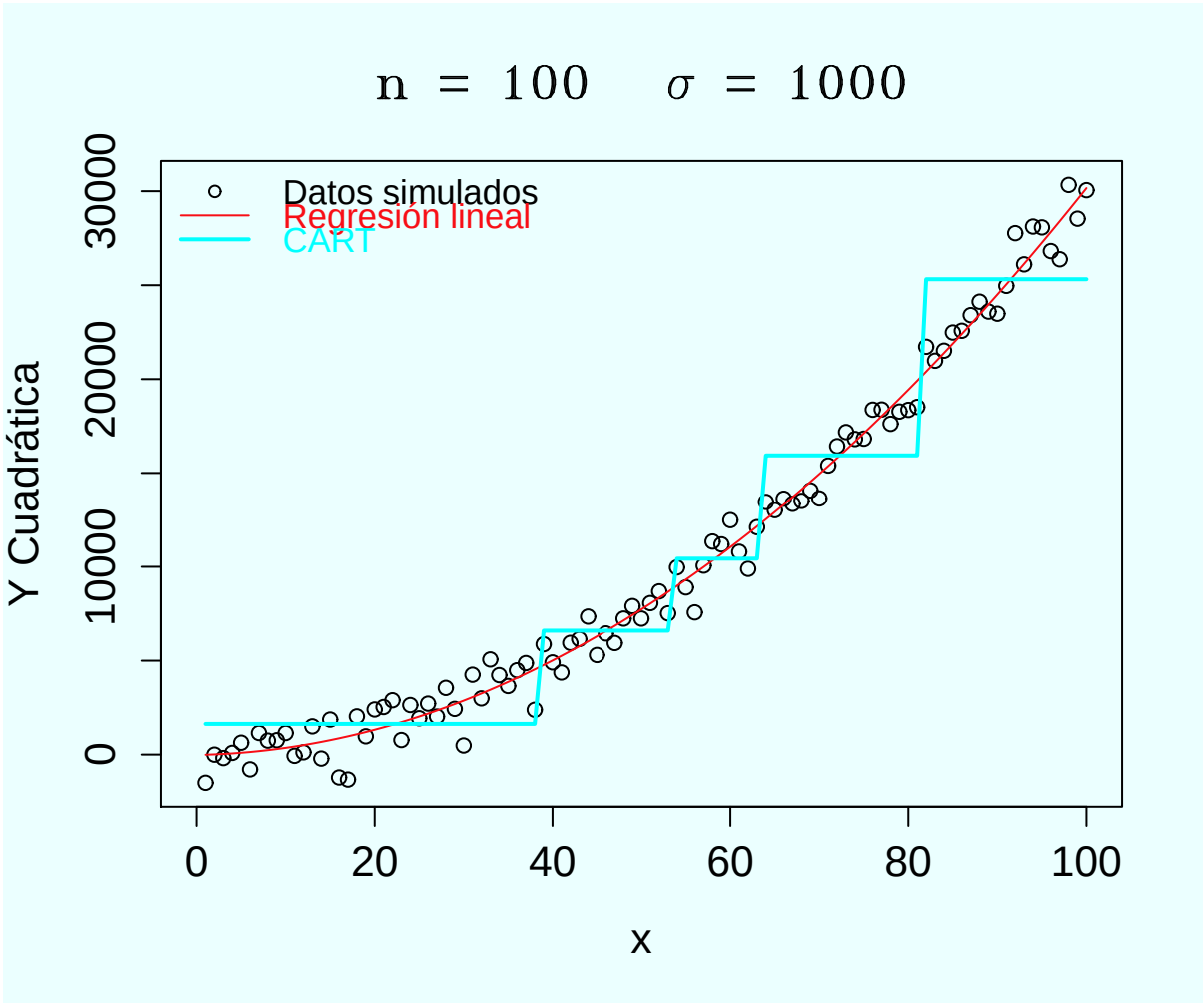

Figura 3-1.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 100$ y $\sigma = 1000$.

3.1.2. Errores de predicción de CART vs Regresión Lineal para el modelo cuadrático 2

El segundo modelo a analizar se obtiene al sustituir $\beta_0 = 680,\ \beta_1 = -22,\ \beta_2 = 0{,}25$ en la ecuación 3-1 y se llamará modelo cuadrático 2.

En la tabla **3-2** nuevamente se observa que para cualquier valor de n fijo, al aumentar la desviación estándar σ de los errores de los datos el error de predicción de la regresión lineal se aproxima al error de predicción de CART, siendo en todos los casos menor el error de predicción de la regresión lineal.

En los gráficos **3-4**, **3-5** y **3-6** se ve de nuevo como las predicciones de CART describen la forma del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ cuando $n = 100$ o $n = 1000$, pero, el modelo de regresión lineal describe mejor los datos que CART. Nótese

como el aumento de la desviación estándar no influye en la forma de las predicciones de CART para un n en particular en este modelo cuadrático.

Tabla 3-2.: Comparación de los errores de predicción para el modelo cuadrático 2.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
50	1	0.0000	0.1241	272582.3246	5.4355	0.1241
	5	0.0000	0.1197	10967.2505	4.0401	0.1196
	10	0.0000	0.1196	2824.7887	3.4510	0.1196
	25	0.0003	0.1246	395.7086	2.5974	0.1243
	50	0.0017	0.1357	78.4926	1.8948	0.1340
	100	0.0123	0.1640	13.3349	1.1250	0.1517
100	1	0.0000	0.1158	542250.4272	5.7342	0.1158
	5	0.0000	0.1015	18381.1063	4.2644	0.1015
	10	0.0000	0.0940	3772.3103	3.5766	0.0939
	25	0.0002	0.0919	482.3731	2.6834	0.0917
	50	0.0013	0.0915	70.1867	1.8463	0.0902
	100	0.0114	0.1021	8.9246	0.9506	0.0907
500	1	0.0000	0.0483	1110127.0781	6.0454	0.0483
	5	0.0000	0.0467	36489.6872	4.5622	0.0467
	10	0.0000	0.0465	7948.3349	3.9003	0.0465
	25	0.0001	0.0465	579.7854	2.7633	0.0464
	50	0.0009	0.0482	53.1405	1.7254	0.0473
	100	0.0105	0.0615	5.8495	0.7671	0.0510
1000	1	0.0000	0.0491	2035380.9164	6.3086	0.0491
	5	0.0000	0.0476	69433.1726	4.8416	0.0476
	10	0.0000	0.0467	12329.9136	4.0910	0.0467
	25	0.0001	0.0468	678.4437	2.8315	0.0467
	50	0.0009	0.0481	55.7232	1.7460	0.0472
	100	0.0103	0.0613	5.9369	0.7736	0.0510
5000	1	0.0000	0.0504	10249492.8548	7.0107	0.0504
	5	0.0000	0.0483	236674.7004	5.3742	0.0483
	10	0.0000	0.0476	24615.6716	4.3912	0.0476
	25	0.0001	0.0474	833.3753	2.9208	0.0473
	50	0.0008	0.0488	59.3101	1.7731	0.0479
	100	0.0103	0.0622	6.0469	0.7815	0.0519

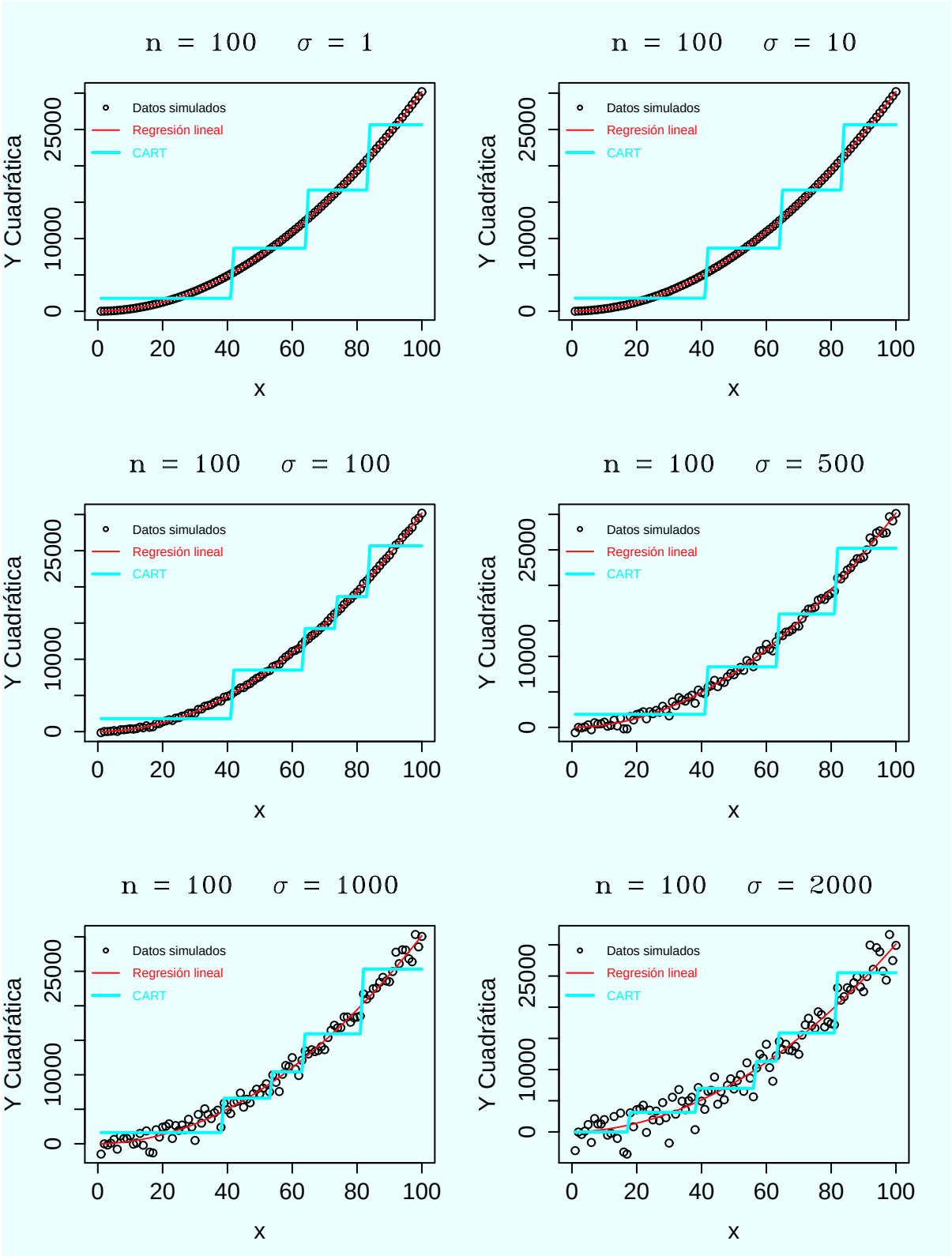

Figura 3-2.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 100$.

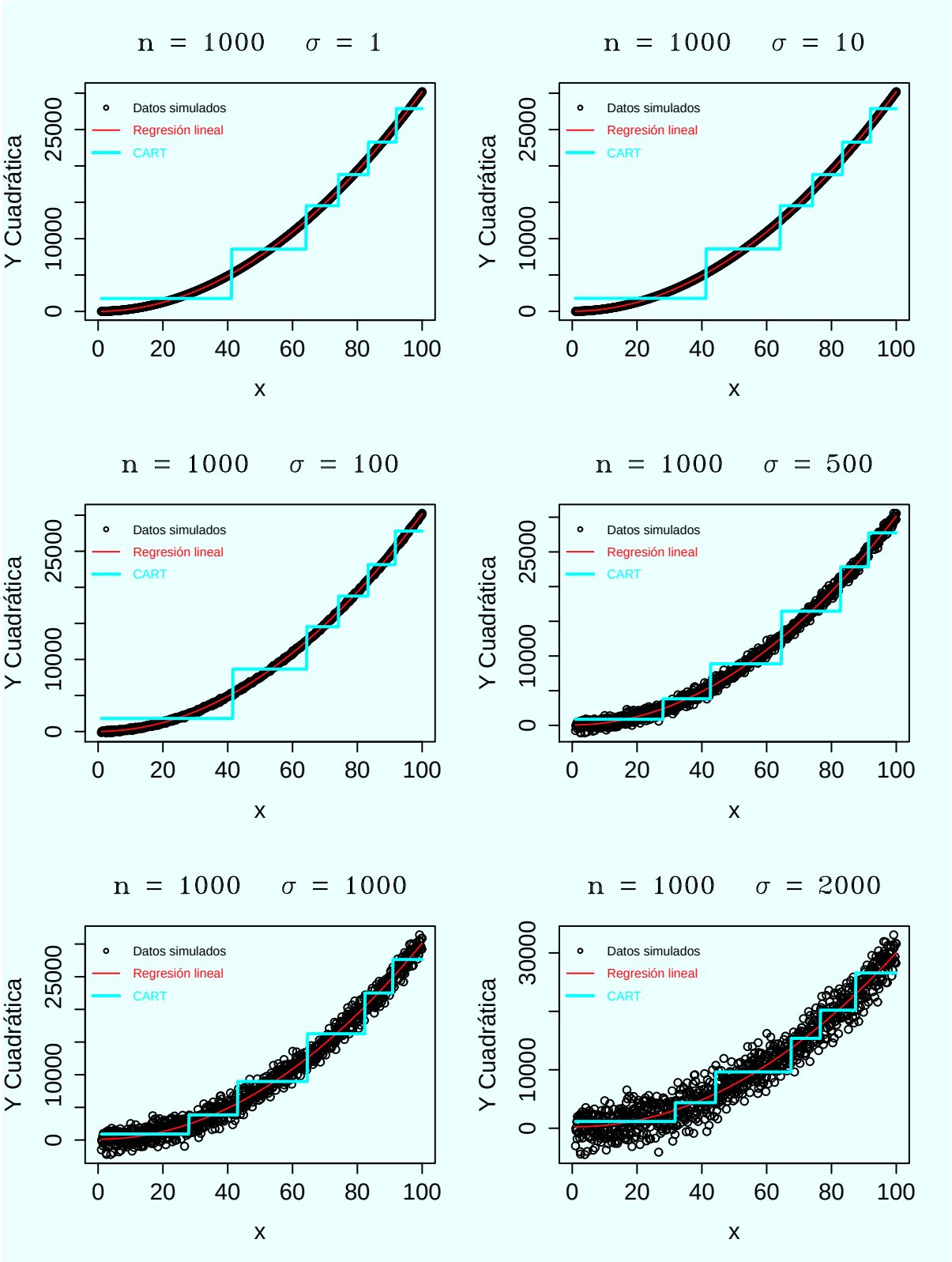

Figura 3-3.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 1000$.

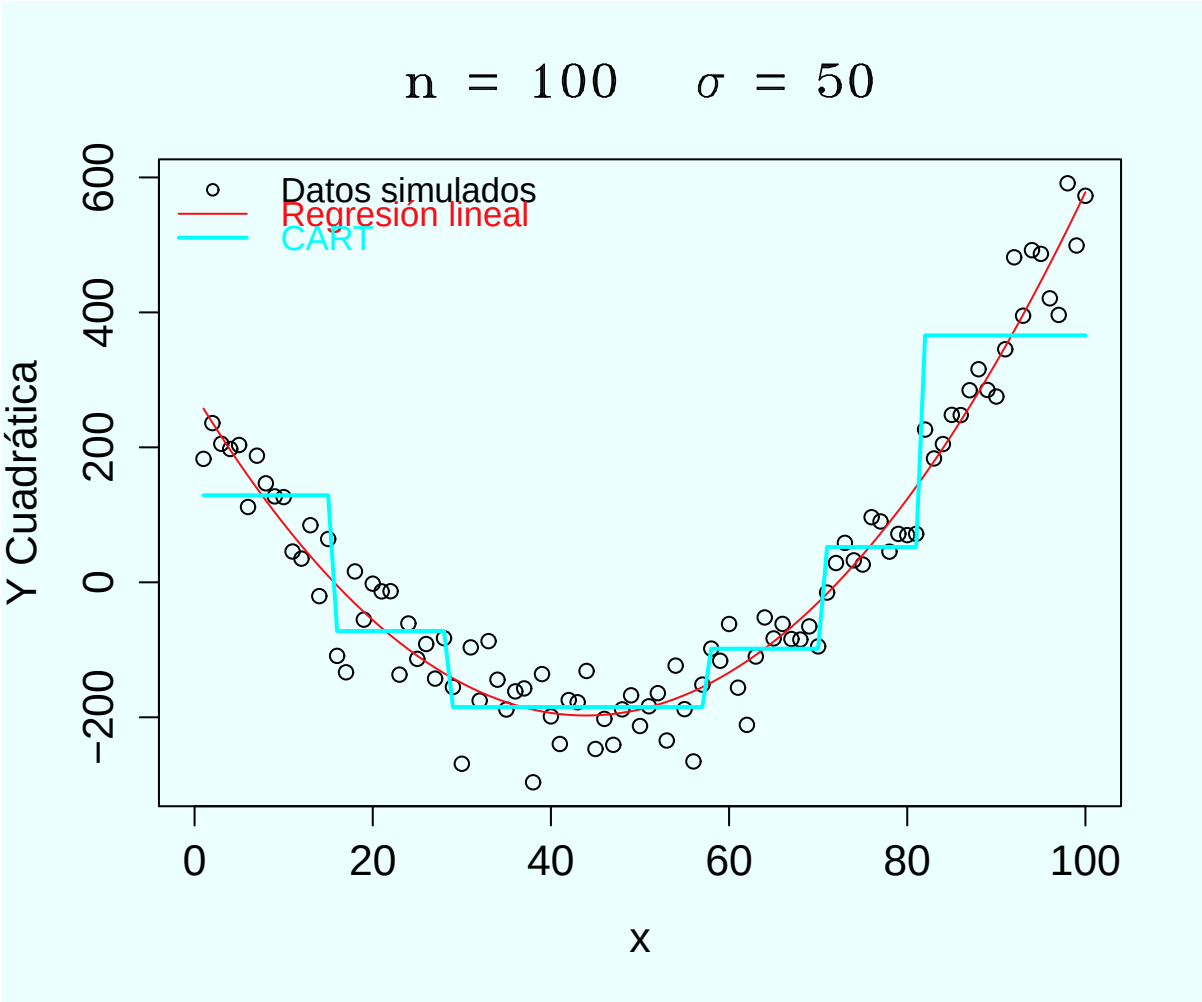

Figura 3-4.: Gráfico de las predicciones para el modelo cuadrático 2 con $n = 100$ y $\sigma = 50$.

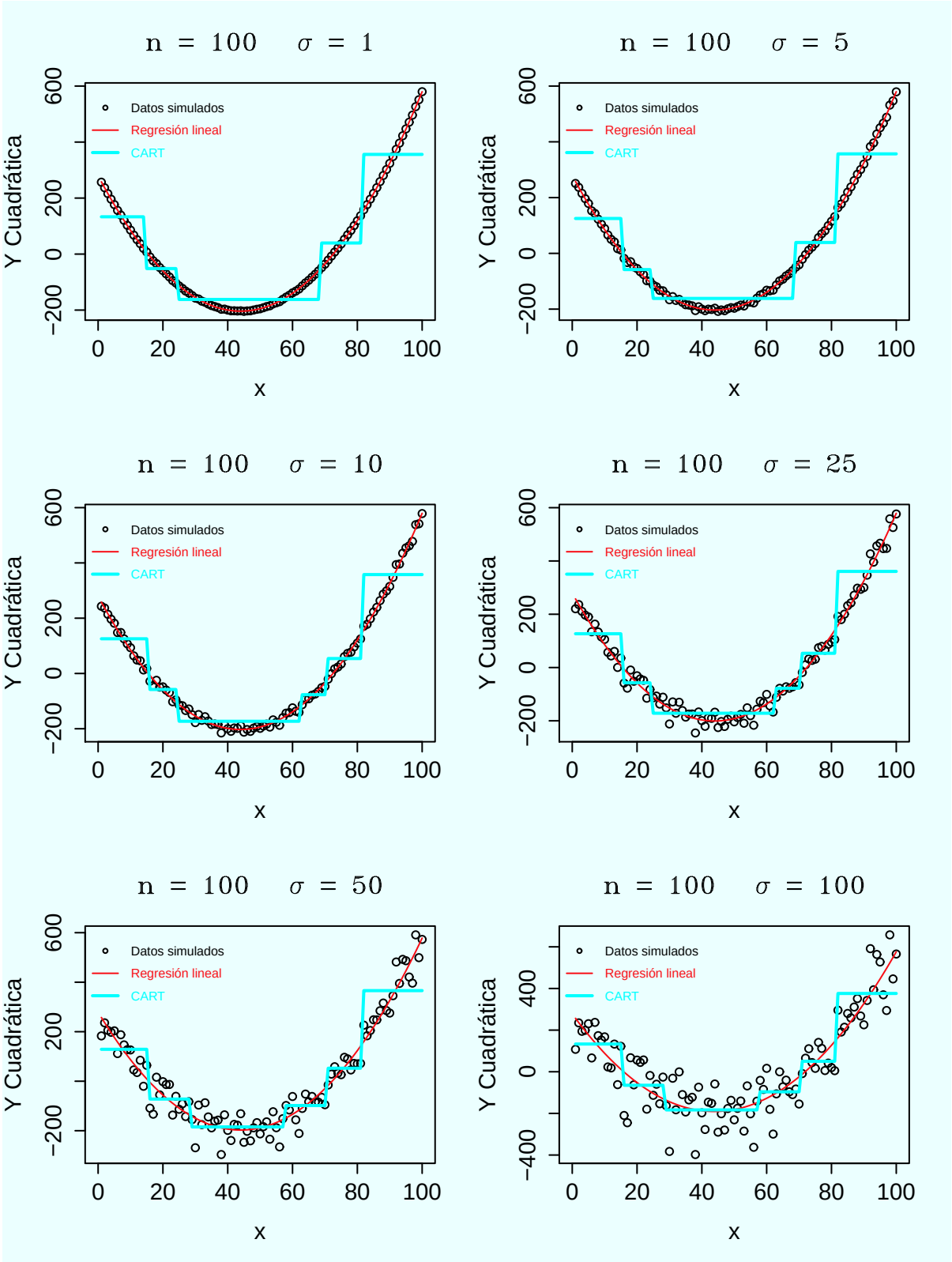

Figura 3-5.: Gráfico de las predicciones para el modelo cuadrático 2 con $n = 100$.

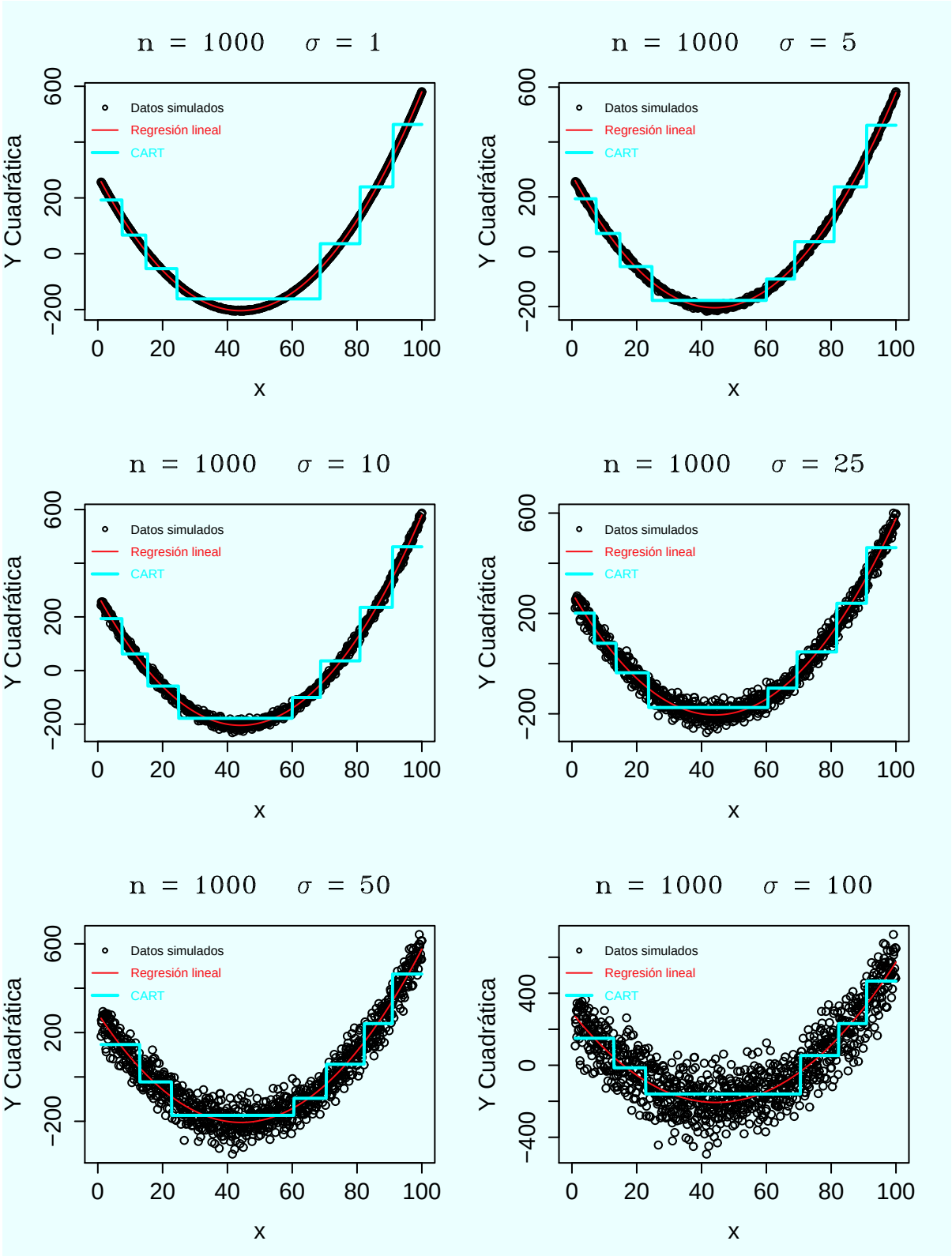

Figura 3-6.: Gráfico de las predicciones para el modelo cuadrático 2 con $n = 1000$.

3.2. Predicción de modelos de regresión lineal trigonométricos

En esta sección se compara los errores de predicción de CART y de regresión lineal para datos que siguen modelos de regresión 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:

$$y = a\sin(bx + c) + d + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2), \tag{3-4}$$

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

$$y_{verd} = E(y) = a\sin(bx + c) + d. \tag{3-5}$$

El modelo 3-4 se puede reescribir como

$$a\sin(bx + c) + d + \varepsilon = a\sin(c)\cos(bx) + a\cos(c)\sin(bx) + d + \varepsilon. \tag{3-6}$$

Para simular los conjuntos de datos se siguen los pasos descritos en la sección 1.8. En el paso 1, se toma $p = 2$, se especifican las funciones

$$g_1(x) = cos(bx), \;\; g_2(x) = sin(bx), \tag{3-7}$$

y se especifican los valores de los parámetros

$$\beta_0 = d, \;\; \beta_1 = a\sin(c), \;\; \beta_2 = a\cos(c). \tag{3-8}$$

Para encontrar a, c y d en términos de β_0, β_1 y β_2, se resuelven las ecuaciones

$$a = \sqrt{\beta_1^2 + \beta_2^2}, \;\; c = \arctan(\beta_1/\beta_2), \;\; d = \beta_0. \tag{3-9}$$

3.2.1. Errores de predicción de CART vs Regresión Lineal para el modelo trigonométrico 1

El tercer modelo a analizar se obtiene al sustituir $a = 10,\ b = 0{,}1,\ c = 1,\ d = 12$ en la ecuación 3-4 y se llamará modelo trigonométrico 1.

De igual manera que para los modelos cuadráticos, en la tabla **3-3** se puede observar que para cualquier valor de n fijo, al aumentar la desviación estándar σ de los errores de los datos el error de predicción de la regresión lineal se aproxima al error de predicción de CART, siendo en todos los casos menor el error de predicción de la regresión lineal.
En los gráficos **3-7**, **3-8** y **3-9** se puede ver como las predicciones de CART describen la forma del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ cuando

$n = 100$ o $n = 1000$, pero, el modelo de regresión lineal describe mejor los datos que CART. Nótese como el aumento de la desviación estándar no influye en la forma de las predicciones de CART para un n en particular en este modelo trigonométrico.

Tabla 3-3.: Comparación de los errores de predicción para el modelo trigonométrico 1.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	0.1	0.0000	0.1979	48657.3866	4.6871	0.1979
	0.3	0.0000	0.1993	5840.1653	3.7664	0.1993
50	0.5	0.0001	0.2023	2003.8434	3.3019	0.2022
	0.8	0.0003	0.2059	809.9077	2.9084	0.2056
	1	0.0004	0.2087	473.4253	2.6753	0.2083
	2	0.0026	0.2181	82.8109	1.9181	0.2155
	0.1	0.0000	0.0831	45553.5482	4.6585	0.0831
	0.3	0.0000	0.0809	5060.1587	3.7042	0.0808
100	0.5	0.0001	0.0803	1531.4987	3.1851	0.0803
	0.8	0.0002	0.0813	520.0218	2.7160	0.0811
	1	0.0003	0.0822	307.7489	2.4882	0.0820
	2	0.0020	0.0891	45.5595	1.6586	0.0872
	0.1	0.0000	0.0546	135665.9892	5.1325	0.0546
	0.3	0.0000	0.0531	12920.3603	4.1113	0.0531
500	0.5	0.0000	0.0527	3326.4887	3.5220	0.0527
	0.8	0.0001	0.0529	859.9178	2.9345	0.0528
	1	0.0001	0.0536	420.9710	2.6243	0.0535
	2	0.0015	0.0564	38.6943	1.5876	0.0550
	0.1	0.0000	0.0547	258728.5504	5.4128	0.0547
	0.3	0.0000	0.0536	21440.0530	4.3312	0.0536
1000	0.5	0.0000	0.0533	5015.1407	3.7003	0.0533
	0.8	0.0000	0.0533	1079.6402	3.0333	0.0533
	1	0.0001	0.0538	496.0412	2.6955	0.0537
	2	0.0014	0.0563	40.5576	1.6081	0.0549
	0.1	0.0000	0.0549	1115370.5524	6.0474	0.0549
	0.3	0.0000	0.0547	50363.0570	4.7021	0.0547
5000	0.5	0.0000	0.0543	8138.7161	3.9106	0.0543
	0.8	0.0000	0.0537	1350.1608	3.1304	0.0537
	1	0.0001	0.0540	576.3095	2.7607	0.0539
	2	0.0013	0.0565	42.0921	1.6242	0.0551

3.2.2. Errores de predicción de CART vs Regresión Lineal para el modelo trigonométrico 2

El cuarto modelo a analizar se obtiene al sustituir $a = 10,\ b = 0{,}5,\ c = 1,\ d = 12$ en la ecuación 3-4 y se llamará modelo trigonométrico 2.

Como en los modelos anteriores, de la tabla **3-4** se puede observar que para cualquier valor de n fijo, al aumentar la desviación estándar σ de los errores de los datos el error de predicción de la regresión lineal se aproxima al error de predicción de CART, siendo en todos los casos menor el error de predicción de la regresión lineal.
En los gráficos **3-10**, **3-11**, con $n = 100$, se puede ver como las predicciones de CART intentan describir el verdadero modelo, con poco éxito, pues, hay máximos y mínimos relativos que no logra

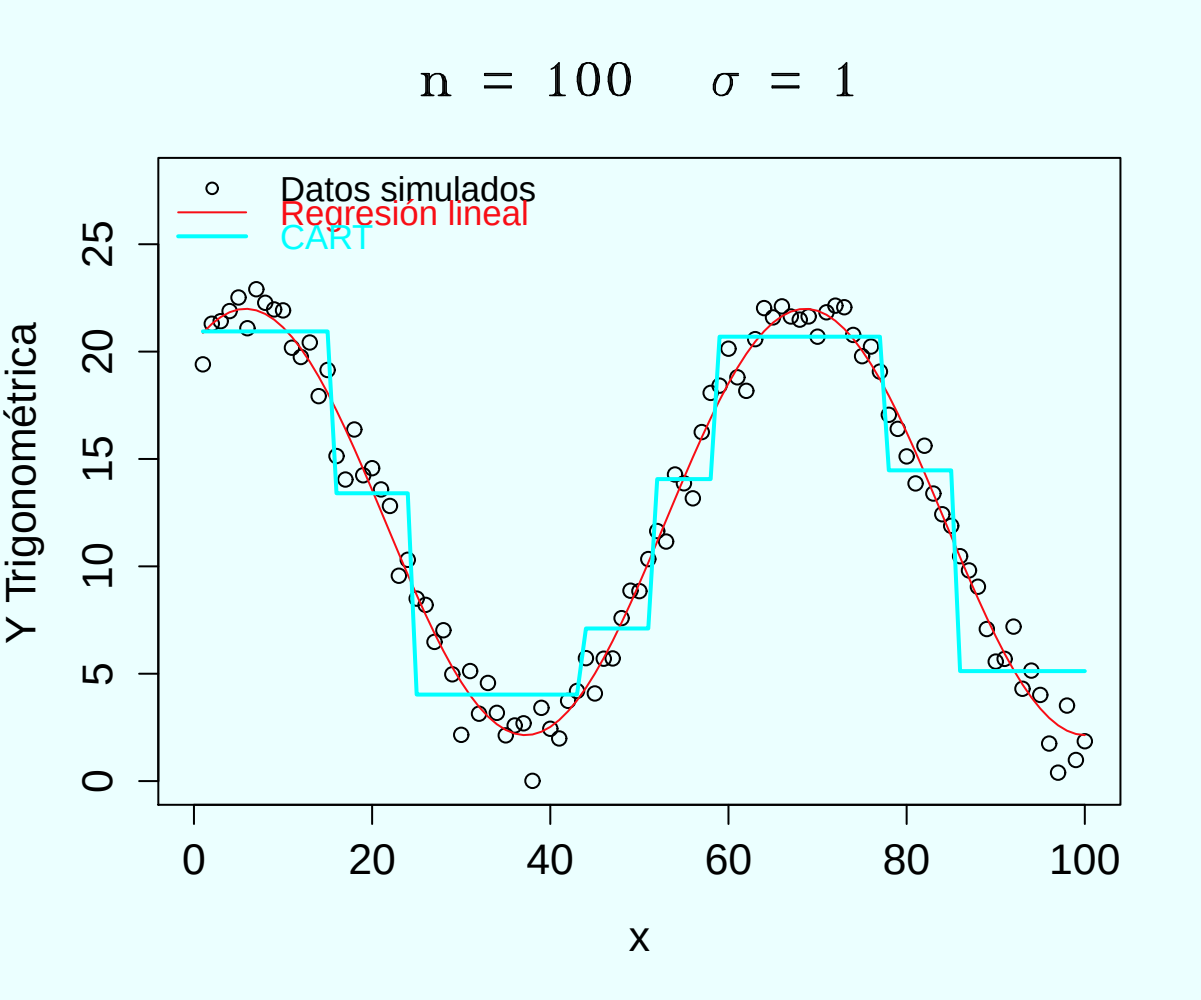

Figura 3-7.: Gráfico de las predicciones para el modelo trigonométrico 1 con $n = 100$ y $\sigma = 1$.

identificar, pero, el gráfico **3-12**, para $n = 1000$, muestra que estas predicciones si logran describir todos los máximos y mínimos relativos del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ. Es evidente que el modelo de regresión lineal describe mejor los datos que CART. Nótese como el aumento de la desviación estándar no influye en la forma de las predicciones de CART para un n en particular en este modelo.

3.2.3. Errores de predicción de CART vs Regresión Lineal para el modelo trigonométrico 3

El quinto y último modelo a analizar se obtiene de sustituir $a = 10,\ b = 1,\ c = 1,\ d = 12$ en la ecuación 3-4 y se llamará modelo trigonométrico 3.

De nuevo se observa en la tabla **3-5** que para cualquier valor de n fijo, al aumentar la desviación

Tabla 3-4.: Comparación de los errores de predicción para el modelo trigonométrico 2.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	0.1	0.0000	0.9175	236952.2511	5.3747	0.9175
	0.3	0.0000	0.9176	23872.9828	4.3779	0.9176
50	0.5	0.0001	0.9184	8866.1871	3.9477	0.9183
	0.8	0.0003	0.9222	3232.6696	3.5096	0.9219
	1	0.0005	0.9253	2024.8350	3.3064	0.9248
	2	0.0028	0.9411	330.7283	2.5195	0.9383
	0.1	0.0000	0.7380	390014.8209	5.5911	0.7380
	0.3	0.0000	0.7309	38653.3488	4.5872	0.7308
100	0.5	0.0001	0.7269	13181.7562	4.1200	0.7269
	0.8	0.0002	0.7223	4403.0087	3.6437	0.7221
	1	0.0003	0.7194	2547.0277	3.4060	0.7191
	2	0.0021	0.7158	339.8076	2.5312	0.7137
	0.1	0.0000	0.1215	301895.0195	5.4799	0.1215
	0.3	0.0000	0.1246	27763.9475	4.4435	0.1246
500	0.5	0.0000	0.1309	8078.2633	3.9073	0.1309
	0.8	0.0001	0.1377	2087.4320	3.3196	0.1376
	1	0.0001	0.1422	1048.6576	3.0206	0.1421
	2	0.0016	0.1594	101.1005	2.0048	0.1578
	0.1	0.0000	0.1180	507389.0555	5.7053	0.1180
	0.3	0.0000	0.1167	42378.3253	4.6271	0.1167
1000	0.5	0.0000	0.1239	11037.0202	4.0429	0.1239
	0.8	0.0001	0.1309	2451.5137	3.3894	0.1309
	1	0.0001	0.1359	1156.4414	3.0631	0.1357
	2	0.0015	0.1537	100.7044	2.0030	0.1522
	0.1	0.0000	0.1043	2011453.3424	6.3035	0.1043
	0.3	0.0000	0.1085	90440.3748	4.9564	0.1085
5000	0.5	0.0000	0.1127	15483.1287	4.1899	0.1127
	0.8	0.0000	0.1216	2793.2496	3.4461	0.1216
	1	0.0001	0.1270	1237.3779	3.0925	0.1269
	2	0.0015	0.1520	103.8684	2.0165	0.1506

estándar σ de los errores de los datos el error de predicción de la regresión lineal se aproxima al error de predicción de CART, siendo en todos los casos menor el error de predicción de la regresión lineal.

En los gráficos **3-13**, **3-14**, con $n = 100$, se ve que las predicciones de CART no describen la forma verdadera de los datos ya que no logran identificar ningún máximo ni mínimo relativo del verdadero modelo, pero, en el gráfico **3-15**, para $n = 1000$, se puede ver como estas predicciones si logran describir todos los máximos y mínimos relativos del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ. 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. Es claro que el modelo de regresión lineal describe mejor los datos que CART. Nótese como el aumento de la desviación estándar no influye en la forma de las predicciones de CART para un n en particular en este modelo.

Tabla 3-5.: Comparación de los errores de predicción para el modelo trigonométrico 3.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	0.1	0.0000	0.9800	251680.9225	5.4009	0.9800
	0.3	0.0000	0.9800	29737.9099	4.4733	0.9800
50	0.5	0.0001	0.9800	9494.2984	3.9775	0.9799
	0.8	0.0003	0.9799	3653.2469	3.5627	0.9796
	1	0.0005	0.9798	2140.4358	3.3305	0.9794
	2	0.0028	0.9809	355.7183	2.5511	0.9781
	0.1	0.0000	0.9900	488980.1428	5.6893	0.9900
	0.3	0.0000	0.9900	51391.8963	4.7109	0.9900
100	0.5	0.0001	0.9900	16779.3161	4.2248	0.9899
	0.8	0.0002	0.9894	6467.5834	3.8107	0.9893
	1	0.0003	0.9870	3541.4006	3.5492	0.9867
	2	0.0021	0.9713	465.4537	2.6679	0.9692
	0.1	0.0000	0.2872	663703.3186	5.8220	0.2872
	0.3	0.0000	0.2857	66182.7558	4.8207	0.2857
500	0.5	0.0000	0.2867	18004.8571	4.2554	0.2867
	0.8	0.0001	0.2898	4470.7214	3.6504	0.2897
	1	0.0001	0.2917	2126.2575	3.3276	0.2916
	2	0.0016	0.3063	193.3307	2.2863	0.3047
	0.1	0.0000	0.2841	1364043.3433	6.1348	0.2841
	0.3	0.0000	0.2841	106227.3501	5.0262	0.2841
1000	0.5	0.0000	0.2848	24822.4726	4.3948	0.2848
	0.8	0.0001	0.2852	5368.5978	3.7299	0.2852
	1	0.0001	0.2864	2451.6576	3.3895	0.2863
	2	0.0015	0.2901	190.9615	2.2809	0.2886
	0.1	0.0000	0.2821	5371231.6251	6.7301	0.2821
	0.3	0.0000	0.2824	236465.7511	5.3738	0.2824
5000	0.5	0.0000	0.2828	38896.0696	4.5899	0.2828
	0.8	0.0000	0.2835	6522.0896	3.8144	0.2834
	1	0.0001	0.2840	2782.6053	3.4445	0.2839
	2	0.0015	0.2874	196.4706	2.2933	0.2859

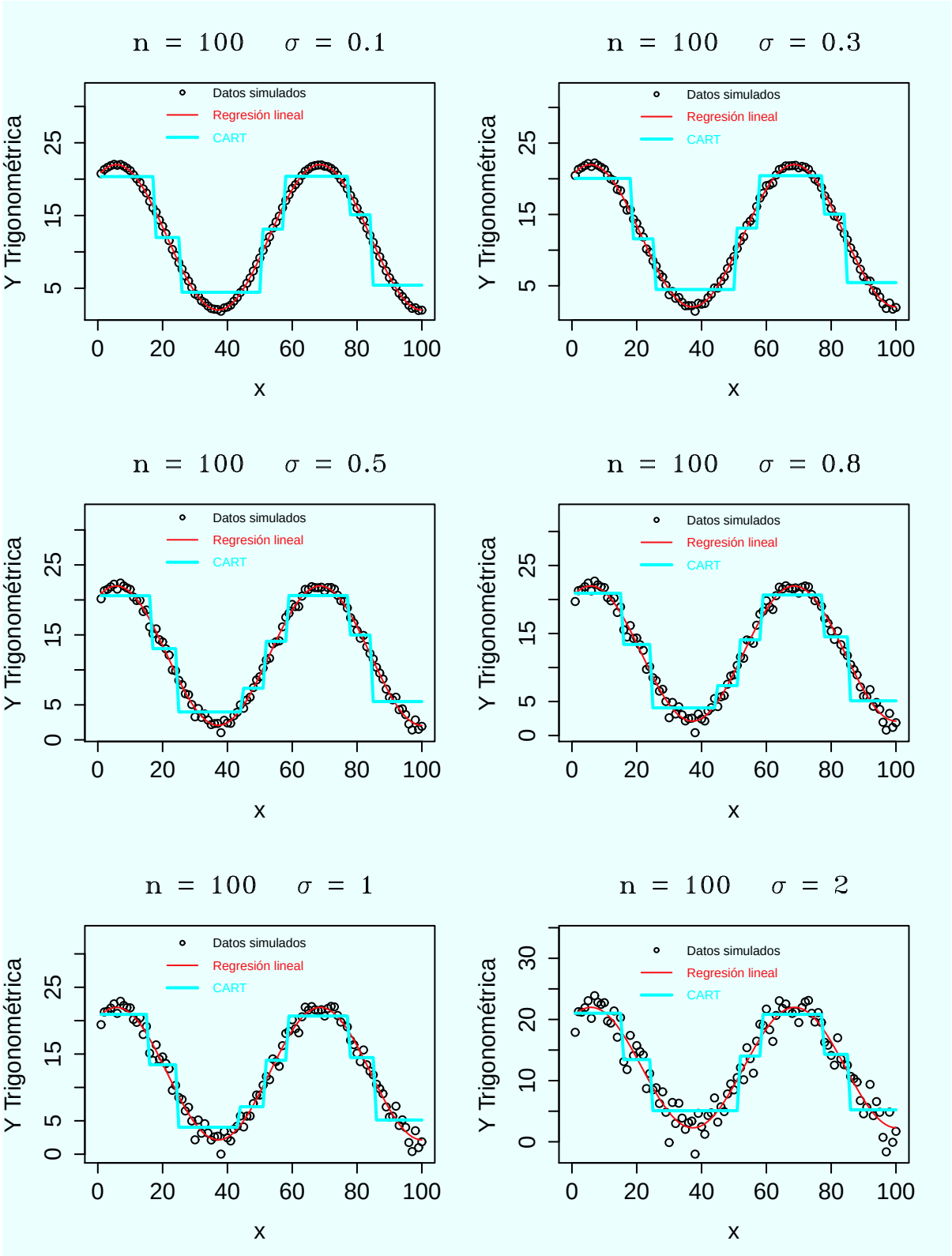

Figura 3-8.: Gráfico de las predicciones para el modelo trigonométrico 1 con $n = 100$.

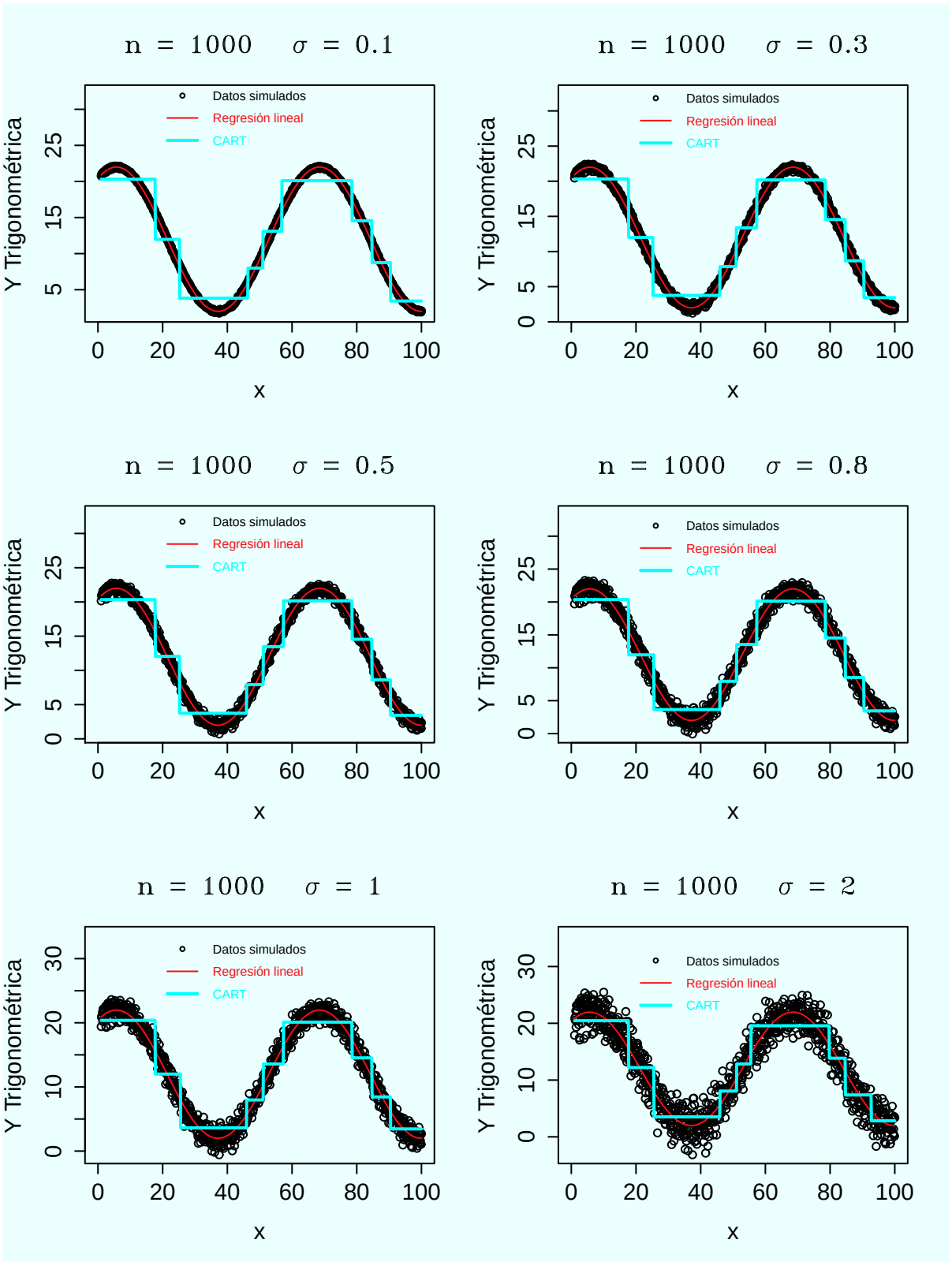

Figura 3-9.: Gráfico de las predicciones para el modelo trigonométrico 1 con $n = 1000$.

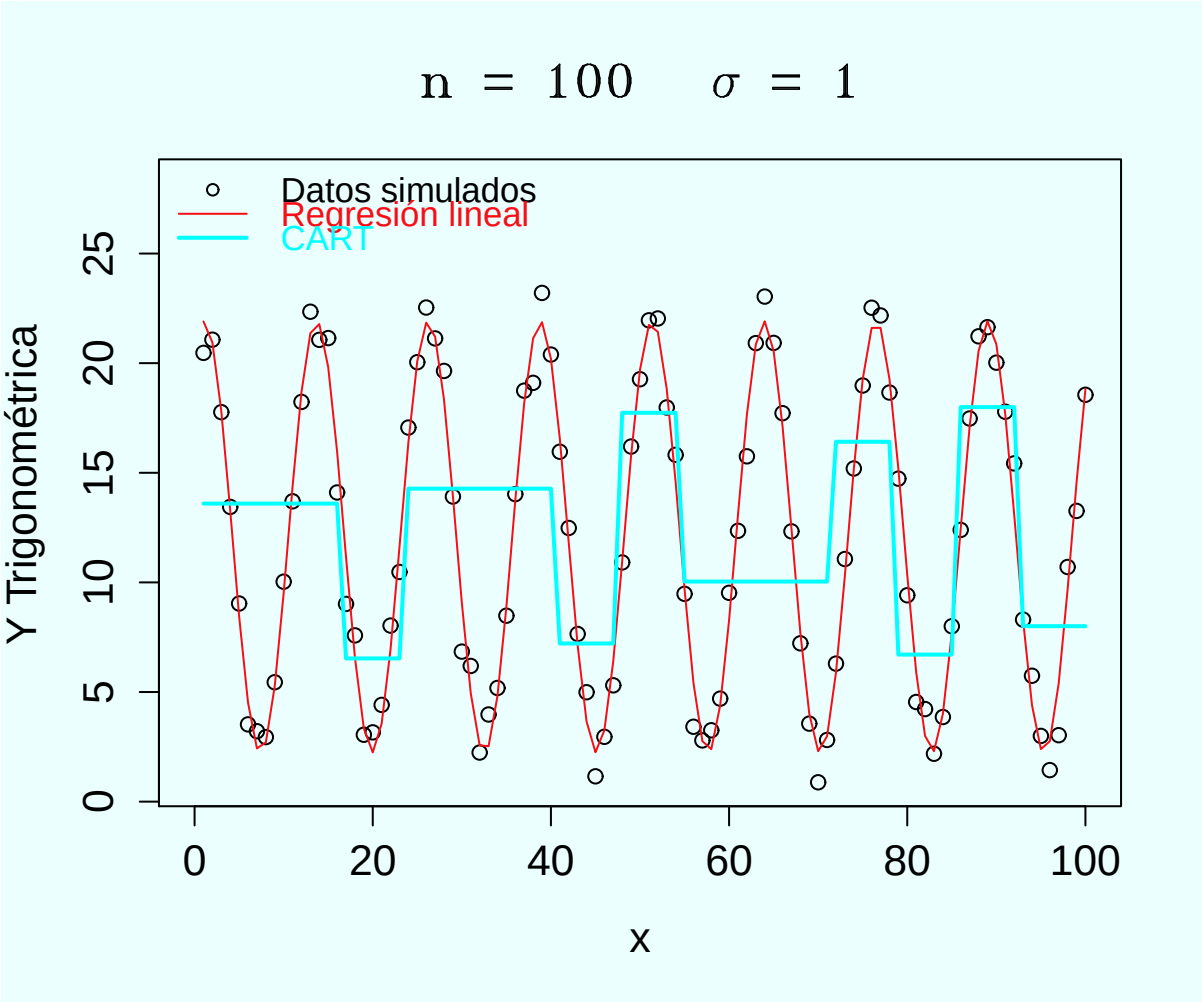

Figura 3-10.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 100$ y $\sigma = 1$.

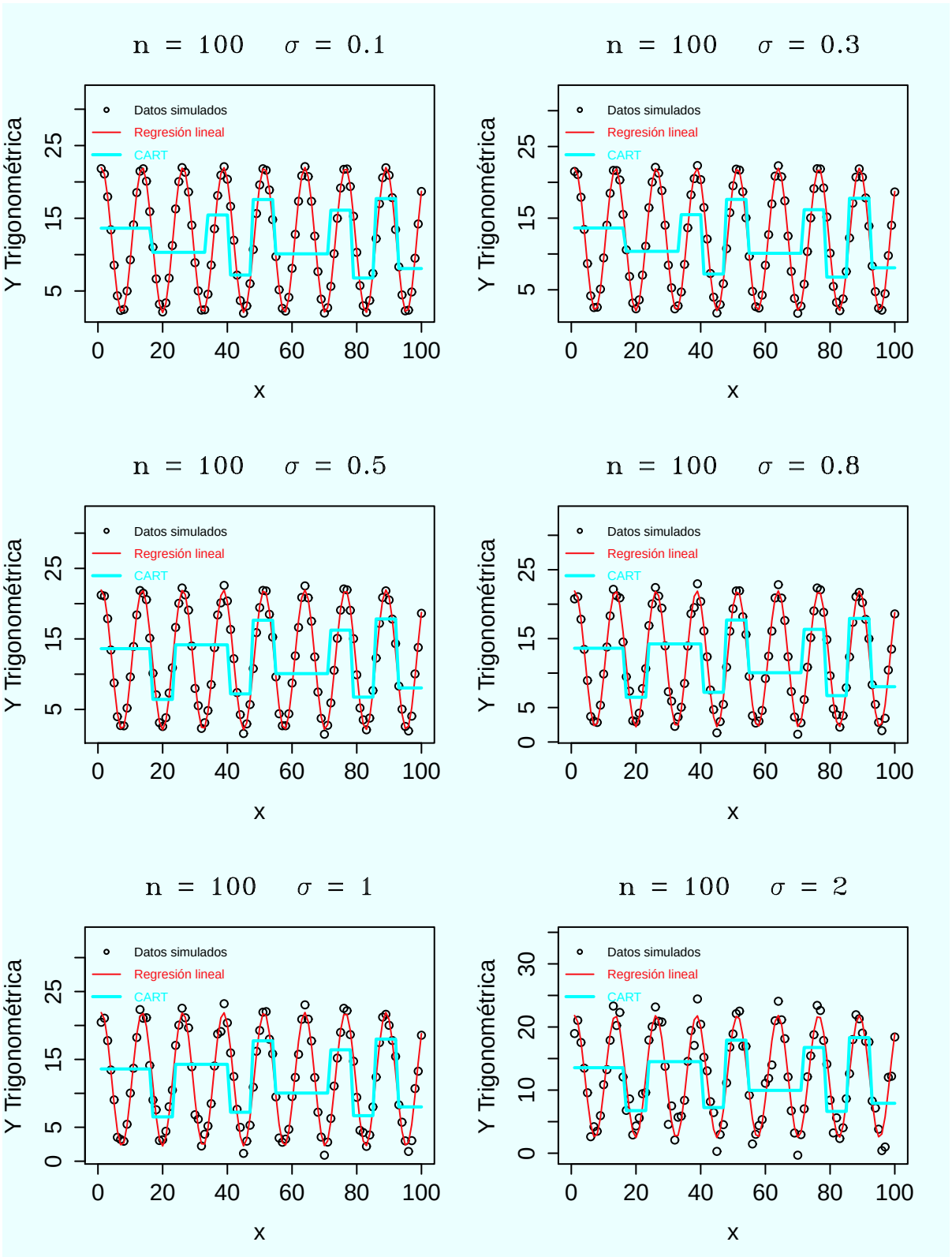

Figura 3-11.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 100$.

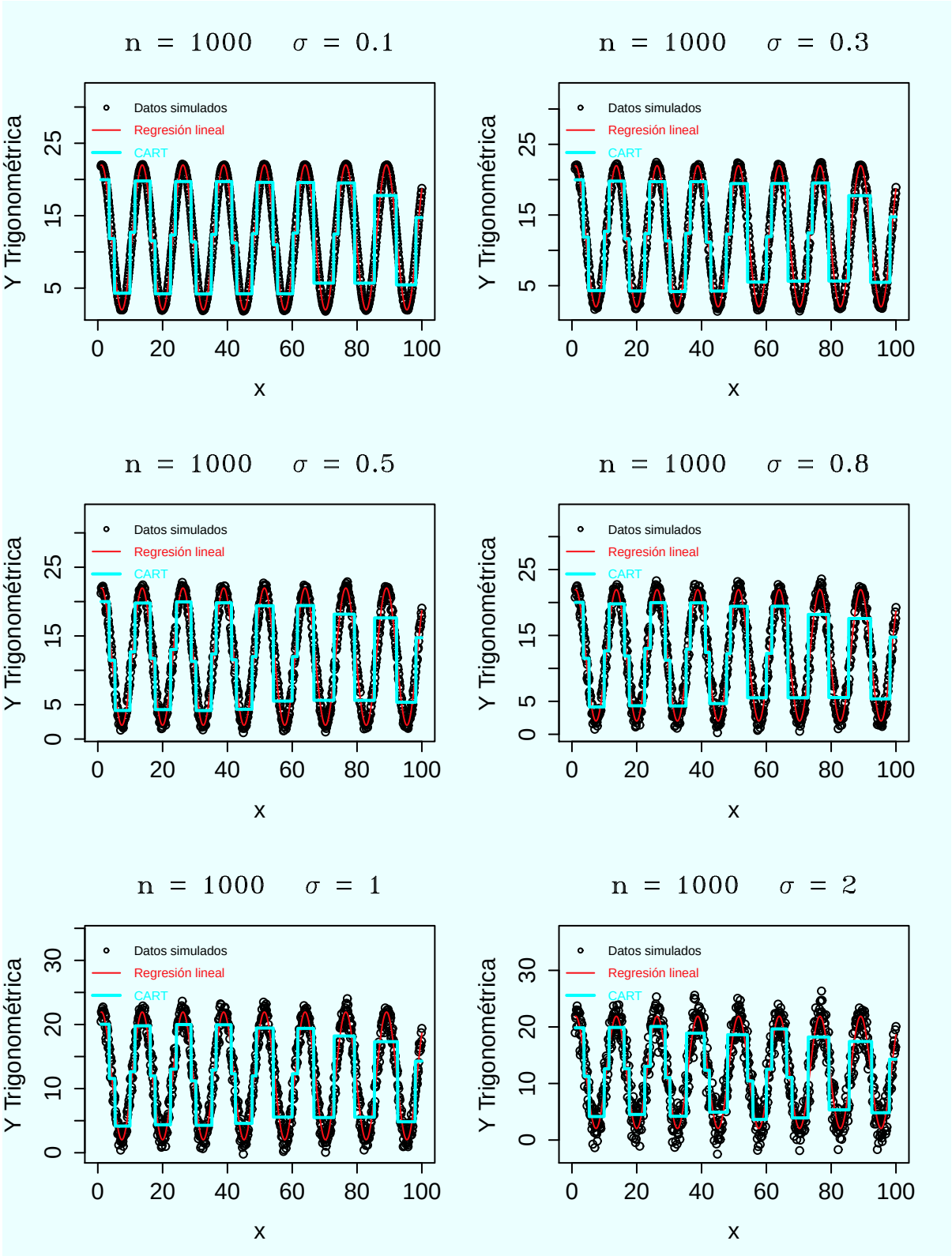

Figura 3-12.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 1000$.

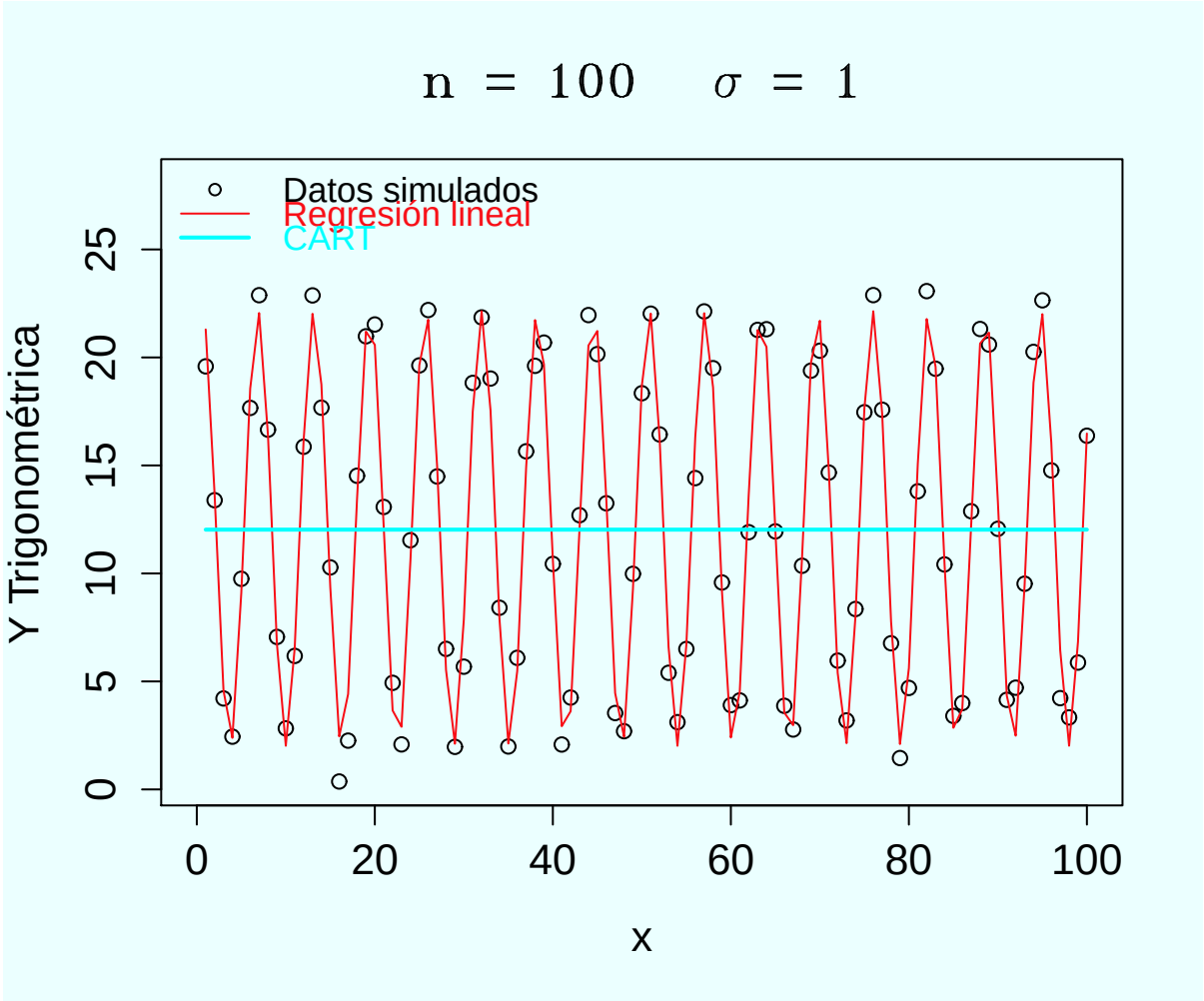

Figura 3-13.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 100$ y $\sigma = 1$.

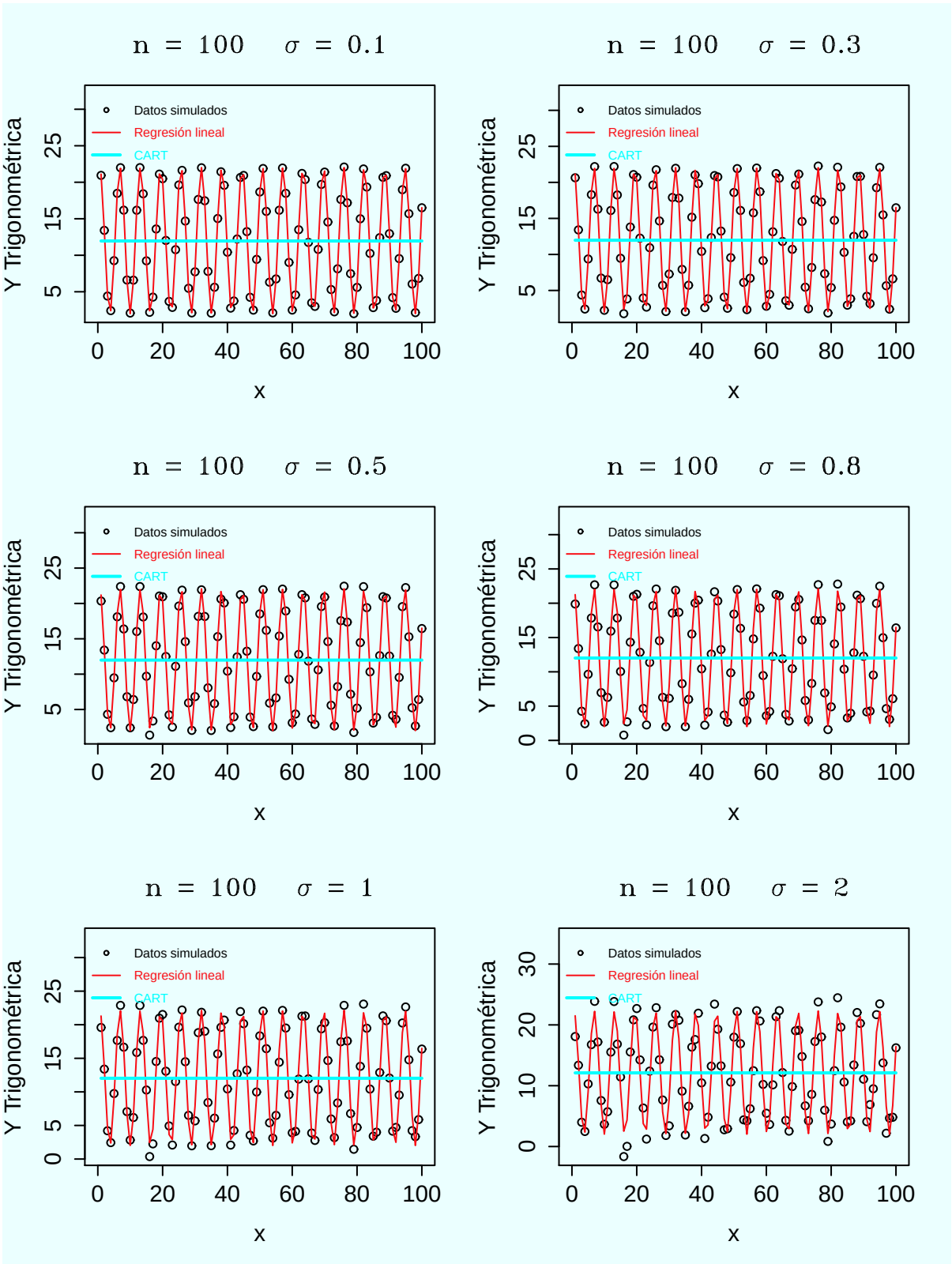

Figura 3-14.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 100$.

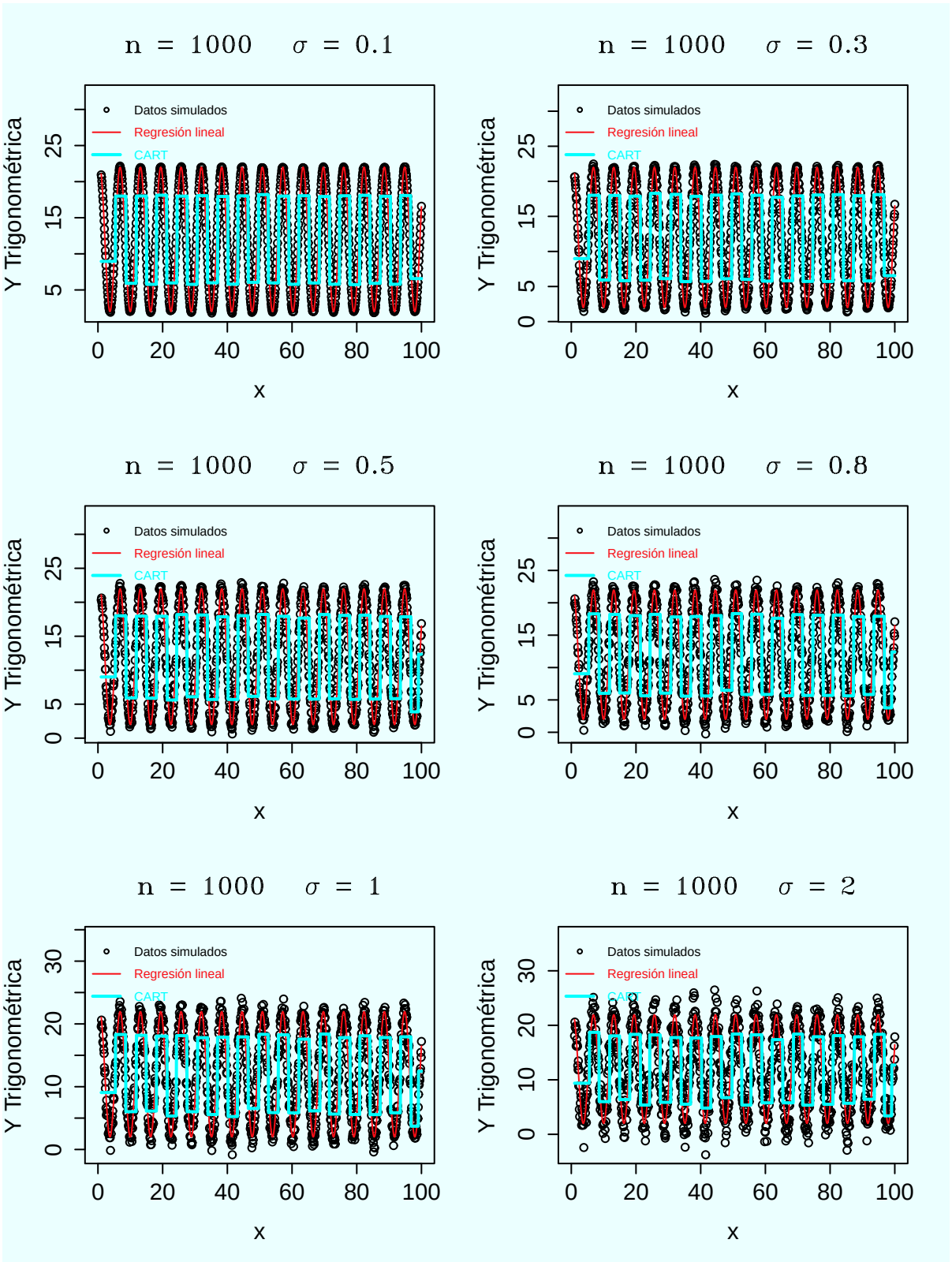

Figura 3-15.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 1000$.

4. Comparación de las predicciones de CART y modelos de regresión lineal ajustados incorrectamente

A continuación se tomarán tres modelos de regresión lineal de los descritos en el capítulo 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 debido a que 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.

4.1. Predicción de un modelo de regresión cuadrático utilizando una recta de regresión y CART

En esta sección se ajustan rectas de regresión a conjuntos de datos cuyo verdadero modelo de regresión es el modelo cuadrático 1. Como se dijo anteriormente, se escogió este modelo debido a que cuando $n = 50$ el estudio de simulación muestra que una recta de regresión lo predice mejor que los árboles de regresión, pero, cuando $n = 100$ o mayor, los árboles de regresión predicen mejor el modelo que la recta de regresión.

4.1.1. Errores de predicción de CART vs recta de regresión para el modelo cuadrático 1

En la tabla **4-1** se puede observar 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. Si bien no existe evidencia que el aumento de n implica un aumento en la precisión de las predicciones de CART con respecto a la recta de regresión (disminución del cociente de errores en la tabla), se puede observar globalmente que esta precisión para $n = 50$ y $n = 100$ es menor que para $n = 500$, $n = 1000$ y $n = 5000$ donde el cociente de los errores se estabiliza con una cifra decimal significativa alrededor de 0,5.
En los gráficos **4-1**, **4-2** y **4-3** se puede ver como las predicciones de CART describen la forma del verdadero modelo de los datos simulados para cualquier valor de la desviación estándar σ cuando $n = 100$ o $n = 1000$. Nótese que este modelo tiene una forma funcional suave, sin máximos

ni mínimos relativos, y que CART en todos los casos describe mejor los datos que la recta de regresión, incluso cuando se tienen desviaciones estándar grandes.

Tabla 4-1.: Comparación de los errores de predicción para el modelo cuadrático 1.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	1	0.0604	0.1388	2.2962	0.3610	0.0783
	10	0.0604	0.1376	2.2773	0.3574	0.0772
50	100	0.0604	0.1353	2.2383	0.3499	0.0748
	500	0.0604	0.1305	2.1599	0.3344	0.0701
	1000	0.0605	0.1198	1.9811	0.2969	0.0593
	2000	0.0609	0.1109	1.8195	0.2600	0.0499
	1	0.0600	0.0531	0.8856	-0.0528	-0.0069
	10	0.0600	0.0531	0.8856	-0.0528	-0.0069
100	100	0.0600	0.0442	0.7366	-0.1328	-0.0158
	500	0.0600	0.0388	0.6470	-0.1891	-0.0212
	1000	0.0600	0.0375	0.6254	-0.2038	-0.0225
	2000	0.0605	0.0389	0.6435	-0.1915	-0.0216
	1	0.0596	0.0318	0.5344	-0.2721	-0.0277
	10	0.0596	0.0318	0.5343	-0.2722	-0.0277
500	100	0.0596	0.0299	0.5020	-0.2993	-0.0297
	500	0.0596	0.0303	0.5090	-0.2933	-0.0292
	1000	0.0596	0.0307	0.5149	-0.2883	-0.0289
	2000	0.0601	0.0329	0.5481	-0.2611	-0.0272
	1	0.0595	0.0319	0.5359	-0.2709	-0.0276
	10	0.0595	0.0319	0.5354	-0.2713	-0.0276
1000	100	0.0595	0.0303	0.5095	-0.2929	-0.0292
	500	0.0595	0.0300	0.5044	-0.2972	-0.0295
	1000	0.0595	0.0310	0.5212	-0.2830	-0.0285
	2000	0.0600	0.0331	0.5510	-0.2588	-0.0270
	1	0.0595	0.0319	0.5364	-0.2705	-0.0276
	10	0.0595	0.0319	0.5364	-0.2705	-0.0276
5000	100	0.0595	0.0315	0.5296	-0.2761	-0.0280
	500	0.0595	0.0303	0.5097	-0.2927	-0.0292
	1000	0.0595	0.0312	0.5243	-0.2804	-0.0283
	2000	0.0600	0.0344	0.5736	-0.2414	-0.0256

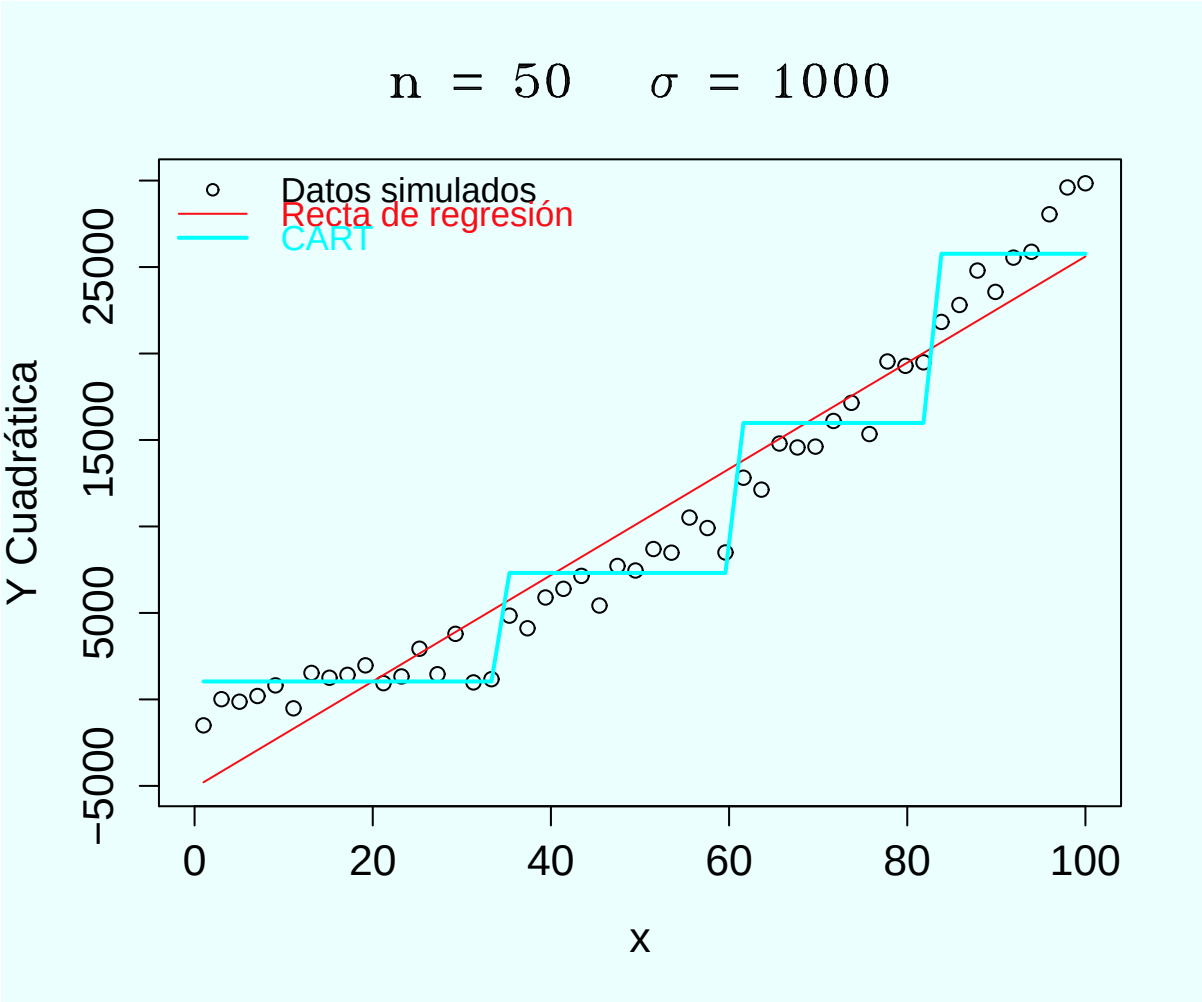

Figura 4-1.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 50$ y $\sigma = 1000$.

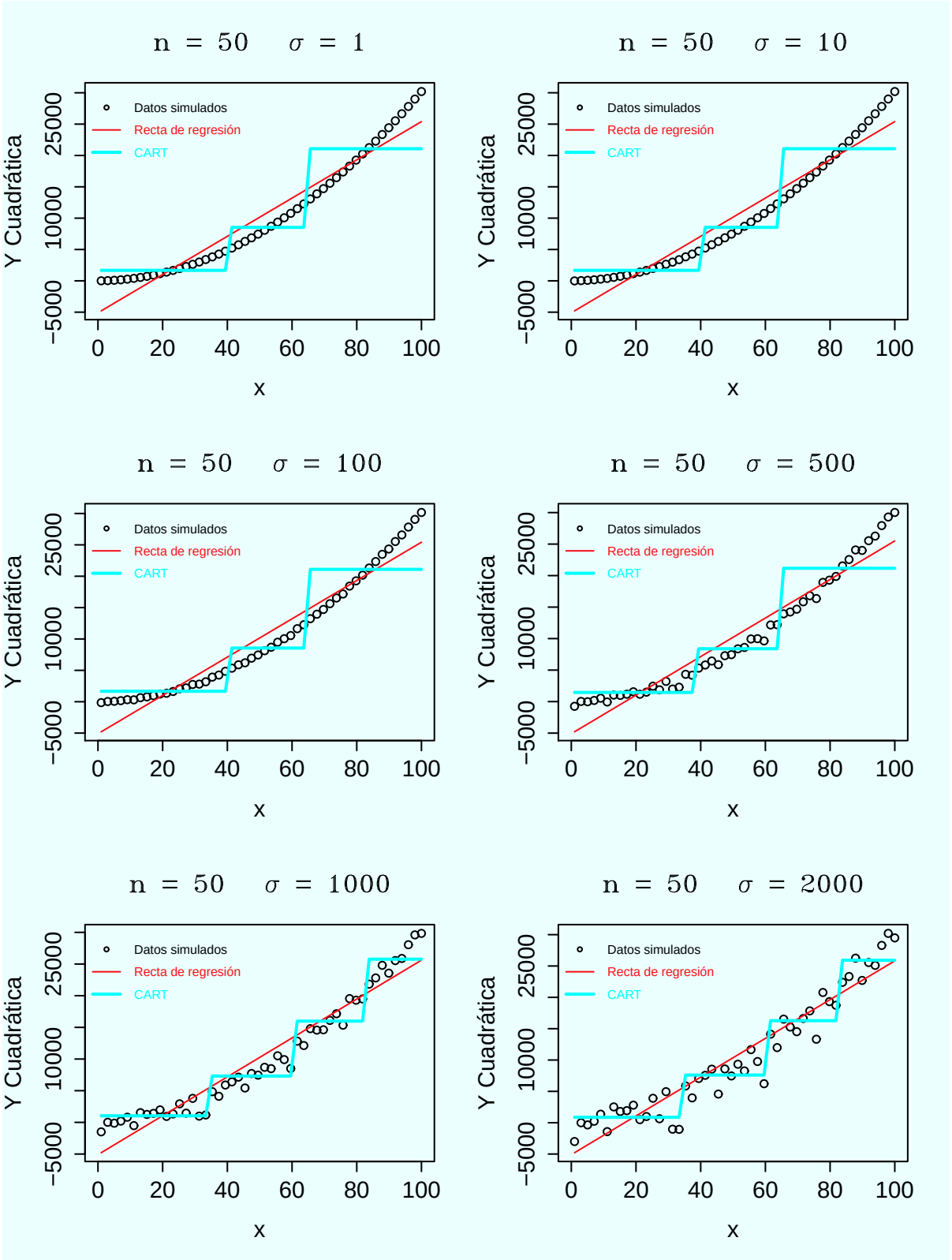

Figura 4-2.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 50$.

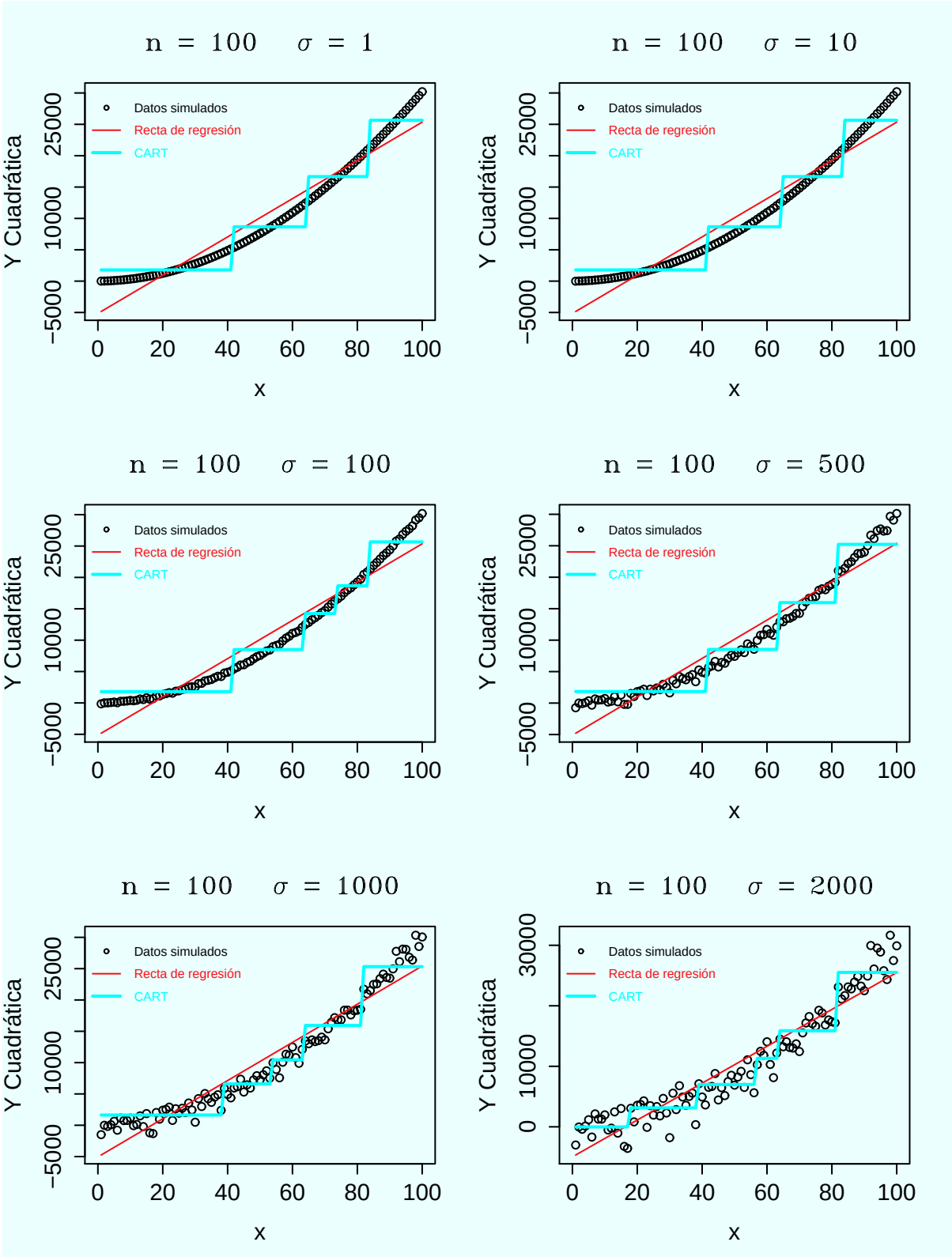

Figura 4-3.: Gráfico de las predicciones para el modelo cuadrático 1 con $n = 100$.

4.2. Predicción de un modelo de regresión trigonométrico utilizando una recta de regresión y CART

En esta sección se ajustan rectas de regresión a conjuntos de datos cuyo verdadero modelo de regresión son los modelos trigonométricos 2 y 3. Se escogió el modelo trigonométrico 2 debido a que cuando $n = 50$ el estudio de simulación muestra que el error de predicción de los árboles de regresión es cercano al de la recta de regresión, pero, cuando $n = 100$ o mayor, los árboles de regresión muestran errores de predicción mucho menores que la recta de regresión. Se escogió el modelo trigonométrico 3 debido a que cuando $n = 50$ y $n = 100$ el estudio de simulación muestra casos en que una recta de regresión lo predice mejor que los árboles de regresión, pero, cuando $n = 500$ o mayor, los árboles de regresión predicen mejor el modelo que la recta de regresión en todos los casos.

4.2.1. Errores de predicción de CART vs recta de regresión para el modelo trigonométrico 2

En la tabla **4-2** se observa 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. Si bien no existe evidencia que el aumento de n implica un aumento en la precisión de las predicciones de CART con respecto a la recta de regresión (disminución del cociente de 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 $n = 5000$ donde el cociente de los errores se estabiliza con una cifra decimal significativa alrededor de 0,1.
En los gráficos **4-4** **4-5** se puede observar como las predicciones de CART descubren patrones en los datos que pueden no notarse a simple vista. Aunque se puede decir de los gráficos **4-4**, **4-5** y **4-6** 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. En el gráfico **4-6** es más clara la verdadera forma del modelo debido a que se tiene más cantidad de datos para describirlo.

4.2.2. Errores de predicción de CART vs recta de regresión para el modelo trigonométrico 3

En la tabla **4-3** se observa 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. Si bien no existe evidencia que el aumento de n implica un aumento en la precisión de las predicciones de CART con respecto a la recta de regresión (disminución del cociente de errores en la tabla), se puede observar en este caso que esta precisión para $n = 50$ y $n = 100$ es bastante menor que para $n = 500$, $n = 1000$ y $n = 5000$.
En los gráficos **4-7** **4-8** se observa que las predicciones de CART aparentemente forman una recta, es decir, CART no es capaz de captar la verdadera forma del modelo con $n = 100$ datos, al igual

Tabla 4-2.: Comparación de los errores de predicción para el modelo trigonométrico 2.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
50	0.1	0.9781	0.9175	0.9380	-0.0278	-0.0606
	0.3	0.9782	0.9177	0.9382	-0.0277	-0.0605
	0.5	0.9782	0.9186	0.9390	-0.0273	-0.0597
	0.8	0.9784	0.9224	0.9428	-0.0256	-0.0560
	1	0.9786	0.9262	0.9465	-0.0239	-0.0523
	2	0.9795	0.9407	0.9604	-0.0175	-0.0388
100	0.1	0.9881	0.7380	0.7468	-0.1268	-0.2502
	0.3	0.9882	0.7316	0.7404	-0.1305	-0.2566
	0.5	0.9882	0.7262	0.7349	-0.1338	-0.2620
	0.8	0.9883	0.7222	0.7307	-0.1363	-0.2661
	1	0.9883	0.7195	0.7280	-0.1379	-0.2688
	2	0.9889	0.7170	0.7250	-0.1397	-0.2719
500	0.1	0.9963	0.1213	0.1218	-0.9144	-0.8749
	0.3	0.9963	0.1253	0.1257	-0.9007	-0.8710
	0.5	0.9963	0.1299	0.1304	-0.8847	-0.8664
	0.8	0.9963	0.1364	0.1369	-0.8636	-0.8599
	1	0.9963	0.1409	0.1415	-0.8492	-0.8554
	2	0.9964	0.1586	0.1592	-0.7981	-0.8378
1000	0.1	0.9973	0.1174	0.1177	-0.9292	-0.8799
	0.3	0.9973	0.1174	0.1177	-0.9292	-0.8799
	0.5	0.9973	0.1230	0.1233	-0.9090	-0.8743
	0.8	0.9973	0.1326	0.1329	-0.8765	-0.8648
	1	0.9973	0.1364	0.1367	-0.8642	-0.8609
	2	0.9974	0.1538	0.1542	-0.8119	-0.8436
5000	0.1	0.9981	0.1041	0.1043	-0.9817	-0.8940
	0.3	0.9981	0.1083	0.1085	-0.9646	-0.8898
	0.5	0.9981	0.1129	0.1131	-0.9465	-0.8852
	0.8	0.9981	0.1217	0.1220	-0.9136	-0.8764
	1	0.9981	0.1273	0.1276	-0.8941	-0.8708
	2	0.9981	0.1518	0.1521	-0.8179	-0.8463

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. Del gráfico **4-9** 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.

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.

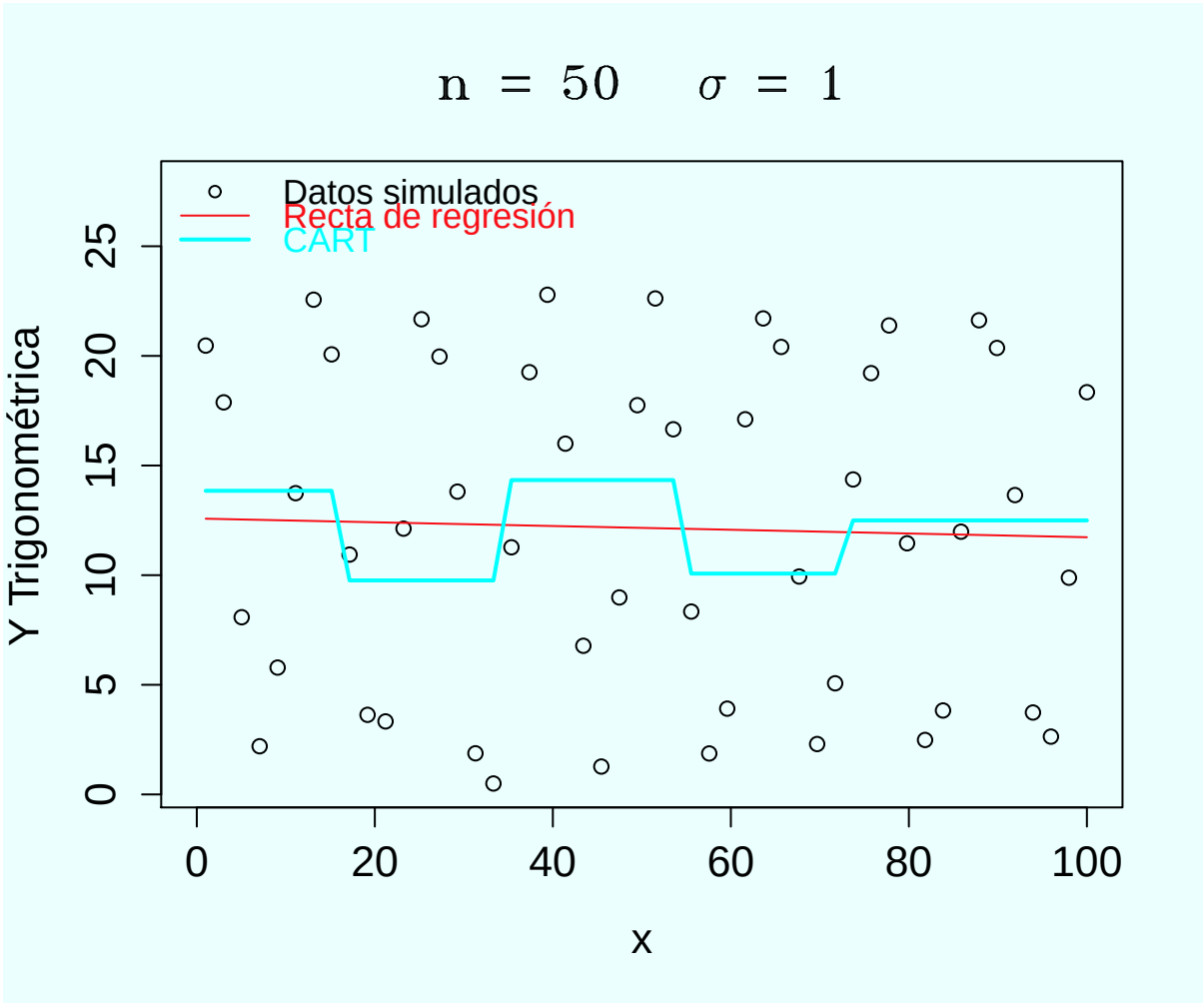

Figura 4-4.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 50$ y $\sigma = 1$.

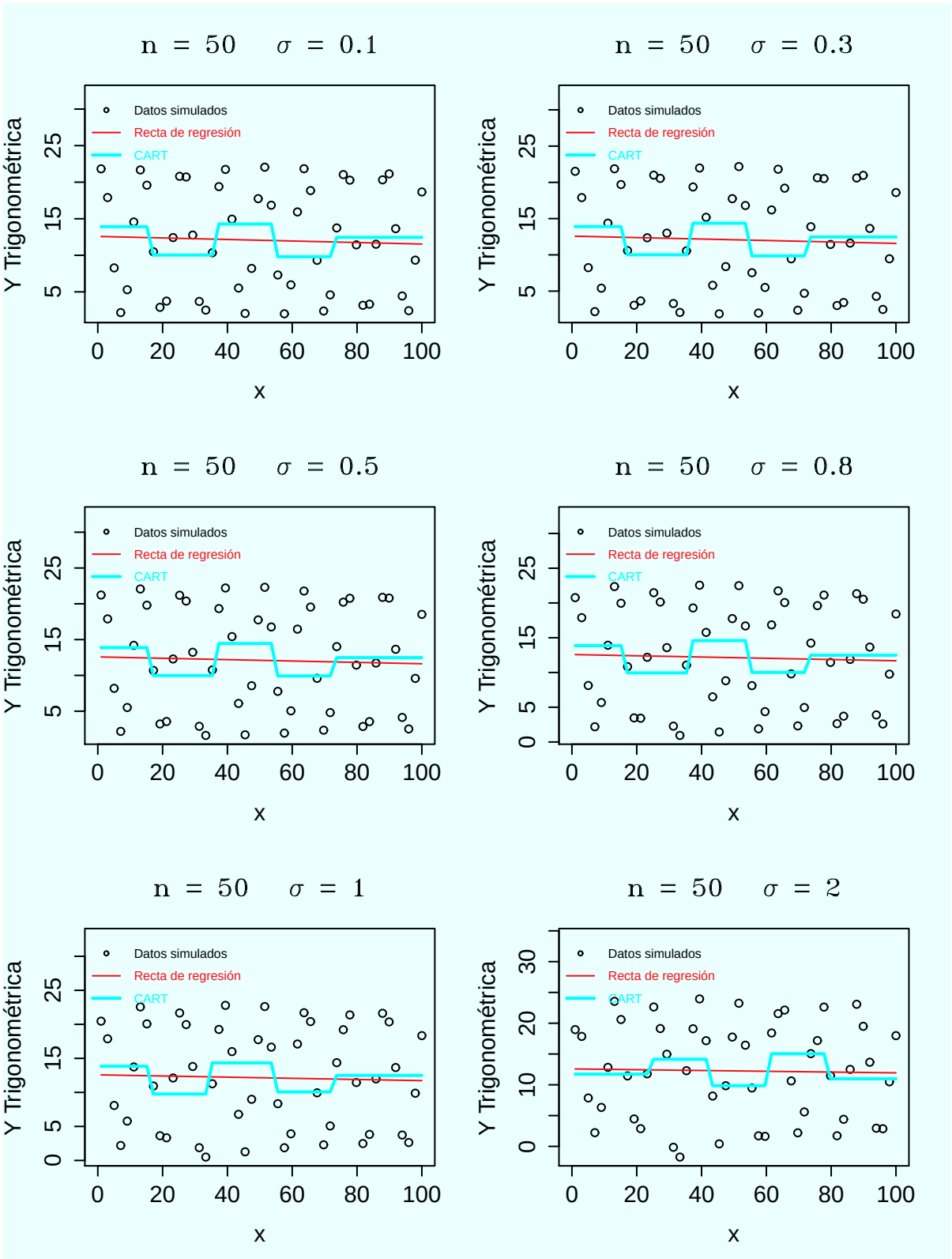

Figura 4-5.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 50$.

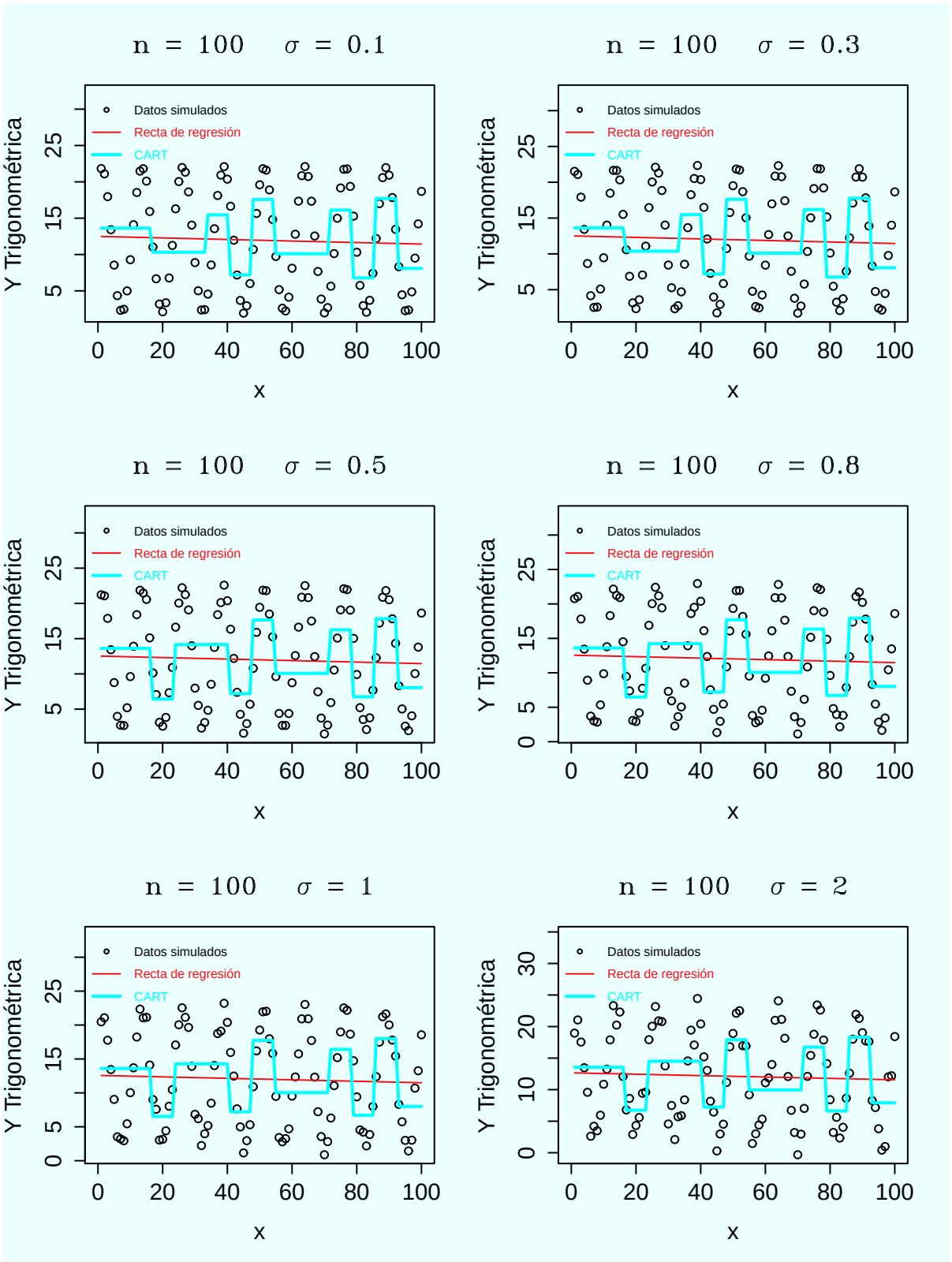

Figura 4-6.: Gráfico de las predicciones para el modelo trigonométrico 2 con $n = 100$.

Tabla 4-3.: Comparación de los errores de predicción para el modelo trigonométrico 3.

n	σ	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
	0.1	0.9797	0.9800	1.0003	0.0001	0.0003
	0.3	0.9797	0.9800	1.0003	0.0001	0.0003
50	0.5	0.9798	0.9800	1.0002	0.0001	0.0002
	0.8	0.9799	0.9800	1.0000	0.0000	0.0000
	1	0.9801	0.9798	0.9997	-0.0001	-0.0003
	2	0.9811	0.9810	0.9999	0.0000	-0.0001
	0.1	0.9897	0.9900	1.0003	0.0001	0.0003
	0.3	0.9898	0.9900	1.0003	0.0001	0.0002
100	0.5	0.9898	0.9900	1.0002	0.0001	0.0002
	0.8	0.9899	0.9891	0.9992	-0.0003	-0.0007
	1	0.9899	0.9878	0.9979	-0.0009	-0.0021
	2	0.9905	0.9712	0.9805	-0.0086	-0.0193
	0.1	0.9978	0.2872	0.2879	-0.5408	-0.7106
	0.3	0.9978	0.2858	0.2864	-0.5430	-0.7121
500	0.5	0.9978	0.2869	0.2875	-0.5414	-0.7109
	0.8	0.9979	0.2899	0.2905	-0.5369	-0.7079
	1	0.9979	0.2920	0.2927	-0.5336	-0.7058
	2	0.9980	0.3065	0.3071	-0.5127	-0.6915
	0.1	0.9988	0.2842	0.2845	-0.5459	-0.7147
	0.3	0.9988	0.2842	0.2845	-0.5459	-0.7147
1000	0.5	0.9988	0.2845	0.2848	-0.5455	-0.7144
	0.8	0.9989	0.2857	0.2860	-0.5436	-0.7132
	1	0.9989	0.2863	0.2867	-0.5426	-0.7125
	2	0.9989	0.2905	0.2908	-0.5364	-0.7084
	0.1	0.9997	0.2821	0.2822	-0.5494	-0.7175
	0.3	0.9997	0.2824	0.2825	-0.5490	-0.7173
5000	0.5	0.9997	0.2827	0.2828	-0.5485	-0.7170
	0.8	0.9997	0.2832	0.2833	-0.5478	-0.7164
	1	0.9997	0.2837	0.2838	-0.5470	-0.7159
	2	0.9997	0.2873	0.2874	-0.5415	-0.7124

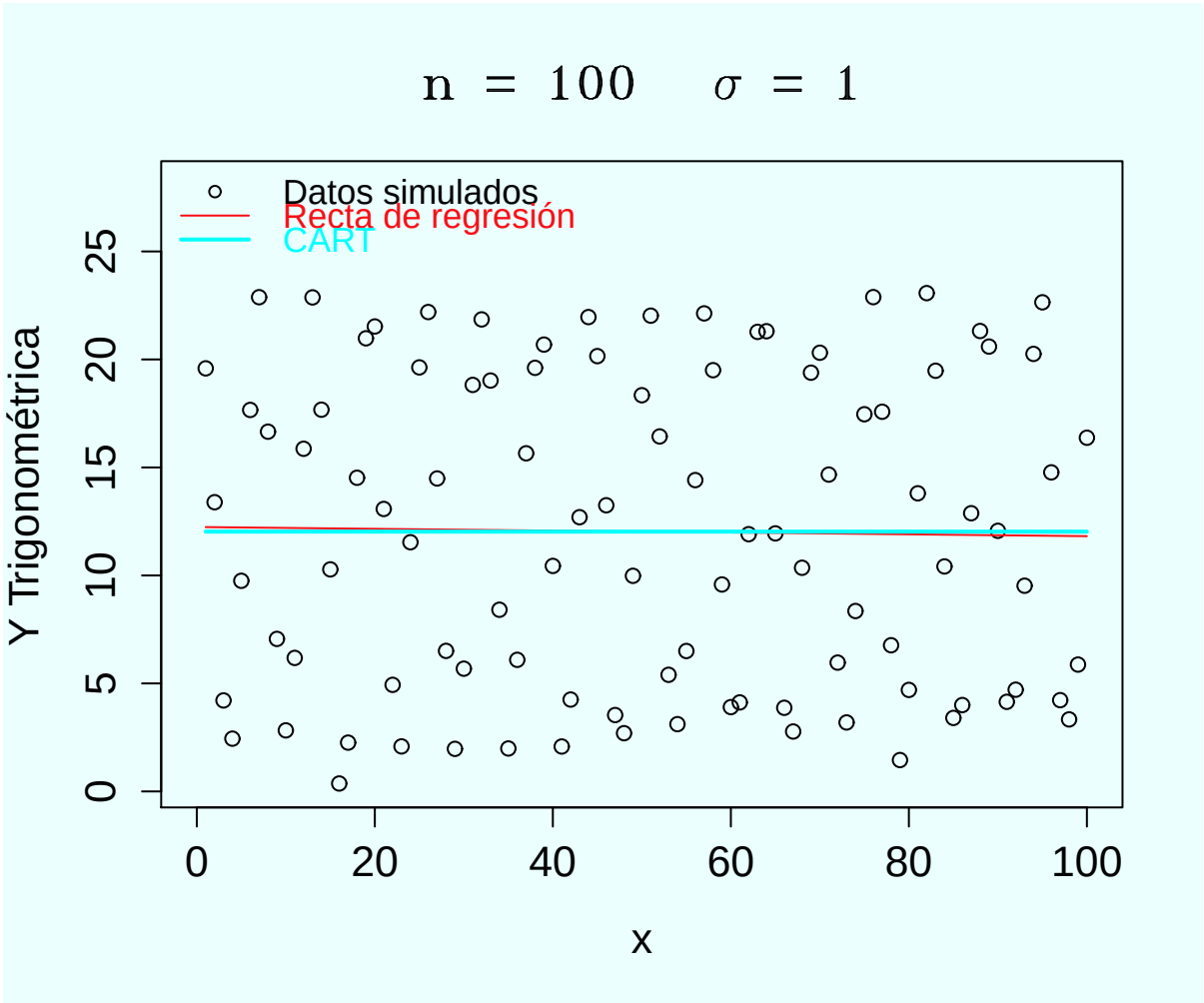

Figura 4-7.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 100$ y $\sigma = 1$.

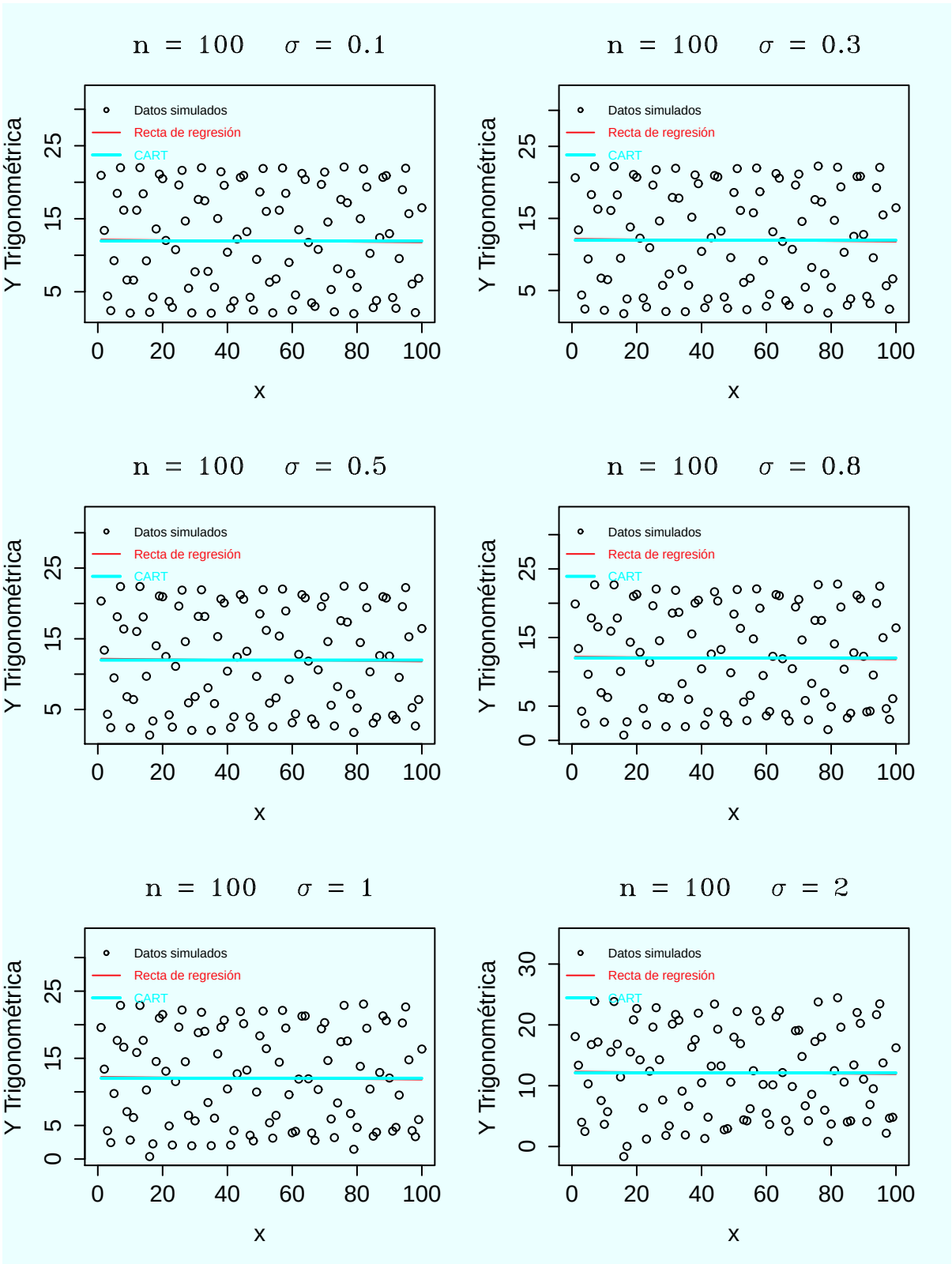

Figura 4-8.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 100$.

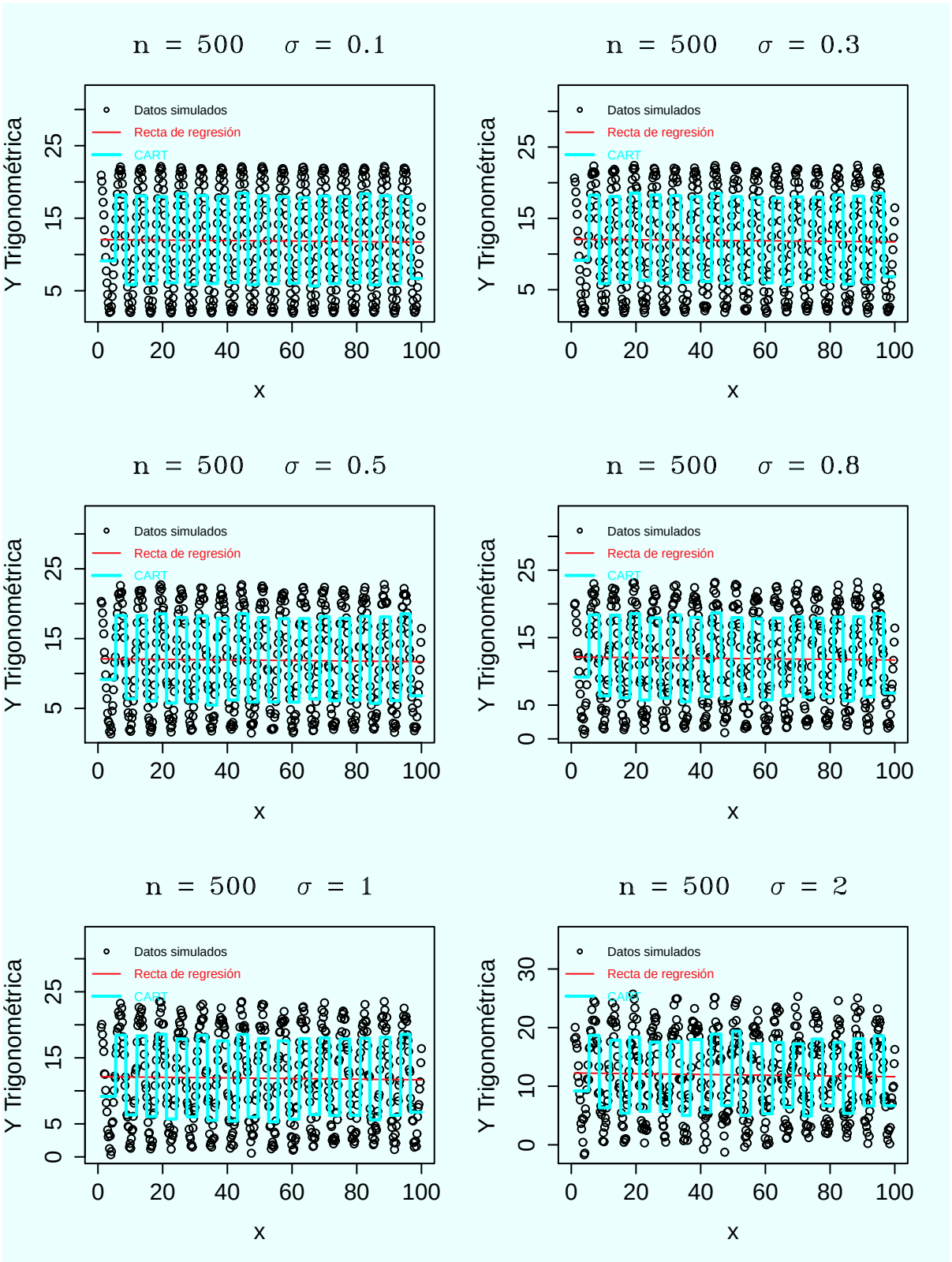

Figura 4-9.: Gráfico de las predicciones para el modelo trigonométrico 3 con $n = 500$.

5. Predicción de un modelo lineal en presencia de observaciones atípicas con CART

Los datos simulados provienen de un modelo de regresión lineal de la forma

$$y = \beta_0 + \beta_1 x + \varepsilon, \text{ donde } \varepsilon \sim N(0, \sigma^2),\ \beta_0 = 10,\ \beta_1 = 1,$$

donde para cada conjunto de n datos se reemplazan aleatoriamente un porcentaje de $100\alpha\,\%$ de los n errores por los de una distribución $N(0, 10^2)$ y con la condición de que su valor absoluto sea mayor que 3σ. Se toman valores de $\sigma = 1, \sqrt{2}$ y $\sqrt{3}$.

La regresión lineal en presencia de outliers muestra errores de predicción más pequeños que los árboles de regresión para cualquier valor de n, σ y α. Se observa en la tabla **5-1**, la cual muestra los errores de predicción para $\sigma = 1$, que para un valor fijo del tamaño muestral n, a medida que aumenta el porcentaje de outliers α en la muestra, los errores de predicción de CART y regresión lineal se van aproximando entre sí. Lo mismo se puede observar en las tablas **5-2** y **5-3** cuando $\sigma = \sqrt{2}$ y $\sigma = \sqrt{3}$, respectivamente. De las tablas **5-1**, **5-2** y **5-3** se puede ver que, para valores fijos de n y α, los errores de predicción de CART y regresión lineal se van aproximando entre sí a medida que aumenta la desviación estándar σ de los datos. Esto es de esperarse ya que en la sección 3 se mostró que el aumento de la varianza en un modelo de regresión hace que el error de predicción de la regresión lineal se aproxime al error de predicción de CART.

Tabla 5-1.: Comparación de los errores de predicción para el modelo con outliers y $\sigma = 1$.

n	α	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
50	0.01	0.0000	0.0619	188501.6065	5.2753	0.0619
	0.02	0.0000	0.0626	13358.6214	4.1258	0.0626
	0.05	0.0000	0.0636	3879.0213	3.5887	0.0635
	0.1	0.0001	0.0655	1036.2753	3.0155	0.0655
100	0.01	0.0000	0.0242	11734.3554	4.0695	0.0242
	0.02	0.0000	0.0251	4831.4962	3.6841	0.0251
	0.05	0.0000	0.0276	1346.1014	3.1291	0.0276
	0.1	0.0001	0.0297	429.4206	2.6329	0.0296
500	0.01	0.0000	0.0198	10189.2473	4.0081	0.0198
	0.02	0.0000	0.0212	4501.0751	3.6533	0.0212
	0.05	0.0000	0.0236	1147.2886	3.0597	0.0236
	0.1	0.0001	0.0266	375.7712	2.5749	0.0265
1000	0.01	0.0000	0.0182	9165.6624	3.9622	0.0182
	0.02	0.0000	0.0193	4006.7962	3.6028	0.0193
	0.05	0.0000	0.0217	1042.8642	3.0182	0.0217
	0.1	0.0001	0.0246	351.1162	2.5455	0.0246
5000	0.01	0.0000	0.0160	8100.5657	3.9085	0.0160
	0.02	0.0000	0.0165	3457.3570	3.5387	0.0165
	0.05	0.0000	0.0178	867.3114	2.9382	0.0178
	0.1	0.0001	0.0201	285.7821	2.4560	0.0200

Tabla 5-2.: Comparación de los errores de predicción para el modelo con outliers y $\sigma = \sqrt{2}$.

n	α	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
50	0.01	0.0000	0.0625	48565.0742	4.6863	0.0625
	0.02	0.0000	0.0632	9027.7419	3.9556	0.0632
	0.05	0.0000	0.0643	3128.3221	3.4953	0.0643
	0.1	0.0001	0.0664	757.2668	2.8792	0.0663
100	0.01	0.0000	0.0259	6631.3039	3.8216	0.0259
	0.02	0.0000	0.0266	3145.1065	3.4976	0.0266
	0.05	0.0000	0.0283	952.8859	2.9790	0.0282
	0.1	0.0001	0.0307	336.8287	2.5274	0.0306
500	0.01	0.0000	0.0216	4909.8478	3.6911	0.0216
	0.02	0.0000	0.0227	2550.8033	3.4067	0.0227
	0.05	0.0000	0.0251	801.1935	2.9037	0.0250
	0.1	0.0001	0.0270	277.9961	2.4440	0.0269
1000	0.01	0.0000	0.0195	4431.4996	3.6466	0.0195
	0.02	0.0000	0.0206	2302.1115	3.3621	0.0206
	0.05	0.0000	0.0231	745.2245	2.8723	0.0231
	0.1	0.0001	0.0254	265.3830	2.4239	0.0253
5000	0.01	0.0000	0.0165	3755.3765	3.5747	0.0165
	0.02	0.0000	0.0169	1898.2280	3.2783	0.0169
	0.05	0.0000	0.0190	609.4933	2.7850	0.0190
	0.1	0.0001	0.0209	213.4979	2.3294	0.0208

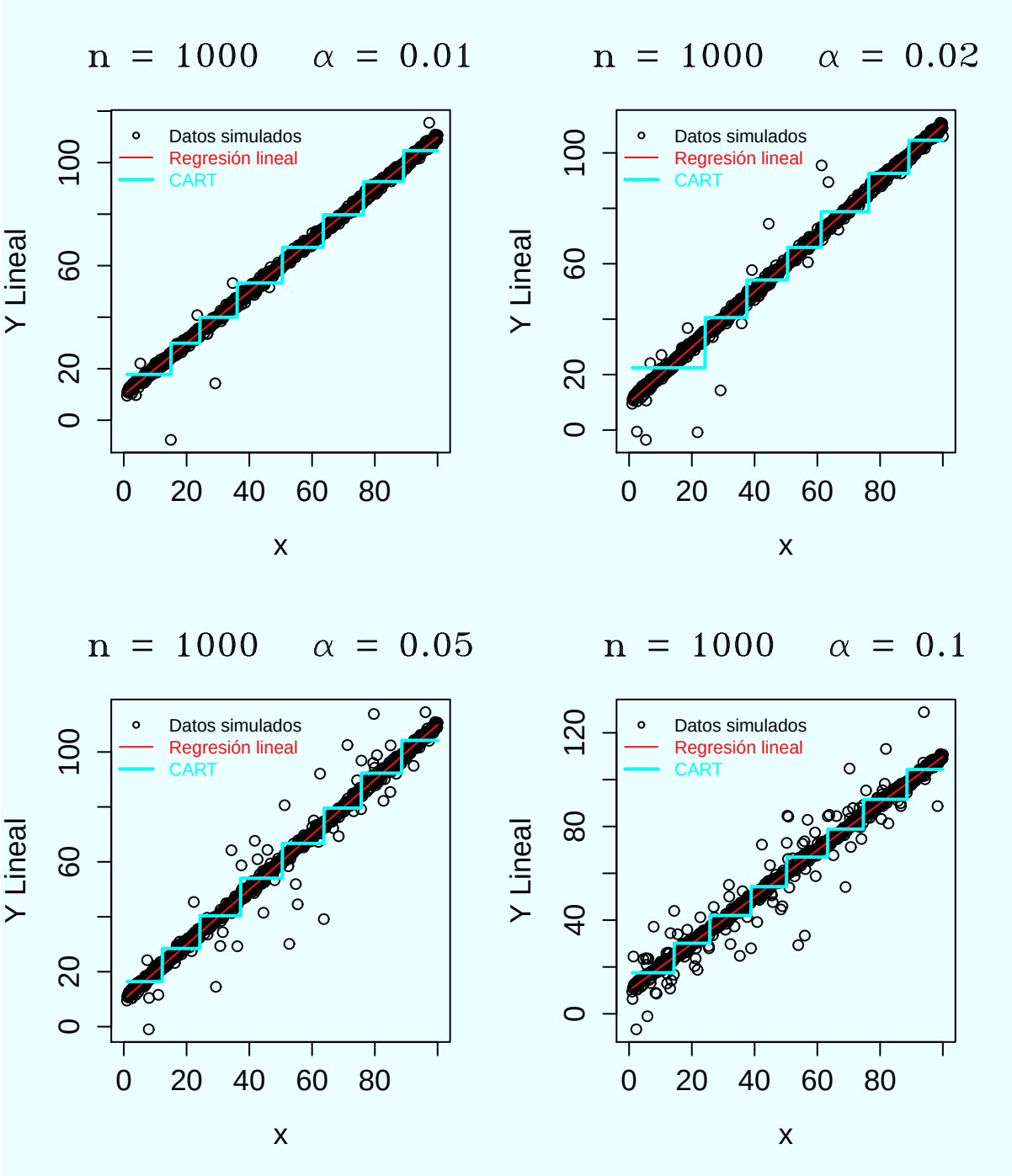

Figura 5-1.: Gráfico de las predicciones para el modelo con outliers y $\sigma = 1$.

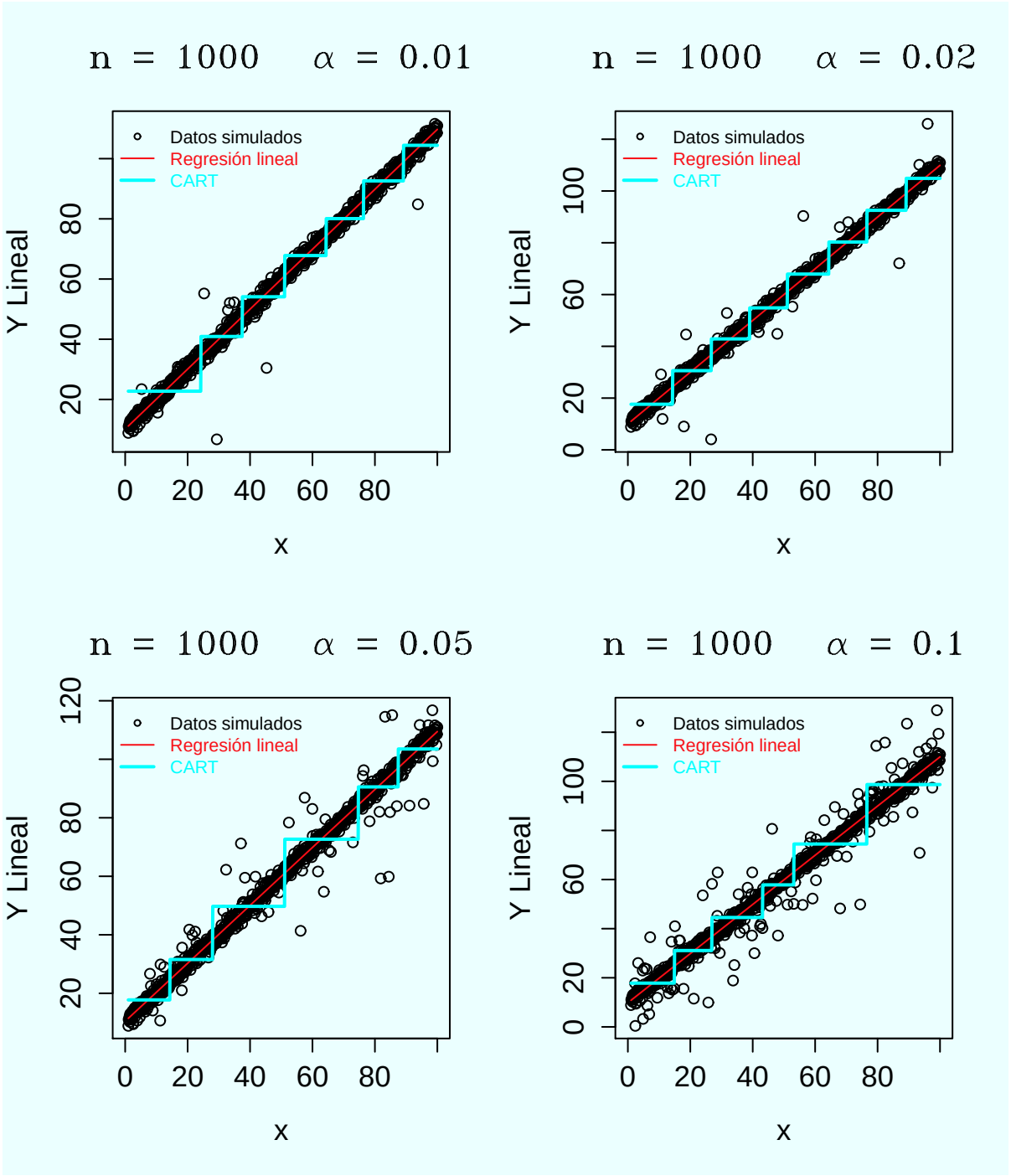

Figura 5-2.: Gráfico de las predicciones para el modelo con outliers y $\sigma = \sqrt{2}$.

Tabla 5-3.: Comparación de los errores de predicción para el modelo con outliers y $\sigma = \sqrt{3}$.

n	α	EPRL	EPCART	EPCART/EPRL	Log(EPCART)-Log(EPRL)	EPCART-EPRL
50	0.01	0.0000	0.0629	22236.6442	4.3471	0.0629
	0.02	0.0000	0.0637	6634.9284	3.8218	0.0637
	0.05	0.0000	0.0643	2503.5472	3.3986	0.0643
	0.1	0.0001	0.0666	633.7381	2.8019	0.0665
100	0.01	0.0000	0.0268	4347.1052	3.6382	0.0268
	0.02	0.0000	0.0275	2251.3956	3.3525	0.0275
	0.05	0.0000	0.0293	734.2634	2.8659	0.0293
	0.1	0.0001	0.0312	268.7805	2.4294	0.0311
500	0.01	0.0000	0.0229	3050.4694	3.4844	0.0229
	0.02	0.0000	0.0237	1733.7209	3.2390	0.0237
	0.05	0.0000	0.0256	598.6982	2.7772	0.0256
	0.1	0.0001	0.0275	220.2307	2.3429	0.0274
1000	0.01	0.0000	0.0205	2680.7361	3.4283	0.0205
	0.02	0.0000	0.0214	1565.2344	3.1946	0.0214
	0.05	0.0000	0.0237	554.6588	2.7440	0.0236
	0.1	0.0001	0.0261	207.2414	2.3165	0.0259
5000	0.01	0.0000	0.0166	2141.6769	3.3308	0.0166
	0.02	0.0000	0.0175	1254.3966	3.0984	0.0175
	0.05	0.0000	0.0192	442.7225	2.6461	0.0191
	0.1	0.0001	0.0216	170.5976	2.2320	0.0215

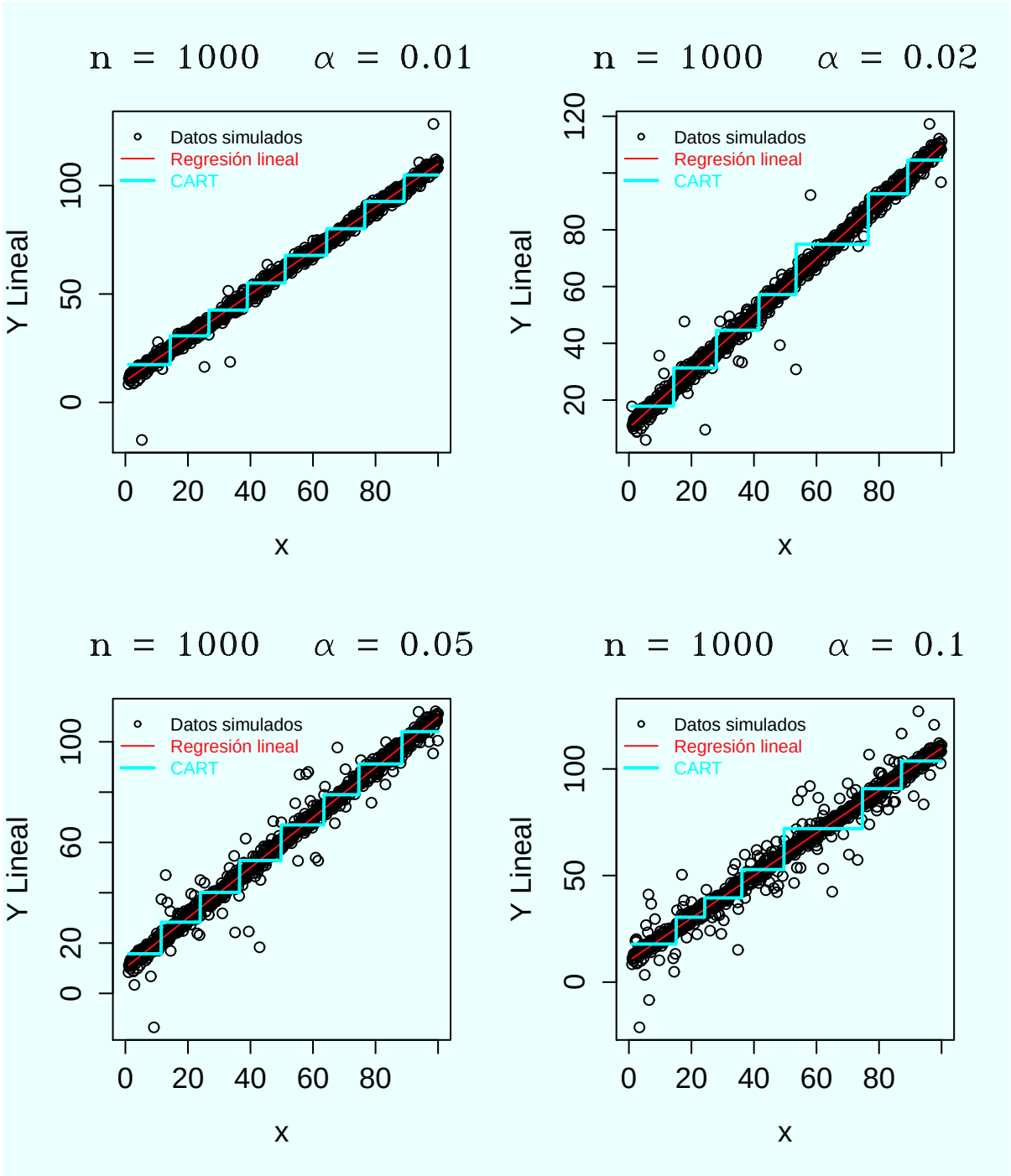

Figura 5-3.: Gráfico de las predicciones para el modelo con outliers y $\sigma = \sqrt{3}$.

6. Aplicación: Predicción de la temperatura en el aeropuerto Olaya Herrera de Medellín

En este capítulo se presenta una aplicación utilizando datos de temperaturas, la cual es una variable meteorológica que se ha mostrado en estudios que sigue un modelo de regresión trigonométrico. Para tal fin, Barrera [3] describe el método de modelización senoidal para variables meteorológicas, el cual se ilustra a continuación.

6.1. Modelización senoidal

La modelización senoidal consiste en considerar que los valores mensuales de las variables meteorológicas siguen un comportamiento senoidal del tipo:

$$y = A\sin(\omega m + \phi) + B, \tag{6-1}$$

donde A y B son dos constantes a ajustar; m, el mes del año en cuestión; y, el valor medio de una variable en el mes en cuestión; , ω la pulsación de la señal, es decir, la inversa del periodo de la señal multiplicado por 2π radianes y ϕ, el desfase. Este tipo de comportamiento solo es válido para aquellas variables que tienen un comportamiento intermensual oscilatorio o estacional, es decir, que presenten un único máximo y mínimo anual. Este tipo de comportamiento debe ser independiente de las zonas climáticas a las que pertenezcan los observatorios. El ajuste de los datos a la función 6-1 se divide en dos partes:

1. Cálculo de los parámetros ω y ϕ: Una variable con un comportamiento estacional (periódico con un único máximo y mínimo anual) tiene como periodo el de 12 meses. Con lo que:

$$\omega = \frac{2\pi}{12meses} = \frac{\pi}{6}meses^{-1}. \tag{6-2}$$

   Teniendo en cuenta como es la gráfica de la función seno y asignando en el eje de abscisas los siguientes valores para cada uno de los meses de un año hidrológico (de octubre a septiembre): 0 = octubre; 1 = noviembre; 2 = diciembre; 3 = enero; 4 = febrero; 5 = marzo; 6 = abril; 7 = mayo; 8 = junio; 9 = julio; 10 = agosto; 11 = septiembre, se tiene que el máximo de la función seno cae en x = 3. Con todo esto se tiene que el desfase entre la función que se quiere ajustar y la función seno será la diferencia de posición (en radianes) entre el máximo de la función seno y el máximo de la función que se quiere ajustar:

$$\phi = \frac{\pi(y_{max} - 3)}{6}$$

2. Ajuste de las constantes A y B: Cuando se conocen ω y ϕ en la ecuación 6-1, las constantes A y B se estiman por regresión lineal.

En su estudio, con datos de temperaturas medidas en España, Barrera concluyó que la temperatura es una de las variables meteorológicas que presentan un claro comportamiento estacional, con máximos en verano y mínimos en invierno.

6.2. Aplicación de la modelización senoidal

Para el ejemplo, se toman datos de la temperatura media por día desde octubre 1 de 2011 hasta septiembre 30 de 2012 (último año hidrológico a la fecha) en el aeropuerto Olaya Herrera de Medellín (Datos accesibles en la página web del portal TuTiempo.net:
$http://www.tutiempo.net/clima/Medellin_Olaya_Herrera/12\text{-}2011/801100.htm$).

A estos datos se les ajustan dos modelos: un árbol de regresión CART y un modelo de regresión trigonométrico.

Para el ajuste del árbol de regresión se utiliza la librería *rpart* del paquete estadístico R.

Para la modelización senoidal por día de la temperatura en este año hidrológico, se tienen 366 días (2012 año bisiesto), por tanto, el periodo es de 366 para esta variable estacional. Con lo que

$$\omega = \frac{2\pi}{366\text{días}} = \frac{\pi}{183}\text{días}^{-1}. \tag{6-3}$$

A diferencia de lo propuesto por Barrera [3] en la sección anterior, el desfase ϕ se estima por regresión lineal como se ilustró en la sección 3,2.

El ajuste de las predicciones de ambos modelos se muestra gráficamente en la figura **6-1**. Se puede observar que el modelo de regresión lineal no es un modelo apropiado para la temperatura en Medellín. Esto se debe a que los datos para la modelización senoidal aplicada por Barrera son de temperaturas en España, donde el clima es bastante regular, con máximos en verano y mínimos en invierno por las cuatro estaciones climáticas. Colombia, a diferencia de España, es un país con clima tropical, donde las condiciones climáticas pueden ser distintas en cualquier época del año. La suma de cuadrados del error para el modelo senoidal es 445.4044 y para el modelo CART es 222.6729, lo cual sugiere que el árbol de regresión está explicando mejor la temperatura diaria. En la figura **6-1** se puede ver que el ajuste de las predicciones por CART parece más apropiado para estos datos ya que tiene en cuenta el comportamiento climático de Medellín, y por tanto, puede explicar más fácil el comportamiento de la temperatura en el aeropuerto Olaya Herrera de Medellín.

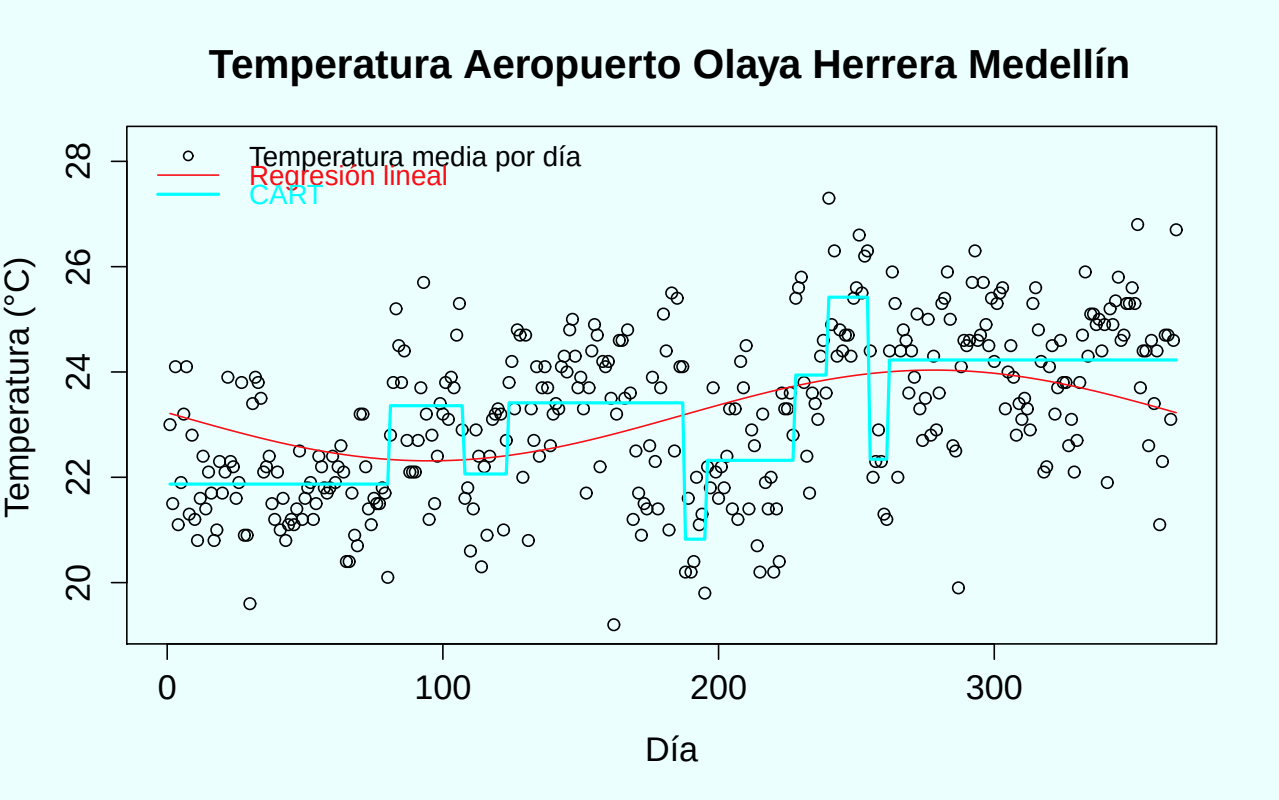

Figura 6-1.: Ajuste por modelización senoidal y por CART para la temperatura diaria.

Como el ajuste por modelización senoidal propuesto por Barrera no parece ser apropiado para los datos de la temperatura media en el aeropuerto Olaya Herrera de Medellín, se podría pensar en ajustar a estos datos un modelo de serie de tiempo. La ACF y PACF para los datos vistos como una serie de tiempo se muestran en el gráfico **6-2**.

Aunque esta ACF y PACF sugieren un modelo de series de tiempo SARIMA, se ajustará un proceso AR(2) debido a que este tipo de modelos es bueno para describir la periodicidad de muchos fenómenos (Giraldo N., comunicación personal). En el gráfico **6-3** se muestra el proceso AR(2) y el árbol de regresión CART ajustados. Aunque se puede observar que el proceso AR(2) describe mejor los datos que el modelo senoidal, la suma de cuadrados del error para este modelo es 281.0147 y para el modelo CART es 222.6729, lo cual sugiere que el árbol de regresión sigue explicando mejor la temperatura diaria en el aeropuerto Olaya Herrera de Medellín.

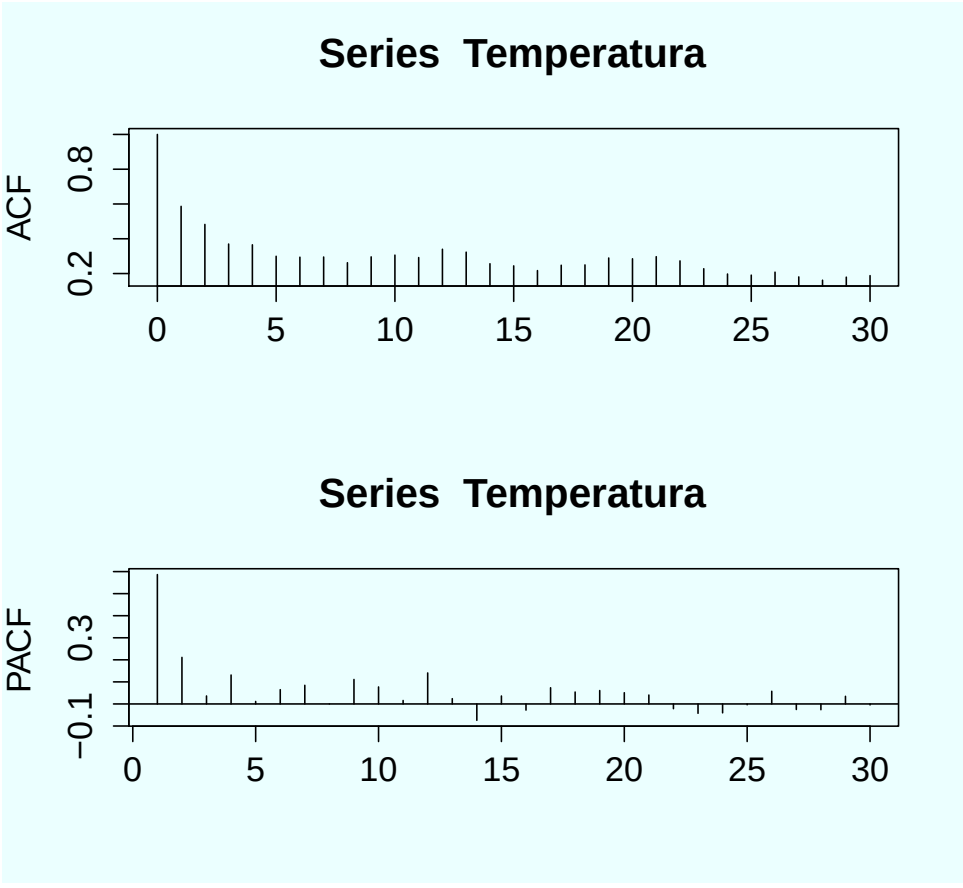

Figura 6-2.: ACF y PACF para la temperatura diaria.

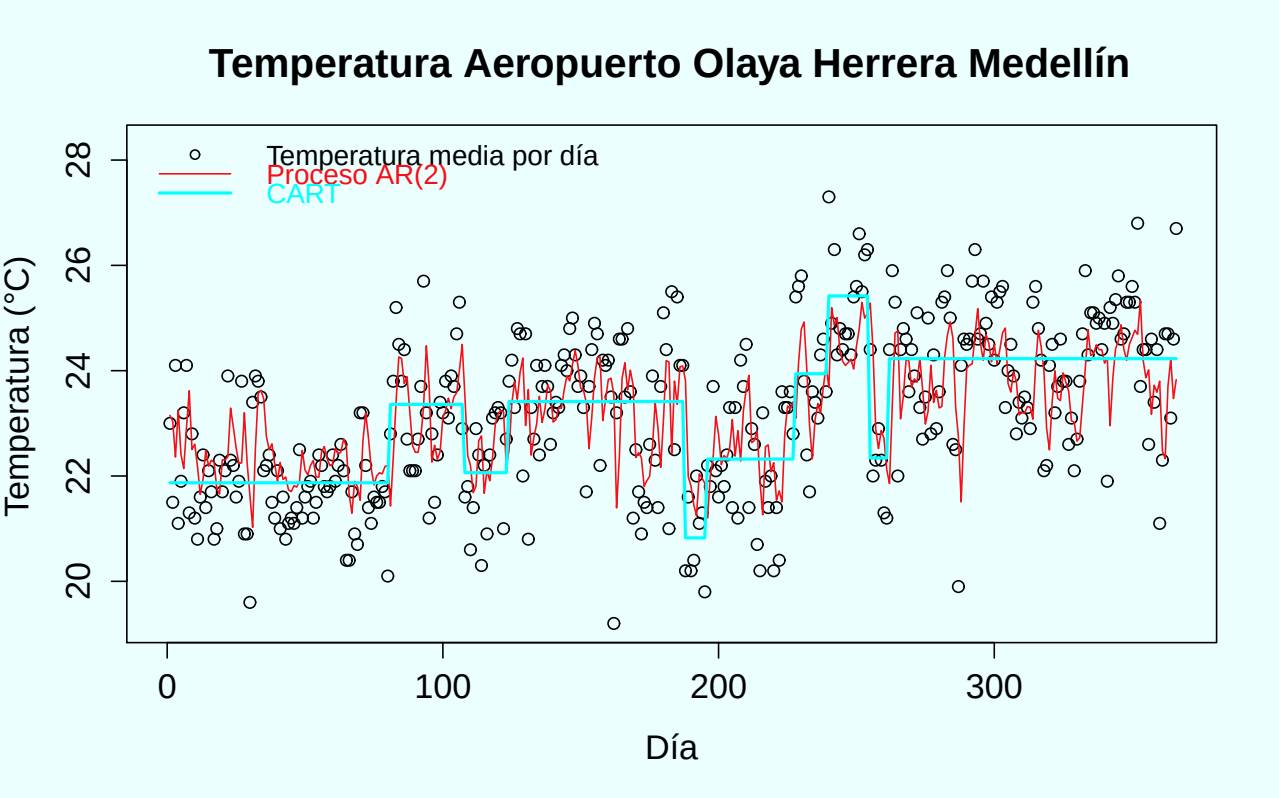

Figura 6-3.: Ajuste por series de tiempo y por CART para la temperatura diaria.

7. Conclusiones y recomendaciones

7.1. 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 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.
Cuando se comparan las predicciones de los árboles de regresión y las de regresión lineal al predecir la respuesta de un modelo lineal contaminado por observaciones atípicas, 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 o el porcentaje de observaciones atípicas 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.

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. Si el usuario está seguro de cuál es la forma funcional de su modelo, entonces CART no es una opción viable.

7.2. Recomendaciones

Cuando no se conoce la forma funcional del verdadero modelo se recomienda utilizar CART.
Como una primera etapa en la parte exploratoria en modelación se recomienda considerar un modelo CART.

A. Programa R

```
## PREDICCIONES CART VS REGRESIÓN LINEAL PARA EL MODELO CUADRÁTICO 1
library(MASS)
library(rpart)
simula.y.compara <- function(n, desvest, y.verd){
t <- sapply(1:1000, function(x){
# Valores de x
x <- seq(1, 100, length.out = n)
# Modelo verdadero estandarizado
y.verd.est <- (y.verd - mean(y.verd))/sd(y.verd)
# Modelo estadístico
e <- rnorm(n, mean = 0, sd = desvest)
y <- y.verd + e
y.est <- (y - mean(y))/sd(y)
# Predichos de las vbles estandarizadas
predicho.cuadrat.est <- predict(lm(y.est ~ x + I(x^2)))
predicho.cart.est <- predict(rpart(y.est ~ x))
s1 <- sum((predicho.cuadrat.est - y.verd.est)^2)/length(x)
s2 <- sum((predicho.cart.est - y.verd.est)^2)/length(x)
matrix(c(s1, s2), nrow = 2, byrow = T)
})
m <- apply(t, 1, mean)
# Errores con variables estandarizadas
EPRL <- m[1]
EPCART <- m[2]
coc.EP <- EPCART/EPRL
dif.EP <- EPCART - EPRL
errores <- c(EPRL, EPCART, coc.EP, dif.EP)
names(errores) <- c('EPRL', 'EPCART', 'EPCART/EPRL', 'EPCART - EPRL')
return(errores)
}
## Función n fijo sigma variable
n.fijo.sigma.vble <- function(DesVest, n){
RES <- NULL
for(i in DesvEst){
RES <- rbind(RES, c(i, simula.y.compara(n = n, desvest = i, y.verd = funcion(n))))
```

```
}
colnames(RES) <-c('DesvEst','EPRL','EPCART','EPCART/EPRL','EPCART-EPRL')
rownames(RES) <-NULL
print(list(n = n, Tabla = RES))
#Para exportar a LaTeX
library(xtable)
print(xtable(RES, digits = 4, display = c('d','d','f','f','f','f')))
}
## función para evaluar los y verdaderos
funcion <-function(n){
# Valores de x
x <-seq(1, 100, length.out = n)
# Modelo verdadero
y.verd <-1 + 2 * x + 3 * x^2
# Devolviendo el vector de valores calculados con la función
return(y.verd)
}
# Ejecucion
DesvEst <-c(1, 10, 100, 500, 1000, 2000)
n <-c(50, 100, 500, 1000, 5000)
for(i in n){
n.fijo.sigma.vble(DesvEst, n = i)
}
```

Nota: Para simular los valores predichos de los otros modelos se utiliza el mismo programa cambiando la función para evaluar los y verdaderos, $y.verd$.

Bibliografía

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

[2] Balac, N. ; Gaines, D.M. ; Fisher, D.: Using Regression Trees to Learn Action Models. En: *IEEE Systems, Man and Cybernetics Conference*, 2000

[3] Barrera, A.: *Técnicas de completado de series mensuales y aplicación al estudio de la influencia de la NAO en la distribución de la precipitación en España.* Barcelona, Universidad de Barcelona, Trabajo para la obtención del Diploma de Estudios Avanzados (DEA). Programa de doctorado de Astronomía y Meteorología (Bienio 2002-2004), 2004

[4] Breiman, L. ; Friedman, J.H. ; Olshen, R.A. ; Stone, C.J.: *Classification And Regression Trees.* Boca Raton : CHAPMAN & HALL/CRC, 1984

[5] Capelli, C. ; Mola, F.: The STP Procedure as Overditting Avoidance Tool in Classification Trees. En: *Advances in Multivariate Data Analysis.* Berlín : Springer - Verlag, 2004, p. 3–13

[6] Cappelli, C. ; Mola, F. ; Siciliano, R.: A statistical approach to growing a reliable honest tree. En: *Computational Statistics & Data Analysis* 38 (2002), p. 285–299

[7] Cappelli, C. ; Reale, M.: Detecting multiple structural breaks in the mean with atheroretical regression trees. En: *Proceedings of the 20th International Workshop on Statistical Modelling.* Sydney, 2004, p. 131–134

[8] Carmack, P.S. ; Sain, S.R. ; Schucany, W.R.: Permutation Testing in Multivariate Regression Trees, 2002, p. 397–402

[9] Chaudhuri, P. ; Lo, W.D. ; Loh, W.Y. ; Yang, C.C.: Generalized regression trees. En: *Statistica Sinica* 5 (1995), p. 641–666

[10] Chaudhuri, P. ; Loh, W.Y.: Nonparametric estimation of conditional quantiles using quantile regression trees. En: *Bernoulli* 8 (2002), p. 561–576

[11] De Carvalho, F. ; De Souza, R. ; Verde, R.: A Modal Symbolic Pattern Classifier. En: *Advances in Multivariate Data Analysis.* Berlín : Springer - Verlag, 2004, p. 15–25

[12] Dudoit, S. ; Gentleman, R. ; Van der Laan, M. J.: Tree-based Multivariate Regression and Density Estimation with Right-Censored Data. En: *Journal of Multivariate Analysis* 90 (2003), p. 154–177

[13] Engle-Warnick, J.: Inferring Strategies from Observed Actions: A Nonparametric, Binary Tree Classification Approach. En: *Journal of Economic Dynamics and Control* 27 (2003), p. 2151–2170

[14] He, Y.: *Missing Data Imputation for Tree-Based Models*. Los Angeles, University of California, Tesis de Doctorado, 2006

[15] Hothorn, T. ; Hornik, K. ; Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. En: *Journal of Computational and Graphical Statistics* 15 (2006), p. 651–674

[16] Huang, W.: *Methods to Extract Rare Events*. Los Angeles, University of California, Tesis de Doctorado, 2005

[17] Izenman, A.J.: *Modern Multivariate Statistical Techniques*. New York : Springer, 2008

[18] Izrailev, S. ; Agrafiotis, D.: A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. En: *J. Chem. Inf. Comput. Sci.* 41 (2001), p. 176–180

[19] Kramer, S. ; Widmer, G. ; Pfahringer, B. ; DeGroeve, M.: Prediction of ordinal classes using regression trees. En: *Fundamenta Informaticae* 47 (2001), p. 1–13

[20] Larsen, D.R. ; Speckman, P.L.: Multivariate Regression Trees for Analysis of Abundance Data. En: *Biometrics* 60 (2004), p. 543–549

[21] Lewis, R.J.: An Introduction to Classification and Regression Tree (CART) Analysis; presented at Annual Meeting of the Society for Academic Emergency Medicine. En: *Annual Meeting of the Society of Academic Emergency Medicine*, 2000

[22] Li, K.C. ; Lue, H.H. ; Chen, C.H.: Interactive Tree-structured Regression via Principal Hessian Directions. En: *Journal of the American Statistical Association* 95 (2000), p. 547–560

[23] Loh, W.Y.: Regression Trees With Unbiased Variable Selection and Interaction Detection. En: *Statistica Sinica* 12 (2002), p. 361–386

[24] Miglio, R. ; Soffritti, G.: Proximity Measures Between Classification Trees. En: *Advances in Multivariate Data Analysis*. Berlín : Springer - Verlag, 2004, p. 27–37

[25] Piccarreta, R.: Ordinal Classification Trees Based on Impurity Measures. En: *Advances in Multivariate Data Analysis*. Berlín : Springer - Verlag, 2004, p. 39–51

[26] Scott, C.D. ; Willett, R.M. ; Nowak, R.D.: CORT: Classification Or Regression Trees. En: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)* 6 (2003), p. 153–156

[27] Struyf, J. ; Dzeroski, S.: Constraint based induction of multi-objective regression trees. En: *proceedings of the 4th International Workshop on Knowledge Discovery in Inductive Databases*, Springer, 2005, p. 110–121

[28] Tamminen, S. ; Laurinen, P. ; Roning, J. *Comparing Regression Trees With Neural Networks In Aerobic Fitness Approximation.* 1999

[29] Therneau, T.M. ; E.J., Atkinson: An Introduction to Recursive Partitioning Using the Rpart Routine. En: *Technical Report 61, Mayo Clinic, Section of Statistics*, 1997

[30] Torgo, L. *Computationally Efficient Linear Regression Trees.* 2002

[31] Vens, C. ; Blockeel, H.: A Simple Regression Based Heuristic for Learning Model Trees. En: *Journal of Intelligent Data Analysis* 10 (2006), p. 215–236

[32] Zhang, H. ; Singer, B.H.: *Recursive Partitioning and Applications.* New York : Springer, 2010