
# Comparación de la metodología BART con otros métodos no paramétricos en la construcción de intervalos de predicción

José Arturo Osorio Londoño
Maestría en Ciencias-Estadística

Isabel Cristina Ramírez Guevara, PhD
Profesora Asociada Escuela de Estadística
Directora de Tesis

Universidad Nacional de Colombia
Facultad de Ciencias,Escuela de Estadística
Medellín, Colombia
2023

# Comparación de la metodología BART con otros métodos no paramétricos en la construcción de intervalos de predicción

**Resumen**

En los últimos años, el uso de algoritmos de aprendizaje automático ha experimentado un rápido crecimiento en una amplia variedad de aplicaciones prácticas, así como un gran interés en la investigación teórica. Estas aplicaciones se centran en gran medida en problemas de predicción, donde el valor desconocido de una variable se estima en función de variables conocidas vinculadas a través de alguna función. Estos modelos se han vuelto cruciales en diversos campos, desde la gestión de calidad y el control industrial de procesos hasta la gestión de riesgos y la detección de enfermedades en el ámbito de la salud. A pesar de sus propiedades ventajosas y su popularidad, estos modelos sufren de una desventaja significativa: solo producen predicciones puntuales sin proporcionar ninguna medida de incertidumbre a estás predicciones. En esta investigación, evaluamos la capacidad de los Árboles de Regresión Aditivos Bayesianos (BART) frente a técnicas diseñadas para modelos de Random Forest y Gradient Boosting, así como heurísticas (método conformacional) y modelos clásicos como la regresión lineal y la regresión cuantílica,para generar intervalos de predicción. Se realizó un estudio de simulación bajo diferentes escenarios, y los métodos fueron validados utilizando un conjunto final de datos de aseguramiento de calidad. Los estudios de simulación revelaron que BART puede proporcionar intervalos de predicción (con una cobertura del 95 % y 90 % ) que engloban correctamente el verdadero valor predicho en la mayoría de los casos. En el caso de estudio, BART fue el mejor modelo en la generación de intervalos de predicción y en la precisión de las predicciones. Estos resultados resaltan el potencial de BART como una alternativa significativa para tareas de regresión en áreas críticas, donde predicciones precisas, modelamiento flexible y medidas de confianza en las predicciones son necesarias.

**Palabras clave:** Árboles de regresión aditivos bayesianos, BART, modelos de ensamble, intervalos de predicción, estudios de simulación.

# Comparison of BART methodology with other nonparametric methods in the construction of prediction intervalsn

**Abstract**

In recent years, the use of machine learning algorithms has rapidly expanded across a wide variety of practical applications as well as garnered significant interest in theoretical research. These applications largely focus on prediction problems, where the unknown value of a variable is estimated based on known variables linked through some function. Machine learning algorithms have become crucial in diverse domains, ranging from quality management and process control performance in industrial settings to risk management and disease detection in healthcare. Despite their advantageous properties and popularity, these models suffer from a significant drawback: they only produce point predictions without any measure of prediction uncertainty. In this research, we assess the capability of Bayesian Additive Regression Trees (BART) compared to techniques designed for Random Forest, Gradient Boosting ensemble models, heuristics (conformal prediction) and classic models as linear regression and quantile regression when generating prediction intervals. A simulation study was conducted under various scenarios, and the methods were validated using a final dataset from quality assurance. The simulation studies revealed that BART demonstrates an impressive ability to generate prediction intervals (at the 95 % and 90 % coverage) that correctly encompass the true predicted value in most of the cases. In the case study, validation BART was the best model in the prediction interval generation and in prediction accuracy. These results highlight BART's potential as a significant alternative for regression tasks in critical areas, where accurate predictions, flexible modeling, and confidence measures on the predictions are imperative.

**Keywords:** Bayesian Additive Regression Trees, BART, ensemble models, prediction intervals, simulation study.

# Índice

# Índice de figuras

# 1. Introducción

En los últimos años, el uso de algoritmos de aprendizaje automático se ha expandido rápidamente en una amplia variedad de aplicaciones prácticas y ha capturado el interés en la investigación teórica. Las aplicaciones se concentran en gran parte en los problemas de predicción, es decir, en la estimación del valor desconocido de una variable cuando se conocen variables vinculadas mediante alguna función. En este sentido, los algoritmos de aprendizaje automático se han vuelto determinantes en una amplia gama de dominios prácticos como en la gestión de sistemas de calidad y el desempeño de los procesos de control en áreas industriales (Bertolini et al., 2021) hasta la gestión del riesgo y la detección de enfermedades en salud (Shehab et al., 2022).

Como un subconjunto de los algoritmos de aprendizaje automático, los modelos de ensamble se han posicionado como una de las opciones más populares: su eficiencia en problemas de alta dimensionalidad, su flexibilidad para modelar procesos no-lineales y su capacidad para lograr niveles adecuados de predicción sin afinar demasiados parámetros explican en parte esta preferencia. Dentro de la familia de modelos de ensamble, aquellos basados en árboles, como Random Forest y Gradient Boosting Machine[1] , son de los más utilizados en la actualidad y conservan su relevancia frente a modelos más complejos principalmente en aquellos dominios de aplicación sobre datos estructurados (Grinsztajn et al., 2022).

A pesar de las propiedades ventajosas que poseen y la popularidad que los rodea estos modelos, presentan un problema importante: solo producen predicciones puntuales sin ninguna medida de incertidumbre vinculada a la predicción. Esta falta de información sobre la incertidumbre asociada a las predicciones puede resultar problemática en dominios donde es necesario tener una idea clara de la precisión de las predicciones para tomar decisiones como es el caso de las aplicaciones médicas, los procesos de gestión de riesgo y el control de calidad de sistemas industriales.

Para abordar esta limitación, se han propuesto diversas estrategias para generar intervalos de predicción a partir de modelos de ensamble basados en árboles. Estas estrategias se basan en variaciones de la técnica de bootstrap (métodos conformacionales), la reformulación de los modelos para predecir cuantiles condicionales de la distribución en lugar del valor esperado, y, como veremos en esta tesis, la reformulación en un sentido bayesiano de los modelos de ensamble basados en árboles.

[1]En el aprendizaje automatico competitivo los modelos de Gradient Boosting y Random Forest siguen siendo dominantes en problemas de datos tabulares: https://mlcontests.com/state-of-competitive-machine-learning-2022/

La presente tesis propone explorar y comparar las heurísticas y métodos derivados para los modelos de ensamble basados en árboles y los modelos de ensamble bayesianos al momento de generar intervalos de predicción.
A continuación, exponemos en primer lugar el estado del arte actual referente al uso comparado de los modelos que nos interesan. En segundo lugar se abordan los principios teóricos para la generación de los intervalos de predicción enmarcados dentro del problema de regresión y se profundiza en los diferentes métodos utilizados para la construcción de los mismos. En tercer lugar se expone el diseño experimental y se validan los diferentes métodos en relación a las métricas de evaluación definidas mediante un estudio de simulación. En cuarto lugar, se muestra una aplicación práctica del problema. Finalmente se presentan las conclusiones.

# 2. Estado del Arte

En lo que respecta al calculo de los intervalos de predicción se encuentran dos grandes lineas de investigación:

Una primera línea de investigación se da en el contexto de los **modelos lineales** (entendemos por un modelo lineal aquel donde la función de aproximación toma la forma de una combinación lineal de parámetros y variables: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$ ). En este caso, el cálculo de intervalos de predicción se da en dos direcciones:

1. A partir de supuestos sobre la distribución de $\epsilon$ como normalidad (por ejemplo, se asume que $\epsilon(x) \sim N(0, \sigma^2)$) y se hace mediante derivaciones de forma cerrada como en Seber y Lee (2012).

2. Mediante métodos menos restrictivos como el uso de bootstrap (Stine (1985)) o el uso de los residuales empíricos y las propiedades límite de los modelos (Schmoyer (1992).

La segunda gran línea de investigación se ha dado en el campo de las **redes neuronales** donde los métodos que se han utilizado son variados: método delta, prediciendo directamente la varianza de la variable objetivo (vía máxima verosimilitud y bootstrap) o mediante la estimación con regresiones cuantílicas (Su et al. (2018)).

En cuanto a los modelos analizados en este estudio, tanto los modelos de referencia (Quantile Random Forest -QRF-, Gradient Boosting Machine -GBM-) como el principal objeto de investigación (Bayesian additive regression trees -BART-) si bien son motivo de interés académico y práctico: el artículo original que introduce Random Forest (Breiman, 2001) acumula más de 60,000 citaciones, la implementación más conocida de Gradient Boosting (Chen y Guestrin, 2016) supera las 7,000 citaciones y cuenta con implementaciones en sistemas distribuidos como Spark y en el caso de BART ha surgido una creciente cantidad de aplicaciones en diversas áreas (Tan y Roy, 2019) estos modelos han sido evaluados tradicionalmente mediante métricas de error asociadas a las predicciones construidas, generalmente en un conjunto de prueba, como la raíz del error cuadrático medio, el error medio absoluto, entre otros. Por ejemplo, este es el enfoque que adopta el artículo original de BART al compararlo con modelos existentes (Chipman et al., 2010).

Fuera de las dos líneas de investigación mencionadas (sobre modelos lineales o redes neuronales) son pocos los trabajos que traten, entre otras cosas, de evaluar la

consistencia o asociar medidas de incertidumbre a las predicciones producidas por estos modelos es escasa. De hecho, esto se extiende más allá de estos modelos a todos aquellos modelos del aprendizaje de máquinas supervizado (de las máquinas de soporte vectorial a los métodos de vecinos más cercanos, entre otros) y donde, de nuevo, solo en el caso de las redes neuronales han existido esfuerzos consistentes en crear medidas de incertidumbre (Zapranis y Livanis, 2005) y en los demás casos han sido incursiones parciales como vemos en De Brabanter et al. (2010) para máquinas de soporte aplicadas a regresión o en Zhang et al. (2019) para el caso de *Random Forest*. En relación al uso de intervalos de predicción, de forma más general, para crear medidas de incertidumbre en los problemas de predicción los mismos han tenido poca relevancia. Esto lo podemos ver, por ejemplo, a través de una recopilación de sobre la teoría y práctica de la medición de incertidumbre como en Li et al. (2012). Su popularización se ha dado más recientemente mediante heurísticas y el uso de las nuevas capacidades computaciones disponibles.
En lo que respecta a la evaluación y comparación de los intervalos de predicción obtenidos con los modelos de este estudio encontramos los siguientes artículos:

- En He et al. (2017) se utilizan QRF y BART para la predicción de cortes en la distribución de electricidad asociados a tormentas. Aquí, se evalúan no solo las predicciones puntuales sino los intervalos de predicción que pueden ser producidos para cada modelo. El artículo utiliza los métodos de QRF y BART en la versión que ya hemos reseñado. Las conclusiones a las que se llega es que BART produce mejores predicciones puntuales, medidas mediante el error cuadrático medio y el error medio absoluto, pero QRF genera intervalos de predicción menos anchos.

- En Ehsan et al. (2019) se utilizan QRF y BART para predecir la velocidad del viento en 127 tormentas convectivas. El estudio evalúa los dos modelos en términos de predicciones puntuales e intervalos de predicción. En este caso, QRF presentó los mejores resultados tanto en predicciones puntuales como en la generación de intervalos de predicción.

- Bogner et al. (2019) lleva a cabo una evaluación de diferentes modelos de aprendizaje de máquinas: redes neuronales cuantílicas (QNN), regresión por kernel cuantílica (KQR), splines adaptativos multivariados (MARS) y, dos de los modelos que nos interesan: QRF y los modelos de boosting cuantílicos (Quantile Boosting Machine,QBM). La aplicación concreta de estos modelos se hizo sobre la predicción de consumo y producción de energía. La evaluación general de las

predicciones, en tanto su relación con el valor real, se llevó a cabo mediante el coeficiente de determinación y para evaluar los modelos en términos de su capacidad para construir medidas de incertidumbre se evaluó directamente el rango de las predicciones a través del *Continuous Rank Predictive Score* (CRPS). Los resultados indican como mejor modelo, para el caso del consumo, al QRF tanto en términos de predicciones puntuales como en el CRPS. En los modelos usados para predecir la producción el mejor desempeño se encontró en los modelos MARS seguidos muy de cerca por QBM. Es importante aclarar que estas evaluaciones se hicieron sobre un conjunto de prueba.

De estos artículos hay que destacar que no representan un esfuerzo completo de investigación sobre los intervalos de predicción generados por los modelos en al menos dos sentidos. En primer lugar, la evaluación está directamente ligada a casos de aplicación concretos, por tanto, no se profundiza en las propiedades de los datos usados y su posible impacto sobre la construcción de los intervalos de predicción. En segundo lugar, no hay una definición de base o puntos de partida adecuados para evaluar los modelos: no hay un punto concreto de comparación y cualquier conclusión es sólo relativa a los modelos y datos usados. Por estas dos razones, las conclusiones que pueden ser extraídas de estos artículos son a lo sumo parciales y vinculadas únicamente a los casos de aplicaciones mencionados como los eventos climáticos.
Por último, es importante mencionar que han surgido métodos de generación de intervalos de predicción más generales, en últimas, con la promesa de poderse extender a casi cualquier modelo de regresión en un contexto predictivo. Estos métodos se han construido dentro de la literatura sobre *conformal prediction* (Shafer y Vovk, 2008) donde las predicciones se asumen parte de un proceso generador intercambiable de tal manera que mediante la construcción de medidas de no conformidad junto con estimaciones aumentadas se derivan regiones de predicción. Más detalles sobre estas técnicas y el desarrollo de un marco de trabajo completo se pueden encontrar en Lei et al. (2018).En este trabajo, se incluye una de las implementaciones del método conformacional como punto adicional de comparación.

# 3. Marco Teórico

## 3.1. El problema de regresión

Los intervalos de predicción se hacen relevantes dentro del marco general de los problemas de regresión: la predicción de observaciones futuras de una variable cuantitativa $y$ mediante el uso de los valores observados tanto de $y$ como de un conjunto de covariables $x$. La construcción de un modelo con esta capacidad predictiva se conoce como el problema de regresión y se puede representar, sin pérdida de generalidad, de la siguiente forma:

$$y = f(x) + \epsilon(x) \tag{1}$$

donde se busca, a partir de un conjunto de datos $\{(x_i, y_i) \in \mathbb{R}^{n+1} \mid i < n\}$ con $n$ observaciones independientes e idénticamente distribuidas ($iid$) estimar de la mejor forma posible una aproximación $\widehat{f(x)}$ a $f(x)$ con el objetivo de predecir valores no observados de $y$. En este sentido, $\epsilon(x)$ representa el ruido presente e irreductible que no puede ser capturado por $f(x)$ ni aproximado mediante $\widehat{f(x)}$.
La aproximación $\widehat{f(x)}$, en la mayor parte de las aplicaciones, no es más que la media condicional como se muestra a continuación.
Supongamos que $y$ es una variable aleatoria con distribución conjunta $P(x, y)$ y buscamos una función $f(x)$ que nos permita predecir los valores de $y$. Adicionalmente, definimos una estrategia para evaluar la capacidad de aproximación de la función $f(x)$. En este sentido, se define una **función de pérdida** $L(y, f(x))$ que pondera y penaliza los errores cometidos en la predicción por $f(x)$. Es decir, buscamos la manera de determinar que tanto se aleja $f(x)$ de los verdaderos valores de $y$.
En el caso de los problemas de regresión la **función de pérdida** más común suele ser el error cuadrático : $L(Y, f(X)) = (Y - f(X))^2$ (Hastie et al. (2009))
De esta manera, tenemos como criterio para escoger $\hat{f}$ la minimización del error esperado. A continuación podemos probar que la media condicional minimiza el error cuadrático esperado:
Buscamos probar lo siguiente:

$$\arg\min_{f(X)} \mathbb{E}[(Y - f(X))^2] = \mathbb{E}[Y|X] \tag{2}$$

que puede ser escrito de forma alternativa como:

$$\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] \leq \mathbb{E}\left[(Y - f(X))^2\right] \tag{3}$$

para cualquier otra función de las covariables: $f(X)$.

Ahora bien, si partimos de

$$\mathbb{E}[(Y - f(X))^2] \tag{4}$$

Sumamos y restamos $\mathbb{E}[Y|X]$

$$\mathbb{E}\left[\{(Y - \mathbb{E}[Y|X]) - (f(X) - \mathbb{E}[Y|X])\}^2\right] \tag{5}$$

Expandiendo el cuadrático se obtiene:

$$\mathbb{E}\left[\{(Y - \mathbb{E}[Y|X])^2 + (f(X) - \mathbb{E}[Y|X])^2 - 2(Y - \mathbb{E}[Y|X])(f(X) - \mathbb{E}[Y|X])\}\right] \tag{6}$$

Recordemos la ley de las esperanzas iteradas (LEI):

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] \tag{7}$$

Examinemos de cerca el último término usando la LEI y el hecho de que $\mathbb{E}[(f(X) - \mathbb{E}(Y|X))|X]$ es una constante:

$$\begin{aligned}
\mathbb{E}[(Y - \mathbb{E}[Y|X])(f(X) - \mathbb{E}[Y|X])] &= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])(f(X) - \mathbb{E}[Y|X])|X]] \\
&= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}(Y|X))|X] \cdot \mathbb{E}[(f(X) - \mathbb{E}(Y|X))|X]] \\
&= \mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y|X])(f(X) - \mathbb{E}[Y|X])] \\
&= 0
\end{aligned}$$

Además, el segundo término es nuevamente un componente cuadrático y por tanto no negativo. El cual será igual a cero si y solo si $f(X) = \mathbb{E}[Y|X]$. Por tanto:

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \text{algo no negativo} \tag{8}$$

Y así:

$$\mathbb{E}[(Y - f(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] \tag{9}$$

con igualdad estricta si y solo si $f(X) = \mathbb{E}[Y|X]$.
Es claro que el $\mathbb{E}[(Y - f(X))^2]$ solo es mínimo si $f(x) = \mathbb{E}[Y|X]$. ■

## 3.2. Intervalos de predicción

### 3.2.1. Intervalos de confianza e intervalos de predicción

Los intervalos de confianza y los intervalos de predicción tienen propósitos diferentes en el análisis estadístico, aunque ambos están vinculados al proceso de cuantificar la incertidumbre en las estimaciones Heskes (1996).

Un intervalo de confianza se vincula a la incertidumbre en la estimación de un parámetro como puede ser la media o la pendiente en una regresión lineal y están vinculados a los procesos de pruebas de hipótesis. Su interés reside en determinar el rango de valores posibles para un parámetro. Por otro lado, un intervalo de predicción tiene en cuenta tanto la incertidumbre en la estimación de los parámetros como la variabilidad intrínseca de los datos al predecir un nuevo valor observado en la población, Agresti (2015, chap. 3).

En relación al problema de regresión los intervalos de confianza se pueden entender como una herramienta para evaluar la calidad de la aproximación $\widehat{f(x)}$ a la verdadera pero desconocida función $f(x)$. Su importancia reside en capturar la distribución de la cantidad $(f(x) - \hat{y})$ con $\hat{y} = \widehat{f(x)}$. Cuando asumimos, adicionalmente, dentro del problema de regresión el objetivo de la minimización del error cuadrático los intervalos de confianza corresponden a una evaluación de las desviaciones alrededor de la media condicional.

Sin embargo, en muchas aplicaciones es más importante tener información de la calidad de las predicciones alrededor de las observaciones realizadas: $y$ y no unicamente de la media condicional. Los intervalos de predicción se encargan de la calidad de la predicción respecto a las observaciones realizadas, es decir, de la distribución de la cantidad $y - \hat{y}$.

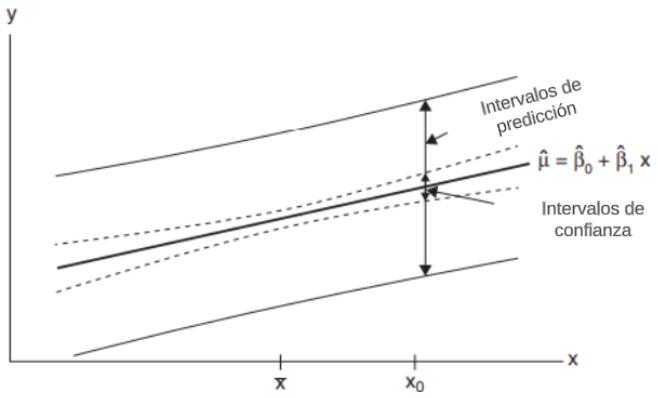


Figura 1: Intervalos de predicción y confianza (Agresti, 2015)

De la ecuación de regresión base (1) podemos observar que:

$$y - \hat{y} = f(x) + \epsilon(x) - \hat{y} = [f(x) - \hat{y}] + \epsilon(x) \tag{10}$$

Y es claro que los intervalos de predicción contienen los intervalos de confianza y a su vez son más informativos: nos dan información alrededor de la calidad de la media condicional y de los valores predichos observados **en si** (Figura 1).

### 3.2.2. Derivación principal

Supongamos que hemos construido el modelo de regresión $\widehat{f(x)}$ usando alguna técnica de estimación y el conjunto de datos $\{(x_i, y_i) \in \mathbb{R}^n \mid i < n\}$ con $n$ observaciones *iid*. Los intervalos de predicción son un rango que contiene la nueva observación $y_{n+1} = f(x_{n+1})$. En particular, para un nivel de significancia $\alpha$ definimos los intervalos de predicción como el rango:

$$I_{1-\alpha} = (L_\alpha(x_{n+1}), U_{1-\alpha}(x_{n+1})), \tag{11}$$

de tal forma que

$$Pr[y_{n+1} \in I_{1-\alpha}(x_{n+1})] \approx 1 - \alpha \tag{12}$$

Adicionalmente, para cualquier modelo de regresión se puede descomponer una nueva predicción $y_{n+1}$ como :

$$y_{n+1} - \hat{y}_{n+1} = f(x_{n+1}) + \epsilon(x_{n+1}) - \hat{y}_{n+1} = [f(x_{n+1}) - \hat{y}_{n+1}] + \epsilon(x_{n+1}) \tag{13}$$

Y reordenando tenemos:

$$y_{n+1} = \hat{y}_{n+1} + [f(x_{n+1}) - \hat{y}_{n+1}] + \epsilon(x_{n+1})$$

$$y_{n+1} = \hat{y}_{n+1} + \nu(x_{n+1}) + \epsilon(x_{n+1}) \tag{14}$$

Donde $\nu(x_{n+1}) = [f(x_{n+1}) - \hat{y}_{n+1}]$ es el error del modelo, es decir, el error de la función de regresión y $\epsilon(x_{n+1})$ es el error irreductible. Dado que la derivación del intervalo de predicción pasa por la distribución $y_{n+1} - \hat{y}_{n+1} = \nu(x_{n+1}) + \epsilon(x_{n+1})$, la mayoría de los métodos de estimación de intervalos de predicción se enfocan en estimar de una forma u otra (o de manera implícita) estos dos componentes.

### 3.2.3. Criterios de evaluación

Para evaluar los intervalos generados nos basaremos en tres métricas, donde la tercera, es una derivación que de las dos primeras.

La probabilidad de cobertura del intervalo de predicción (PICP, Prediction Interval Coverage Probability): se define como la proporción de observaciones de validación para las cuales la predicción esta contenida en el intervalo. En principio, los intervalos de predicción deben, en promedio, encapsular las predicciones en un $(1-\alpha)\,\%$ de los casos (usualmente en el $90\,\%$, $95\,\%$, $99\,\%$). PICP nos da una medida de la validez de la técnica usada para generar los intervalos, formalmente:

$$PICP(I_\alpha) = \frac{1}{N}\sum_{i=1}^{N} t_i \quad \text{donde} \begin{cases} t_i = 1 \text{ si } \; L_\alpha(x_{n+1}) \leq y_{n+1} \leq U_\alpha(x_{n+1})) \\ t_i = 0 \;\; \text{en otro caso} \end{cases} \tag{15}$$

La segunda métrica es el promedio de longitud del intervalo (MPIW, Mean Prediction Interval Width). La longitud del intervalo es una medida de optimalidad, dentro del conjunto de técnicas que produzcan intervalos validos son deseables (menos inciertas) aquellas que producen intervalos con menor longitud.

$$MPIW(I_\alpha) = \frac{1}{N}\sum_{i=n+1}^{N} |L_\alpha(x_{n+1}) - U_\alpha(x_{n+1})| \tag{16}$$

En general, se busca minimizar MPIW bajo la restricción de que se cumpla PICP (Pevec y Kononenko (2015)). Por tanto, se considera una tercera métrica que busca un compromiso entre la validez y la certeza.
Definimos

$$CWC = \text{MPIW}\left(1 + \gamma(\text{PICP})e^{-\eta(\text{PICP}-(1-\alpha))}\right) \tag{17}$$

con

$$\begin{cases} \gamma = 0 \text{ si } \; \text{PICP} \geq (1-\alpha) \\ \gamma = 1 \;\; \text{en otro caso} \end{cases} \tag{18}$$

La métrica CWC (Coverage width criterion, Khosravi et al. (2011)) combina la información sobre el ancho del intervalo de predicción (MPIW) y la proporción de valores verdaderos que caen dentro de dicho intervalo (PICP) para proporcionar una medida comprensiva de los intervalos. El parámetro $\gamma$ penaliza el ancho de los intervalos (se aplica un modificador cuando no se puede garantizar la cobertura, si se garantiza la cobertura, el ancho permanece sin modificación) $\eta$ controla la fuerza de la

penalización: un valor menor de $\eta$ penaliza más fuerte la distancia entre la cobertura alcanzada y la cobertura esperada $1-\alpha$.

## 3.3. Métodos para construir intervalos de predicción

### 3.3.1. Intervalos de predicción con modelos paramétricos.

Los intervalos de predicción con modelos paramétricos son aquellos que se construyen a tráves del uso de los supuestos estadísticos y distribucionales derivados del modelo utilizado. En este sentido, es importante presentar la derivación de intervalos de predicción en el caso clásico del modelo de regresión lineal normal ya que es el ejemplo más sobresaliente de este tipo de modelos.

#### 3.3.1.1. Modelo de regresión lineal normal

En Agresti (2015, chap. 3) se muestra que en el modelo lineal normal el valor futuro de $y$ satisface:

$$y = x_{n+1}\beta + \epsilon, \text{ donde } \epsilon \sim N(0, \sigma^2). \tag{19}$$

Aquí $\epsilon$ es el error aleatorio. Además, del modelo estimado tenemos que el valor futuro de $y$ es $\hat{\mu} = x_{n+1}\hat{\beta}$. Por tanto, $y$ también cumple que

$$y = x_{n+1}\hat{\beta} + e, \ e = y - \hat{\mu}, \tag{20}$$

donde $e$ es el residual para la observación. Dado que se supone que $y$ es independiente de las observaciones $y_1, y_2, \ldots, y_n$ usadas para estimar $\hat{\beta}$ y por tanto de $\hat{\mu}$,

$$var(e) = var(y - \hat{\mu}) = var(y) + var(\hat{\mu}) = \sigma^2[1 + x_{n+1}(\mathbf{X}^\top\mathbf{X})x_{n+1}^\top], \tag{21}$$

donde $\mathbf{X}$ es la matriz de diseño con $n$ observaciones y $p$ covariables. De aquí, se sigue que

$$\frac{y - \hat{\mu}}{\sigma\sqrt{[1 + x_{n+1}(\mathbf{X}^\top\mathbf{X})x_{n+1}^T]}} \sim N(0, 1) \tag{22}$$

Y por tanto, al usar el estimador clásico de la varianza $\hat{s}^2 = \frac{\hat{\epsilon}\hat{\epsilon}^\top}{n-p}$ como estimador de $\sigma^2$, tenemos

$$\frac{y - \hat{\mu}}{\hat{s}\sqrt{[1 + x_{n+1}(\mathbf{X}^\top\mathbf{X})x_{n+1}^T]}} \sim t_{n-p} \text{ ya que } \epsilon \sim N(0, \sigma^2) \tag{23}$$

Basta con invertir (24) para generar el intervalo de predicción para la observación futura $y$:

$$\hat{\mu} \pm t_{\alpha/2, n-p}\hat{s}\sqrt{1 + x_{n+1}(\mathbf{X}^\top\mathbf{X})x_{n+1}^T} \tag{24}$$

Está derivación es paramétrica ya que depende de la distribución asumida y de una correcta estimación de la varianza del sesgo $(y - \hat{\mu})$. Se hace más claro cuando observamos la ecuación (22): los intervalos dependen de la descomposición de la varianza de la estimación del valor esperado y de la varianza asociada a $y$.

### 3.3.2. Intervalos de predicción con modelos no paramétricos

La técnicas que se utilizarán para construir los intervalos de predicción en este trabajo son no paramétricas. Con esto nos referimos al hecho de que no se hace ningún supuesto respecto a la distribución de la variable de interés (Mayr et al., 2012)
En el método tradicional para derivar intervalos de predicción visto en la sección anterior, se procede modelando el valor esperado de la nueva observación y, partir de allí, se deriva un intervalo alrededor de este valor esperado $\hat{\mu}$ haciendo uso de la descomposición del sesgo $(y - \hat{\mu})$. En este trabajo se evalúan múltiples estrategias para obtener intervalos de predicción con una idea en común: el uso de cuantiles en lugar de la derivación explicita a tráves de supuestos distribucionales (Meinshausen, 2006). En este sentido, se busca

$$\text{IP}_{1-\alpha}(x_{n+1}) = [q_{\alpha/2}(x_{n+1}), q_{1-\alpha/2}(x_{n+1})] \tag{25}$$

Así, el ejercicio está orientado a la obtención de $q_\alpha$ y $q_{1-\alpha/2}$ que no son más que los cuantiles condicionales: $q_\alpha = \hat{f}_\alpha(Y|X_{new})$ generados por el método respectivo. En este trabajo se exploran diferentes formar de obtener las bandas del intervalo, es decir, $q_\alpha$ y $q_{1-\alpha/2}$.

#### 3.3.2.1. Regresion cuantílica

En el caso de la regresión lineal normal nos enfocamos en predecir la media condicional de la variable de respuesta. Por otro lado, en la regresión cuantílica los cuantiles de la distribución condicional de la variable de respuesta son estimados directamente. Es decir, en contraste con la regresión lineal normal, en la regresión cuantílica no se modela el valor esperado $E[Y|X = \mathbf{x}^\top]$ sino el $\alpha - ésimo$ cuantil $f_\tau(Y|X = \mathbf{x}^\top)$ de la distribución condicional.
Dado un conjunto de datos $(x_i, y_i)$ para $i = 1, \dots n$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$. La regresión cuantílica busca encontrar una función $Q_\tau(x)$ que estime el valor del cuantil $\tau$ de $y$ dado $x$. Es decir, $Q_\tau(x)$ satisface:

$$P(y \leq Q_\tau(x)|x) = \tau \tag{26}$$

Para estimar $Q_\tau(x)$, utilizamos un enfoque de optimización que minimiza la función de pérdida cuantílica sobre todas las observaciones, que es una función no simétrica que penaliza los errores de predicción según su posición relativa al cuantil $\tau$

$$\underset{\beta_\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta_\tau) \text{ donde } \rho_\tau = \begin{cases} \mathbf{x}_i^T \beta_\tau \tau & \text{si } \mathbf{x}_i^T \beta_\tau \geq 0 \\ \mathbf{x}_i^T \beta_\tau (\tau - 1) & \text{si } \mathbf{x}_i^T \beta_\tau < 0 \end{cases} \tag{27}$$

donde $\rho_\tau(u)$ es la función de pérdida cuantílica (tambien conocida como función *check* ó *pinball* (Koenker, 2005, chap. 2)), definida como:

$$\rho_\tau(u) = u(\tau - 1\{u < 0\}), \tag{28}$$

y $1\{.\}$ es la función indicadora.
En la Figura 2 se muestra el resultado de la estimación para el modelo de regresión lineal simple (en color rojo), es decir, la media condicional $E(Y|X)$ y las diferentes estimaciones cuantílicas para $\tau$. La regresion cuantílica se desplaza sobre la distribución condicional $f_\tau(Y|X = \mathbf{x}^\top)$. Este desplazamiento es el que le permite capturar el comportamiento en diferentes puntos de la distribución.

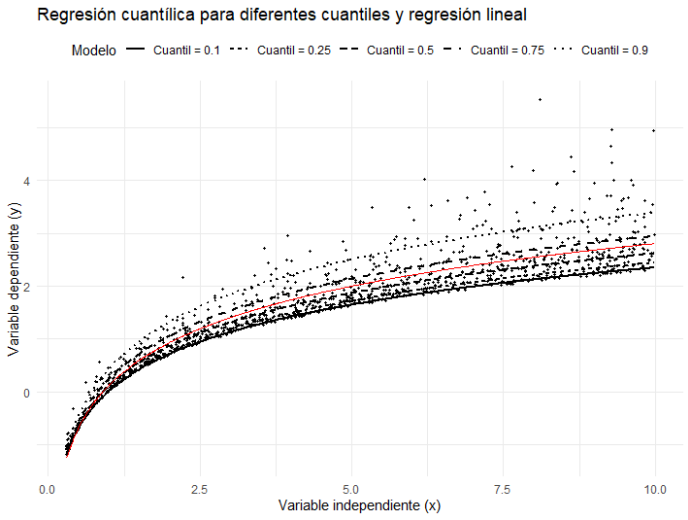


Figura 2: Regresión Lineal vs Regresión Cuantílica

La función *check*, ecuación (29), no es diferenciable en 0 por lo que en el proceso de optimización se utilizan técnicas de programación lineal. Estás se pueden encontrar implementadas en la libreria **quantreq** de R (Koenker et al., 2012).
A su vez, existen métodos de estimación adicionales como pueden ser por quasi-maxima verosimilitud (Geraci y Bottai, 2007) o mediante inferencia bayesiana (Yu y Moyeed,

2001).

#### 3.3.2.2. Predicción conformacional

La predicción conformacional o conforme (Angelopoulos y Bates, 2021) busca ser un marco general para la construcción de intervalos de predicción ($I_{1-\alpha}$) usando el supuesto de intercambiabilidad y las propiedades de la transformación de ranking de variables continuas. El método posee las siguientes características:

- Cobertura en muestras finita garantizada (exacta)
- Independiente de supuestos distribucionales
- Independiente de supuestos vinculados a modelos.

Para comprender el alcance de los métodos conformacionales podemos partir de su versión más sencilla o "naive". El método de predicción conforme hace uso de los estadísticos de orden y de ranking de la siguiente manera:
En primer lugar, necesitamos la propiedad de intercambialidad. En este sentido, si $Y$ es un conjunto de variables *iid* entonces las variables $Y$ son intercambiables, es decir, la distribución de probabilidad conjunta de las variables no cambia con modificaciones en la posición en la que aparecen las variables. En este sentido, $Y_1, Y_2, Y_3$ tiene la misma distribución conjunta que $Y_2, Y_3, Y_1$.
Veamos, dado que las variables aleatorias $Y$ son independientes, la función de distribución acumulada (cdf) conjunta es igual al producto de las funciones marginales individuales. Por lo tanto, para cualquier permutación $\pi$ ($\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$)
,

$$F_{Y_{\pi(1)},\ldots,Y_{\pi(n)}}(y_1, \ldots, y_n) = \prod_{i=1}^{n} F_{Y_{\pi(i)}}(y_i). \tag{29}$$

Y puesto que las variables aleatorias están idénticamente distribuidas, las funciones de distribución acumuladas marginales son todas iguales.De forma que, $F_{Y_{\pi(1)}} = \cdots = F_{Y_{\pi(n)}} =: F$ y

$$F_{Y_{\pi(1)},\ldots,Y_{\pi(n)}}(y_1, \ldots, y_n) = \prod_{i=1}^{n} F(y_i), \tag{30}$$

lo que indica que la distribución es invariante bajo operaciones de permutación, ya que el lado derecho no depende de $\pi$.
En segundo lugar, necesitamos conocer la distribución de la transformación ranking de la variable $Y$.

Sea $Y_{(i)}$ el $i$-esimo estadístico de orden tal que $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. Definimos el ranking como $R_i \in \{1, \ldots, n\}$ de tal forma que $Y_i = Y_{(R_i)}$. Es decir, $R$ representa la posición de los estadísticos de orden. Si además asumimos que las $Y_i$s son continuas (y por tanto $P(Y_{(i)} = Y_{(j)}) = 0$) entonces el ranking asignado a $Y_i$ esta únicamente definido: no existen empates.
Ya hemos visto que las variables $Y_1, \ldots, Y_n$ son *iid* y por tanto intercambiables. En este sentido, para cualquier permutación la distribución de $Y$ permanece igual: $[Y_1, \ldots, Y_n] \stackrel{d}{\equiv} [Y_{\pi(1)}, \ldots, Y_{\pi(n)}]$ donde $\pi$ es una función de permutación.
Ahora bien, dado que las variables son intercambiables tenemos que $R_1 = r$ tiene la misma probabilidad que $R_{\pi(1)} = r$ y, en principio, $\pi(1)$ puede ser cualquier número en $\{1, \ldots, n\}$. Por lo tanto:

$$P(R_1 = r) = P(R_2 = r) = P(R_n = r) \tag{31}$$

Dado el supuesto de continuidad estos son eventos mutuamente exclusivos y exhaustivos que suman 1. De aquí se deduce que cada uno de ellos debe tener probabilidad $\frac{1}{n}$. Luego,

$$\mathrm{P}(R_i = r) = 1/n \tag{32}$$

Con estos resultados presentes podemos derivar la construcción de intervalos de predicción de la siguiente manera:
Sean $Y_1, \ldots, Y_n$ muestras *iid* de una variable aleatoria continua. Para un nivel de no cobertura dado $\alpha \in (0, 1)$, y otra muestra *iid* $Y_{n+1}$, observamos que

$$P(Y_{n+1} \leq q_{1-\alpha}) \geq 1 - \alpha, \tag{33}$$

donde definimos el cuantil muestral $\hat{q}_{1-\alpha}$ en función de $Y_1, \ldots, Y_n$ mediante

$$\hat{q}_{1-\alpha} = \begin{cases} Y\left(\lceil (n+1)(1-\alpha) \rceil\right) & \text{si } \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty & \text{en otro caso,} \end{cases} \tag{34}$$

Aquí $Y(1) \leq \ldots \leq Y(n)$ definen los estadísticos de orden de $Y_1, \ldots, Y_n$. La propiedad de cobertura en la ecuación (34) es fácil de verificar: por intercambiabilidad el ranking de $Y_{n+1}$ entre $Y_1, \ldots, Y_n$, se distribuye uniformemente en el conjunto $\{1, \ldots, n+1\}$ y por tanto, la ecuación (34) recuerda la definición del valor p en una prueba de hipótesis: validamos que tan cercana es la observación $Y_{n+1}$ a $Y_1, \ldots, Y_n$ .
Para el caso de los modelos de regresión donde tenemos muestras *iid* $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$, y heredamos las propiedades de intercambialidad, podemos considerar una primera versión para construir un intervalo de predicción para $Y_{n+1}$ en

el nuevo valor $x_{n+1}$. A partir de lo expuesto anteriormente, una forma para construir el intervalo de predicción es la siguiente:

$$I_{1-\alpha}(\mathbf{x_{n+1}}) = \left[\hat{\mu}(\mathbf{x_{n+1}}) - \hat{\epsilon}_n^{-1}(1-\alpha), \hat{\mu}(\mathbf{x_{n+1}}) + \hat{\epsilon}_n^{-1}(1-\alpha)\right], \tag{35}$$

donde $\hat{\mu}$ es un estimador de la función de regresión subyacente, $\hat{\epsilon}_n$ es la distribución empírica de los residuos obtenidos desde la función de validación $|Y_i - \hat{\mu}(X_i)|, i = 1, \ldots, n$ y $\hat{\epsilon}_n(1-\alpha)$ representa el $(1-\alpha)$ cuantil de $\hat{\epsilon}_n$. Este método requiere que los cuantiles de los residuales estimados sean cercanos a los de los residuales poblacionales y por tanto que la función $\hat{\mu}$ sea precisa. En este sentido, se requiere de un modelo correctamente especificado y condiciones de regularidad en la distribución conjunta de $X, Y$ (Lei et al., 2018).

**3.3.2.3. Predicción comformal completa** En general, el método naive anterior puede subestimar la cobertura ya que la distribución de los residuales ajustados está sesgada. En este sentido, Lei y Wasserman (2014) propone una versión del método conformal que permite garantizar la cobertura sin hacer supuestos sobre la distribución conjunta de $X, Y$ ni sobre la función de regresión $\hat{\mu}$ (con excepción de que $\hat{\mu}$ sea una función simétrica respecto a los residuales)
El método propuesto consiste de los siguientes pasos:

1. La reestimación del modelo $\hat{\mu}$ con un conjunto de datos aumentando
2. La construcción de una función de scoring que evalúa la cercanía de $X, Y$
3. La construcción de una prueba de hipótesis sobre la pertenencia de la observación aumentada usando los estadísticos de ranking.
4. La inversión de la prueba de hipótesis para derivar intervalos de predicción.

Podemos proceder de la siguiente manera: para cada valor $y \in \mathbb{R}$, se construye un modelo de regresión $\hat{\mu}_y$, que se ajusta sobre el conjunto de datos incremental dado por $Z_1, \ldots, Z_n, (X_{n+1}, y)$. Definimos entonces las funciones de scoring o evaluación:

$$S_{y,i} = |Y_i - \hat{\mu}_y(X_i)|, \quad i = 1, \ldots, n \quad \text{y} \quad S_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})|, \tag{36}$$

y evaluamos el ranking de $S_{y,n+1}$ entre los residuos ajustados originales $S_{y,1}, \ldots, S_{y,n}$, mediante

$$\Gamma(y) = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\{S_{y,i} \leq S_{y,n+1}\} = \frac{1}{n+1} + \frac{1}{n+1}\sum_{i=1}^{n} 1\{S_{y,i} \leq S_{y,n+1}\} \tag{37}$$

Es decir, cual es la proporción de puntos en la muestra aumentada cuyos residuos ajustados son menores que el último, $S_{y,n+1}$. Aquí, $1\{.\}$ es la función indicadora. Por el supuesto de intercambiabilidad y la simetría de $\hat{\mu}$ cuando evaluamos $y = Y_{n+1}$ vemos que el estadístico construido $\Gamma(Y_{n+1})$ está distribuido uniformemente en el conjunto $\left\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, 1\right\}$, lo que implica

$$P\left((n+1)\,\Gamma(Y_{n+1}) \leq \lfloor (1-\alpha)(n+1) \rfloor\right) \geq 1-\alpha. \tag{38}$$

La ecuación anterior tiene la estructura del cálculo de un valor p. Por tanto, estamos construyendo una prueba de hipótesis que tiene como hipótesis nula que $H_0 : Y_{n+1} = y$. Los intervalos de predicción se derivan, como en el caso clásico de la construcción de intervalos de confianza para pruebas de hipótesis, en la inversión de la prueba:

$$C_{\text{conf}}(X_{n+1}) = \{y \in \mathbb{R} : (n+1)\Gamma(y) \leq \lfloor (1-\alpha)(n+1) \rfloor\} \tag{39}$$

Las ecuaciones (38),(39) y (40) deben repetirse cada vez que buscamos un intervalo de predicción para una observación nueva ($X_{n+1}$).Además, en la práctica, debemos restringir el número de posibles valores de $y$ a evaluar.
Una versión algorítmica del método construida en dos pasos se presenta a continuación: (Lei et al., 2018)

1. Definimos los valores iniciales para la tasa de no cobertura $\alpha \in (0,1)$, un modelo de regresión genérico $A$, los puntos de evaluación $X_{\text{trial}} = \{X_{n+1}, X_{n+2}, \ldots\}$ sobre los cuales construir el intervalo de predicción y los valores $Y_{\text{trial}} = \{y_1, y_2, \ldots\}$ que sirven de prueba.

2. Calculamos los intervalos para cada $x$ de la siguiente manera:

   - Para cada $x \in X_{\text{trial}}$ calculamos:
     - Para cada $y \in Y_{\text{trial}}$ calculamos:
       - $\hat{\mu}_y \leftarrow A\{(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)\}$
       - $S_{y,i} \leftarrow |Y_i - \hat{\mu}_y(X_i)|$, para $i = 1, \ldots, n$ y $S_{y,n+1} \leftarrow |y - \hat{\mu}_y(x)|$
       - $\Gamma(y) \leftarrow \frac{1+\sum_{i=1}^{n} 1\{S_{y,i} \leq S_{y,n+1}\}}{n+1}$
     - Calculamos: $C_{\text{conf}}(x) \leftarrow \{y \in Y_{\text{trial}} : (n+1)\Gamma(y) \leq \lfloor (1-\alpha)(n+1) \rfloor\}$
     - Repetimos para las demás $x$

**3.3.2.4. Predicción conformal partida** La generación de intervalos mediante el método conformacional reseñado anteriormente es computacionalmente intensiva: para cada punto nuevo $X_{n+1}$ se debe reestimar el modelo en un conjunto de datos aumentado con el nuevo punto $(X_{n+1}, y_{trial})$ y recalcular los rankings de los resultados de las funciones de scoring. Por estás razones, se ha desarrollado derivaciones del método conformacional que separan los procesos de estimación y ranking generando eficiencias en el costo computacional como vemos en Lei et al. (2015). Esta versión que separa en dos momentos la estimación y el ranking se observa en el siguiente algoritmo.

1. Definimos los valores iniciales para la tasa de no cobertura $\alpha \in (0, 1)$ y un modelo de regresión genérico $A$.

2. Separamos aleatoriamente las observaciones ($\{1, \ldots, n\}$) en dos subconjuntos iguales $I_1, I_2$

   - Estimamos el modelo $\hat{\mu} \leftarrow A(\{(X_i, Y_i) : i \in I_1\})$
   - Calculamos los scores $S_i \leftarrow |Y_i - \hat{\mu}(X_i)|$, para $i \in I_2$
   - Calculamos $k \leftarrow \left\lfloor \frac{n}{2} + 1 \right\rfloor (1 - \alpha)$
   - Obtenemos $d$ como el $k$-esimo valor más pequeño en $\{S_i : i \in I_2\}$
   - Construimos el intervalo de predicción como $C_{\text{split}}(x) \leftarrow [\mu_b(x) - d, \mu_b(x) + d]$, para todo $x \in \mathbb{R}^d$

**3.3.2.5. Ensambles** Un ensamble se puede considerar simplemente como la agregación de modelos individuales. Más allá de detallar el proceso de agregación, de los cuales se hablará más adelante, es importante explicar su motivación: la principal idea detrás de los ensambles es que la agregación de un número $M$ de modelos individuales puede generalizar -predecir sobre datos no observados- de mejor forma que cualquiera de los modelos constituyentes del ensamble, es decir, un ensamble puede alcanzar mayor precisión y ser más estable. Los sistemas de ensamble, por lo tanto, consisten en crear muchos modelos y combinar sus resultados de tal manera que la combinación mejore el rendimiento de un solo modelo. Sin embargo, esto requiere que los modelos individuales cometan errores en diferentes direcciones. Intuitivamente si cada modelo comete errores en diferentes direcciones (estiman distintas partes de $f(x)$) entonces una combinación inteligente de estos modelos puede reducir el error total. El principio general en los sistemas de ensamble es construir modelos lo más diferentes entre si que sea posible de modo que los errores que comentan sean diferentes. Una discusión

completa sobre las diferentes formas de construir ensambles y las implicaciones de su construcion se pueden encontrar en Polikar (2006).

**3.3.2.6. Descomposición Sesgo - Varianza** La descomposición sesgo-varianza es un concepto fundamental ya que proporciona una perspectiva sobre el compromiso entre el sesgo (la calidad del ajuste) y la varianza (la estabilidad) en el rendimiento de un modelo. Además, como veremos más adelante, los modelos de ensamble considerados explotan de forma diferente la descomposición del error en estos componentes.
Consideremos el problema de regresión donde $y = f(x)+\epsilon$, con $E[\epsilon] = 0$ y $\text{Var}(\epsilon) = \sigma_\epsilon^2$, a partir de aquí podemos derivar la expresión del error de predicción esperado de un modelo ajustado de regresión $\widehat{f(X)}$ en un punto $X = x_0$ utilizando como función de perdida el error cuadrático.
Supogamos que estimamos $f$ usando $\hat{f}$. Entonces, el error cuadrático medio esperado (MSE) para un nuevo $y$ en $x_0$ será igual a:

$$E\left[(y - \hat{f}(x_0))^2\right] \tag{40}$$

Definiremos $f = f(x_0)$ y $\hat{f} = \hat{f}(x_0)$ y dado que $f$ es determinista, $E[f] = f$ y $\text{Var}[f] = 0$.
Siendo así, expresamos el MSE (al sumar y restar $f$)

$$\begin{aligned} E\left[(y - \hat{f})^2\right] &= E\left[(y - f + f - \hat{f})^2\right] \\ &= E\left[(y - f)^2 + (f - \hat{f})^2 + 2(y - f)(f - \hat{f})\right] \\ &= \sigma^2 + E\left[(\hat{f} - f)^2\right] + 2E\left[(y - f)(f - \hat{f})\right] \end{aligned} \tag{41}$$

En primer lugar, mostramos que el tercer término es cero:

$$\begin{aligned} E\left[(y - f)(f - \hat{f})\right] &= E\left[yf - f^2 - y\hat{f} + f\hat{f}\right] \\ &= f^2 - f^2 - E\left[y\hat{f}\right] + fE\left[\hat{f}\right] \\ &= -E\left[(f + \epsilon)\hat{f}\right] + fE[\hat{f}] \\ &= -E[f\hat{f}] - E[\epsilon\hat{f}] + fE[\hat{f}] \\ &= 0 \end{aligned} \tag{42}$$

Ahora, para el segundo término, al sumar y restar $\hat{f}$ tenemos:

$$E\left[(\hat{f} - f)^2\right] = E\left[(\hat{f} - E[\hat{f}] + E[\hat{f}] - f)^2\right]$$

$$
\begin{aligned}
&= E\left[(\hat{f} - E[\hat{f}])^2\right] + (E[\hat{f}] - f)^2 + 2E\left[(\hat{f} - E[\hat{f}])(E[\hat{f}] - f)\right] \\
&= \text{Var}[\hat{f}] + \text{Bias}^2[\hat{f}] + 2E\left[(\hat{f} - E[\hat{f}])(E[\hat{f}] - f)\right]
\end{aligned} \tag{43}
$$

Donde nuevamente podemos descartar el tercer término ya que:

$$
\begin{aligned}
E\left[(y - f)(f - \hat{f})\right] &= E\left[yf - f^2 - y\hat{f} + f\hat{f}\right] \\
&= f^2 - f^2 - E\left[y\hat{f}\right] + fE\left[\hat{f}\right] \\
&= -E\left[(f + \epsilon)\hat{f}\right] + fE[\hat{f}] \\
&= -E[f\hat{f}] - E[\epsilon\hat{f}] + fE[\hat{f}] \\
&= 0
\end{aligned} \tag{44}
$$

Y finalmente, tenemos

$$
E\left[(y - \hat{f}(x_0))^2\right] = \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) + \sigma^2 \tag{45}
$$

donde

$$
\text{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0) \tag{46}
$$

El sesgo es la diferencia entre la predicción promedio del modelo y el valor objetivo real. Representa el error sistemático introducido por las suposiciones del modelo. Un modelo con un alto sesgo tiende a ser demasiado simple y no puede capturar la estructura subyacente de los datos. De igual modo, tenemos

$$
\text{Var}(\hat{f}(x_0)) = E\left[\left(\hat{f}(x_0) - E[\hat{f}(x_0)]\right)^2\right] \tag{47}
$$

La varianza es una medida de la variabilidad de las predicciones del modelo. Representa la sensibilidad del modelo a pequeñas fluctuaciones en el conjunto de datos de entrenamiento. Un modelo con alta varianza tiende a ser demasiado complejo y puede sobreajustar los datos de entrenamiento.

**3.3.2.7. Árboles de regresión** Los tres métodos principales que se mencionan a continuación se construyen -o pueden construirse- sobre árboles de regresión, por lo que es necesario mencionar cómo están compuestos.
Siguiendo el desarrollo de Friedman et al. (2001) un árbol de regresión es un método de regresión de aprendizaje supervisado que se utiliza para predecir una variable de respuesta cuantitativa $Y$ a partir de una serie de variables predictoras $X_1, X_2, ..., X_p$.

El proceso de construcción de un árbol de regresión implica la división sucesiva del espacio generado por las variables en regiones independientes que no se sobreponen. Estos son los pasos para construir un árbol de regresión:

- División del espacio generado por las variables: El espacio se divide en $J$ regiones distintas, llamadas nodos terminales u hojas, denotadas por $R_1, R_2, ..., R_J$. Cada región se define mediante un conjunto de reglas que involucran las variables predictoras.

- Predicción de la variable de respuesta: Para predecir la variable de respuesta en una región específica $R_j$, se toma el promedio de los valores observados de la variable de respuesta en las observaciones usadas, para construir el modelo, que pertenecen a esa región, denotado como $\hat{y}_{R_j}$:

$$\hat{y}_{R_j} = \frac{1}{N_j} \sum Y_{i \in R_j} \tag{48}$$

  donde $N_j$ es el número de observaciones en la región $R_j$.

El proceso de construcción del árbol de regresión sigue un enfoque recursivo y divide el espacio de variables en regiones de forma sucesiva. Para llevar a cabo estas divisiones, se utiliza un proceso de dos etapas:
Selección de la variable predictora y el punto de división: En cada etapa, se selecciona la variable predictora $X_p$ y el punto de división $s$ que generan la mayor reducción en la suma de errores cuadráticos. La división se realiza de tal manera que las observaciones que caen a la izquierda y a la derecha del punto de división tengan la menor variabilidad en la variable de respuesta (creación de grupos homogéneos en cada rama). Matemáticamente, buscamos minimizar:

$$\sum_{i:x_i \in R_1(p,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(p,s)} (y_i - \hat{y}_{R_2})^2 \tag{49}$$

donde $R_1(j, s)$ y $R_2(p, s)$ son las regiones definidas por la variable predictora $X_p$ y el punto de división $s$.
El proceso de división se repite en cada una de las dos regiones resultantes y se detiene cuando se alcanza un criterio de terminación, como un número mínimo de observaciones en cada nodo terminal o una profundidad máxima del árbol.
Los árboles de regresión se caracterizan por ser modelos de bajo sesgo y alta varianza: representan correctamente los datos usados en entrenamiento pero son incapaces de generalizar más allá de los datos de entrenamiento y son inestables respecto a los

datos usados. Por esta razón, raramente los árboles de regresión son usados de forma individual y se agregan mediante diferentes estrategias para construir modelos más robustos.

En la Figura 3 vemos una representación de un árbol de regresión. En la parte superior derecha vemos como se generan las regiones de partición a partir de las covariables $x_1$ y $x_2$. En la parte superior izquierda se muestra una partición en regiones que no puede ser generada mediante los árboles dado el proceso binario secuencial de partición el cual produce regiones que no se sobreponen. En la parte inferior izquierda se evidencia como las regiones de la parte superior derecha pueden expresarse mediante nodos de separación y nodos terminales en forma de árboles y finalmente en la parte inferior derecha vemos un gráfico de la superficie de la variable de respuesta generada por las particiones.

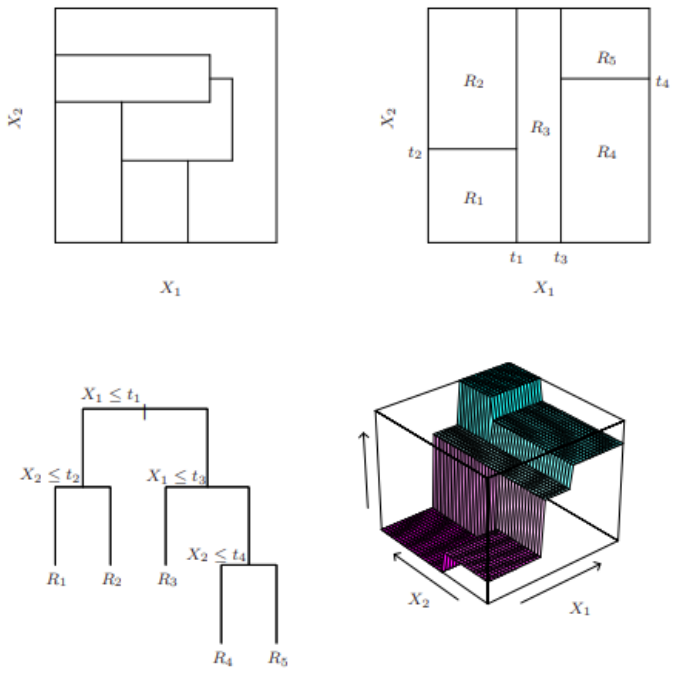


Figura 3: Árboles de regresión Hastie et al. (2009)

**3.3.2.8.** ***Quantile Random Forest.*** El modelo *Random Forest* es un ensamble de árboles de regresión. Cada árbol de regresión se construye sobre un subconjunto de los datos tomado de forma aleatoria con reemplazo (Boostrap). A su vez, cada árbol solo considera un subconjunto aleatorio de todas las covariables disponibles. Estos dos procesos están orientados a generar diversidad y producir submodelos no

correlacionados entre si. La predicción del modelo final, en el caso de regresión, es el promedio de la predicción de todos los árboles. Al promediar las predicciones de múltiples árboles, la varianza se reduce porque el ensamble es menos sensible a las fluctuaciones individuales de cada árbol. Esto ayuda a evitar el sobreajuste y mejora la estabilidad y el rendimiento del modelo en datos no vistos. Por tanto, las estrategias de boostrap de *Random Forest* buscan minimizar el error cuadrático medio a partir de minimizar la varianza. Más detalles se pueden encontrar en (Breiman, 2001).
Si partimos de un árbol individual, como hemos visto, se puede considerar la predicción como un promedio ponderado de las observaciones de la variable de respuesta que caen en los nodos terminales luego del proceso de partición. Así para un árbol y un número de observaciones $n$ comenzamos definiendo:

$$w_i(x;T) = \frac{1_{X_i \in R_\ell(x;T)}}{(\#j : X_j \in R_\ell(x;T))} \tag{50}$$

Donde:
$1_{X_i \in R_\ell(x;T)}$ es una función indicadora que vale 1 si la observación $X_i$ pertenece al nodo terminal $R_\ell(x,T)$ y 0 en caso contrario, $\#j : X_j \in R_\ell(x,T)$ cuenta el número total de observaciones en el nodo terminal $R_\ell(x,T)$.Por último, $T$ es el vector de parámetros aleatorio que representa cómo fue construido el árbol(variables usadas en los nodos, puntos concretos de corte, etc). Luego, es claro que la predicción de un árbol individual puede expresarse como:

$$\hat{\mu} = \sum_{i}^{n} w_i(X;T) Y_i \tag{51}$$

En *Random Forest*, modelamos la media condicional $E[Y|X = x]$ mediante un promedio de $K$ árboles individuales cada uno con un vector *iid* $T_t$, $t = 1, 2, \ldots, K$. Sea $w_i(X)$ el promedio de $w_i(X;T)$ sobre esta colección de árboles:

$$w_i(X) = \frac{\sum_{t=1}^{k} w_i(X;T_t)}{k} \tag{52}$$

Y de aquí se deriva que la predicción del *Random Forest* no es más que:

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(X) Y_i. \tag{53}$$

Es decir, una suma ponderada sobre todas las observaciones.
La construcción del modelo *Quantile Random Forest* en Meinshausen (2006) parte del hecho de que si el bosque aleatorio puede aproximar la media condicional $E[Y|X = x]$ a través de un promedio ponderado de la variable de respuesta entonces también posee

la capacidad de aproximar la distribución condicional completa. En este sentido, la distribución condicional de $Y|X = x$ es:

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{Y \leq y}|X = x) \tag{54}$$

y podemos usar los pesos $w_i(x)$ definidos en la ecuación (53) para aproximar la distribucion condicional como un promedio ponderado:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(X) 1_{Y_i \leq y} \tag{55}$$

A nivel práctico y de implementación, la principal diferencia entre *Random Forest* y *Quantile Random Forest* es que el último conserva, en cada árbol, todos los valores de $Y$ que caen dentro de un nodo terminal; en tanto el primero solo conserva el promedio y descarta las demás observaciones. Al usar todas las observaciones de la variable respuesta y los pesos $w_i(X)$ se construye la aproximación a la distribución condicional. Se cuenta con una implementación en la librería **quantregForest** para R (Meinshausen, 2007).

**3.3.2.9. *Grandient boosting.*** *Random Forest* se define como el promedio ponderado de los modelos individuales del ensamble (árboles) construidos en paralelo con muestras boostrap de los datos. En el caso de *boosting* el ensamble se construye de manera secuencial, entrenando en cada iteración, un modelo "débil" o base, por lo general un árbol, usando como variable dependiente el error del ensamble hasta la iteración anterior (Schapire, 2003). La formulación más estudiada de *boosting* es *gradient boosting* donde los nuevos modelos base se construyen de tal forma que estén correlacionados con el gradiente negativo de la función de pérdida asociada al ensamble en la iteración anterior.

Dado que los árboles se construyen de manera secuencial, y cada árbol se ajusta para corregir los errores cometidos por el modelo anterior en el ensamble, tenemos que en *gradient boosting* se mejora el rendimiento al explorar aquellas áreas del conjunto de datos de entrenamiento donde el modelo actual comete errores. En este sentido, el énfasis de los modelos de boosting es en la reducción del sesgo.

Dado un conjunto de datos $D = (X, y)_i^N$ donde $y$ es la variable de respuesta y $X = (x_1, x_2, \ldots, x_p)$ las variables predictoras, el objetivo del modelo de *boosting* es reconstruir una forma funcional desconocida $f(X)$ donde una función de pérdida $Ł(y, f)$ es minimizada.

$$\underset{f(X)}{\text{argmin}} Ł(y, f(X)) \tag{56}$$

En la versión de Friedman (2001) se procede inicializando $\widehat{(X)}$ a un valor constante y se siguen los pasos de iteración consecutiva:

1. Inicializamos $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
2. Calculamos secuencialmente para $m = 1$ hasta $M$:
   - Para $i = 1, 2, , \ldots, N$ calculamos los pseudo residuales:
   $$r_{im} = - \left. \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right|_{f=f_{m-1}} \tag{57}$$
   - Ajustamos un árbol de regresión a los residuos $r_{im}$, generando regiones de nodos terminales $R_{jm}$, con $j = 1, 2, \ldots, J_m$.
   - Calculamos los coeficientes $\gamma_{jm}$ que minimizan la función de pérdida dentro de cada región terminal:
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{58}$$
   - Actualizamos el modelo añadiendo el nuevo árbol:
   $$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{59}$$
3. Definimos el modelo final como:
$$\hat{f}(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{60}$$

En la práctica, la forma para construir la estimación de cuantiles condicionales a partir de *boosting* es mediante la función de pérdida: minizamos la función check que se vio en el caso de la regresión cuantílica. Un ejemplo de lo anterior se encuentra en Fenske et al. (2011) donde la derivada de la función check necesaria para calcular los pseudo-residuales puede expresarse como:

$$\frac{\partial L}{\partial u} = -\tau \cdot 1\{y \geq u\} + (1 - \tau) \cdot 1\{y < u\} \tag{61}$$

La librería **gbm** para R ofrece una implementación de GBM con función de perdida check como se ha mencionado aquí.

### 3.3.3. *Bayesian additive regression trees (BART).*

BART es una suma de árboles de regresión. Recordemos que tanto *Random Forest* como *Gradient Boosting* son igualmente sumas de árboles. En el caso de *Random Forest* los árboles se construyen en paralelo sobre muestras aleatorias de los datos. En *Gradient Boosting*, por el contrario, los árboles se construyen secuencialmente sobre los residuales del modelo agregado (la suma de árboles) hasta la iteración anterior. BART como suma de árboles hereda partes de estás dos aproximaciones: cada árbol se construye de manera aleatoria como en *Random Forest* y el desarrollo del modelo en sí es secuencial como en *Gradient Boosting*. A continuación, seguimos el desarrollo de Hernández et al. (2018).

Dada una matriz $n$ x $p$ de covariables $X$, sea $x_k = [x_{k1}, \ldots, x_{kn}]$ la k-ésima observación, es decir, la k-ésima fila de $X$. Entonces BART se define como:

$$Y_k = \sum_{j=1}^{K} f(x_k|T_j, M_j) + \epsilon_k \tag{62}$$

donde $f(x_k|T_j, M_j)$ es el j-ésimo árbol de regresión con $T_j$ la estructura del árbol (variables y puntos de separación para cada nodo interno) y con $M_j$ el vector de las predicciones en los nodos terminales asociados a $T_j$. Como en la versión reseñada de los árboles de regresión, este vector $(M_j)$ esta compuesto por los promedios de las observaciones de $Y$ que caen dentro de cada nodo terminal. Por último, se asume que $\epsilon_k \sim N(0, \sigma^2)$.

La construcción de BART comparte semejanzas a la de *Gradient Boosting* y *Random Forest*: en la primera iteración, todos los árboles se inicializan en paralelo, semejante a *Random Forest*, con un único nodo terminal donde la predicción se da por $\hat{f}_m^1(x) = \frac{1}{nM}\sum_{i=1}^{n} y_i$($m$ es el $k$-esimo árbol para la iteración 1) la media de los valores de respuesta dividida por el número total de árboles. Por lo tanto $\hat{f}^1(x) = \sum_{k=1}^{K} \hat{f}_k^1(x) = \frac{1}{n}\sum_{i=1}^{n} y_i$. En iteraciones posteriores, BART actualiza cada uno de los $K$ árboles, uno a la vez, como en *Gradient Boosting* utilizando los residuales obtenidos hasta la iteración anterior.Así, en la iteración $b$-ésima, para actualizar el $k$-ésimo árbol, restamos de cada valor de respuesta las predicciones de todos los árboles excepto el $k$-ésimo, para obtener un residual parcial:

$$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^b(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i) \tag{63}$$

para la i-ésima observación, $i = 1, \ldots, n$. Ahora bien, a diferencia de *Gradient Boosting* en lugar de ajustar un árbol nuevo a este residual parcial, BART elige aleatoriamente una perturbación al árbol de la iteración anterior $(\hat{f}_k^{b-1})$ de un conjunto de posibles

perturbaciones, favoreciendo aquellas que mejoren el ajuste al residual parcial. La clase de perturbaciones posibles están gobernadas sobre los priors que definen la forma que toman los árboles y se estudian a continuación.

### 3.3.4. BART: distribuciones a priori.

La idea detrás de la eficiencia en los modelos de ensamble está en la agregación de un número $m$ de modelos simples o débiles, cómo árboles de regresión con pocos nodos. Las máquinas de *boosting* basadas en árboles y *Random Forest* generan estos modelos débiles, por ejemplo, a través de la fijación de la profundidad de los árboles a un valor pequeño -que se suele escoger por validación cruzada-. En BART se utiliza el mismo principio; sin embargo, se logra de una forma más natural al asociar un prior que sirve de regularización sobre la profundidad (perturbación sobre la forma del árbol) y otro sobre el ajuste (perturbación sobre las predicciones en cada nodo terminal) de cada árbol $(T_j, M_j)$. Por tanto, cada modelo individual contribuye solo marginalmente al ajuste global.
En la derivación se asume que los árboles $T_j$ y los parámetros de los nodos terminales $M_j = (\mu_{11}, \ldots, \mu_{ij})$ donde $i = 1 \ldots b_j$ indexa los nodos terminales del árbol $T_j$ son independientes e idénticamente distribuidos *iid*. Con esto, se puede mostrar que

$$P(M_1, \ldots, M_m, T_1, \ldots, T_m, \sigma) \propto [\prod_j \prod_i P(\mu_{i,j}|T_j)P(T_j)]P(\sigma) \tag{64}$$

**3.3.4.1. Prior para $\mu_{i,j}$** $\mu_{i,j}$ representa la media calculada en los nodos terminales en los árboles. Dentro de la construcción de BART se hacen perturbaciones a $\mu_{i,j}$. Estás perturbaciones están gobernadas por la distribución prior para $\mu_{i,j}$ definida como $P(\mu_{i,j}|T_j) \sim N(0, \sigma_0^2)$ con $\sigma_0^2 = \frac{0.5}{e\sqrt{m}}$. Aquí $e$ es considerado un hiperparámetro que debe ser fijado. La razón detrás del prior de $\mu_{i,j}$ es que $E[Y|\boldsymbol{X}]$ es una suma de $m$ $\mu_{i,j}$ y por tanto, si centramos y escalamos $Y$ entre (-0.5, 0.5), $E[Y|\boldsymbol{X}]$ queda normalmente distribuida con media cero y desviación $\frac{0.5}{e}$, en este sentido, $e$ es el número de desviaciones estándar entre 0 y 0.5. Por ejemplo, con $e = 2$, tenemos una probabilidad del 95 % de que $E[Y|X]$ este entre (-0.5, 0.5). Esto no es más que un prior empírico bayesiano tal que $E[Y|\boldsymbol{X}]$ se encuentre dentro del rango de valores de $Y$ con alta probabilidad (Chipman et al., 2010).

**3.3.4.2. Prior para $T_j$** La estructura de los árboles definida por $P(T_j)$ es otro de los componentes que sufre perturbaciones durante la reconstruccion de uno de los árboles en el proceso de ajuste secuencial. El prior sobre los árboles para un nodo

concreto de un árbol se define como $\alpha(1+d_i)^{-\beta}$. El sentido detrás del prior es que sirve para regularizar el árbol ya que se desestimula su crecimiento con la profundidad $d_i$, y la fórmula corresponde a la probabilidad de que un nodo terminal sea divido en dos. $\alpha$ y $\beta$ son hiperparámetros que deben fijarse. Así, para todo el árbol podemos definir $P(T_j) = \prod_{i=1}^{\mu_j} \alpha(1+d_i)^{-\beta}$ con $\alpha \in (0,1)$ y $\beta \in (0,\infty)$.

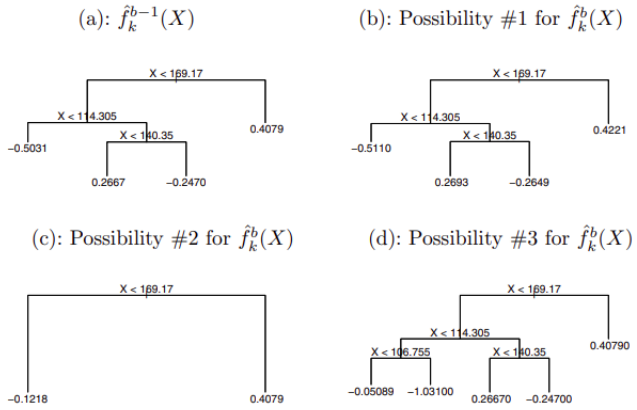


Figura 4: Perturbaciones en BART Hastie et al. (2009)

La Figura 4 esquematiza el proceso de transformación en los árboles individuales. En la Figura (a) tenemos la estructura de un árbol $k$ hasta la iteración $b-1$ a partir de allí se proponen transformaciones para el árbol: en la Figura (b) vemos transformaciones sobre las predicciones de los nodos terminales ($\mu_{i,j}$) sin transformar la estructura ($T$); en la Figura (c) encontramos una transformación sobre la forma ($T$) que se traduce en la eliminación de nodos y en la Figura (d) vemos otra transformación que implica la adición de nodos nuevos.

Finalmente, $\sigma^2$ tiene un prior conjugado dado por una inversa-$\chi^2(\lambda)$ calibrada sobre los datos. La calibración, es decir, la selección de los grados de libertad ($\lambda$), se hace sobre el supuesto de que si el verdadero modelo para $E[Y|X]$ no es lineal ni aditivo, una estimación vía regresión lineal sobreestimará la desviación estándar. Bajo esta consideración, se escogen los grados de libertad de forma que la estimación de la desviación estándar mediante BART sea inferior a la obtenida por mínimos cuadrados ordinarios (MCO) con cierta probabilidad. Chipman et al. (2010) encontraron que un buen default es escoger $\lambda$ tal que $P(\sigma < \hat{\sigma}_{MCO}) < 0.9$.

### 3.3.5. BART: Construcción.

$$P(T, M, \sigma|\boldsymbol{X}, Y) \propto [\prod_j \prod_i P(\mu_{i,j}|T_j)P(T_j)]P(\sigma)] \times P(Y|\boldsymbol{X}, T, M, \sigma) \tag{65}$$

donde $P(Y|\boldsymbol{X}, T, M, \sigma)$ es la verosimilitud para una suma de árboles. La construcción esquematica de BART se puede representar de la siguiente manera:

1. Inicializamos los árboles en paralelo:

$$\hat{f}_1^1(x) = \hat{f}_2^1(x) = \cdots = \hat{f}_K^1 = \frac{1}{nK}\sum_{i=1}^{n} y_i \tag{66}$$

2. Calculamos el modelo para la primera iteración como:

$$\hat{f}^1(x) = \sum_{k=1}^{K} \hat{f}_k^1(x) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{67}$$

3. Para las iteraciones de $b = 2, \ldots, B$:

   - Para los árboles $k = 1, 2, \ldots, K$:

     *a*) Para las observaciones $i = 1, \ldots, n$ calculamos los residuales parciales

$$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^b(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i) \tag{68}$$

     *b*) Estimamos un nuevo árbol, $\hat{f}_k^b(x)$, a los residuales $r_i$ mediante la perturbación aleatoria del $k$-ésimo árbol de la iteración anterior $\hat{f}_k^{b-1}(x)$ En el proceso de perturbación se favorecen las perturbaciones que mejoran el ajuste y son controlados por los priors ya reseñados.

   - Calculamos el modelo para la iteración como:

$$\hat{f}^b(x) = \sum_{k=1}^{K} \hat{f}_k^b(x) \tag{69}$$

4. Finalmente, luego de $L$ muestras de convergencia, calculamos la predicción como:

$$\hat{f}(x) = \frac{1}{B-L}\sum_{b=L+1}^{B} \hat{f}^b(x) \tag{70}$$

El proceso anterior puede ser visto como una Cadena de Markov Monte Carlo. El proceso preciso de estimación se sale del alcance de este trabajo pero puede encontrarse con detalle en el artículo original de Chipman et al. (2010).

### 3.3.6. BART: intervalos de predicción

En el caso de BART las muestras construidas mediante iteraciones de Monte Carlo Cadenas de Markov proveen una aproximación a la distribución posterior de $f(Y|X)$. En este sentido, se pueden contruir intervalos asociados a la incertidumbre de una predicción mediante los cuantiles de las muestras tomadas de la posterior. Estos son los intervalos alrededor de $E[Y|X]$ o intervalos de credibilidad en la literatura bayesiana. Para el caso de predicciones individuales, $f(x_{n+1})$, se pueden construir intervalos de predicción combinando la incertidumbre alrededor de la función de esperanza condicional con el error asumido en el modelo $\epsilon$. En la práctica esto se hace simulando $B$ muestras de la distribución predictiva posterior y tomando los cuantiles deseados. Esto se da en el últimos paso del proceso de predicción de BART donde en lugar de calcular un promedio tomamos percentiles de las muestras $\hat{f}^{1+L}(x), \ldots, \hat{f}^{B}(x)$
Encontramos la librería **bartMachine** (Kapelner y Bleich, 2013) para R que lleva a cabo precisamente este procedimiento en la estimación de intervalos de predicción.

# 4. Estudio de simulación

A continuación se desarrolla el estudió de simulación. El objetivo es analizar la capacidad de los diferentes métodos en la construcción de intervalos de predicción bajo los criterios de evaluación definidos en las ecuaciones (16),(17) y (18). Adicionalmente se valida la robustez de los métodos en su capacidad predictiva mediante el uso del *Rsq* definido como el cuadrado de la correlación lineal (Pearson) entre los valores predichos y los valores verdaderos. Se tomó como criterio del valor predictivo está métrica ya que permite observar la fuerza de la relación de las predicciones y los valores reales sin tener en cuenta las diferencias en la escala de los datos. Todas las simulaciones fueron desarrollados usando el Software estadístico R.

## 4.1. Metodología

Se consideran 3 tipos de escenarios principales. Estos escenarios se definen por el tipo de variación en la generación de los datos: funcional, de volumen y del componente aleatorio. Es decir, los métodos presentados buscan evaluar la capacidad de generar intervalos de predicción bajo variaciones en el volumen de datos, efectos de la forma funcional (presencia de interacciones fuertes) y la participación del componente aleatorio (heterocedasticidad). La escala del estudio de simulación se define en relación a la escala de estudios anteriores donde se evalúa BART en relación a los demás modelos de ensamble como vemos en Chipman et al. (2010); a su vez, se evita el uso de muestras pequeñas ya que los modelos utilizados tienen sentido sobre muestras de datos mayores: esto tiene sentido en relación al uso de estrategias de construcción como boosting o bagging. En este sentido, la combinación total de escenarios contempla 4 variaciones funcionales, 3 variaciones en la forma del componente aleatorio y 3 variaciones en el tamaño de muestra (100, 1000 y 10000 observaciones). Para un total de 36 escenarios contemplados por método. Las especificaciones funcionales escogidas parten tomando como guías los trabajos de simulación para la evaluación de intervalos de predicción de Pevec y Kononenko (2015) y Kumar y Srivistava (2012).

## 4.2. Escenarios

A continuación se describen los escenarios desde la construcción de los datos.

1. **Modelo Lineal**: Conjunto de datos generado a partir de una función lineal. La función lineal es de la forma

$$y = 2x + 1 \tag{71}$$

2. **Lineal a tramos**: Conjunto de datos generado a partir de una función lineal a tramos con tres segmentos. La función se define como:

$$y = \begin{cases} 3x, & \text{si } x \leq \frac{1}{3} \\ -3x + 2, & \text{si } \frac{1}{3} < x \leq \frac{2}{3} \\ x + 1, & \text{si } \frac{2}{3} < x \end{cases} \tag{72}$$

3. **Modelo No Lineal**: Conjunto de datos generado a partir de una función no lineal. La función se define como

$$y = e^{x_1} + x_2 x_3^2 + \log(|x_4 + x_5|) \tag{73}$$

donde $x = (x_1, x_2, x_3, x_4, x_5)$ se toman de una normal multivariada con vector de medias con $\mu = [0.1, 0.2, 0, 0.05, 1.2]$ y con matriz de covarianzas dada por

$$\Sigma = \begin{pmatrix} 1.00 & 0.43 & 0.45 & -0.29 & -0.69 \\ 0.43 & 1.00 & 0.25 & -0.36 & -0.36 \\ 0.45 & -0.36 & 1.00 & -0.91 & -0.36 \\ -0.29 & -0.36 & -0.91 & 1.00 & 0.49 \\ -0.69 & -0.36 & -0.36 & 0.49 & 1.00 \end{pmatrix}$$

4. **La función de prueba de Friedman de 5 dimensiones**: Conjunto de datos generado a partir de la función no lineal de Friedman para 5 dimensiones:

$$y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \tag{74}$$

### 4.2.1. Transformaciones

Para cada uno de los escenarios la función fue modificada bajo las siguientes transformaciones:

1. La función incorpora errores aditivos normales estándar: $y = f(X) + \epsilon$ con $\epsilon \sim N(0, 1)$

2. La función incorpora errores aditivos uniformes: $y = f(X) + \epsilon$ con $\epsilon \sim U(-1, 1)$

3. La función incorpora errores aditivos exponenciales de la forma: $y = f(X) + \epsilon$ con $\epsilon \sim exp(1) - 1$

## 4.3. Resultados

El código utilizado se puede encontrar en el siguiente enlace: Implementación y análisis.Las lineas rojas de los gráficos de cobertura representan el porcentaje de cobertura deseado.

### 4.3.1. Modelo Lineal

**4.3.1.1. Evaluación del Desempeño** En el caso del conjunto de datos lineal podemos observar, en la Figura 5, que el desempeño entre los modelos es semejante y que sigue un patrón compartido en relación al efecto del término de error: en todos los modelos, el error impacta negativamente cuando pertenece a la familia exponencial o normal lo cual es de esperarse dado el rango de la variable $Y$. A su vez, es claro que el modelo que presenta un peor desempeño es el *Random Forest*. Esto tiene sentido ya que para el caso de una sola covariable $x$ *Random Forest* termina compuesto de árboles de regresión que solo crean funciones a tramos sobre $x$ (las cuales, cuando el verdadero modelo es lineal, son ineficientes). En tanto BART, GBM, y QR tienen un comportamiento semejante al modelo de regresión lineal, que para este caso, representa correctamente la función usada para modelar los datos.

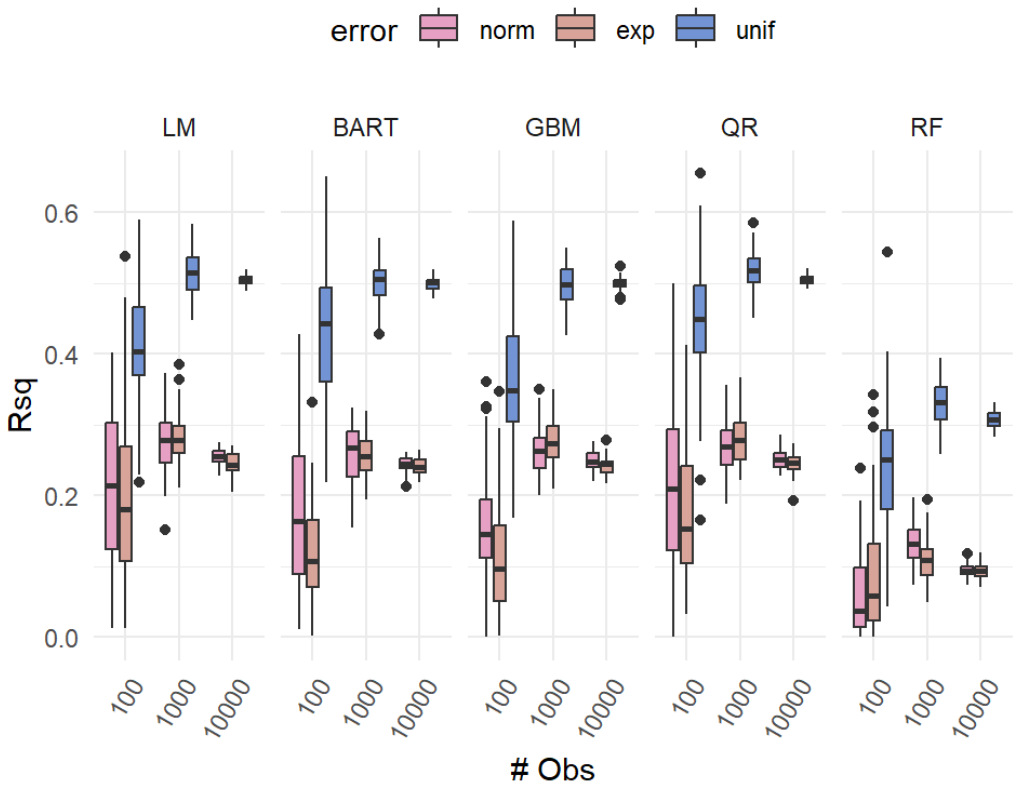


Figura 5: RSQ para el conjunto de datos lineal

**4.3.1.2. Evaluación de Cobertura** En el caso de la cobertura al 90 %, únicamente los modelos BART y LM alcanzan a cubrir el intervalo de predicción deseado en la mayoría de los casos. Los errores normales y exponenciales impactan negativamente el PIPC. En la mayoría de los modelos hay mayor precisión en la construcción del PIPC a medida que aumenta el volumen de datos usado. El comportamiento del *Random Forest* se explica en la medida de que la construcción de los intervalos está ligada un proceso predictivo, por tanto, la incapacidad de representar correctamente la función objetivo impacta negativamente en los intervalos construidos. Este comportamiento es de esperarse para los modelos y difiere en el caso de la combinación entre en *Random Forest* y el método conformacional ya que el proceso de construcción de intervalos en este caso no se deriva directamente del proceso de construcción del modelo en sí. Es importante destacar que para el método conformacional hay una garantía en el método de dar intervalos correctos, en cobertura, de forma explicita.

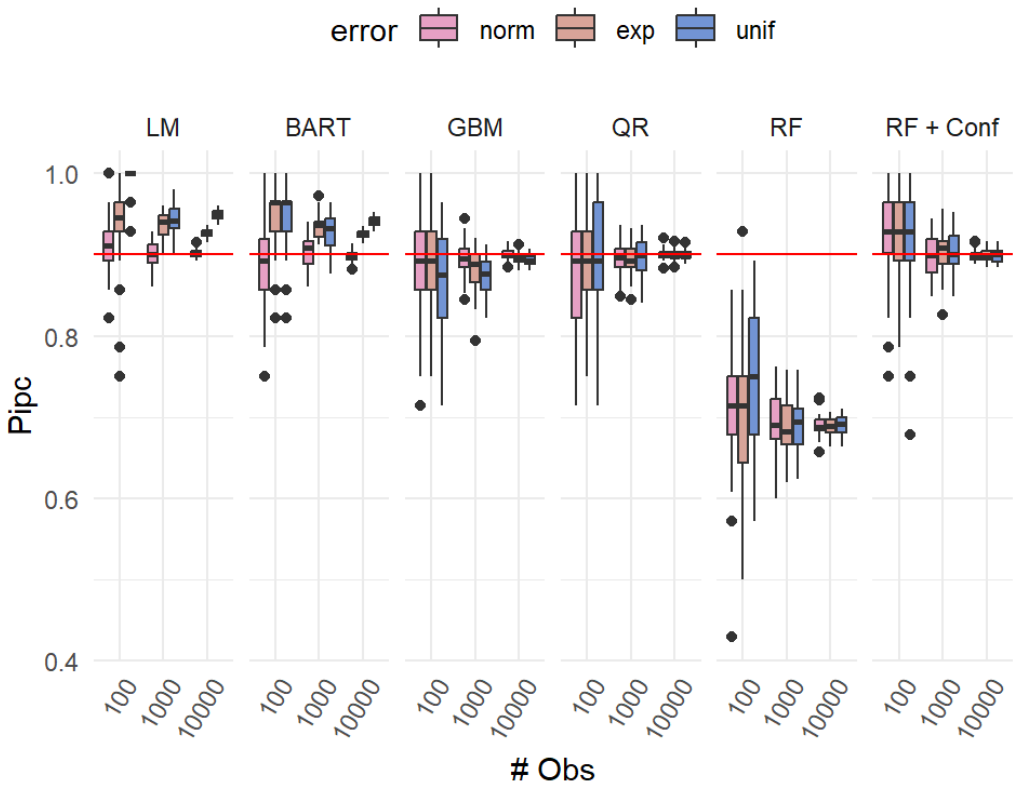


Figura 6: Cobertura al 90 % para el conjunto de datos lineal

En el caso de la cobertura al 95 % vemos que se repiten los patrones para la cobertura al 90 %. Sin embargo, hay un deterioro en los modelos GBM y QR.

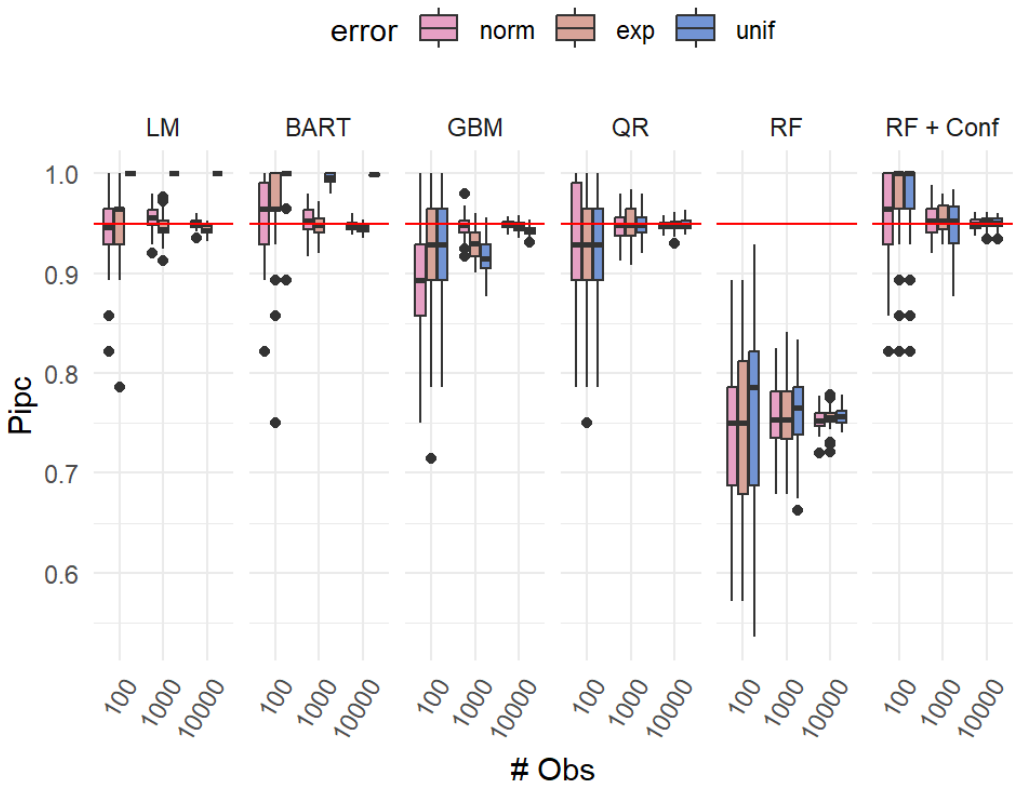


Figura 7: Cobertura al 95 % para el conjunto de datos lineal

**4.3.1.3. Evaluación ancho de Intervalos** El ancho de los intervalos fue medido mediante la métrica $CWC$ y la posición relativa de cada modelo bajo esta métrica. Como $CWC$ es función de $\eta$ observamos la evolución de la posición relativa de cada modelo respecto a $CWC$ a medida que cambia el valor de $\eta$. Recordemos que $\eta$ es una penalización sobre la cobertura: a medida que $\eta$ crece preferimos intervalos que garanticen la cobertura y penalizamos aquellos que no la cumplen en mayor medida. La posición relativa es el lugar que ocupa un modelo entre los demás a medida que $\eta$ se hace más restrictiva. Es decir, indica qué modelos mantienen el mejor ancho al subir el nivel de exigencia frente la cobertura, después de todo, es trivial tener un ancho corto al no exigir, por ejemplo, ninguna cobertura. En este caso, podemos observar como el modelo BART se mantiene entre los mejores tres modelos (con excepción de la combinación de 10000 observaciones y errores distribuidos normales) para el caso de los conjuntos de datos generados linealmente. Compite en este sentido, con el modelo lineal y la regresión cuantílica.

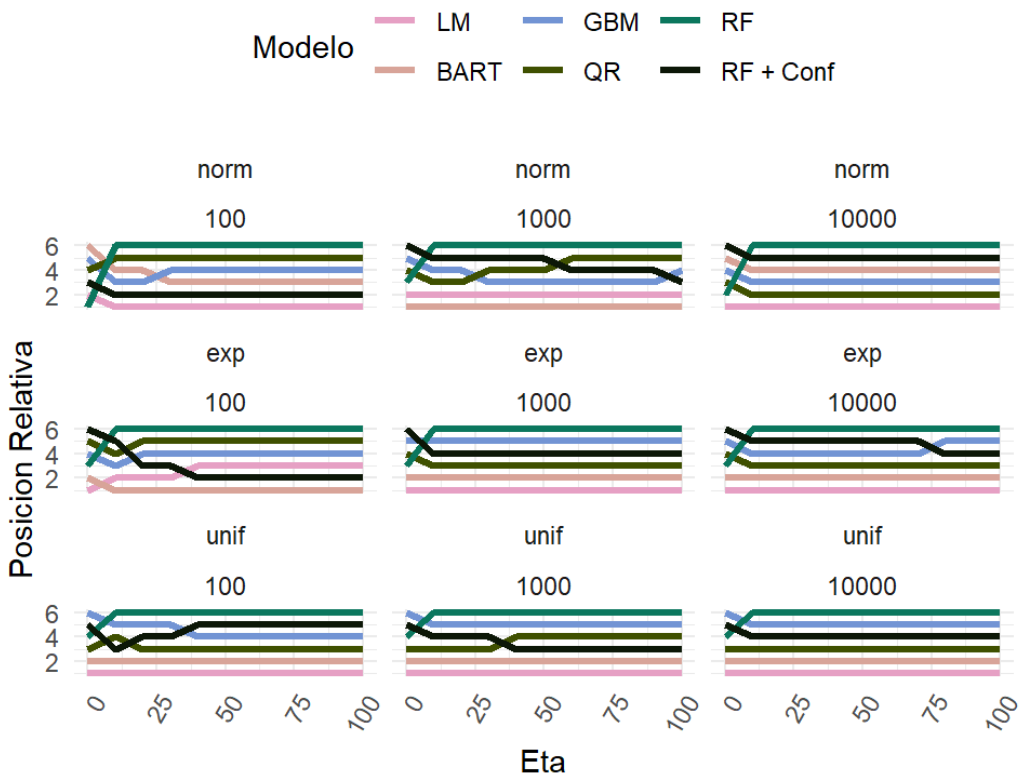


Figura 8: Ancho intervalos al 90 % para el conjunto de datos lineal

Cuando observamos al 95 % las conclusiones se mantienen idénticas.

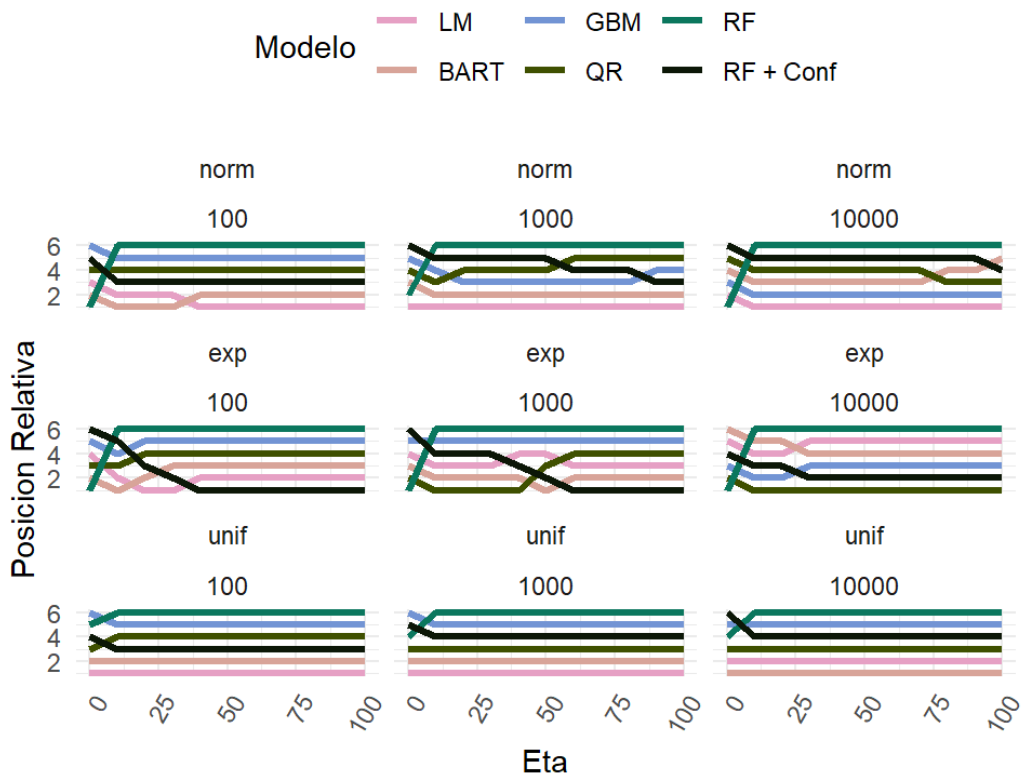


Figura 9: Ancho intervalos al 95 % para el conjunto de datos lineal

### 4.3.2. Modelo lineal a tramos

**4.3.2.1. Evaluación del Desempeño** Los modelos que presentaron mejor desempeño en el caso del conjunto de datos generado con la función lineal a trozos fueron aquellos basados en árboles: BART, RF, GBM. La estructura de los modelos base de estos ensambles está, como se ha referenciado anteriormente, especialmente diseñada para explorar el espacio de covariables mediante particiones a trozos. En el caso de la regresión lineal y la regresión cuantifica el modelo queda subespecificado: bastaría con una covariable a tramos adicional sobre el rango de $x$ para mejorar su desempeño. Adicionalmente, los modelos como BART y GBM (ambos procesos de boosting) tienen mejores resultados de forma consistente con el incremento en el volumen de datos. Se observa un impacto negativo de la presencia de errores aditivos de forma exponencial o normal.

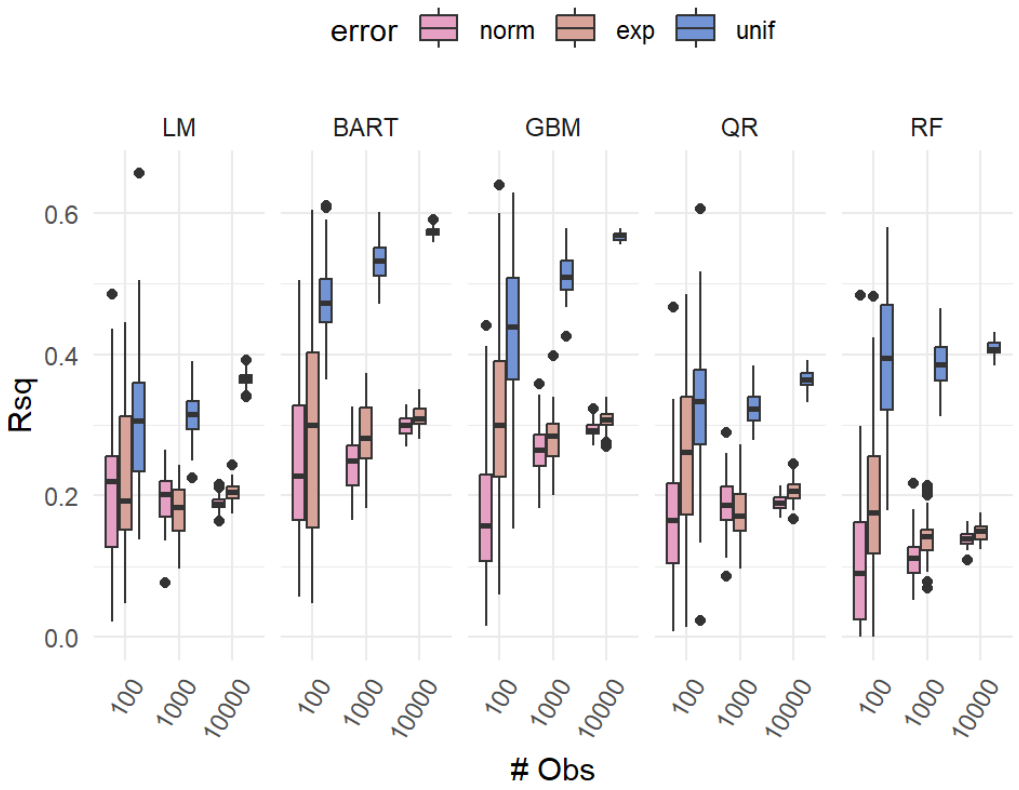


Figura 10: RSQ para el conjunto de datos a tramos

**4.3.2.2. Evaluación de Cobertura** En el caso de la cobertura al 90 %, los modelos BART y LM y la combinación del modelo RF con el método conformacional alcanzan a cubrir el intervalo de predicción deseado en la mayoría de los casos. Los errores uniformes y exponenciales impactan negativamente el PIPC en el caso del modelo lineal y el GBM pero su efecto parece atenuado en el caso del modelo BART donde las diferencias apreciables sobre el impacto de los errores están vinculadas al volumen de datos. En todos los modelos hay mayor precisión en la construcción del PIPC a medida que aumenta el volumen de datos usado.

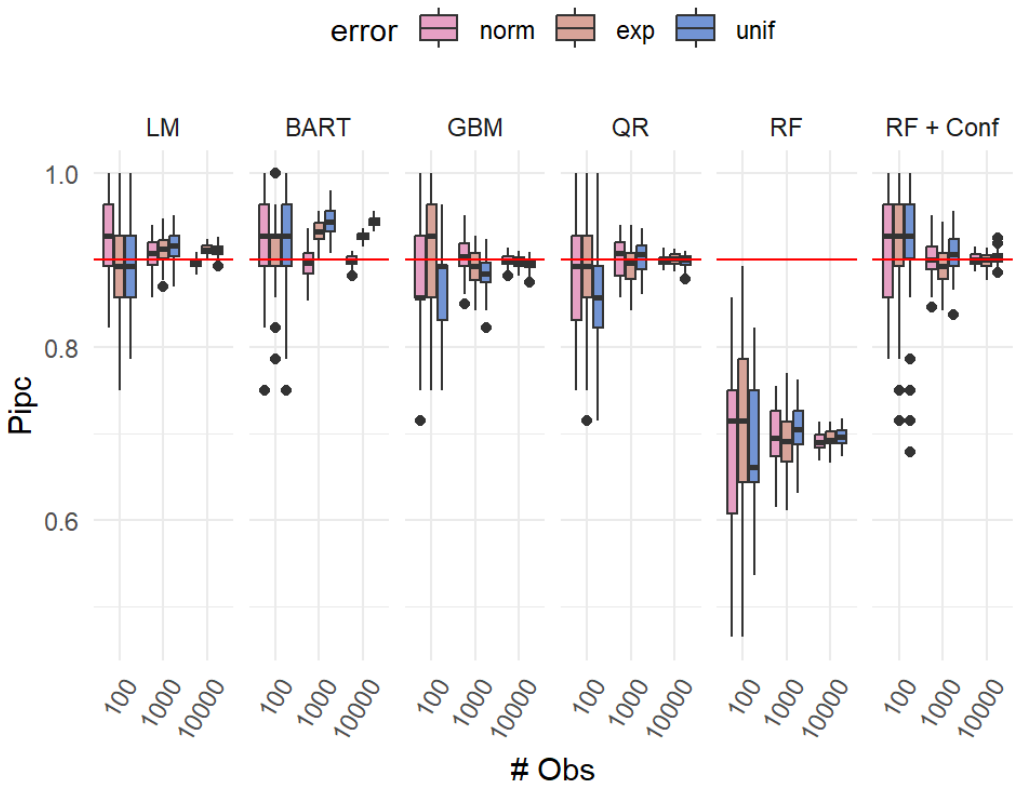


Figura 11: Cobertura al 90 % para el conjunto de datos a tramos

En el caso de la cobertura al 95 % vemos que se repiten los patrones para la cobertura al 90 %. Sin embargo, hay un deterioro en los modelos y solo el método comformacional garantiza la cobertura al 95 % en la mayoría de los casos.

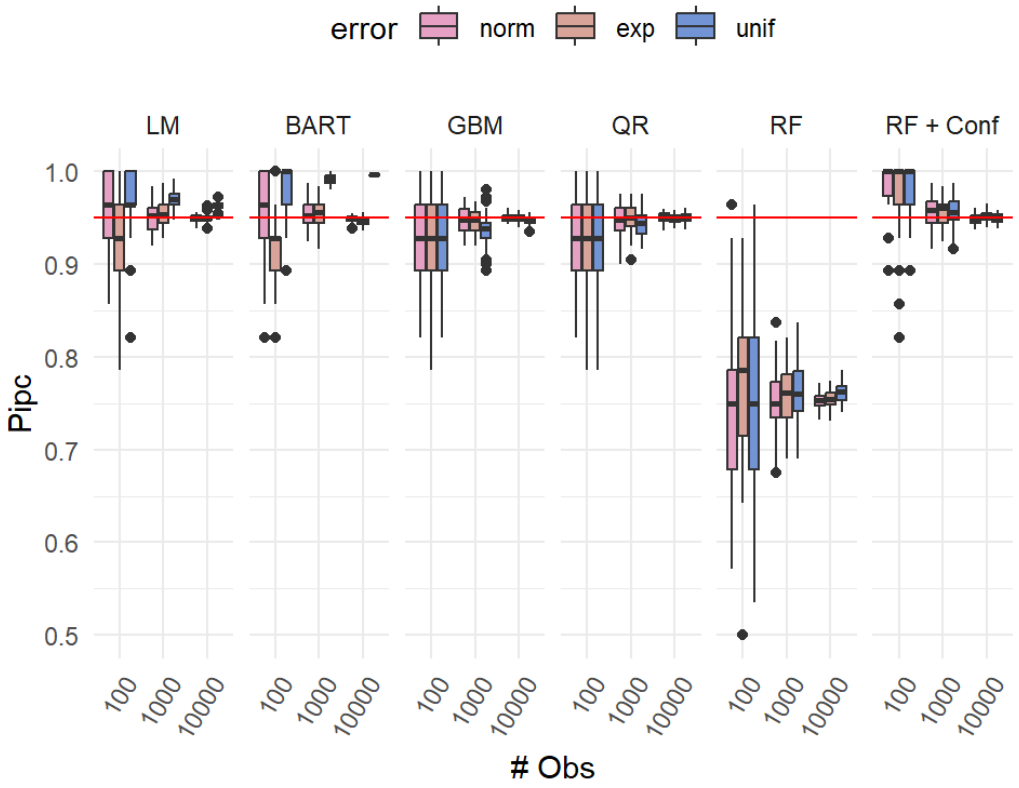


Figura 12: Cobertura al 95 % para el conjunto de datos a tramos

**4.3.2.3. Evaluación ancho de Intervalos** En el caso de los intervalos al 90 % podemos observar como el modelo BART se mantiene entre los mejores tres modelos para el caso de los conjuntos de datos generados a tramos. Es el mejor modelo bajo cualquier valor posible de $\eta$ y destaca principalmente bajo la errores aditivos de forma exponencial y uniforme. De cerca encontramos al modelo lineal. Recordemos que $CWC$ penaliza el ancho de los intervalos (certeza) y la validez, por ejemplo, la posición del método comformacional se puede deber a esto: intervalos validos (vistos anteriormente en la cobertura) pero cuyo ancho es mucho mayor que los generados por BART o LM.

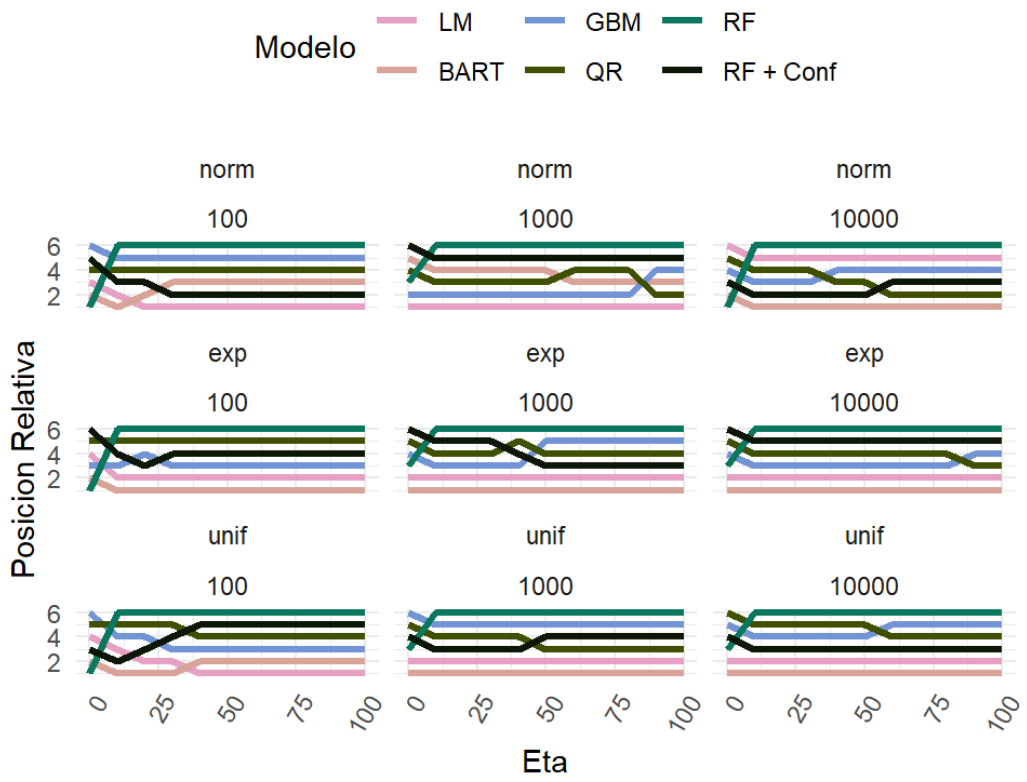


Figura 13: Ancho intervalos al 90 % para el conjunto de datos a tramos

En el caso de la cobertura buscada al 95 % el modelo GBM es superior en los casos de 10000 observaciones y errores con distribuidos normales o exponenciales. Sin embargo, BART se mantiene entre los mejores modelos para distribuciones uniformes bajo cualquier combinación de observaciones.

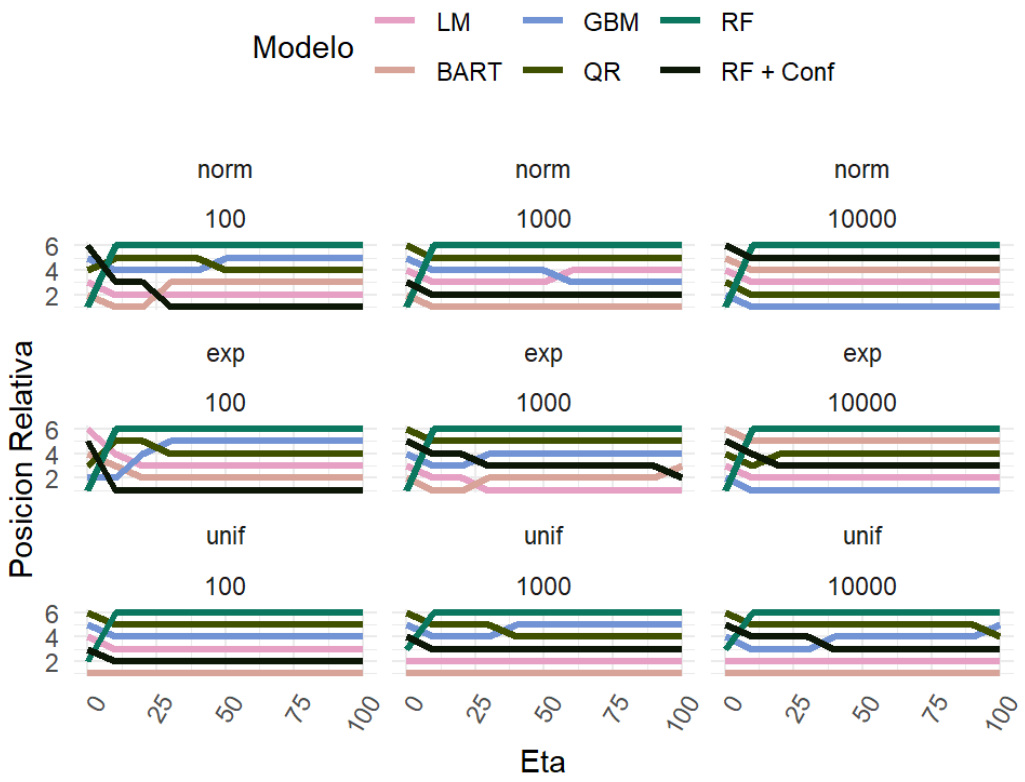


Figura 14: Ancho intervalos al 95 % para el conjunto de datos a tramos

### 4.3.3. Modelo no Lineal

**Evaluación del Desempeño** Para el conjunto de datos multivariado no lineal los modelos con mejor desempeño fueron BART y RF. La capacidad predictiva de los modelos aumenta con el numero de observaciones tanto para BART como RF, incrementa pero luego se satura para GBM, y carece de efecto sobre LM y QR. En los modelos el error impacta negativamente cuando pertenece a la familia exponencial y su efecto es casi indistinguible cuando pertenece a las familias uniforme o normal. El desempeño de los modelos QR y LM se debe a un problema de subespecificación: una aproximación lineal simple a una función no lineal.

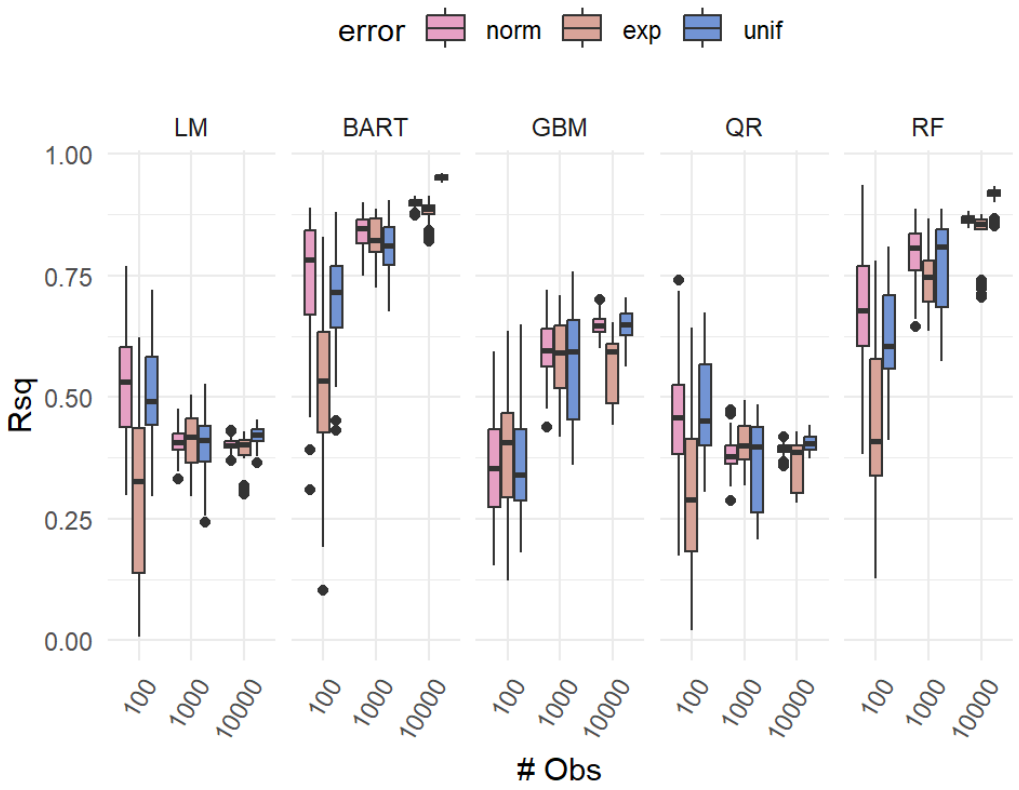


Figura 15: RSQ para el conjunto de datos no lineal

**4.3.3.1. Evaluación de Cobertura** En el caso de la cobertura al 90 %, los modelos BART, LM, RF y la combinación del modelo RF con el método comformacional alcanzan a cubrir el el intervalo de predicción deseado en la mayoría de los casos. Vemos un deterioro para el modelo QR en el caso de datos de las 100 observaciones. Los errores uniformes y exponenciales impactan negativamente el PIPC en el caso del modelo lineal y BART pero tiene un efecto imperceptible en modelos como GBM, QR y RF con el método comformacional. En todos los casos, a medida que aumenta el volumen de datos la precisión en la construcción del PIPC aumenta.

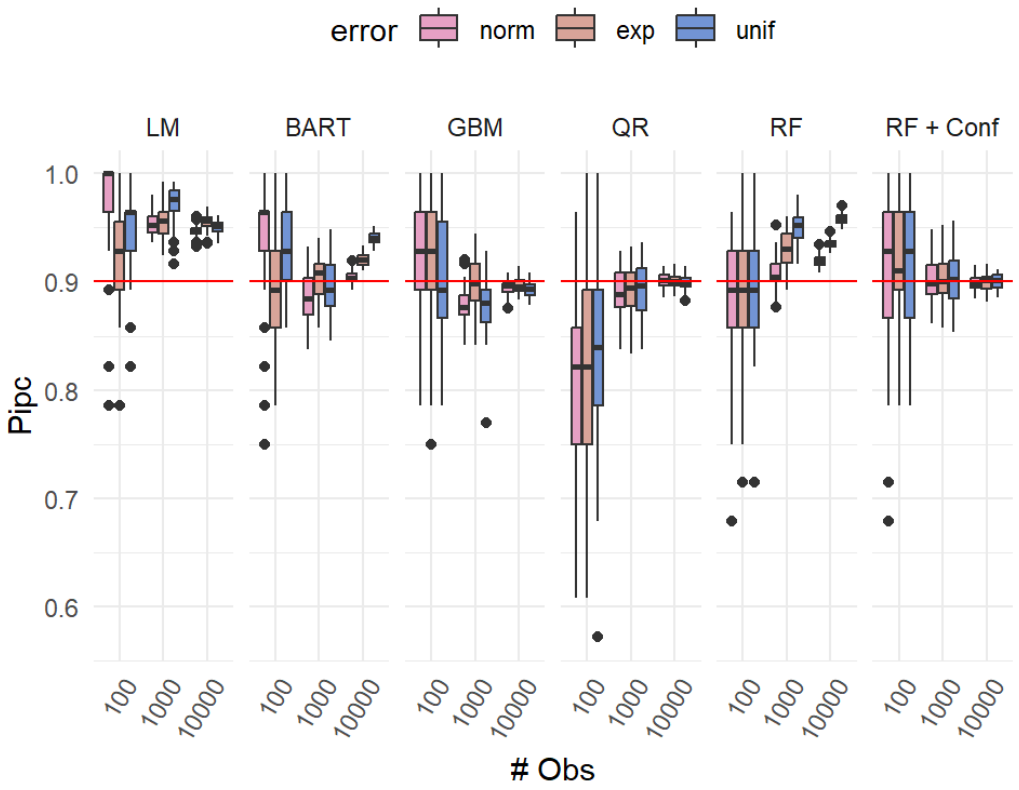


Figura 16: Cobertura al 90 % para el conjunto de datos no lineal

Para la cobertura al 95 % vemos que se repiten los patrones de la cobertura al 90 % respecto al volumen de datos.El efecto de los errores se diluye en el caso del modelo BART y los modelos mantienen su desempeño general con deterioros en el caso de GBM y RF.

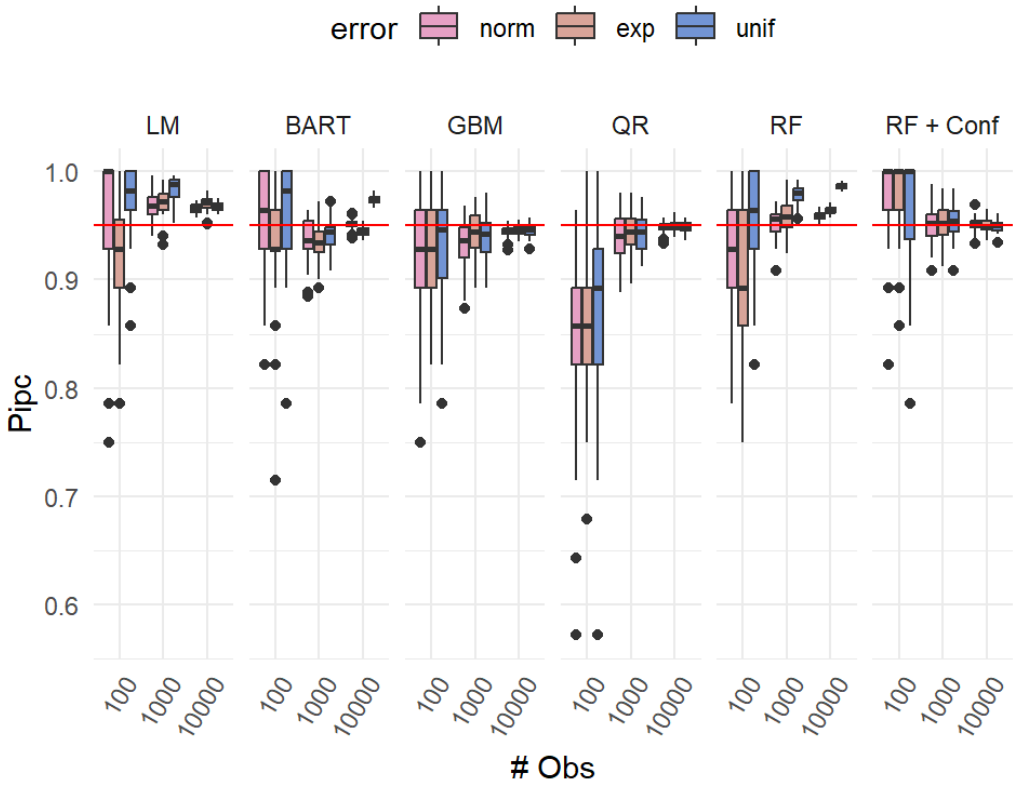


Figura 17: Cobertura al 95 % para el conjunto de datos no lineal

**4.3.3.2. Evaluación ancho de Intervalos** Para una cobertura del 90 % el LM presenta de los mejores desempeños cuando tenemos pocas observaciones para los errores normales, exponenciales y uniformes. Sin embargo, cuando el numero de observaciones aumenta los dos mejores modelos con los que contamos son BART y RF. Esto coincide con el hecho de que la generación de intervalos correctos y precisos se vincula a la capacidad de aproximar correctamente la función subyacente, la cual mejora en el caso de modelos como RF y BART a medida que aumenta el tamaño de muestra

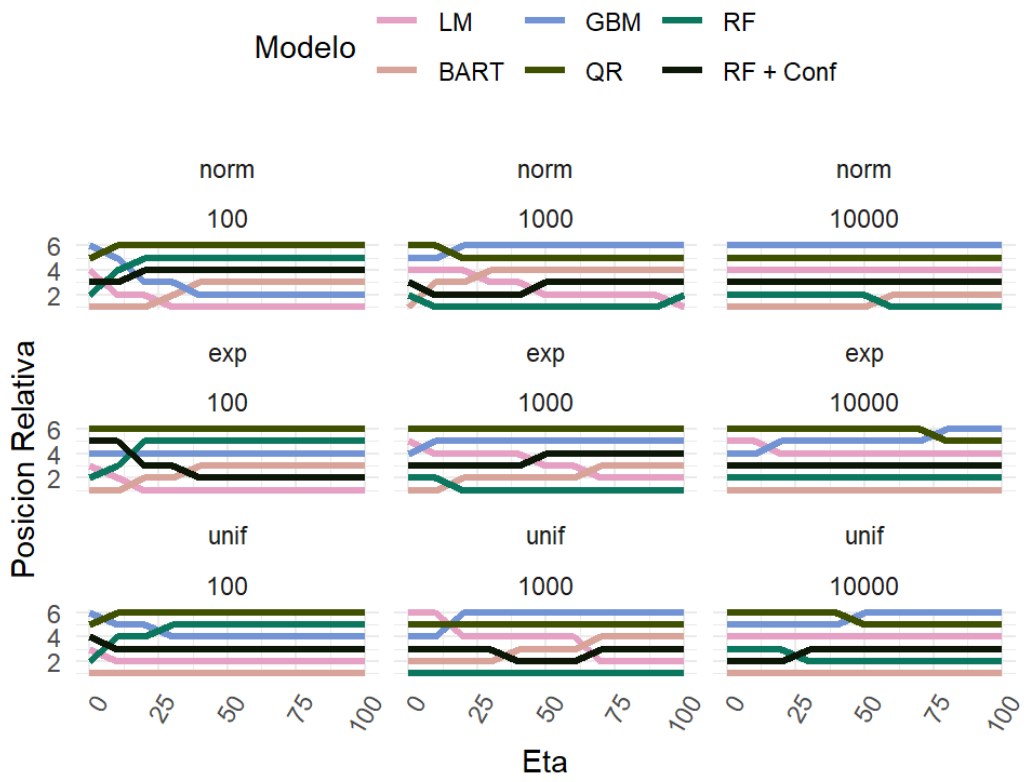


Figura 18: Ancho intervalos al 90 % para el conjunto de datos no lineal

Cuando la cobertura es del 95 % el LM pierde frente al modelo comformacional para casos de pocas observaciones con errores normales y exponenciales. El modelo RF sigue siendo dominando en el casos de 1000 y 10000 observaciones de forma independiente al error. BART se mantiene entre los mejores 3 modelos cuando se observan en conjunto las diferentes combinaciones especialmente para 10000 observaciones.

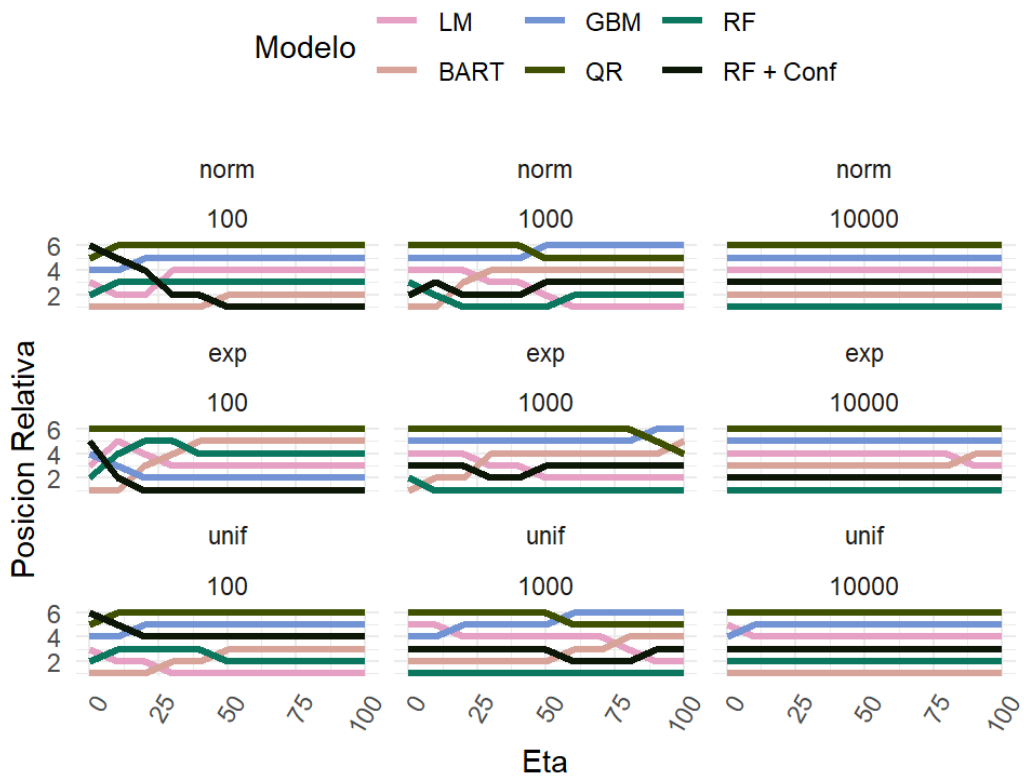


Figura 19: Ancho intervalos al 95 % para el conjunto de datos no lineal

### 4.3.4. Ecuación de Friedman

**4.3.4.1. Evaluación del Desempeño** Los modelos con mejor desempeño fueron BART y RF. La capacidad predictiva, tanto en precisión como en desempeño, de los modelos aumenta con el numero de observaciones tanto para BART y RF; incrementa pero luego se satura para GBM, y carece de efecto sobre LM y QR. En los modelos, el error impacta negativamente cuando pertenece a la familia exponencial o normal. El desempeño de los modelos QR y LM nuevamente obedece a un problema de subespecificación.

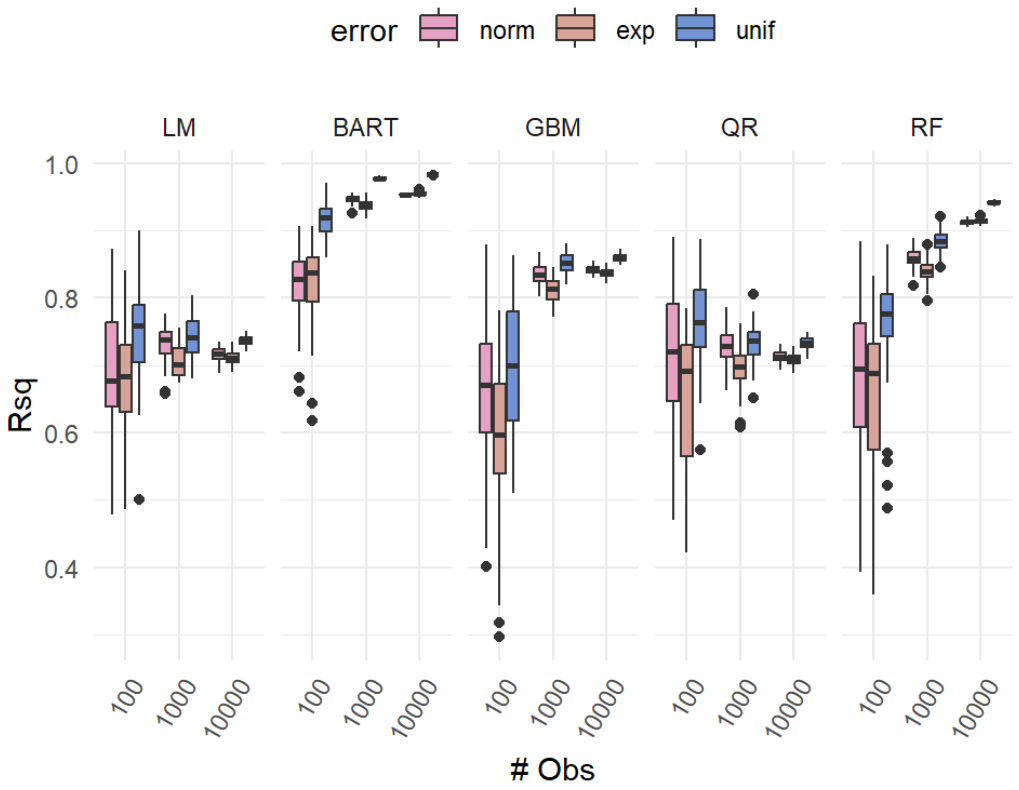


Figura 20: RSQ para el conjunto de datos de Friedman

**4.3.4.2. Evaluación de Cobertura** En el caso de la cobertura al 90 %, los modelos LM, BART, RF y la combinación del modelo RF con el método conformacional alcanzan a cubrir el intervalo de predicción deseado en la mayoría de los casos. El efecto de la distribución de los errores es mixto en los modelos BART y LM y despreciable en los demás modelos. En todos los casos, a medida que aumenta el volumen de datos la precisión de en la construcción del PIPC aumenta y en el caso de RF vemos que a medida que aumenta el volumen de datos también aumenta el promedio de cobertura como tal.

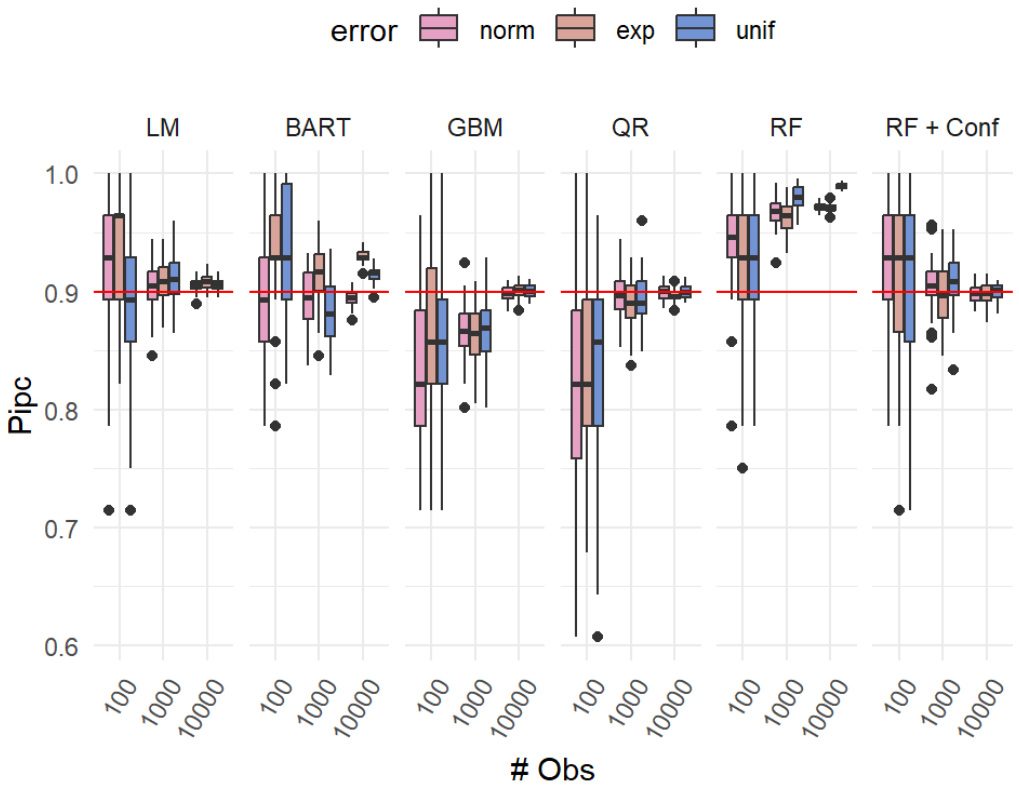


Figura 21: Cobertura 90 % para el conjunto de datos de Friedman

En el caso de la cobertura al 95 % los mejores modelos son el LM, BART y el metodo comformacional. El efecto de los errores es mixto en los modelos LM y BART e indiferente en los demás casos.

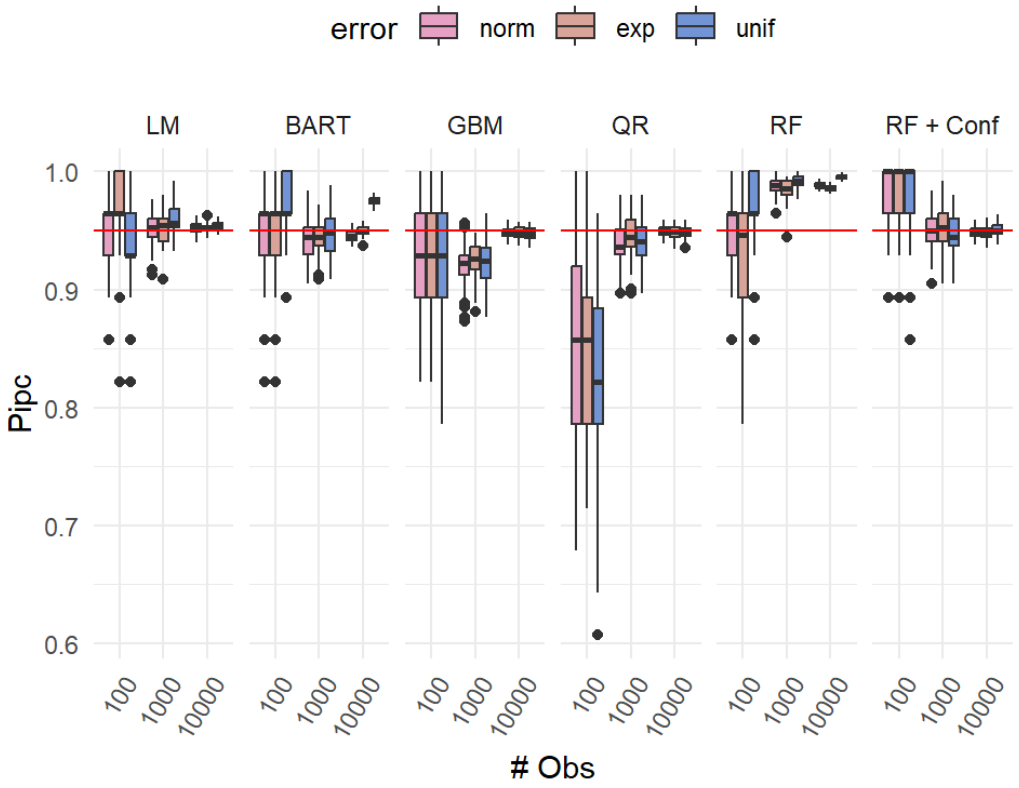


Figura 22: Cobertura al 95 % para el conjunto de datos de Friedman

**4.3.4.3. Evaluación ancho de Intervalos** Al evaluar para la cobertura al 90 % vemos que los mejores modelos son BART y RF que se mantienen entre los mejores 2 modelos en la mayoría de la combinaciones. Los modelos QR y GBM presentan los peores resultados.

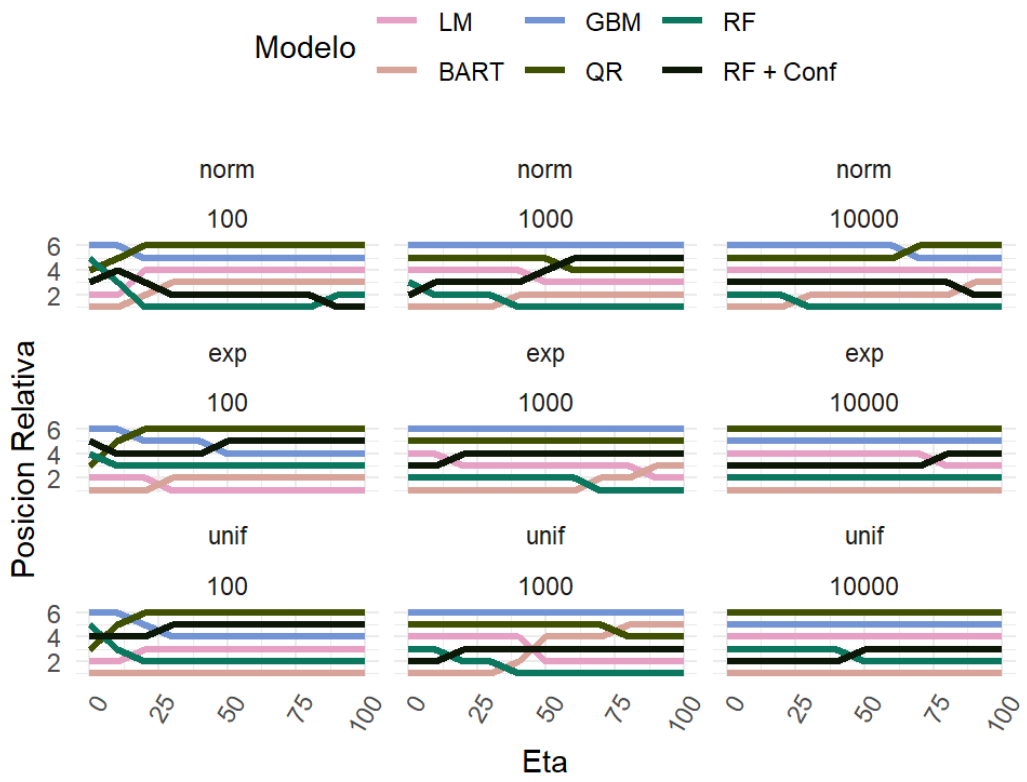


Figura 23: Ancho intervalos al 90 % para el conjunto de datos de Friedman

En el caso del 95 % de cobertura los resultados son semejantes donde BART y RF dominan en la mayoría de los escenarios pero vemos una fuerte mejora del modelo RF con el método comformacional que pasa a estar entre los primeros 3 modelos en casi cualquier caso.

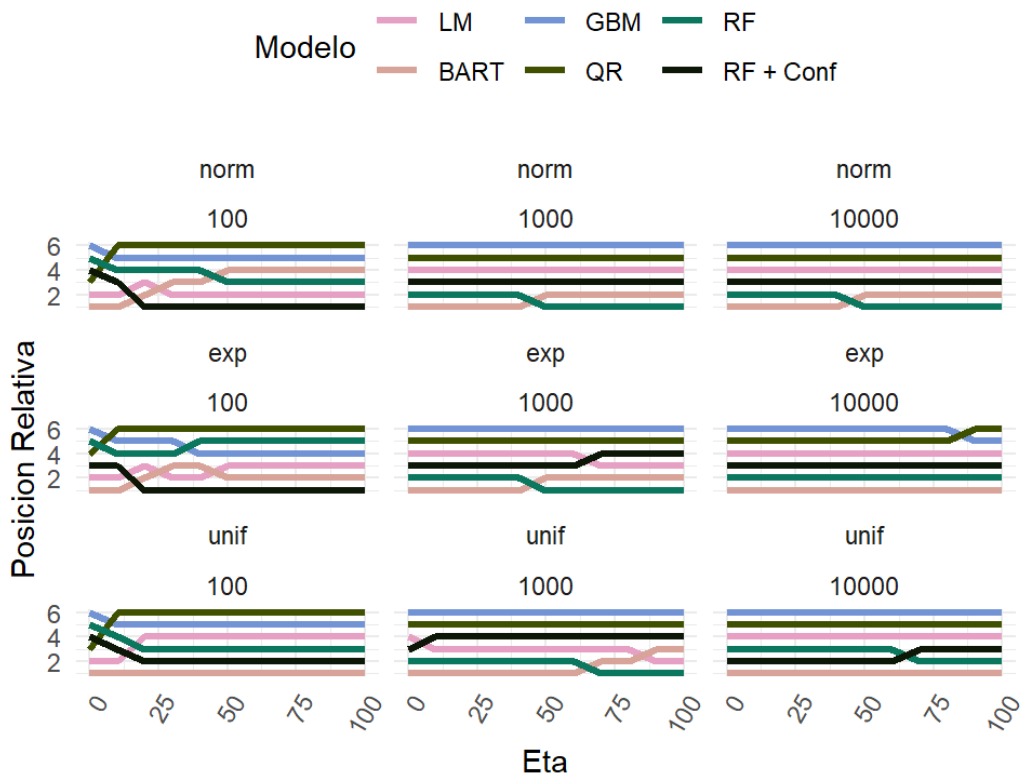


Figura 24: Ancho intervalos al 95 % para el conjunto de datos de Friedman

# 5. Caso aplicado

A continuación presentamos un caso de uso como validación complementaria al ejercicio de simulación. Las conclusiones se presentan al final del apartado haciendo uso de los resultados de los dos momentos.
El caso de uso está vinculado a la modelación de la resistencia a la compresión del cemento, es decir, a la resistencia a fuerzas que tratan de reducir su tamaño o compactarlo. La resistencia a la compresión es la medida más importante de calidad en la producción de concreto ya que determina la capacidad del material para soportar cargas y mantener la integridad de las estructuras. Además, el concreto es un material complejo donde el modelamiento de su comportamiento ha demostrado ser una tarea complicada por la presencia de relaciones no lineales e interacciones complejas en su construcción. Esfuerzos para modelar la resistencia a la comprensión del cemento mediante métodos del aprendizaje de maquinas los encontramos en Yeh (1998) con el uso de redes neuronales y mas recientemente vemos una comparación de diferentes modelos en Chou et al. (2011). Sin embargo, estos métodos solo se han evaluado en su capacidad predictiva y no hay discusiones vinculadas a las medidas de incertidumbre de las predicciones que puede ser críticas en estos dominios de aplicación.
Los datos usados son los mismos que los artículos referenciados obtenidos inicialmente por Yeh (1998). Se accedió a ellos directamente del repositorio de datos de la Universidad de California en Irvine donde fueron donados inicialmente por I-Cheng Yeh: Conjunto de Datos

## 5.1. Descripción de los datos utilizados

El conjunto de datos utilizado es un conjunto de datos diseñado (mediante recolección experimental) para predecir la resistencia a la compresión del concreto en función de la mezcla de sus componentes y la edad o tiempo de secado. Este conjunto de datos consta de 1.030 registros y 9 variables, incluida la variable objetivo que representa la resistencia a la compresión. Las 8 variables restantes son variables independientes que describen la composición y la edad del concreto.
Las variables independientes en el conjunto de datos son las siguientes:

- Cemento (cement): cantidad de cemento (componente principal del concreto) en la mezcla, en kg/m$^3$.
- Escoria de alto horno (blast_furnace_slag): cantidad de escoria de alto horno en la mezcla, en kg/m$^3$. Es un subproducto de la producción de hierro que se utiliza como sustituto del cemento.

- Cenizas volantes (fly_ash): cantidad de cenizas volantes en la mezcla, en kg/m$^3$. Es un subproducto de la combustión del carbón en centrales eléctricas que se utiliza como sustituto del cemento.
- Agua (water): cantidad de agua en la mezcla, en kg/m$^3$. El agua es esencial para la hidratación del cemento y la formación de la matriz de concreto.
- Superplastificante (superplasticizer): cantidad de superplastificante en la mezcla, en kg/m$^3$. Los superplastificantes se utilizan para mejorar la trabajabilidad del concreto sin aumentar su contenido de agua.
- Árido grueso (coarse_aggregate): cantidad de árido grueso en la mezcla, en kg/m$^3$. Los áridos gruesos son partículas grandes de material inerte, como grava o piedra triturada, que proporcionan resistencia y estabilidad al concreto.
- Árido fino (fine_aggregate): cantidad de árido fino en la mezcla, en kg/m$^3$. Los áridos finos son partículas más pequeñas de material inerte, como arena, que llenan los espacios vacíos entre las partículas de árido grueso y mejoran la trabajabilidad del concreto.
- Edad (age): edad del concreto en días cuando se midió la resistencia a la compresión. La resistencia del concreto generalmente aumenta con la edad debido al proceso de hidratación y endurecimiento.

Como variable objetivo tenemos la resistencia a la compresión (compressive_strength) del concreto, en megapascales (MPa).
El la siguiente tabla encontramos un resumen estadístico de los datos:

Tabla 1: Descripción de los datos

| Variables | Media | SD | MIN | MAX | n |
|---|---|---|---|---|---|
| age | 45.662136 | 63.169912 | 1.000000 | 365.00000 | 1030 |
| blast_furnace_slag | 73.895485 | 86.279104 | 0.000000 | 359.40000 | 1030 |
| cement | 281.165631 | 104.507142 | 102.000000 | 540.00000 | 1030 |
| coarse_aggregate | 972.918592 | 77.753818 | 801.000000 | 1145.00000 | 1030 |
| compressive_strength | 35.817836 | 16.705679 | 2.331808 | 82.59922 | 1030 |
| fine_aggregate | 773.578883 | 80.175427 | 594.000000 | 992.60000 | 1030 |
| fly_ash | 54.187136 | 63.996469 | 0.000000 | 200.10000 | 1030 |
| superplasticizer | 6.203112 | 5.973492 | 0.000000 | 32.20000 | 1030 |
| water | 181.566359 | 21.355567 | 121.750000 | 247.00000 | 1030 |

## 5.2. Resultados

El proceso de moldeamiento consistió en un proceso de separación de los datos en dos conjuntos de entrenamiento y validación. El conjunto de validación se utiliza para evaluar la capacidad predictiva y de construcción de intervalos de predicción. No sé aplicaron métodos de búsqueda de hiperparámetros ni optimizaciones adicionales sobre los modelos. Los datos ingresaron sin ninguna transformación (escalamiento, normalización, creación de factores o variables dummies, etc).

La tabla a continuación presenta los resultados de los cinco modelos estudiados: Regresión Lineal, Random Forest, Gradient Boosting Machine, Regresión Cuantil y Bayesian Additive Regression Trees donde se comparan utilizando cuatro métricas de rendimiento: PIPC, MPIW, RMSE (definido como la raíz del error cuadrático medio) y el RSQ. Dado que todos los modelos cumplen la cobertura esperada del 95 % o están muy cerca de hacerlo el $CWC$ es casi igual al $MPIW$ El modelo BART muestra el mejor rendimiento en términos de capacidad predictiva (RSQ más alto, RMSE más bajo), así como intervalos de predicción más estrechos (MPIW más bajo). Aunque el modelo Random Forest tiene una mejor cobertura del intervalo de predicción (PIPC más alto), su precisión y capacidad predictiva son inferiores a las de BART. En este sentido, BART es el mejor modelo para este problema. Incluso podemos afirmar que los resultados obtenidos son superiores a los encontrados en Chou et al. (2011) donde el mejor modelo (MART) alcazaba un RMSE de 4.94 en tanto BART llega a un 4.47.

Tabla 2: Resultados de los Modelos

| PIPC | MPIW | RMSE | RSQ | Modelo |
|---|---|---|---|---|
| 0.9346154 | 40.66640 | 10.775206 | 0.5696083 | LM |
| 0.9769231 | 34.53415 | 5.144564 | 0.9159259 | RF |
| 0.9500000 | 42.49003 | 7.715760 | 0.8103230 | GBM |
| 0.9538462 | 36.90663 | 10.508474 | 0.6272432 | QR |
| 0.9384615 | 16.05723 | 4.477250 | 0.9327370 | BART |

Gráficamente comparamos la predicción generada contra el valor real de la observación en Figura 25. Vemos como la mayoría de modelos, a diferencia de BART, construyen intervalos de predicción consistentes pero mucho más anchos. Sacrifican la certeza por la validez y la cobertura.

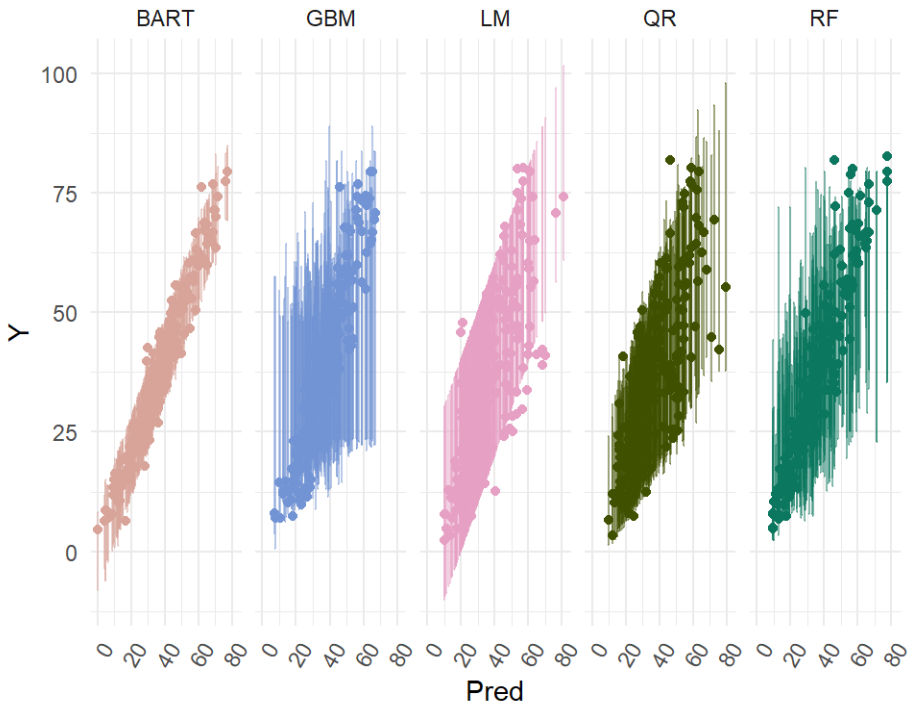


Figura 25: Valores predichos y reales con sus respectivos intervalos de predicción

# 6. Conclusiones

Del proceso de simulación y el caso aplicado tenemos las siguientes conclusiones:

- En el caso de los datos generados linealmente BART tuvo un desempeño al menos tan bueno como los demás modelos en RSQ, es superior a GBM, QR y RF en cobertura y genera los intervalos más pequeños, luego del modelo lineal, en casi todos los casos. Es importante destacar que el efecto del volumen de datos es indiferente en las métricas de error para este problema y no afecta a ninguno de los modelos.

- En el caso de los datos generados por un modelo a tramos BART fue superior a todos los modelos presentando un poco más de dispersión en el RSQ (para la combinación 100 observaciones - errores exponenciales) que los demás modelos. En este caso, los modelos GBM y BART se benefician y aproximan mejor $f(x)$ a medida que aumenta el volumen de datos. El efecto de los errores es semejante en todos los modelos. En tanto a cobertura, BART presenta un comportamiento superior o al menos tan bueno a los demás modelos en todos los casos. Cuando observamos los intervalos generados, la peor posición alcanzada por BART es la tercera y es el mejor modelo en las combinaciones de errores uniformes.

- En los datos generados de forma no lineal y la función de Friedman tenemos conclusiones semejantes: BART obtiene un resultado al menos tan bueno como RF y superior a los demás modelos en desempeño predictivo y una cobertura valida en un porcentaje superior a modelos como RF y LM. Respecto a la posición relativa de los intervalos generados, BART llega hasta a la cuarta posición, pero es superior a medida que el volumen de datos incrementa. Además, variaciones de $\eta$ para estos dos casos modifican fuertemente la posición relativa.

- En el caso aplicado, BART fue el mejor modelo tanto en desempeño predictivo como en la generación de los intervalos (una cobertura mejor en 2 % a un modelo como GBM, pero con un ancho del intervalo dos veces más pequeño). Además, se alcanzaron errores inferiores a los referenciados en la literatura sin pasar por procesos de optimización de parámetros.

- BART ofrece una alternativa al uso de métodos conformacionales que, si bien presentan una garantía de cobertura en la mayoría de los casos, generan intervalos más anchos y son computacionalmente más ineficientes.

Este trabajo constituye uno de los primeros esfuerzos de evaluar metodologías para la generación de intervalos de predicción en lo que respecta a modelos de ensamble. De los resultados obtenidos podemos deducir que BART es un modelo capaz de predecir y generalizar sobre diferentes tipos de datos y, además, genera intervalos de predicción al menos tan buenos como los demás modelos posibles sin la necesidad de modificar de alguna manera el modelo base (como es el caso en RF o GBM) o usar una combinación de métodos adicionales y modelos (como vemos con RF y el método conformacional). Esto lo convierte es una buena alternativa en lo que respecta a modelos basados en ensambles cuando, además de capacidad predictiva, se requieren medidas de incertidumbre sobre las predicciones al momento de construir modelos predictivos.

Es importante que destacar que los modelos evaluados se utilizaron en el contexto del problema de regresión cuantitativo de corte transversal, y por tanto, las conclusiones extraídas se restringen a este caso; sin embargo, los modelos de ensamble puede extenderse, y son frecuentemente usados, en problemas de clasificación, análisis de supervivencia o predicción de series de tiempo temporales. En estos dominios también es importante la cuantificación de la incertidumbre por lo que una extensión natural de esta investigación puede dirigirse en esa lineas.

# Referencias


Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Angelopoulos, A. N. & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, 175:114820.

Bogner, K., Pappenberger, F., & Zappa, M. (2019). Machine learning techniques for predicting the energy consumption/production and its uncertainties driven by meteorological observations and forecasts. *Sustainability*, 11(12):3328.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In

*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chipman, H. A., George, E. I., & McCulloch, R. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chou, J.-S., Chiu, C.-K., Farfoura, M., & Al-Taharwa, I. (2011). Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques. *Journal of Computing in Civil Engineering*, 25(3):242–253.

De Brabanter, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2010). Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, 22(1):110–120.

Ehsan, B. M. A., Begum, F., Ilham, S. J., & Khan, R. S. (2019). Advanced wind speed prediction using convective weather variables through machine learning application. *Applied Computing and Geosciences*, 1:100002.

Fenske, N., Kneib, T., & Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, 106(494):494–510.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Geraci, M. & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., & Frediani, M. E. (2017). Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3):441–458.

Hernández, B., Raftery, A. E., Pennington, S. R., & Parnell, A. C. (2018). Bayesian additive regression trees using bayesian model averaging. *Statistics and computing*, 28(4):869–890.

Heskes, T. (1996). Practical confidence and prediction intervals. *Advances in neural information processing systems*, 9.

Kapelner, A. & Bleich, J. (2013). bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*.

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. (2012). Package 'quantreg'.

Kumar, S. & Srivistava, A. N. (2012). Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. In *The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, number ARC-E-DAA-TN6188.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Rinaldo, A., & Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43.

Lei, J. & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96.

Li, Y., Chen, J., & Feng, L. (2012). Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2463–2482.

Mayr, A., Hothorn, T., & Fenske, N. (2012). Prediction intervals for future bmi values of individual children-a non-parametric approach by quantile boosting. *BMC Medical Research Methodology*, 12(1):6.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Meinshausen, N. (2007). Quantregforest: quantile regression forests. *R package version 0.2-2*.

Pevec, D. & Kononenko, I. (2015). Prediction intervals in supervised learning for model evaluation and discrimination. *Applied Intelligence*, 42(4):790–804.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

Schmoyer, R. L. (1992). Asymptotically valid prediction intervals for linear models. *Technometrics*, 34(4):399–408.

Seber, G. A. & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Shafer, G. & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031.

Su, D., Ting, Y. Y., & Ansel, J. (2018). Tight prediction intervals using expanded interval minimization. *arXiv preprint arXiv:1806.11222*.

Tan, Y. V. & Roy, J. (2019). Bayesian additive regression trees and the general bart model. *Statistics in medicine*, 38(25):5048–5069.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808.

Yu, K. & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zapranis, A. & Livanis, E. (2005). Prediction intervals for neural network models. In *Proceedings of the 9th WSEAS International Conference on Computers*, page 76. World Scientific and Engineering Academy and Society (WSEAS).

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*, 74(4):392–406.