ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

Análisis estadístico de la evolución cósmica de los agujeros negros masivos usando variables fotométricas

PROYECTO INTEGRADOR

Previo a la obtención del Título de:

Matemático

Presentado por:

Alexander Felipe Palma Pozo

GUAYAQUIL - ECUADOR

Año: 2022

DEDICATORIA

A Dios por la inteligencia, sabiduría y fortaleza que me brindó para seguir adelante. A mis padres y hermanos porque han sido los pilares fundamentales de mi formación.

Alexander Palma Pozo

AGRADECIMIENTOS

Mi más sincero agradecimiento a mi querido amigo el estadístico Christian Galarza, por sus enseñanzas, revisión y corrección del presente trabajo, y a su esposa Silvia por sus amables consejos. Al físico Erick Lamilla por la revisión de los conceptos de la física teórica. A la matemática Elimar Marchán por sus minuciosas correcciones y al conjunto de profesores que permitieron mi formación como matemático.

Alexander Palma Pozo

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; *Alexander Felipe Palma Pozo,* y doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual”

Alexander Felipe Palma Pozo

EVALUADORES

Firmado electrónicamente por:
LUZ ELIMAR MARCHAN MENDOZA

Luz Marchán Mendoza, Ph.D.

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
CHRISTIAN EDUARDO GALARZA MORALES

Christian Galarza Morales, Ph.D.

PROFESOR TUTOR

RESUMEN

La edad del universo es una de las principales interrogantes de la humanidad, la cual puede ser abordada desde la Física mediante la medición del desplazamiento al rojo emitido por los objetos más distantes en el universo. Sin embargo, la relación entre el desplazamiento al rojo y la edad del universo no es lineal, siendo esta compleja y difícil de calcular sin establecer hipótesis sobre la topología del universo. Así, los modelos de regresión en media no resultan adecuados para estudiar esta relación, pues modelar la media del desplazamiento al rojo no será equivalente a modelar la edad media de los objetos del universo. Para los cuantiles, lo anterior sí se satisface debido a su propiedad de invariancia bajo transformaciones. En este trabajo se estudia el tiempo histórico del universo al modelar la relación entre desplazamiento al rojo y características fotométricas medidas en 46420 cuásares por medio de un modelo de regresión cuantílica. En particular, un modelo de regresión cuantílica aditiva parcialmente lineal fue ajustado sobre el percentil 90, permitiendo incluir efectos lineales y no lineales para modelar el desplazamiento al rojo del 10% de los cuásares más antiguos del universo conocido. Métodos clásicos estadísticos fueron usados para la clasificación y selección de variables. Finalmente, se interpretan los efectos de las variables fotométricas significativas del modelo final, permitiendo hacer predicciones más robustas sobre el desplazamiento al rojo, lo que ofrece un enfoque nuevo en el estudio de la evolución cósmica.

Palabras Clave: Desplazamiento al Rojo, Cuásar, Agujero Negro Masivo, Regresión Cuantílica Aditiva Parcialmente Lineal, LASSO.

ABSTRACT

The age of the universe is one of the main questions of humanity, which can be approached from physics by measuring the redshift emitted by the most distant objects in the universe. However, the relationship between the redshift and the age of the universe is not linear, being complex and difficult to calculate without establishing hypotheses about the topology of the universe. Thus, mean regression models are not suitable for studying this relationship, since modeling the mean redshift will not be equivalent to modeling the mean age of the objects in the universe. For quantiles, the above is satisfied due to their property of invariance under transformations. In this work, we study the historical time of the universe by modeling the relationship between redshift and measured photometric characteristics in 46420 quasars by means of a quantile regression model. In particular, a partially linear additive quantile regression model was fitted over the 90th percentile, allowing linear and nonlinear effects to be included to model the redshift of 10% of the oldest quasars in the known universe. Classical statistical methods were used for classification and variable selection. Finally, the effects of significant photometric variables in the final model are interpreted, allowing more robust redshift predictions to be made, providing a new approach to the study of cosmic evolution.

Keywords: *Redshift, Quasar, Massive Black Hole, Partially Linear Additive Quantile Regression, LASSO.*

ÍNDICE GENERAL

ABREVIATURAS

SDSS	Sloan Digital Sky Survey
CERN	Organización Europea para la Investigación Nuclear
ALD	Distribución Asimétrica de Laplace
SKD	Clase General de Distribuciones Asimétricas
PLAQR	Regresión Cuantílica Aditiva Parcialmente Lineal
SSE	Suma Residual de Cuadrados
MSE	Error Cuadrático Medio
LASSO	Least Absolute Shrinkage and Selection Operator
MCO	Mínimos Cuadrados Ordinarios
FIRST	Observatorio Nacional de Radioastronomía de Imágenes Débiles
VIF	Índice de Inflación de la Varianza

SIMBOLOGÍA

$\mathcal{F}(\mathbb{F})$	Espacio de funciones definido sobre un campo escalar $\mathbb{F}$
$\mathcal{C}^{\omega}(\mathbb{R})$	Espacio de funciones analíticas definidas sobre los reales
$L^2([0,1];\mathbb{R})$	Espacio de funciones de valor real cuadrado integrables sobre $[0,1]$
$\mathbb{R}^{m\times n}$	Conjunto de las matrices de orden $m \times n$ con entradas reales
$\mathbb{R}^n$	n-vector con entradas reales
$\|\cdot\|_{L_1}$	Norma en L_1
$\mathbf{I}_n$	Matriz identidad de orden $n \times n$
$\mathbf{1}_n$	n-vector de unos
∞	Infinito

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

En este capítulo, se empieza por plantear el problema que surge al analizar con métodos tradicionales características de naturaleza fotométrica y la importancia de un método robusto que pueda ser implementado en un laboratorio de óptica en Ecuador para el análisis de este tipo de datos. Luego, se introducen las bases matemáticas que sustentan un modelo de regresión cuantílica y finalmente se adapta el modelo al problema del presente trabajo.

1.1 Descripción del problema

La edad del universo es una de las incógnitas más grandes que aqueja al ser humano. La edad de las galaxias, unidades básicas del universo, puede ser estimada a través de la cantidad de energía (por ejemplo, la radiación de Hawking) que emiten los agujeros negros masivos, usualmente, localizados en el centro de las mismas. Estas fuentes de radiación intensa se conocen como cuásares, donde su cantidad de energía expedida, medida a través del desplazamiento al rojo, es utilizada para medir su evolución cósmica. Por otro lado, la astronomía utiliza técnicas fotométricas para diferentes propósitos como medir distancias a objetos, estudiar su composición y atmósfera, y en este caso en particular, determinar la edad. Sin embargo, variables fotométricas como el brillo de la banda ultravioleta o infrarroja, rayos X, entre otros, presentan características como tendencias no lineales, valores atípicos y censuras, las cuales resultan difíciles de modelar usando métodos predictivos tradicionales. El problema a ser abordado afecta directamente a los laboratorios de astrofísica dedicados a estudiar el tiempo histórico del universo, e indirectamente a otras áreas de la ingeniería con aplicaciones en la óptica.

1.2 Justificación del problema

El ser humano siempre ha estado intrigado por su propia naturaleza y la del universo. En este proyecto, se desea abordar la segunda cuestión con el objetivo de proveer nuevos hallazgos sobre el tema a través del estudio de datos astronómicos usando métodos predictivos robustos. Este enfoque es pertinente ya que los métodos usuales, como mínimos cuadrados, no son lo suficientemente flexibles para poder capturar relaciones complejas que pueden existir en relación a las variables fotométricas. Estas últimas características son fácilmente medidas en los observatorios que se dedican al estudio de agujeros negros masivos; que son singularidades en el espacio tiempo donde las leyes de la física aún están siendo estudiadas. Como resultado, se obtendrá un modelo predictivo, a través de un modelo de regresión cuantílica que permitirá estimar la edad de las galaxias más antiguas del universo; en particular, aquellas que se encuentran en el 10% de las más antiguas. A lo mejor de nuestro conocimiento, no existe otra propuesta similar para el estudio de la evolución cósmica de los agujeros negros masivos en nuestro país, estableciendo un punto de partida para futuros análisis estadísticos especializados en laboratorios de investigación cosmológica y de óptica.

1.3 Objetivos

1.3.1 Objetivo General

Adaptar un modelo de regresión por cuantiles a la evolución cósmica de los agujeros negros con base en sus características fotométricas para describir la cronología del universo.

1.3.2 Objetivos Específicos

1. Estimar el efecto de las variables fotométricas relevantes para los agujeros negros masivos más antiguos del universo por medio de un análisis de regresión estadístico.

2. Evaluar un modelo supervisado que permita predecir la evolución cósmica de nuevos agujeros negros a partir de sus características fotométricas para conocer su tiempo histórico en el universo.

1.4 Marco teórico

1.4.1 El origen del universo físico

Una de las cuestiones más grandes que el ser humano constantemente se ha planteado es la naturaleza que subyace alrededor del concepto de universo o estrictamente universo físico y su origen. Siguiendo el concepto propuesto por Penrose (2005) el universo se define como el espacio-tiempo y todo lo que contiene. Desde la perspectiva de la física teórica moderna y la teoría del Big Bang, hace aproximadamente $t \approx 10^{-43}$s la temperatura del universo era cerca de 10^{32} K, dando lugar a la existencia del universo físico. Es así, como una de las principales razones por la que físicos y científicos dedicados a esta área estudian el comportamiento de las partículas expuestas a altas energías es para entender la estructura del universo momentos después de su origen.

A pesar que hasta el momento, se tienen teorías sólidas sobre el origen del universo y su evolución temporal, aún resulta extremadamente complejo formar una idea clara y precisa de la forma en la que el universo ha ido evolucionando a lo largo del tiempo, debido a que la teoría del Big Bang en su forma clásica sobre la formulación del universo en expansión es incompleta como se indica en Biswas et al. (2012) puesto que no puede explicar situaciones como las condiciones iniciales de su creación.

A lo largo del tiempo, la teoría del Big Bang ha tenido modificaciones importantes como la hipótesis agregada por Penzias and Wilson (1965) quienes observaron mientras probaban un receptor sensible de microondas que se usaba en comunicaciones, que se producía un silbido de fondo que no cambiaba de intensidad, sin importar la orientación de la antena del microondas; esto era lo que en 1978 les haría ganadores del Premio Nobel de Física por el descubrimiento de la radiación cósmica de fondo de microondas que representa la luz que ha estado viajando a través del universo justo momentos después de que el universo comenzara su existencia y expansión hace miles de millones de años. Esto permite la inflación cósmica que es explicada en Buen-Abad et al. (2022).

Como se mencionó al inicio de esta sección, el universo físico contenía en sus origenes particulas fundamentales a temperaturas extremadamentes altas en un espacio diminuto. Luego, a medida que el tiempo transcurría positivamente como puede verse en Barrau (2022) el universo empezó a expandirse y se enfriaba a temperaturas cada vez más bajas, hasta eventualmente alcanzar la temperatura y tamaño que se puede estimar actualmente.

Sin embargo, el universo físico actualmente observable es de hecho el pasado histórico del universo. Para ver esto, imagine que usted está en un punto p de un cono de luz, que marca los límites de causalidad del universo físico, entonces cuando observa el universo actual usted realmente está mirando hacia atrás en el tiempo (cono del pasado) porque la luz presente en el universo ha tardado mucho tiempo en llegar hasta el punto p donde usted se encuentra teóricamente.

La observación anterior, permite tener un esquema de cómo estudiar la evolución cósmica: mirando hacia el pasado histórico del espacio-tiempo en búsqueda de los objetos más lejanos que se puedan detectar. Así, un enfoque prominente para estudiar la cronología del universo físico es a través del desplazamiento hacia el rojo (redshift) de la longitud de onda de la luz y otras características fotométricas de unos objetos del universo denominados agujeros negros masivos ubicados en los cuásares.

1.4.2 Agujeros negros masivos

Considere el espacio-tiempo de la relatividad general que posiciona la velocidad de la luz de manera fundamental como la velocidad límite. Así, en este marco de referencia para el universo físico pueden existir, teóricamente, situaciones donde la velocidad de escape supere la velocidad de la luz, dando lugar a un objeto extraño en el universo físico conocido como agujero negro.

De acuerdo a Penrose (2005) los agujeros negros son lugares del espacio-tiempo que han resultado de un colapso gravitacional con tan alta gravedad que nada puede escapar de su tirón gravitacional, ni siquiera la luz. En el pasado, estos objetos eran

considerados únicamente quimeras teóricas encontradas dentro de la teoría de la relatividad general de Albert Einstein en 1915 y fue fruto de muchos trabajos posteriores de físicos y matemáticos como Schwarzschild, Oppenheimer, Chandarsekaer, Wheeler, Hawking y Penrose, que se lograron asentar las bases para el estudio de los agujeros negros.

En la actualidad, el primer agujero negro en ser detectado fue Cygnus X1 en la Constelación del Cisne. Hoy en día, se cuenta con una imagen de uno de estos extraños objetos tomada por la Organización Europea para la Investigación Nuclear (CERN), eliminando así anteriores dudas sobre su existencia.

Una forma en la que estos objetos extraños se pueden formar es por la “muerte” de una estrella masiva cuando se acaba su combustible nuclear. Esto se sigue de un análisis realizado sobre el radio de Schwarzschild en situaciones críticas observando que, si una estrella llega a colapsar por debajo de ese radio, entonces nada puede escapar de su tirón gravitatorio.

Adicionalmente, el trabajo del físico Hawking muestra que los agujeros negros no son completamente negros, sino que desprenden radiación que actualmente se conoce como radiación de Hawking, lo que implica que un agujero negro va perdiendo energía hasta eventualmente desaparecer. Esto da apertura a la paradoja de la información que es fuertemente estudiada en mecánica cuántica y además provee otra forma de estudiar un agujero negro que es a través de características como las fotométricas. (He et al., 2022)

Los agujeros negros se pueden clasificar según su naturaleza espaciotemporal en:

1. Agujeros negros estelares: se producen por el colapso de una estrella.
2. Agujeros negros masivos (supermasivos): su origen sigue siendo un misterio y se pueden encontrar en el centro de las galaxias.

3. Agujeros negros primordiales: se supone se crearon en los primeros instantes del universo.

Una suposición fuertemente aceptada es que los agujeros negros primordiales dieron lugar a los agujeros negros masivos (o supermasivos). De esta forma, el estudio de los agujeros negros supermasivos provee una forma teórica de estudiar el tiempo histórico del universo pues permite acercarse, en términos de tiempo, cada vez más a objetos mucho más primitivos; esto es, se tiene un método para estudiar la cronología del universo (Gaztanaga, 2022).

El estudio de los agujeros negros supermasivos se suele realizar a través de su radiación como se mencionó anteriormente, permitiendo explicar los complejos procesos de acreción de estos objetos y además encontrarlos; sin embargo, otra forma es detectarlos a través de los cuásares.

Un cuásar es esencialmente un objeto celeste extremadamente masivo y muy lejano que es parecido a una estrella, emite una señal de radio de acuerdo a Sou et al. (2022) y contiene un agujero negro supermasivo. Estos son los objetos que se estudiarán en el presente trabajo, a través de un análisis estadístico de sus características fotométricas.

Sin embargo, para estudiar los cuásares y otros objetos astronómicos que se encuentran a millones de años luz es necesario tener telescopios cada vez más sofisticados para su detección. Es por esta razón que como resultado del trabajo conjunto de grandes científicos por estudiar la naturaleza del cosmos, se han construido telescopios con una gran capacidad de sondear el universo; por ejemplo, el Telescopio Espacial Hubble y actualmente el Telescopio Espacial James Web. En la figura 1.1 se puede ver una simulación de un cuásar y la galaxia asociada usando los Telescopios Webb y Hubble en longitud de onda infrarroja 1.5 y 1.6 micras. El Telescopio Webb está compuesto por espejos y se espera que el espejo más grande de estos entregue más de cuatro veces la resolución que se tenía con el Hubble, de darse este caso, esto permitiría a los físicos separar la luz emitida por las galaxias de la luz emitida por el

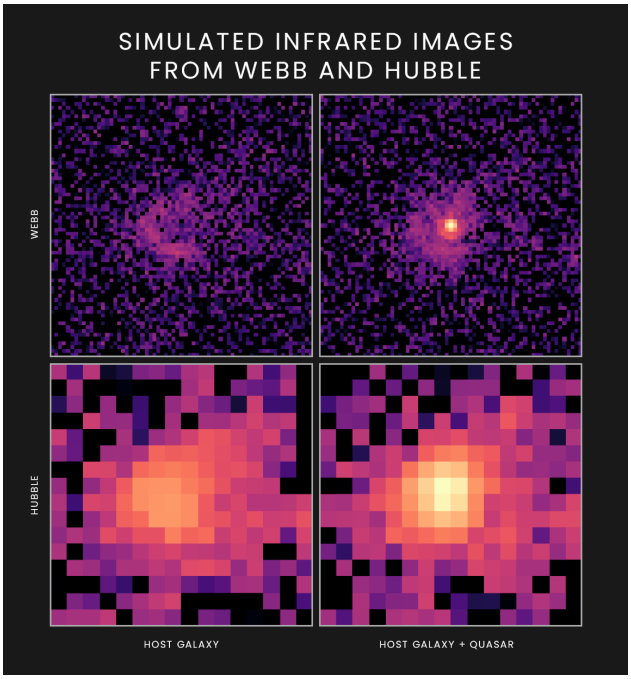

Figura 1.1. Simulación de imágenes de un cuásar y su galaxia anfitriona usando el Telescopio Espacial James Webb de la NASA (arriba) y el Telescopio Espacial Hubble (abajo) en longitudes de onda infrarrojas de 1.5 y 1.6 micras, respectivamente. Fuente: Marshall (2020).

cuásar asociado. Las simulaciones mostradas abarcan aproximadamente dos segundos de arco en el cielo, esto es una distancia de aproximadamente 36.000 años luz con un desplazamiento al rojo de siete.

1.4.3 Fotometría de un cuásar

La idea que subyace en el concepto y medibilidad de la luz, tiene grandes repercusiones en la física teórica, por ejemplo, en la mecánica cuántica es posible interpretar la luz a través de la función de onda. Sin embargo, el presente trabajo se considera desde la cosmovisión de la astrofísica estadística, por lo que, se está considerando la luz con comportamiento dual en onda y partícula, esto es, la luz se considera como una onda electromagnética que únicamente se permite en ciertos cuantos y estos cuantos se denominan fotones.

La interpretación dual de la luz permite introducir la noción de espectro

Figura 1.2. Desplazamiento al rojo como medida del tiempo desde un observador (izquierda) hasta un objeto astronómico (derecha). Fuente: NASA and Feild (2018)

electromagnético, esto es, radio, infrarrojo, visible, ultravioleta, rayos X, rayos γ, de manera simple, puesto que permite interpretar la luz como radiación electromagnética y esto último es un elemento crucial para estudiar el universo físico, analizando características visibles en el espectro electromagnético como el brillo en varias bandas espectrales, desplazamientos al rojo y luminosidades de un objeto astronómico en particular; en este caso los cuásares. En términos de la naturaleza de los cuásares cada una de estas características permite realizar diferentes tipos de análisis, por ejemplo el brillo de las bandas espectrales permite trabajar con técnicas de clasificación estadística y el desplazamiento al rojo permite la aplicación de técnicas predictivas, como la regresión cuantílica o regresión en media.

Al viajar las partículas de fotones a través de un universo en expansión, los fotones se estiran causando que las ondas cortas se alarguen. Este fenómeno es intuitivo y es conocido como desplazamiento al rojo (o redshift, en inglés). Inicialmente se puede "distinguir" entre tres tipos de desplazamiento al rojo:

1. Desplazamiento al rojo Doppler: observadores que se mueven relativamente entre sí, miden diferentes longitudes de ondas de las partículas fotónicas.

2. Desplazamiento al rojo gravitacional: observadores en distintos lugares dentro de un campo gravitacional miden diferentes longitudes de onda de las partículas fotónicas.

3. Desplazamiento al rojo cosmológico: observadores intercambiando partículas fotónicas a través de una gran distancia cosmológica en un universo físico en expansión miden diferentes longitudes de onda de las partículas fotónicas. Una ilustración de este fenómeno puede verse en la figura 1.2, en la cual se encuentra un observador desde la tierra (a la izquierda de la imagen) y un objeto astronómico observable (a la derecha de la imagen) y el correspondiente desplazamiento al rojo cosmológico.

Los tres desplazamientos al rojo mencionados anteriormente aunque son gobernados por diferentes ecuaciones y parecen representar eventos físicos diferentes; de hecho describen situaciones físicas idénticas. Lo que se observa como desplazamiento al rojo no es algo que le sucede a la partícula fotónica en sí mismo, sino que depende de lo que está sucediendo con los observadores en el punto de emisión y recepción de ese fotón. Es así como, los tres desplazamientos únicamente se ven distintos, pero es un único desplazamiento al rojo que es determinado por un marco matemático que subyace a los tres. Esto está fuertemente relacionado con la relatividad de la simultaneidad.

El término fotometría no tiene su origen directamente en el campo de la astrofísica. La fotometría fue acuñada primordialmente en el campo de la óptica para describir mediciones de energía que estaban estrechamente relacionadas con las sensaciones visuales. Por otro lado, está la radiometría que es esencialmente el estudio de la radiación electromagnética en cualquier longitud de onda. En este trabajo se seguirá la definición dada en Léna et al. (2013) que unifica estas dos nociones para referirse a la fotometría o estrictamente la espectrofotometría. Esta definición es mucho más fina ya que muchas estrellas y galaxias emiten emiten fracciones significativas de su energía de manera que son detectables en el espectro electromagnético.

Para el presente trabajo se consideran características fotométricas, puesto que

permiten medir propiedades de los fotones como la intensidad de radiación en los cuásares y esto a su vez permite estudiar como su población evoluciona con el tiempo.

1.4.4 Sloan Digital Sky Survey (SDSS)

La elección de los datos astronómicos con los que se realiza algún tipo de análisis en astrofísica estadística es importante y no es una tarea simple, desde que la astronomía actual cuenta con un crecimiento exponencial de datos genuinos. Este crecimiento se debe principalmente a que la astronomía galáctica y la cosmología son disciplinas que dependen principalmente de las observaciones.

Por esta razón, para la elección de datos con los que se trabaja en el presente proyecto se siguió el esquema planteado en Léna et al. (2013) considerando parámetros como el campo observacional cubierto, sensibilidad (o profundidad), exhaustividad (o completitud), el límite de confusión y el objetivo del análisis.

Dos grandes estudios sobre espectrofotometría cosmológica han adquirido una cantidad considerable de datos que han sido recolectados desde el 2001, el Two Degree Field Galaxy Redshift Survey (2dFGRS) y el Sloan Digital Sky Survey (SDSS). El 2dFGRS es un conjunto de datos de un proyecto del observatorio Anglo-Australiano con el 2dF espectrógrafo multi-objeto que funciona con apoyo de una máquina que trabaja a través de fibras ópticas y placas fotográficas digitalizadas y el SDSS de manera similar es un espectrógrafo multi-objeto que permite obtener 640 espectros de manera simultánea, esto es enormemente poderoso a nivel computacional puesto que permite obtener espectros de una mayor cantidad de galaxias.

Considerando los parámetros expuestos anteriormente y el fin del presente trabajo, se ha elegido la base de datos propuesta por Sloan Digital Sky Survey (SDSS) en Schneider et al. (2005) que es una tercera versión del conjunto de datos propuesto en York et al. (2000) que contiene una muestra de $46,400$ cuásares con 23 características, entre ellas, las de naturaleza fotométrica.

1.4.5 Desplazamiento al rojo como una medida del tiempo

El desplazamiento al rojo (redshift) es definido como

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}}, \tag{1.1}$$

donde λ_{em} es el desplazamiento en longitud de onda de un fotón emitido por una galaxia distante en un tiempo anterior t_{em} y λ_{obs} es el desplazamiento en longitud de onda de un fotón de una galaxia distante que es observada desde la tierra en un tiempo t_{obs}. La expresión (1.1) puede escribirse como $1 + z = a_0/a(t_{\text{em}})$, donde a_0 denota el valor presente de un factor de escala. Usando esta última expresión, se puede observar que existe una correspondencia uno a uno entre z y t_{em}. De esto, además se sigue que el redshift z puede ser utilizado en lugar del tiempo t para parametrizar la historia evolutiva del universo. Un z dado corresponde a un tiempo en el que el universo físico actual era $1 + z$ más pequeño que en el presente.

Conseguir una expresión para $a(t)$ es extremedamente complicado y no se puede invertir directamente $1 + z = a_0/(a_{t_{\text{em}}})$ para expresar el tiempo de evolución cósmica $t = t_{\text{em}}$ en términos del redshift. Una forma de abordar este problema es a través de la diferenciación de $1 + z = a_0/(a_{t_{\text{em}}})$ dando $\mathrm{d}z = -(1 + z)H(t)\,\mathrm{d}t$ donde $H(\cdot)$ es el parámetro de Hubble, de donde se sigue que

$$t = \int_z^\infty \frac{1}{H(z)(1+z)}\,\mathrm{d}z, \tag{1.2}$$

donde se ha elegido una constante de integración de modo que $z \to \infty$ corresponde al momento inicial del tiempo $t = 0$. Entonces, para determinar $z \mapsto t(z)$, primero se debe encontrar la densidad de energía $z \mapsto \epsilon(z)$ que puede ser encontrada a través de la ley de conservación $\mathrm{d}\epsilon = -3(\epsilon + p)\,\mathrm{d}\ln a$, esto es,

$$\int_{\epsilon_0}^{z\mapsto\epsilon(z)} \frac{1}{\epsilon + p(\epsilon)}\,\mathrm{d}\epsilon = 3\ln(1+z),$$

y posteriormente sustituir

$$H(z) = H_0\left((1-\Omega_0)(1+z)^2 + \Omega_0\frac{\epsilon(z)}{\epsilon_0}\right)^{1/2}$$

en (1.2) y realizar la integración, esto es, resolver

$$t = \int_z^{\infty} \frac{1}{H_0 \left((1-\Omega_0)(1+z)^2 + \Omega_0 \frac{\epsilon(z)}{\epsilon_0} \right)^{1/2} (1+z)} \, \mathrm{d}z.$$

En un universo físico con curvatura espacial no nula, esto es, $k \neq 0$ el actual factor de escala puede ser escrito como $a_0^{-1} = |\Omega_0 - 1|^{1/2} H_0$, con esta consideración y tomando un universo dominado por el polvo con densidad de energía $\epsilon(z) = \epsilon_0(1+z)^3$ y tomando $H(z) = H_0(1+z)\sqrt{1+\Omega_0 z}$ para un universo plano, la integral en (1.2) puede ser sencillamente reducida a $z \mapsto t(z) = 2/(3H_0(1+z)^{3/2})$.

1.5 Regresión cuantílica

Los modelos de regresión tienen como objetivo modelar la tendencia general de una variable de interés o repuesta en función de otras variables independientes también conocidas como variables predictoras. Cuando se piensa en una tendencia general, este modelo considera la media. Sin embargo, esta medida de centralidad resulta inadecuada si la distribución de los datos no es simétrica. Además, esta es sensible a los valores atípicos, a la especificación errada de la distribución de errores, así como a la aplicación de transformaciones a las variables del modelo, pudiendo causar a menudo efectos de ímpetu, o destacando efectos que no existen para las variables originales. Más aún, no siempre se está interesado en predecir una tendencia central para la variable respuesta; puesto que, como en el presente trabajo, algunas veces se está interesado en ciertos cuantiles o valores extremos y es en esto donde la regresión cuantílica toma una relevancia significativa, como se puede apreciar en la figura 1.3.

De la figura 1.3 se puede apreciar cómo las curvas de predicción para la media y mediana condicional de la capacidad pulmonar (FEV) en jóvenes lucen similares, siendo ambas útiles a la hora de estudiar una tendencia central. No es coincidencia que la curva media (línea magenta) se ve empujada hacia arriba por los valores altos observados para individuos entre los 10 y 15 años de edad. Este efecto pasa inadvertido en la curva mediana, gracias a sus propiedades de robustez (Huber, 1981).

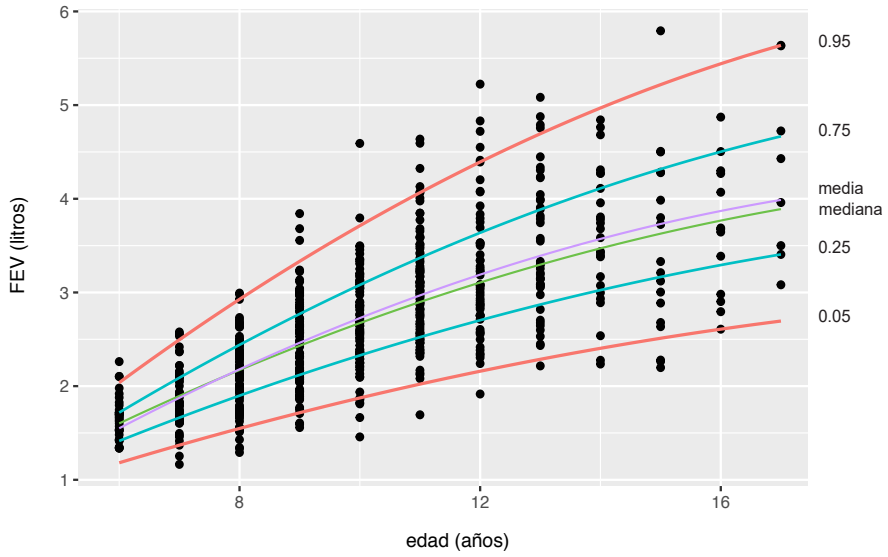

Figura 1.3. Curvas estimadas usando regresión tradicional en media (línea magenta) vs. curvas estimadas para varios cuantiles condicionales usando regresión cuantílica.

La regresión cuantílica no hace suposiciones sobre la distribución de los datos, puede caracterizar toda la distribución condicional de la variable de respuesta y así poder obtener una mejor imagen de los datos. Finalmente, una de las características más importantes es que el método de regresión cuantílica hereda la propiedad de invarianza de los cuantiles. Por ejemplo, la mediana es siempre la mediana independientemente de la transformación que le sea aplicada, esto es, invariante bajo transformaciones continuas. El explotar esta propiedad es lo que justamente nos motiva a utilizar este tipo de modelo para el estudio del redshift, ya que sin importar o conocer la forma en cómo se relaciona matemáticamente el redshift con la antigüedad de los agujeros negros, aquellos cuyo redshift se encuentre en el decil superior pertenecerán necesariamente al grupo del 10% más antiguo.

1.5.1 Formulación general

Sea y_i con $i = 1, \ldots, n$ una variable de respuesta totalmente observada, $\boldsymbol{x}_i$ un vector de covariables de dimensión $q \times 1$ para la i-ésima observación y $Q_p(y_i \mid \boldsymbol{x}_i)$ (con $0 < p < 1$) la función cuantil de y_i dado $\boldsymbol{x}_i$ para un cuantil p a modelar. Suponga que la relación entre

este cuantil y $\boldsymbol{x}_i$ puede ser modelada como $Q_p(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_p$, donde $\boldsymbol{\beta}_p$ es un vector no conocido de parámetros de interés de dimensión $q \times 1$. Entonces, el modelo de regresión cuantílica puede ser formulado de la siguiente manera

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_p + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.3}$$

donde ϵ_i es el término del error cuya distribución (con densidad, $f_p(\cdot)$) está restringida para tener el p-ésimo cuantil igual a cero, esto es, $\int_{-\infty}^{0} f_p(\epsilon_i)\, \mathrm{d}\epsilon_i = p$, y consecuentemente $\mathbb{P}(y_i \leqslant \boldsymbol{x}_i^\top \boldsymbol{\beta}_p) = p$. La densidad $f_p(\cdot)$ comúnmente se deja sin especificar en la literatura clásica. Por lo tanto, una estimación en regresión cuantílica para $\boldsymbol{\beta}_p$ es obtenida por minimización

$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}_p \in \mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_p), \tag{1.4}$$

donde $\rho_p(\cdot)$ es una función de pérdida definida por

$$\rho_p(u) = u(p - \mathbb{I}\{u < 0\}), \tag{1.5}$$

con $\mathbb{I}\{\cdot\}$ la usual función característica (o identidad) y $\hat{\boldsymbol{\beta}}_p$ es la estimación de la regresión cuantílica de $\boldsymbol{\beta}_p$ para el p-ésimo cuantil. Cuando $p = 0.5$, se tiene una regresión en mediana. Es importante destacar que existe una relación entre la minimización de la suma en (1.4) y la teoría de máxima verosimilitud, ya que minimizar (1.4) equivale a maximizar la verosimilitud cuando los datos siguen una distribución perteneciente a la familia de distribuciones con cuantiles condicionales iguales a cero.

1.5.2 Enfoques de abordaje

Los enfoques para abordar el problema de regresión cuantílica son esencialmente no parámetrico y parámetrico. Ambos métodos permiten hacer inferencia sobre los datos estudiados, es decir, es posible desde estos enfoques construir intervalos de confianza, responder a ciertas hipótesis, etc. No obstante, existen algunos elementos importantes que se deben mencionar de los mismos.

Enfoque no parámetrico. Históricamente el primer enfoque para abordar el problema de regresión cuantílica fue desde la optimización matemática, esencialmente usa la

formulación del problema dado en (1.4) y lo reformula como un modelo de programación lineal

$$\min_{(\boldsymbol{\beta}_p, u, v) \in \mathbb{R}^k \times \mathbb{R}_+^{2n}} \left\{ p\mathbf{1}_n^\top u + (1-p)\mathbf{1}_n^\top v : \boldsymbol{x}^\top \boldsymbol{\beta}_p + u - v = y \right\},$$

donde $\mathbf{1}_n$ representa un n-vector de unos y $\{u_i, v_i : 1, \ldots, n\}$ son variables de holgura que representan la parte positiva y negativa del vector residual $y - \boldsymbol{x}^\top \boldsymbol{\beta}_p$ de acuerdo a Koenker and Bassett (1978), que luego puede ser abordado por los métodos tradicionales de la programación lineal.

Existe una amplia variedad de trabajos asociados al método no paramétrico que consideran métodos como el vecino más cercano (KNN, por sus siglas en inglés) que permite además establecer de manera rigurosa tasas de convergencia y elementos importantes sobre la consistencia de los estimadores no paramétricos y otros enfoques donde se aborda el problema de la regresión cuantílica de manera localmente polinomial.

Una manera de hacer inferencia desde este enfoque a través de la regresión cuantílica es mediante métodos de remuestreo. Uno de los métodos de remuestreo más usados es el bootstrap (Efron, 1979) que consiste en remuestrear a partir de la muestra original, lo que a su vez trae como consecuencia un método costoso computacionalmente. No obstante, existe una gran cantidad de trabajos donde se usa el bootstrap en el problema de regresión cuantílica, como puede verse en Angelis et al. (1993), Buchinsky (1998), Hahn (1995) y muchos otros con técnicas de refinamiento, suavización y submuestreo.

Dado que el problema de regresión cuantílica puede ser formulado como un problema de programación lineal, entonces los enfoques tradicionales como el método Simplex, punto interior y punto exterior son viables para resolver el problema. Sin embargo, es costoso computacionalmente como es mencionado en Portnoy and Koenker (1997).

Enfoque parámetrico. El enfoque parámetrico es puramente probabilístico, permitiendo abordar el problema de regresión cuantílica desde los puntos de vista

Frecuentista o Bayesiano. La inferencia para los modelos de regresión cuantílica bajo este enfoque se basan en que la distribución de los errores sigue una distribución asimétrica de Laplace (ALD). Esta última posee una representación jerárquica que facilita la implementación de algoritmos eficientes, tales como, el muestreador de Gibbs y el algoritmo EM, como se describe en Lachos et al. (2015) y Galarza et al. (2020), respectivamente.

Una variable aleatoria Y tiene una distribución asimétrica de Laplace ALD con parámetro de localización μ, parámetro de escala $\sigma > 0$ y parámetro de asimetría $p \in (0, 1)$, si su función de densidad es dada por

$$f(y|\mu, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{y-\mu}{\sigma}\right)\right\}.$$

Sea $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_p$. Asumiendo que, $y_i \sim ALD(\mu_i, \sigma, p)$, entonces la función de verosimilitud para n observaciones independientes es dada por

$$L(\boldsymbol{\beta}_p, \sigma|y) = \frac{p^n(1-p)^n}{\sigma^n} \exp\left\{-\sum_{i=1}^{n} \rho_p\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_p}{\sigma}\right)\right\}. \tag{1.6}$$

Bajo estos supuestos, se puede escribir el problema de regresión cuantílica, es decir, la minimización de (1.4) como el problema de maximización de la verosimilitud arriba, debido a la relación inversa entre ambas. Nótese que si el cuantil de interés a modelar es la mediana ($p = 0.5$), el problema de regresión cuantílica se reduce a minimizar las desviaciones mínimas absolutas

$$\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \left|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_p\right|,$$

o a maximizar la función de verosimilitud

$$L(\boldsymbol{\beta}_p, \sigma|y) = \frac{p^n(1-p)^n}{\sigma^n} \exp\left\{-\frac{1}{2\sigma}\sum_{i=1}^{n} \left|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_p\right|\right\}.$$

En analogía con la regresión usual, si se tiene una regresión normal (en media) o el método de mínimos cuadrados, esto puede ser resuelto por el problema de minimización,

$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2,$$

cuya equivalencia se da al maximizar la verosimilitud normal

$$L(\boldsymbol{\beta}, \sigma^2|y) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2\right\}.$$

Un inconveniente con la distribución ALD es que no es diferenciable en cero, lo que tiene como consecuencias problemas de inestabilidad numérica. Wichitaksorn et al. (2014) introducen una clase general de distribuciones asimétricas (SKD) útiles para la implementación de la regresión cuantílica, proporcionando soluciones que compiten con la formulación basada en ALD. Esencialmente, definieron una distribución normal asimétrica con parámetro de localización μ, parámetro de escala $\sigma > 0$ y el parámetro de asimetría $p \in (0, 1)$ si esta tiene función de densidad dada por

$$f(y|\mu, \sigma, p) = 2\left[p\,\phi\left(y\left|\mu, \frac{\sigma^2}{4(1-p)^2}\right.\right)\mathbb{I}(y \leqslant \mu) + (1-p)\,\phi\left(y\left|\mu, \frac{\sigma^2}{4p^2}\right.\right)\mathbb{I}(y > \mu)\right], \quad (1.7)$$

donde $\phi(\cdot|\mu, \sigma^2)$ representa la función de densidad de probabilidad de la distribución normal univariada con media μ y varianza σ^2. La clase robusta de distribuciones SKD incluye como casos especiales versiones asimétricas de las distribuciones normal, t de Student, Laplace, entre otras. Esta clase de distribuciones además fue estudiada por Galarza et al. (2017) quién demostró que (1.7) puede ser escrita de manera conveniente como

$$f(y|\mu, \sigma, p) = \frac{4p(1-p)}{\sqrt{2\pi\sigma^2}} \exp\left\{-2\rho_p^2\left(\frac{y-\mu}{\sigma}\right)\right\}, \quad (1.8)$$

quién además introduce un algoritmo eficiente de tipo EM para la estimación en regresión cuantílica considerando (1.8) y que el error sigue a un miembro de la clase SKD.

Bajo el enfoque frecuentista, la inferencia es usualmente asintótica y debe ser tratada con cuidado cuando los tamaños muestrales no son grandes. Esta se basa en la propiedad de que los estimadores de máxima verosimilitud son consistentes y asintóticamente normales. Sin embargo, esto conduce a un segundo problema de optimización que es el cálculo o estimación de la matriz de información de Fisher.

Otra estrategia válida cuando la matemática es completa y el poder computacional abunda es utilizar métodos de remuestreo como el Bootstrap, donde el ajuste por regresión cuantílica debe ser realizado un número grande de veces. Por otro lado, el enfoque Bayesiano no sufre de estos problemas al disponer de toda la información sobre los estimadores a través de las bien conocidas distribuciones a posteriori.

Es importante mencionar que se han propuesto algunas soluciones en la literatura para tratar el problema de la mala especificación de la distribución de errores en la regresión cuantílica. Sin embargo, un problema de este enfoque consiste en la verificación de los supuestos, los que se pueden validar a través de métodos de bondad de ajuste. Finalmente, una forma de suavizar este problema es haciendo uso de métodos robustos que estiman los parámetros con una precisión adecuada aún cuando los supuestos del modelo no son satisfechos.

1.6 Regresión cuantílica aditiva parcialmente lineal

No siempre es razonable suponer linealidad entre las variables independientes y la variable de respuesta. Debido a las posibles relaciones no lineales que comprende la dinámica de la fotometría, nos es de interés usar un modelo de regresión cuantílica aditiva parcialmente lineal (PLAQR, por sus siglas en inglés) que permita estudiar, además de los efectos lineales, los efectos no lineales sobre el redshift. Para ver esto, escriba $\boldsymbol{x}_i = (\boldsymbol{v}_i^\top, \boldsymbol{w}_i^\top)^\top$, con $\boldsymbol{v}_i$ un q-vector de covariables con efectos lineales y $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ir})^\top$ un r-vector de covariables con efectos no lineales. Dada la muestra aleatoria $(\boldsymbol{x}_i^\top, y_i)^\top$ con $i = 1, \ldots, n$, el modelo de regresión cuantílica aditiva parcialmente lineal asume que

$$Q_p(y_i \mid \boldsymbol{x}_i) = \boldsymbol{v}_i^\top \boldsymbol{\beta}_p + \sum_{k=1}^{r} g_k(w_{ik}), \tag{1.9}$$

con $p \in (0, 1)$ y g_k una función no paramétrica suave no conocida. Alternativamente, se puede escribir el modelo como

$$y_i = \boldsymbol{v}_i^\top \boldsymbol{\beta}_p + \sum_{k=1}^{r} g_k(w_{ik}) + \epsilon_i, \tag{1.10}$$

donde los errores $\{\epsilon_i\}_{i=1}^n$ son independientes y satisfacen la restricción sobre el cuantil p, tal que, $\mathbb{P}(\epsilon_i \leqslant 0 \mid \boldsymbol{x}_i) = p$. Dado que, en este enfoque no se supone una distribución paramétrica para los errores ϵ_i, ni la varianza es asumida constante; entonces, este modelo resulta atractivo para estudiar, por ejemplo, datos no normales con presencia de heterocedasticidad. Para la indentificabilidad del modelo, se asume $\mathbb{E}[g_k(w_{ik})] = 0$. Los modelos de regresión cuantílica semiparamétricos considerados por He et al. (2002), He and Shi (1996) y Wang et al. (2009) son útiles para incorporar la no linealidad y evitar problemas de dimensionalidad.

En este estudio en particular, consideraremos un modelo PLAQR que aproxime las funciones no lineales suaves no conocidas g_k como una combinación lineal de funciones base B-splines. Más sobre esta construcción se puede ver en Schumaker (2007). Como es de esperarse, existen muchas funciones bases útiles para la aproximación de estas funciones suaves, como son las ondaletas, las bases de Fourier, entre otras. Estos métodos se estudian a profundidad en una rama de la Estadística: el análisis de datos funcionales.

Sin pérdida de la generalidad, asuma las covariables z_{ik} estandarizadas en el intervalo cerrado $[0, 1]$. Denote $\boldsymbol{\pi}(t) = (b_1(t), \ldots, b_{k_n+l+1}(t))^\top$ un vector normalizado de funciones base B-splines de orden $l + 1$ con k_n nodos de intervalo cuasi-uniforme en $[0, 1]$. Luego, $g_k(w_{ik})$ puede ser aproximado por $\boldsymbol{\pi}(z_{ik})^\top \boldsymbol{\zeta}_k$, donde los $\boldsymbol{\zeta}_k$, $k = 1, \ldots, q$, son parámetros que pueden ser estimados por los datos. La aproximación por B-splines es flexible y además es computacionalmente eficiente. Para simplificar, se utiliza el mismo número de funciones base para todos los componentes no paramétricos, pero esto no es necesario en la práctica.

El estimador para el modelo de regresión cuantílica parcialmente lineal aditivo es dado por

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^s}{\arg\min} \sum_{i=1}^{n} \rho_p \left[y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta} - \sum_{k=1}^{r} \boldsymbol{\pi}(w_{ik})^\top \boldsymbol{\zeta}_j \right], \tag{1.11}$$

donde $\boldsymbol{\theta} = (\boldsymbol{\beta}_p, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_q)^\top$, and $s = q + (k_n + l + 1) \times r$. El problema de optimización

puede ser resuelto mediante programación lineal como se indica en Koenker and d'Orey (1987). En la próxima sección se introduce la regularización, principio fundamental en la determinación de las funciones suaves que describen las tendencias no lineales en el modelo PLAQR.

1.7 Regularización

La regularización en Matemáticas permite introducir cierta información adicional a un modelo con el objetivo de solucionar un problema mal definido, modificando la solución original obtenida a través de una objetivo. En el contexto del modelado estadístico, la regularización permite corregir la multicolinealidad (dependencia lineal entre las variables regresoras), huir del sobreajuste, la selección de variables, entre otros.

1.7.1 Regularización de funciones

Modelos de regresión como el de la ecuación (1.10), cuyo modelado involucra la determinación de funciones suaves, se fundamenta en el Análisis Funcional y el Álgebra matricial para poder construir las mismas a través de las conocidas funciones bases.

Un sistema de funciones base es un conjunto de funciones conocidas $\{\phi_i\}_{i\in I}$ con I alguna familia de índices (por ejemplo, $\mathbb{Z}$ o $\mathbb{N}$) en un espacio funcional $\mathcal{F}$ definido sobre un campo $\mathbb{F}$ (por ejemplo, $\mathbb{R}$ o $\mathbb{C}$) que satisface las siguientes condiciones:

1. Cualquier vector en $\mathcal{F}$ puede ser escrito como una combinación lineal de los ϕ_i;
2. Si $\sum_{i=1}^{n} c_i\phi_{\sigma(i)} = 0$, entonces $c_i = 0$, para todo $i \in I$ (independencia lineal de los ϕ_i).

En un espacio funcional $\mathcal{F}$ con un sistema de funciones base $\{\phi_i\}_{i\in I}$, podemos representar cualquier $f \in \mathcal{F}$ por definición como

$$f(x) = \sum_{i=1}^{n} c_i\phi_i(x), \tag{1.12}$$

en términos de n funciones base conocidas ϕ_i. Algunos ejemplos de sistemas de funciones base, son los siguientes:

- En el espacio de funciones analíticas $\mathcal{C}^\omega$, el conjunto $\{t^n : n \in \mathbb{N}\}$ conforma un sistema de funciones monomiales base para $\mathcal{C}^\omega$.

- En el espacio de funciones de valor real cuadrado integrables definidas sobre $[0,1]$, $L^2([0,1];\mathbb{R})$, el conjunto $\{1\} \cup \{\sqrt{2}\text{sen}(2\pi nt) : n \in \mathbb{N}\} \cup \{\sqrt{2}\cos(2\pi nt) : n \in \mathbb{N}\}$ conforma un sistema de funciones base para $L^2([0,1];\mathbb{R})$.

- Sea una sucesión nodal definida como $\boldsymbol{U} := (u_i)_{i=1}^m = \{u_1 \leqslant u_2 \leqslant \cdots \leqslant u_m\}$, con m un número natural. Suponga para un entero no negativo p y algún entero j tal que $u_j \leqslant u_{j+1} \leqslant \cdots \leqslant u_{j+p+1}$ son $p+2$ números reales tomados de la sucesión $\boldsymbol{U}$. La j-ésima B-Spline función $B_{j,p,\boldsymbol{U}} : \mathbb{R} \longrightarrow \mathbb{R}$ de grado p es idénticamente cero si $u_{j+p+1} = u_j$ y de otra manera es definida recursivamente por

$$B_{j,p,\boldsymbol{U}} := \frac{x - u_j}{u_{j+p} - u_j} B_{j,p-1,\boldsymbol{U}}(x) + \frac{u_{j+p+1} - x}{u_{j+p+1} - u_{j+1}} B_{j+1,p-1,\boldsymbol{U}}(x), \tag{1.13}$$

  empezando con

$$B_{i,0,\boldsymbol{U}}(x) := \begin{cases} 1, & x \in [u_i, u_{i+1}), \\ 0, & \text{otro caso.} \end{cases} \tag{1.14}$$

  El conjunto de funciones nodales o B-Splines conforma un sistema de funciones base.

Las funciones B-Splines tienen un rol importante en el presente trabajo y como se puede ver en (1.13) se pueden calcular por recursividad y así programarlas no es complicado, por lo que es necesario presentar algunas características importantes de este sistema de funciones base.

- Poseen la propiedad de partición de la unidad, esto es, $\sum_{j=1}^{n} B_{j,p,\boldsymbol{U}} = 1$.

- Tienen soporte local, es decir, $B_{j,p,\boldsymbol{U}}(x) = 0$, para todo $x \notin [u_j, u_{j+p+1})$.

- Son no negativas en todas partes y positivas dentro de su soporte, esto se puede escribir formalmente como, $B_{j,p,\boldsymbol{U}}(x)(x) \geqslant 0$, para todo $x \in \mathbb{R}$ y $B_{j,p,\boldsymbol{U}(x)} > 0$ para todo $x \in (u_j, u_{j+p+1})$.

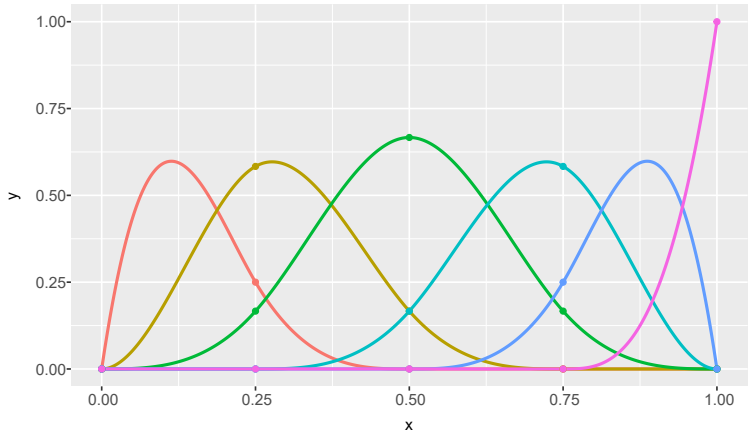

Figura 1.4. Seis bases B-Splines cúbicas en $[0,1]$ con tres nodos interiores igualmente espaciados.

- Una función B-Spline es invariante bajo traslación y/o transformación por escala sobre la sucesión $\boldsymbol{U}$, esto es, $B_{j,p+\alpha\boldsymbol{U}+\beta}(\alpha x+\beta) = B_{j,p,\boldsymbol{U}}(x)$, con $\alpha, \beta \in \mathbb{R}$, $\alpha \neq 0$ y $\alpha\boldsymbol{U}+\beta = (\alpha u_j + \beta, \ldots, \alpha u_{j+p+1} + \beta)$.

La idea es que al comportarse bien las funciones nodales en el sentido que conforman un sistema base, entonces permiten escribir curvas como en (1.12), donde los c_i en particular se denominan puntos de control para el caso de las funciones B-Spline. Un ejemplo de seis funciones bases B-Splines cúbicas definidas en el intervalo $[0,1]$, se puede ver en la figura 1.5.

Por otro lado, sea $\boldsymbol{c}$ un vector de dimensión n de los coeficientes c_i y $\boldsymbol{\phi}$ el vector funcional cuyos elementos pertenecen a las funciones base ϕ_i. Entonces, se puede reescribir (1.12) con la notación introducida como,

$$f(t) = \boldsymbol{c}^\top \boldsymbol{\phi} = \boldsymbol{\phi}^\top \boldsymbol{c}. \tag{1.15}$$

Si se estima una función f minimizando la ecuación que equilibra el ajuste de mínimos cuadrados, y una penalización por rugosidad (falta de suavidad) a través de su k-ésima derivada, se tiene que la solución satisface,

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i)\right)^2 + \lambda \int \left(f^{(k)}(\boldsymbol{x})\right)^2 \, \mathrm{d}x, \tag{1.16}$$

sobre un conjunto apropiado de funciones (por ejemplo, el espacio habitual de Hilbert cuadrado integrables). La solución obtenida es conocida como Smoothing Splines (SS). Vale mencionar que la función de pérdida cuadrática en el primer sumando de (1.16) puede ser substituida por cualquier otra función de pérdida de interés.

Objetivos en competencia en la estimación de funciones. El método implementado por funciones Smoothing Splines estima una curva f desde las observaciones $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ haciendo explícito dos objetivos conflictivos en la estimación de la curva. Por un lado, se desea asegurar que la curva estimada se ajusta bien a los datos, por ejemplo en términos de la suma residual de cuadrados (SSE) $\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2$. Adicionalmente, no se desea que la curva se ajuste demasiado a las observaciones individuales sino que exista un modelamiento de la tendencia. Este problema, más conocido como sobreajuste, da lugar a una curva x que es excesivamente "ondulada" o localmente variable como se ilustra en la figura 1.5. Los dos objetivos en competencia pueden ser compensados a través de una medida fundamental de la teoría de estimación estadística: el error cuadrático medio. Esta se define como

$$\text{Error cuadrático medio} = \text{Sesgo}^2 + \text{Varianza},$$

Por ejemplo, un estimador $\hat{f}$ de f se dice insesgado si esta es una curva que ajuste exactamente a los valores de y_i, i.e., obteniéndose una interpolación y, consecuentemente, una alta variabilidad debido a la rápida variación local de la curva. En el suavizado por Splines, como en otros métodos de suavizado, el error cuadrático medio es una forma de captar lo que solemos entender por la calidad de la ajuste. Vale mencionar que otras funciones de pérdida pueden ser preferibles en determinadas situaciones o por conveniencia matemática. Por ejemplo, se podría utilizar la función de pérdida cuantílica (1.5) o el negativo de una función de verosimilitud para un correspondiente modelo de interés.

Cuantificación de la rugosidad. El cuadrado de la segunda derivada $(D^2 f(x))^2$ de una función es frecuentemente denominada su curvatura en x. Así, una medida natural de la

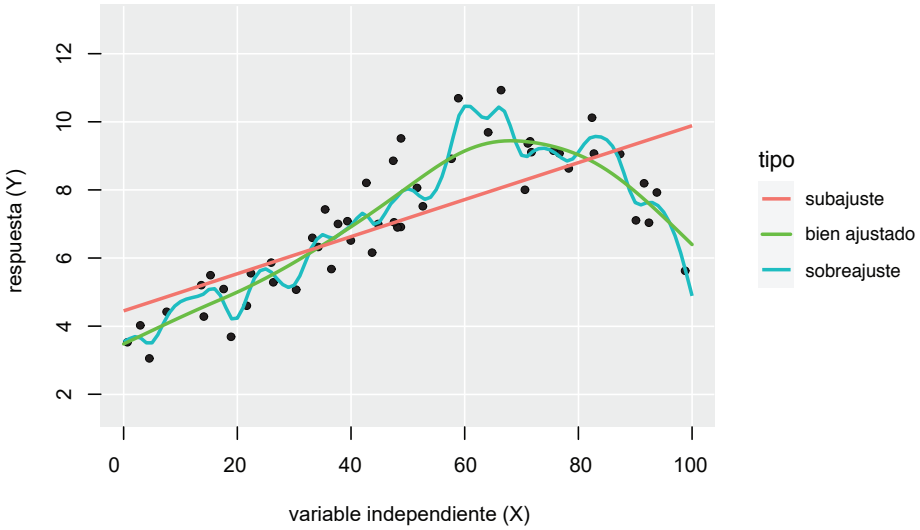

Figura 1.5. De color rojo una curva con subajuste, de color verde una curva bien ajustada a los datos y de color azul una curva con sobreajuste respecto a los datos.

rugosidad de una función es dada por

$$\text{PEN}_2(f(x)) = \int (D^2 f(s))^2 \, \mathrm{d}s.$$

Note que es de esperar que las funciones altamente variables produzcan valores altos de $\text{PEN}_2(f(x))$ porque sus segundas derivadas son grandes en al menos parte del rango de interés. Por otro lado, siendo f una función lineal, su cantidad de rugosidad o curvatura es cero, i.e., $\text{PEN}_2(f(x)) = 0$.

Por otra parte, sea $f(\boldsymbol{x})$ el vector resultante de la función f evaluada en el vector $\boldsymbol{x}$. Una medida que compensa explícitamente la suavidad con el ajuste de los datos, es la suma cuadrática del error penalizada definida como

$$\text{PENSSE}_\lambda(\boldsymbol{x} \mid \boldsymbol{y}) = (\boldsymbol{y} - f(\boldsymbol{x}))^\top \boldsymbol{\Omega} \, (\boldsymbol{y} - f(\boldsymbol{x})) + \lambda \, \text{PEN}_2(f(\boldsymbol{x})),$$

siendo $\boldsymbol{\Omega}$ una matriz de pesos, y donde λ opera como un parámetro de suavización. La estimación de la función es obtenida, encontrando la función f que minimiza $\text{PENSSE}_\lambda(f(x))$ sobre el espacio de funciones f para el cual $\text{PEN}_2(f(x))$ está bien definida. El parámetro λ mide la razón de intercambio entre el ajuste a los datos,

medido por SSE en el primer término, y la variabilidad de la función f, cuantificada por $\text{PEN}_2(f(x))$ en el segundo término. A medida que el parámetro λ se hace más grande, las funciones que no son lineales deben incurrir en una penalización de rugosidad cada vez más importante a través de el término $\text{PEN}_2(f(x))$ y, en consecuencia, el criterio compuesto $\text{PENSSE}_\lambda(f(x))$ debe poner cada vez más énfasis en la suavidad de f y menos en el ajuste de los datos. Por esta razón, haciendo que $\lambda \to \infty$ la curva ajustada f debe acercarse a la solución de mínimos cuadrados ponderados (o MCO para $\boldsymbol{\Omega} = \boldsymbol{I}_n$, una matriz identidad $n \times n$), donde $\text{PEN}_2(f(x)) \to 0$. Por otro lado, para λ suficientemente pequeño la curva tiende a hacerse cada vez más variable ya que cada vez se penaliza menos su rugosidad, y a medida que se hace $\lambda \to 0$ la curva x aproxima una interpolante a los datos, satisfaciendo $f(\boldsymbol{x}_i) = y_i$ para toda observación i. Sin embargo, incluso en este caso límite, la curva interpolante no será arbitrariamente variable sino que es la curva más suave en $\mathcal{F}$ y dos veces diferenciable que se ajuste exactamente a los datos.

Nótese que sin una penalización por rugosidad, el vector de coeficientes $\boldsymbol{c}$ en la expansión $f(\boldsymbol{x}) = \boldsymbol{\phi}^\top \boldsymbol{c}$, donde $\boldsymbol{c}$ es el n-vector de coeficientes y $\boldsymbol{\phi}$ es el n-vector de funciones bases, tiene solución

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{y},$$

donde $\boldsymbol{\Phi}$ contiene los valores de las n funciones base, $\boldsymbol{\Omega}$ es una matriz de pesos (e.g., para establecer una estructura de covarianza entre los residuos), y donde $\boldsymbol{y}$ es el vector de datos a ser suavizado. La expresión correspondiente para obtener los valores estimados $\hat{\boldsymbol{y}}$ está dada por

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{y} = \boldsymbol{S}_\phi \boldsymbol{y},$$

donde $\boldsymbol{S}_\phi$ es el operador proyección,

$$\boldsymbol{S}_\phi = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Omega},$$

correspondiente para el sistema base $\boldsymbol{\phi}$. Esta última matriz tiene la propiedad de ser idempotente y es referida en el contexto de la regresión lineal múltiple como la matriz *hat* o sombrero.

Con respecto a la cuantificación de la rugosidad, de forma general se puede definir matricialmente la rugosidad penalizada de orden m, denotada por $\mathsf{PEN}_m(f(x))$ como

$$\begin{aligned}
\mathsf{PEN}_m(f(x)) &= \int (D^m f(s))^2 \,\mathsf{d}s, \\
&= \int (D^m \boldsymbol{c}^\top \boldsymbol{\phi}(s))^2 \,\mathsf{d}s, \\
&= \int \boldsymbol{c}^\top D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^\top(s) \boldsymbol{c} \,\mathsf{d}s, \\
&= \boldsymbol{c}^\top \left(\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^\top(s) \,\mathsf{d}s \right) \boldsymbol{c}, \\
&= \boldsymbol{c}^\top \boldsymbol{R} \boldsymbol{c}.
\end{aligned}$$

Luego, $\mathsf{PEN}_m(f(x))$ puede ser escrito como

$$\mathsf{PEN}_m(f(x)) = \boldsymbol{c}^\top \boldsymbol{R} \boldsymbol{c},$$

donde

$$\boldsymbol{R} = \int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^\top(s) \,\mathsf{d}s.$$

Sumando la suma cuadrática del error $\mathsf{SSE}(\boldsymbol{y}|\boldsymbol{c})$ y $\mathsf{PEN}_m(f(\boldsymbol{x}))$ multiplicado por λ, se tiene que

$$\mathsf{PENSSE}_m(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c})^\top \boldsymbol{\Omega} (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c}) + \lambda \boldsymbol{c}^\top \boldsymbol{R} \boldsymbol{c}.$$

Tomando la derivada respecto al vector $\boldsymbol{c}$, se obtiene

$$-2\boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{y} + \boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{\Phi} \boldsymbol{c} + \lambda \boldsymbol{R} \boldsymbol{c} = \boldsymbol{0},$$

de donde se obtiene una expresión para el vector de coeficientes estimado,

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{\Phi} + \lambda \boldsymbol{R})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Omega} \boldsymbol{y}, \tag{1.17}$$

donde en el contexto de regresión, el vector $\hat{\boldsymbol{c}}$ representaría los coeficientes estimados de regresión.

Vale mencionar que la expresión (1.17) es conocida como la solución a la regularización de Tikhonov, un método comúnmente usado para problemas que no están bien propuestos en el sentido de Hadamard. Un caso particular de esta regularización (fuera

del contexto funcional) es la regresión Ridge, la cual se obtiene cuando $\boldsymbol{\Omega} = \boldsymbol{I}_n$ y $\boldsymbol{R} = \boldsymbol{I}_p$, respectivamente, y que es ampliamente usada para corregir el problema de multicolinealidad. Finalmente, debe tenerse en cuenta que la expresión arriba es posible pues el término de penalización ha sido expresado de forma cuadrática. Sin embargo, existen otros tipos de penalizaciones de interés que no son de este tipo como la que veremos a continuación.

1.7.2 Regularización para la selección de variables

El método LASSO (Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés), introducido por Tibshirani (1996), es el método más usado en modelos de regresión para realizar la selección de variables con el objetivo de mejorar la exactitud e interpretabilidad del modelo estadístico producido, al seleccionar las características relevantes para el modelo y eliminando aquellas cuya aportación es ínfima. Este método es la base para determinar las características relevantes en casos donde el número de variables es muy grande, en muchos casos siendo mucho mayor al número de observaciones ($n \ll p$), y donde las soluciones tradicionales como mínimos cuadrados no son factibles. Una ilustración que contrasta la regresión de Ridge mencionada en la sección anterior con la regresión por LASSO puede verse en la figura 1.6.

Sin pérdida de generalidad, considerando la función de pérdida cuadrática, la solución estará dada por

$$\hat{\boldsymbol{\beta}}_{L_1,\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\|\boldsymbol{\beta}\|_{L_1},$$

siendo el primer sumando la suma cuadrática del error y el segundo sumando el término de penalización, donde la normal 1 del vector de solución $\boldsymbol{\beta}$ es controlado a través del parámetro de regularización λ. Sabiendo que

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{q}\beta_j x_{ij}, \quad \text{y}, \quad \|\boldsymbol{\beta}\|_{L_1} = \sum_{j=1}^{q}|\beta_j|,$$

se obtiene

$$\hat{\boldsymbol{\beta}}_{L_1,\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\arg\min} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{q}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{q}|\beta_j|. \tag{1.18}$$

Para todo λ, existe un $\boldsymbol{\beta}$ particular. Si $\lambda = 0$, entonces se obtiene la solución de MCO. Si

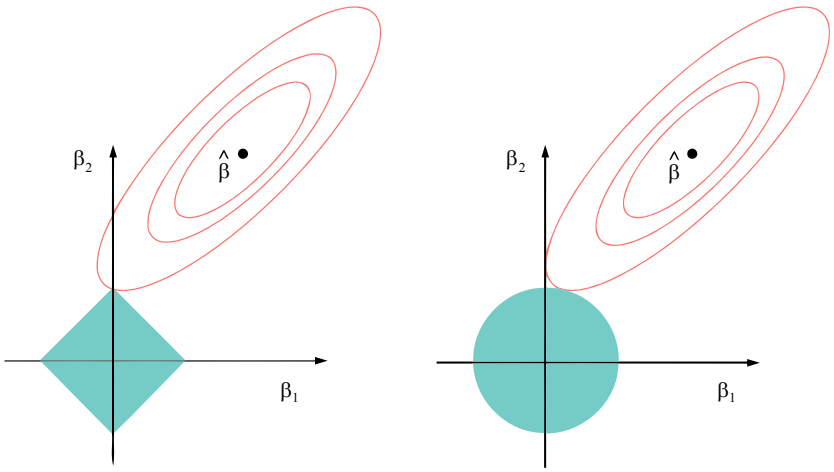

Figura 1.6. Contornos de las funciones de error y restricción para la regresión LASSO (izquierda) y regresión de Ridge (derecha). Las áreas azules sólidas son las regiones de restricción, $|\beta_1| + |\beta_2| \leqslant s$ y $\beta_1^2 + \beta_2^2 \leqslant s$ mientras que las elipses rojas son los contornos de el RSS. Fuente: James et al. (2019).

$\lambda \to \infty$, entonces los coeficientes estimados tenderán a cero. Así, el valor de λ óptimo, que permita destacar las variables importantes del modelo, puede ser determinando a través de métodos de búsqueda tal que se minimice alguna función objetivo que podría estar relacionada con error de predicción o con la bondad de ajuste de un modelo propuesto (léase sección 5.1 de James et al. 2019). Finalmente, la penalización L_1 induce a estimadores de menor varianza que los de MCO, aunque sesgados. La regresión LASSO es en la actualidad una piedra angular en el análisis de grandes volúmenes o más conocido por su término acuñado en inglés, Big Data. Entre sus aplicaciones más exitosas se encuentra la compresión de archivos, el estudio de enfermedades genéticas, la descomposición de señales, el análisis de sentimientos, entre otros.

CAPÍTULO 2

2. METODOLOGÍA

Para el presente trabajo se usó el conjunto de datos “SDSS quasar” que contiene 46420 filas conformadas por observaciones de cuásares, y 23 columnas de variables con las siguientes características:

- `SDSS_J`: designación SDSS;
- `R.A.`: ascensión recta (coordenada celeste equivalente a la longitud en la Tierra, de 0 a 360 grados);
- `Dec.`: declinación (coordenada celeste equivalente a la latitud en la Tierra, de -90 a +90 grados);
- `z`: desplazamiento al rojo (escala con la distancia);
- `u_mag`: brillo en la banda `u` (ultravioleta) en magnitudes. Las magnitudes son una unidad logarítmica invertida de brillo (un cuásar con `u_mag` $= 16$ es 100 veces más brillante que uno con `u_mag=21`);
- `sig_u`: error de medición de `u_mag`. Los errores de medición heteroscedásticos para cada magnitud son determinados por el equipo del SDSS a partir del conocimiento de las condiciones de observación, el fondo del detector y otras consideraciones técnicas;
- `g_mag`: brillo en la banda g (verde);
- `sig_g`: error de medición de `g_mag`;
- `r_mag`: brillo en la banda r (roja);
- `sig_r`: error de la medición de `r_mag`;

- `i_mag`: brillo en la banda i (más roja);
- `sig_i`: error de la medición de `i_max`;
- `z_mag`: brillo en la banda z (aún más roja);
- `sig_z`: error de la medición de `z_mag`;
- `Radio`: brillo en la banda de radio, en magnitudes a escalas a partir de la densidad de flujo medida en el Observatorio Nacional de Radioastronomía de Imágenes débiles del cielo radioeléctrico a 20 cm (o FIRST, por sus siglas en inglés). El 0 indica que el cuásar no fue detectado por FIRST, mientras que -1 indica que no fue observado por FIRST;
- `X.ray`: brillo en la banda de rayos X, en $\log(\text{tasa de recuento})$ del ROSAT All-Sky Survey (RASS) en la banda de $0,2-2,4$ keV. El -9 indica que no ha sido detectado por RASS;
- `J_mag`: brillo en la banda J del infrarrojo cercano, en magnitudes, del catálogo 2MASS Point Source;
- `sig_J`: error de la medición de `J_mag`;
- `H_mag`: brillo en la banda H del infrarrojo cercano;
- `sig_H`: error de la medición de `H_mag`;
- `K_mag`: brillo en la banda K del infrarrojo cercano;
- `sig_K`: error de la medición de `K_mag`;
- `M_i`: la magnitud absoluta en la banda i. Se trata de una medida logarítmica invertida de la luminosidad intrínseca del cuásar, donde un cuásar con $\texttt{M\_i} = -29$ es 100 veces más luminoso que uno con $\texttt{M\_i} = -24$.

La variable de respuesta es el desplazamiento al rojo, `z`. Nótese además que entre las variables independientes, existen algunas que son parcialmente observadas. Unas de ellas presentan errores de medición, censuras, y valores completamente no

observados. Estos tipos de problemas son estudiados por modelos estadísticos específicos; sin embargo, a lo mejor de nuestro conocimiento, no existe en literatura aún métodos de regresión cuantílica parcialmente lineal que incorporen estas características. Por simplicidad, los errores de medida fueron tratados también como variables independientes con el fin de que aporten información respecto a la variabilidad al momento de medir su covariable respectiva.

Previo al análisis estadístico, se realizó una validación y selección de variables potencialmente significativas para predecir la variable de interés. Se pudieron observar dos variables en particular, `X.ray` y `Radio`, que presentan al menos un 90% de valores faltantes, como se puede ver en la figura 2.1.

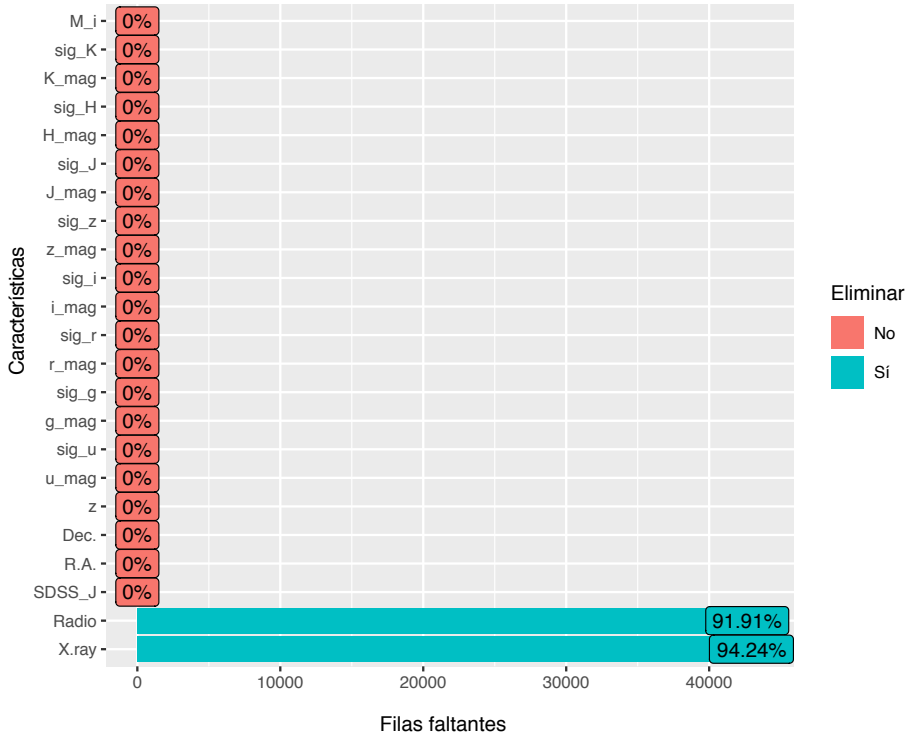

Figura 2.1. Análisis de valores faltantes.

Sin duda, esta es una situación donde no es apropiado aplicar métodos de imputación por lo que las variables fueron descartadas del estudio. Adicionalmente, la variable `SDSS_J` fue eliminada del modelo pues es una etiqueta que no aporta información.

Luego, se usó el Índice de Inflación de la Varianza (VIF, por sus siglas en inglés) junto a la matriz de correlación lineal, para la eliminación de las variables que se encuentren altamente correlacionadas entre sí, evitando problemas de multicolinealidad, concurvidad, y consecuentemente la singularidad de la matriz de diseño.

Como paso intermedio, se ajustó el modelo de regresión cuantílica lineal LASSO, esto es, un método de regularización cuyo término de penalización en L_1 induce la selección de características relevantes. A partir de estos resultados preliminares se obtuvieron las variables que podrían resultar significativas al momento de ajustar el modelo. Además, debido a que el modelo ajustado impone una estructura lineal, la significancia de las variables, a través de sus valores p, fue estudiada para clasificar las variables en dos grupos: aquellas que se relacionan con la respuesta de forma lineal, y las que no (posible relación no lineal, o nula). Luego, un análisis descriptivo exploratorio de las variables fue realizado con el fin de estudiar las relaciones entre las variables de ambos grupos y la respuesta.

Una vez realizado el preprocesamiento de datos y su respectivo análisis, se procedió a ajustar un modelo cuantílico parcialmente lineal para diversos cuantiles; en particular, se modeló el percentil 90 de la variable del desplazamiento al rojo o redshift $\mathtt{z}$, lo que permitió encontrar las características más relevantes, sean estas con relación lineal o no lineal para caracterizar a aquellos agujeros negros más antiguos del universo. Como es usual en la práctica, las variables que resultaron no significativas fueron descartadas del modelo para así obtener lo que llamaremos de aquí en adelante el “modelo final”.

Por último, todo lo descrito anteriormente fue realizado usando el software libre R por medio de la interfaz web RStudio Cloud. Debido a la dimensión del conjunto de datos y a las limitaciones para usuarios gratuitos, se usó el servidor web de la Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia - CEDIA, el que provee acceso al clúster de alto rendimiento a investigadores de universidades ecuatorianas.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

El presente capítulo se divide en dos secciones en las cuales se muestran los resultados obtenidos y su respectivo análisis. La primera sección corresponde a un análisis estadístico exhaustivo con el fin de ajustar el modelo PLAQR propuesto sobre el conjunto de datos del SDSS. Esto involucrará un estudio descriptivo univariado y multivariado de las variables inmersas en el modelo, con principal énfasis en la variable de respuesta, y cómo esta se relaciona con las variables independientes. Finalmente, se mostrarán los resultados y se discutirán los hallazgos obtenidos de los modelos de regresión cuantílica.

3.1 Análisis Estadístico

3.1.1 Análisis Descriptivo

Análisis univariado de la variable de respuesta. Para describir la distribución de la variable de respuesta z, se procedió a realizar gráficos descriptivos; en particular, por medio de un histograma de frecuencias y un diagrama de cajas. La figura 3.1 muestra un histograma de frecuencia relativa con una curva de densidad estimada (línea sólida salmón). Además, en su parte inferior se presenta un diagramas de cajas. De este último se puede observar que el valor mínimo del desplazamiento al rojo es cercano a cero (0.0780, para ser exactos), siendo a su vez el máximo valor observado de 5.4135. La caja luce algo simétrica, concentrando la mitad de las observaciones en un rango de aproximadamente una unidad, i.e., tomando valores entre 0.9415 y 1.9127. Consecuencia de esta concentración de observaciones es la cercanía de las medidas de tendencia central media y mediana, tomando los valores de 1.4717 y 1.5178. Por otro lado, existe una fuerte asimetría en los datos, lo cual es evidente en la cola derecha del histograma de frecuencias. Esto es debido a la gran cantidad de valores atípicos (aquellos cuyo valor de desplazamiento al rojo sea mayor a 3.3695), los cuales han sido

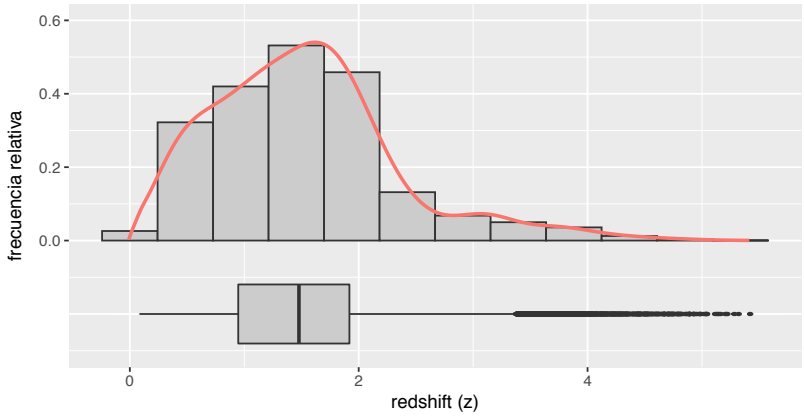

Figura 3.1. Histograma de frecuencia relativa y diagrama de cajas del desplazamiento al rojo. La línea sólida roja representa la curva de densidad ajustada vía kernels.

colocados en evidencia por diagrama de cajas. Respecto a esto último, ayudados del cuartil tres podemos concluir que un 25% de los cuásares presenta un desplazamiento al rojo de al menos 1.9127 unidades. Finalmente, con fines ilustrativos, la curva de densidad ofrece una representación suave, en comparación al histograma de la función de densidad de probabilidad del desplazamiento al rojo o redshift.

Análisis bivariado. Con el objetivo de estudiar la relación entre las variables independientes y la variable de respuesta, se procedió a realizar diagramas de dispersión dos a dos. Estos se muestran en la matriz triangular inferior de la figura 3.2. Para poder destacar las posibles relaciones no lineales, se muestran curvas de tendencia las que han sido calculadas por medio de B-splines. Adicionalmente, esta figura presenta los coeficientes de correlación respectivos por encima de la diagonal, y curvas de densidades en la diagonal principal. Por fines ilustrativos, a más de la variable de respuesta, se consideraron únicamente seis de las variables independientes.

La relación entre la variable de respuesta `z`, y las variables independientes `u_mag`, `sig_u`, `g_mag`, `sig_g`, `r_mag` y `sig_r`, se puede apreciar en la primera columna. De esta última es notorio que existen tendencias no lineales bastante evidentes como es el caso de las variables `sig_u` y `g_mag`, teniéndose coeficientes de correlación positivos, medianamente grandes para las variables `u_mag`, `sig_u` y `g_mag`. Nótese que estas

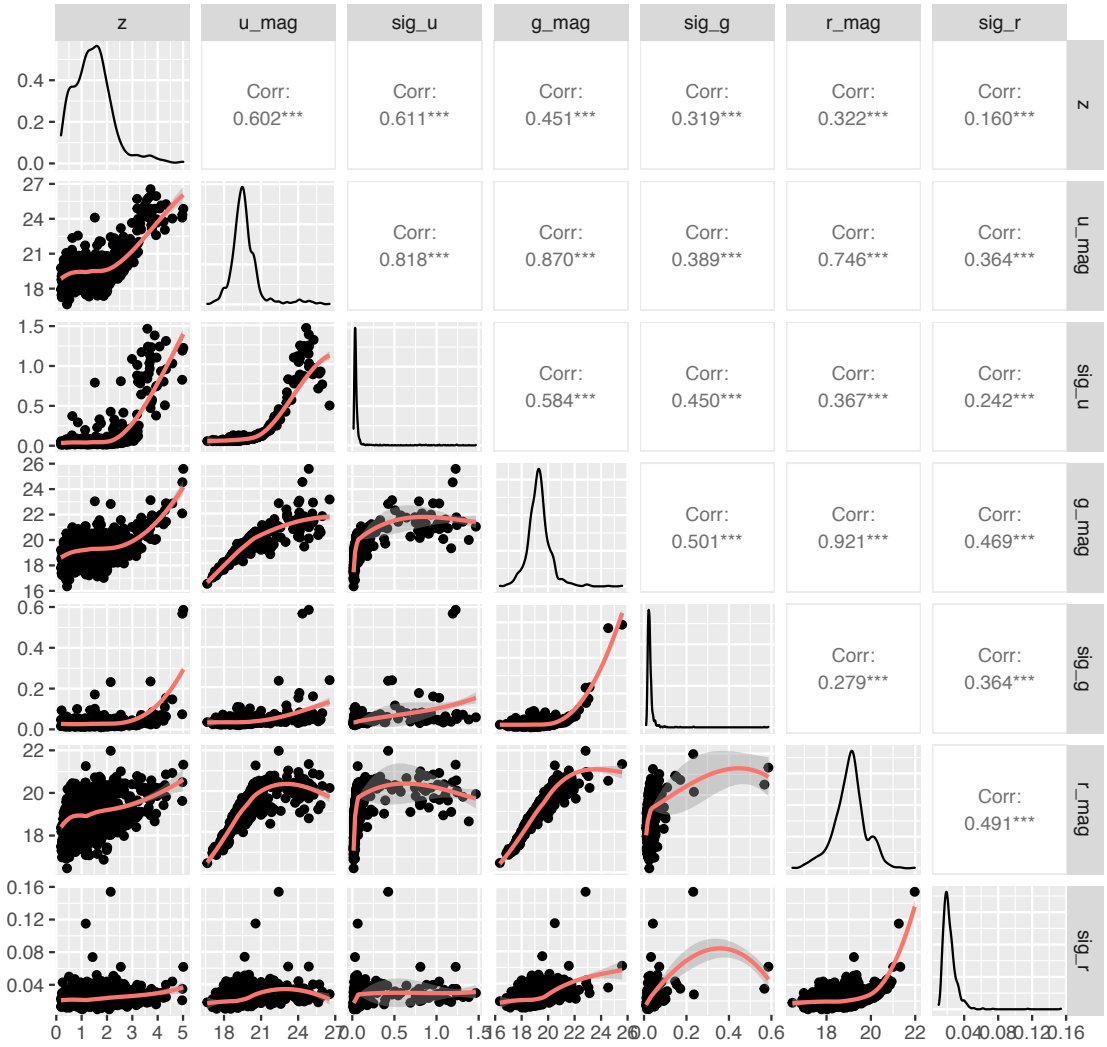

Figura 3.2. Diagramas de dispersión para siete variables del modelo. Las tendencias no lineales son representadas por curvas obtenidas usando B-splines.

relaciones no lineales justifican el uso de un modelo parcialmente lineal, como el modelo PLAQR propuesto. Además, dado que todas las correlaciones que se pueden observar para los pares de variables son positivas, entonces estas se encuentran relacionadas de manera directamente proporcional.

Con respecto a la relación y tendencias entre las variables independientes, se puede observar un grave problema de concurvidad, i.e., la versión no lineal de la colinealidad, la que resulta incluso más nociva que su análoga lineal al momento de realizar operaciones algebraicas durante el proceso de optimización (Morlini, 2006). Por

ejemplo, se puede observar que el brillo en la banda verde (`g_mag`) y el brillo en la banda roja (`r_mag`) presentan una tendencia muy evidente, además de presentar un muy alto coeficiente de correlación de 0.921. Así, para el modelo final en cuestión, bastaría la información aportada por una única de estas variables.

Con el objetivo de analizar la correlación lineal entre todas las variables independientes y la variable de respuesta, se realizó un mapa de correlación lineal que puede observarse en la figura 3.3, las mismas que se encuentran destacadas en las celdas enmarcadas en gris. Vale mencionar que, para observar una correlación en particular, se fija una variable en la fila (o en su defecto, en la columna) y se encuentra la intersección con la columna o fila que se desea analizar. Por ejemplo, en la figura 3.3 se puede observar en la intersección entre la variable de respuesta desplazamiento al rojo `z` y la magnitud de la banda `i`, esto es `M_i`, la cuál es una medida logarítmica invertida de la luminosidad intrínseca del cuásar, una correlación lineal alta y de valor numérico

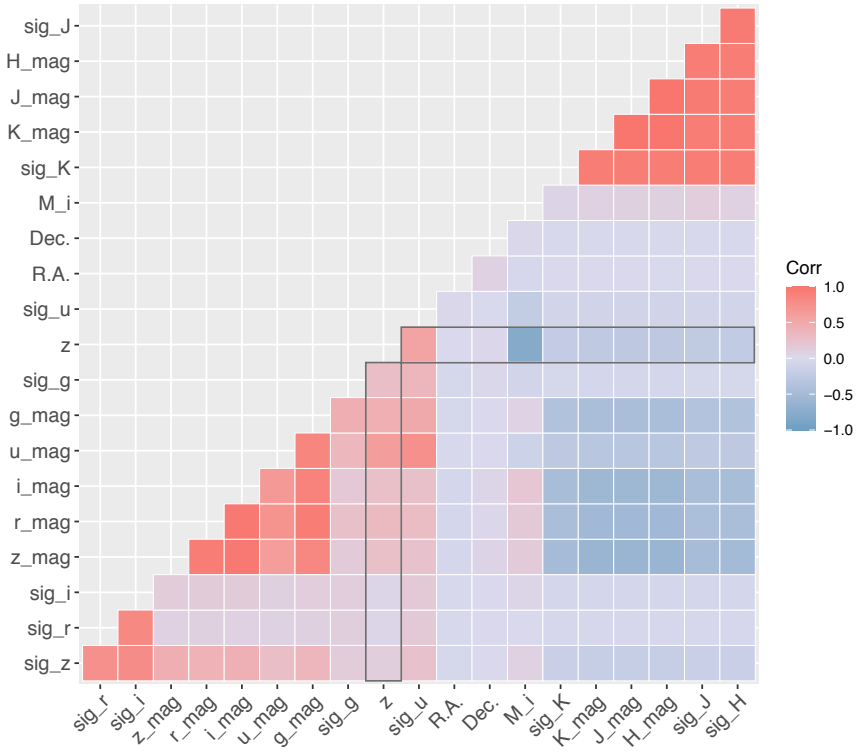

Figura 3.3. Mapa de correlaciones lineales.

negativo de -0.792. La segunda variable independiente más fuerte relacionada (linealmente) con el desplazamiento al rojo que se encontró fue la variable que almacena el brillo de la banda ultravioleta de un cuásar (`u_mag`), con una correlación medianamente alta y positiva de 0.625. Finalmente, del mapa de correlación lineal se logró apreciar la existencia de agrupamientos (colores intensos agrupados fuertemente en secciones del mapa), como los que responden a las variables independientes `sig_H`, `sig_J`, `H_mag`, `J_mag`, `K_mag` y `sig_K`. Estas variables almacenan información física como el error de medición del `H_mag`, error de medición del `J_mag`, el brillo de la banda del infrarrojo cercano y el brillo de la banda `J` del infrarrojo cercano, por lo cual, dada la naturaleza física de estas variables tiene sentido la existencia de dichos agrupamientos.

3.1.2 Modelamiento estadístico

Regresión cuantílica no lineal simple. De forma preliminar, fue de interés modelar la relación entre el desplazamiento al rojo y la variable independiente que presente una mayor relación lineal con la misma. Así, se propuso un modelo que regresión cuantílica simple que explique el p-ésimo cuantil del desplazamiento al rojo como una función no lineal de la magnitud absoluta en la banda. La ecuación del modelo está dada por

$$Q_p(\mathtt{z}) = \beta_{0p} + g_p(\mathtt{M\_i}), \tag{3.1}$$

siendo $p \in (0, 1)$ el cuantil de interés, $g_p(\cdot)$ una función suave obtenida vía B-splines. El subíndice i, correspondiente al cuásar i, ha sido omitido por simplicidad. Este modelo permite ilustrar la aplicación de las técnicas de regresión cuantílica para el presente estudio al considerar una única variable independiente, lo que a su vez permite visualizar las curvas de predicción de cada uno de los cuantiles de interés, lo cual no será posible para el modelo completo ya que este opera en $\mathbb{R}^{q+r}$.

El modelo en (3.1) fue ajustado para los cuantiles 0.025, 0.10, 0.50, 0.90 y 0.975, cuyas curvas de predicción respectivas pueden observarse en la figura 3.4. Se puede apreciar la curva de la mediana condicional en color verde, la que resultaría útil para predecir el desplazamiento al rojo para cuásares “promedios”, esto es, cuásares con un tiempo

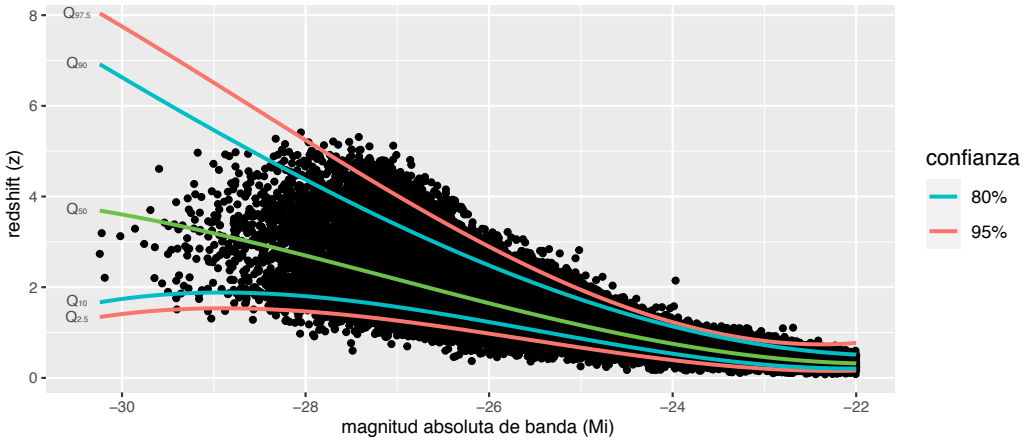

Figura 3.4. Diagrama de dispersión del desplazamiento al rojo vs. la magnitud de la banda y curvas de predicción para los cuantiles 0.025, 0.10, 0.50, 0.90 y 0.975.

histórico esperado (cercano a la media). Nótese que esta curva resulta una alternativa robusta a una curva de predicción media obtenida por una regresión no lineal tradicional. Por otro lado, las curvas en color azul y rojo constituyen intervalos de predicción del 80% y 95%, respectivamente; es más, es la curva para el percentil 90 $Q_{0.9}$ (aquella que separa al 10% de los agujeros negros masivos más antiguos de los restantes) en la que estaremos interesados en el estudio.

De la figura 3.4 podemos destacar además las bondades de la regresión cuantílica para lidiar con modelos heterocedásticos, así como caracterizar completamente la distribución condicional de la variable de respuesta, e.g., únicamente conociendo las curvas de predicción con sus respectivos niveles se podría tener una idea de la distribución condicional de la respuesta.

Corrección de la concurvidad. En los modelos de regresión es sabido que la alta relación entre las variables independientes produce la inflación de la varianza, un aumento de los errores tipo II y por consiguiente valores p asociados poco confiables. Esto en la práctica es el aporte redundante de información, lo que en la teoría conduce a conocidos problemas al realizar las operación atómicas necesarias para realizar el ajuste del modelo.

Variable	R.A	Dec.	u_mag	sig_u	g_mag	sig_g	r_mag	sig_r	i_mag	M_i
VIF	1.02	1.03	8.68	3.29	23.10	1.89	38.98	3.71	46.55	1.56

Variable	sig_i	z_mag	sig_z	J_mag	sig_J	H_mag	sig_H	K_mag	sig_K
VIF	4.13	24.23	4.89	3907.61	22.79	4151.02	17.02	2753.76	15.96

Tabla 3.1. Factor de Inflación de la Varianza para las variables independientes.

Es por esto que se calculó el VIF para las variables potenciales para el presente estudio. Las variables redundantes (cuyo VIF$>$10) fueron desconsideradas en el modelo. Los valores numéricos pueden observarse en la tabla 3.1. Al final, las variables que no fueron eliminadas como resultado de este proceso fueron `R.A`, `Dec.`, `u_mag`, `sig_u`, `sig_g`, `sig_r`, `M_i`, `sig_i` y `sig_z`, las que serán consideradas para el modelo final.

Clasificación de variables según su tendencia. Una vez seleccionadas las variables independientes para el estudio, se ajustó un modelo de regresión cuantílica lineal LASSO para el cuantil 0.90 del del desplazamiento al rojo, cuya solución está dada por

$$\hat{\boldsymbol{\beta}}_{0.9} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathbf{10}}}{\arg\min} \sum_{i=1}^{46420} \rho_{0.9}\left[y_i - \beta_0 - \sum_{j=1}^{9}\beta_j x_{ij}\right] + \lambda\sum_{j=1}^{9}|\beta_j|, \tag{3.2}$$

siendo $\hat{\boldsymbol{\beta}}_{0.9}$ el estimador del vector de coeficientes de regresión, y $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i9})$ el vector de características fotométricas para el i-ésimo cuásar observado. La solución arriba fue obtenida usando la función `rq` de la librería **quantreg** del software R. Nótese que a diferencia de la ecuación (1.18), el primer sumando es la función de pérdida cuantílica definida en (1.5).

Así, gracias a la propiedad de selección de variables de la regresión LASSO, las variables que resultaron significativas para modelar linealmente el percentil 90 del desplazamiento al rojo fueron los errores de medición asociados a los brillos de banda `sig_u`, `sig_g`, `sig_r`, `sig_i` y `sig_z`. Por otro lado, las variables que presentaron una relación no lineal o nula fueron por complemento la ascensión de la recta, declinación, brillos de la banda ultravioleta y la luminosidad intrínseca del cuásar (magnitud absoluta

en la banda) R.A, Dec., u_mag y M_i, respectivamente. Estos resultados muestran además la relevancia de incorporar en los estudios astronómicos, estructuras no lineales para las variables físicas. Siguiendo la notación en la sección 1.6, se tienen los dos grupos de variables, lineales y no lineales, definidos por $\boldsymbol{v} = (\texttt{sig\_u}, \texttt{sig\_g}, \texttt{sig\_r}, \texttt{sig\_i}, \texttt{sig\_z})^\top$ y $\boldsymbol{w} = (\texttt{R.A}, \texttt{Dec.}, \texttt{u\_mag}, \texttt{M\_i})^\top$, donde de aquí en adelante el subíndice i será omitido para suavizar la notación.

Estimación y ajuste del modelo PLAQR. Finalmente, una vez realizado el proceso de depuración, selección y clasificación de las variables independientes, se implementó el modelo de regresión cuantílica aditiva parcialmente lineal para los percentiles 50 y 90. Es justamente este último percentil en el cual estamos interesados, permitiéndonos encontrar y cuantificar las relaciones de la evolución cósmica de los agujeros negros masivos más primitivos por medio de las características fotométricas medidas. El ajuste de la mediana fue realizado con fines de comparación.

El modelo matemático puede ser expresado como

$$\begin{aligned} Q_p(\texttt{z}) = \beta_{0p} + \beta_{1p}\texttt{sig\_u} + \beta_{2p}\texttt{sig\_g} + \beta_{3p}\texttt{sig\_r} + \beta_{4p}\texttt{sig\_i} + \beta_{5p}\texttt{sig\_z} \\ + g_{1p}(\texttt{R.A.}) + g_{2p}(\texttt{Dec.}) + g_{3p}(\texttt{u\_mag}) + g_{4p}(\texttt{M\_i}), \end{aligned} \tag{3.3}$$

para $p \in \{0.5, 0.9\}$. Las variables independientes en la primera línea de la ecuación (3.3) son aquellas que se relacionan linealmente con la respuesta, y aquellas cuya relación es no lineal se encuentran en la segunda línea. El modelo fue ajustado usando la función plaqr de la librería con igual nombre del software libre R, cuyas salidas se muestran en los códigos 3.1 y 3.2. En ambos bloques de código se pueden observar la estimación puntual e inferencia para los parámetros del modelo θ_i's. ($\boldsymbol{\beta}$'s y las bases de los g_k's).

En la última columna se tiene los valores p's correspondientes a la prueba de significancia dada por $H_0 : \theta_i = 0$ vs. $H_1 : \theta_i \neq 0$. Para $p = 0.50$ se tiene que las variables sig_r y sig_i ambas asociadas a errores de medición en el brillo de la roja por intensidad resultaron ser no significativas para modelar la evolución cósmica de los cuásares anfitriones de agujeros negros masivos con edades medianas. De igual forma,

el término cúbico usado en la construcción de la función suave $g_2(\texttt{Dec.})$ resultó no significativo, sugiriendo que bastaría usar bases de orden dos para su modelamiento. Por otro lado, para los agujeros negros masivos más primitivos ($p = 0.9$) observamos que las variables que no son significativas son únicamente el `sig_r` para la parte lineal, y los términos cúbicos de las funciones suaves asociadas a las variables `R.A.` y `Dec.` que están asociados a parámetros de ascensión recta y declinación en el cielo nocturno cuyo valor p es muy cercano al limiar del 0.05 y ha sido eliminado con el fin de perseguir el principio de parsimonia. Es importante notar que el error de medición en el brillo de la banda roja `sig_i` no resulta significativo para modelar los agujeros negros masivos de edad mediana pero sí para aquellos más antiguos ubicados en el percentil 90. Por último, cabe mencionar que al igual que en el análisis de coeficientes de regresión de los modelos de regresión polinomial, se deben preservar todos los términos cuyo orden sea menor o igual al término de mayor orden que resulte significativo.

Modelo final. Una vez eliminados los términos no significativos del modelo, se procedió a ajustar el modelo final para modelar de igual forma los percentiles 50 y 90 del desplazamiento al rojo, cuyos resultados se sintetizan en la tabla 3.2 y la figura 3.5. La tabla 3.2 muestra las estimaciones puntuales e intervalos del 95% de confianza para los parámetros de la regresión cuantílica cuya relación es lineal. Los valores p's han sido omitidos debido a que todas las variables son significativas.

Respecto a las relaciones lineales podemos destacar la diferencia en el efecto, en términos de magnitud y dirección, que presentan los coeficientes de regresión para los diferentes cuantiles. Por ejemplo, las variables `sig_g` y `sig_z`, correspondientes a los errores de medición en el brillo de la banda verde y roja, respectivamente, tienen un efecto positivo para la mediana, es decir, a mayor el error de medición en el brillo de la banda verde `sig_g` o error de medición en el brillo de la banda roja `sig_z`, mayor el desplazamiento al rojo mediano y por lo tanto el cuásar con su respectivo agujero negro masivo será más antiguo. Nótese que el efecto aumenta para el percentil 90, lo cual es concordante con lo encontrado anteriormente. Es más, se puede decir que el efecto del error de medición en el brillo de la banda verde `sig_g` sobre el desplazamiento al rojo

```
> summary(modelo_p50)
Call: plaqr(formula = z ~ sig_u + sig_g + sig_r + sig_i + sig_z,
            nonlinVars = ~R.A. + Dec. + u_mag + M_i,
            tau = 0.5,
            data = SDSS)
Coefficients:
                   Value      Std. Error t value    Pr(>|t|)
    (Intercept)    1.80447    0.04071   44.32902    0.00000
    sig_u          0.07825    0.02018    3.87671    0.00011
    sig_g          1.79138    0.15753   11.37151    0.00000
    sig_r         -0.31224    0.38385   -0.81344    0.41597
    sig_i          0.03560    0.22596    0.15756    0.87480
    sig_z          2.88045    0.07207   39.96955    0.00000
    bs(R.A.)1      0.03255    0.01142    2.85146    0.00435
    bs(R.A.)2      0.11638    0.00846   13.75730    0.00000
    bs(R.A.)3     -0.01593    0.00709   -2.24804    0.02458
    bs(Dec.)1     -0.02601    0.01269   -2.04931    0.04044
    bs(Dec.)2      0.11140    0.00665   16.74087    0.00000
    bs(Dec.)3      0.01323    0.00689    1.91892    0.05500
    bs(u_mag)1     1.49392    0.07833   19.07296    0.00000
    bs(u_mag)2     2.65627    0.02887   92.01974    0.00000
    bs(u_mag)3     2.83462    0.03761   75.37423    0.00000
    bs(M_i)1      -0.34441    0.05037   -6.83792    0.00000
    bs(M_i)2      -3.26383    0.02292 -142.38790    0.00000
    bs(M_i)3      -3.08430    0.03377  -91.34258    0.00000
```

Código 3.1. Código en software R para modelar el percentil 50 del desplazamiento al rojo usando un modelo PLAQR.

aumenta significativamente cuando modelamos los agujeros negros masivos antiguos, dado que sus respectivos intervalos de confianza tienen intersección nula. Así, al modelar los agujeros negros masivos antiguos se debe adicionar 0.61 unidades al desplazamiento al rojo (en comparación con la mediana) por cada unidad que aumente

```
> summary(modelo_p90)
Call: plaqr(formula = z ~ sig_u + sig_g + sig_r + sig_i + sig_z,
            nonlinVars = ~R.A. + Dec. + u_mag + M_i,
            tau = 0.9,
            data = SDSS)
Coefficients:
                   Value      Std. Error t value    Pr(>|t|)
    (Intercept)    3.12554    0.04421   70.70215    0.00000
    sig_u         -0.16223    0.01725   -9.40708    0.00000
    sig_g          2.37105    0.16171   14.66255    0.00000
    sig_r         -0.06991    0.44086   -0.15858    0.87400
    sig_i          3.42687    0.78886    4.34409    0.00001
    sig_z          3.10280    0.09489   32.69991    0.00000
    bs(R.A.)1      0.04914    0.01494    3.29002    0.00100
    bs(R.A.)2      0.08892    0.01146    7.75893    0.00000
    bs(R.A.)3      0.00980    0.01003    0.97702    0.32856
    bs(Dec.)1     -0.00860    0.01632   -0.52723    0.59804
    bs(Dec.)2      0.11600    0.00891   13.01742    0.00000
    bs(Dec.)3      0.01896    0.00962    1.97082    0.04875
    bs(u_mag)1    -1.75980    0.07985  -22.03855    0.00000
    bs(u_mag)2     3.01086    0.03724   80.84410    0.00000
    bs(u_mag)3     1.03052    0.04027   25.59278    0.00000
    bs(M_i)1       0.34498    0.05097    6.76815    0.00000
    bs(M_i)2      -3.55168    0.02448 -145.10665    0.00000
    bs(M_i)3      -2.72026    0.03337  -81.51576    0.00000
```

Código 3.2. Código en software R para modelar el percentil 90 del desplazamiento al rojo usando un modelo PLAQR.

el error de medición en el brillo de la banda verde `sig_g`. Diferencias aún más significativas pueden ser observadas para las variables `sig_u` y `sig_i`, correspondientes a los errores de medición en el brillo de la u-banda e i-banda ultravioleta, respectivamente, donde para la primera se tiene un cambio de signo, y para la segunda

	Percentil 50		Percentil 90	
	Estimación	**IC 95%**	**Estimación**	**IC 95%**
sig_u	0.07716	$(0.031\,,\,0.124)$	-0.163	$(-0.199\,,\,-0.125)$
sig_g	1.76440	$(1.502\,,\,2.027)$	2.366	$(2.234\,,\,2.496)$
sig_i	-	-	3.405	$(1.693\,,\,5.117)$
sig_z	2.84376	$(2.687\,,\,3.001)$	3.121	$(2.901\,,\,3.340)$

Tabla 3.2. Estimaciones puntuales y por intervalos (del 95% de confianza) para los coeficientes lineales de regresión cuantílica para los percentiles 50 y 90.

la variable pasa de ser no importante para el modelo, a ser significativa. De la primera se puede decir que, el error de medición en el brillo de la u-banda ultravioleta `sig_u` afecta positivamente al tiempo de vida de los agujeros negros masivos con edades medianas, pero afecta negativamente y en mayor magnitud a aquellos más antiguos. Por otro lado, se tiene para el error de medición en el brillo de la i-banda ultravioleta `sig_i`, que si bien tiene un efecto nulo para el modelo en mediana, este aumenta en 3.405 unidades el desplazamiento al rojo de los agujeros negros masivos más antiguos.

Finalmente, en la figura 3.5 se muestran las curvas correspondientes a los efectos no lineales de las variables independientes como la ascensión recta en coordenadas del cielo celeste `R.A.`, declinación en coordenadas del cielo celeste `Dec.`, brillo en la banda ultravioleta `u_mag`, y magnitud absoluta en la i-banda `M_i`, esto para los percentiles 50 y 90 del desplazamiento al rojo o redshift z. Nótese que dichos efectos son significativos ya que las funciones toman valores diferentes a cero.

En primer lugar podemos observar que existe un similar comportamiento en los efectos de la declinación en coordenadas del cielo celeste `Dec.` y la magnitud absoluta en la i-banda `M_i` para ambos percentiles, i.e., sean agujeros negros masivos de edad mediana o aquellos antiguos. El efecto de la declinación luce cuadrática, cóncava hacia abajo, influyendo positivamente sobre el desplazamiento al rojo para valores mayores a

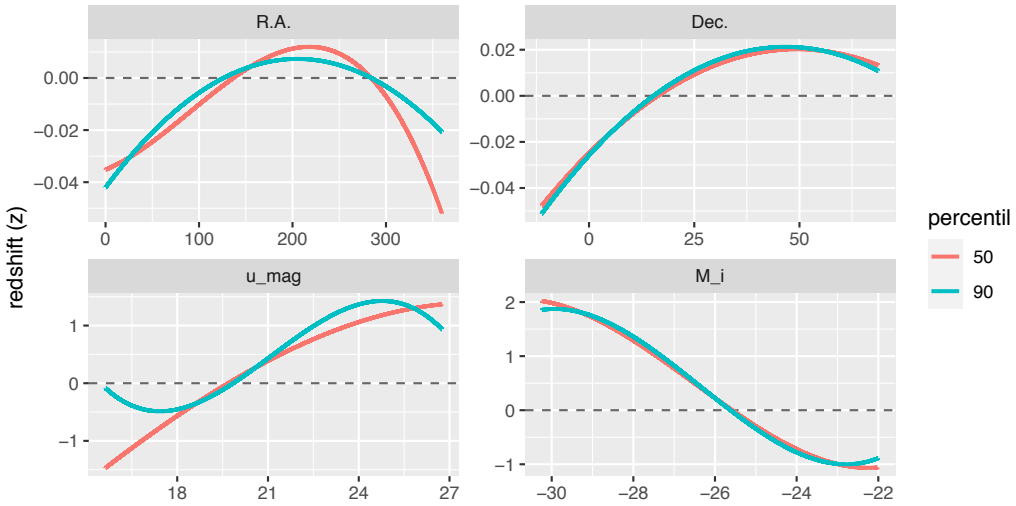

Figura 3.5. Curvas estimadas para los efectos no lineales de las variables independientes `R.A.`, `Dec.`, `u_mag` y `M_i` con respecto al desplazamiento al rojo `z`.

15, pero cuya magnitud comienza a disminuir para declinaciones mayores a 45, aproximadamente. Vale mencionar que estas curvas no son polinomios de segundo orden, sino una construcción de estos últimos. Con respecto a la magnitud absoluta en la i-banda, ambas curvas de predicción para los percentiles 50 y 90 presentan un cambio de concavidad en un entorno suficientemente cercano a -26, donde además el efecto sobre el desplazamiento al rojo pasa de ser positivo a negativo. En general, se puede decir que a mayor la magnitud absoluta en la i-banda `M_i`, más joven será el agujero negro supermasivo.

Por otro lado, los efectos para los percentiles 50 y 90 de la ascensión recta en coordenadas del cielo celeste, presentan ambas una concavidad hacia abajo siendo el valor máximo de la curva mayor para el percentil 50 en comparación al del percentil 90. El efecto sobre el desplazamiento al rojo es principalmente negativo, a excepción de valores entre 140 y 280 donde el efecto se torna positivo. Se observa además que para valores altos de ascensión recta, existe una diferencia en el efecto a ambos percentiles, afectando en mayor magnitud al desplazamiento al rojo de agujeros negros masivos de edades medianas.

Para terminar el presente análisis, observamos los efectos del brillo en la banda ultravioleta sobre los percentiles en estudio. Se puede ver a simple vista una diferencia en el comportamiento de las funciones para los percentiles. Ambas curvas influencian de forma negativa para valores de brillos menores a 20, y de forma positiva una vez superado este limiar. Sin embargo, el efecto del brillo luce monótonamente creciente y cóncavo hacia abajo para la mediana, mientras que el efecto para el percentil 90 presenta dos puntos de inflexión que determinan sus cambios de monotonía, aportando de forma máxima, negativa y positiva, para brillos de 18 y 25, respectivamente.

Por último, el modelo que hemos llamado final puede ser utilizado para fines predictivos, i.e., para estimar el percentil 90 del desplazamiento al rojo basado en las variables fotométricas que resultaron significativas. El valor estimado del percentil 90 $\widehat{Q_{0.9}}(\mathtt{z})$ dado una conjunto de características fotométricas observadas, puede ser calculado usando la expresión matemática

$$\begin{aligned}\widehat{Q_{0.9}}(\mathtt{z}) = 3.126 - 0.163\,\mathtt{sig\_u} + 2.366\,\mathtt{sig\_g} + 3.405\,\mathtt{sig\_i} + 3.121\,\mathtt{sig\_z} \\ + \widehat{g_{1,0.9}}(\mathtt{R.A.}) + \widehat{g_{2,0.9}}(\mathtt{Dec.}) + \widehat{g_{3,0.9}}(\mathtt{u\_mag}) + \widehat{g_{4,0.9}}(\mathtt{M\_i}),\end{aligned}$$

donde las funciones $\widehat{g_{k,0.9}}$, para $k = 1, \ldots, 4$, representan las curvas de efecto estimadas para el percentil 90, las cuales se muestran en la figura 3.5 en color azul. Otros percentiles de interés puedes ser ajustados de igual forma, obteniéndose intervalos de predicción para el desplazamiento al rojo análogos a los mostrados en la figura 3.4, pero que no pueden ser graficados debido a la limitación natural para representar superficies en altas dimensiones.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En el presente trabajo se realizó un análisis estadístico de la edad cósmica del universo a través del modelamiento del desplazamiento al rojo de cuásares como función de variables fotométricas medidas. Se estudiaron cómo algunas de las variables fotométricas afectan su desplazamiento al rojo z, y por correspondencia directa, a su tiempo histórico en el universo físico. Los códigos necesarios para la implementación del presente estudio se encuentran disponibles a los usuarios a través del repositorio de código abierto en Bitbucket (https://bitbucket.org/afpalma/plaqr/src/master/).

Cabe destacar que trabajos como Meshcheryakov et al. (2018) que usan herramientas de Machine Learning, tienen un enfoque netamente predictivo, a diferencia de la propuesta de este trabajo la que ofrece además inferencia, siendo capaz de determinar bajo un grado de incerteza o nivel de confianza, las variables significativas, junto a sus estimaciones puntuales e intervalos de confianza.

Futuros trabajos pueden ser plausibles al considerar en el modelo los efectos de absorción del bosque Lyman-alfa de acuerdo a Rauch (1998), que aparece como una disminución en los brillos de las bandas más azules que presentan un desplazamiento al rojo significativo, pudiendo conducir a hallazgos que podrían ser de interés. A continuación, se listan las conclusiones y recomendaciones más relevantes del trabajo.

4.1 Conclusiones

- El modelo de regresión cuantílica aditiva parcialmente lineal permitió estudiar la evolución cósmica de los cuásares anfitriones de un agujero negro supermasivo en su centro por medio de características fotométricas como brillos en las bandas espectrales y luminosidad, ofreciendo además un modelo predictivo para poder

describir de la cronología de estos objetos base del universo.

- Acorde al modelo implementado, se determinó que las variables fotométricas con mayor significancia para el presente estudio fueron aquellas que estaban relacionadas con los errores de medición asociados a los brillos en las bandas espectrales cuyos efectos son lineales sobre el desplazamiento al rojo, mientras que variables como ascensión recta, declinación en coordenadas del cielo celeste presentaron un efecto no lineal significativo sobre la variable de interés.

- Los resultados obtenidos sugieren que a medida que se incrementa la magnitud absoluta en la i-banda que almacena información lumínica de un cuásar, entonces más cerca del origen del universo físico estuvo el cuásar. Más aún, el modelo final permite observar que el error de medición asociado al brillo en la i-banda ultravioleta, aumenta en 3.4 unidades el desplazamiento al rojo de los cuásares que se encuentran en el percentil 90.

4.2 Recomendaciones

- Determinar familias o grupos de cuásares en función de la distribución de la función de energía lumínica medida por las magnitudes absolutas en las i-bandas usando métodos de clusterización pudiendo ser utilizadas como variables predictoras para futuros modelos.

- Realizar un análisis estadístico usando modelos de regresión complejos que permitan la inclusión de errores de medida en las variables independientes, cuando estos modelos se encuentren disponibles en la literatura.

- Medir el costo computacional del modelo presentado con el objetivo de conocer si es computacionalmente escalable a conjuntos de datos de grandes dimensiones, y su implementación en computación en alto rendimiento (o HPC, por sus siglas en inglés).

BIBLIOGRAFÍA

Angelis, D. D., Hall, P., and Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in l 1 regression. *Journal of the American Statistical Association*, 88(424):1310–1316.

Barrau, A. (2022). The holographic space-time and black hole remnants as dark matter. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*, 829.

Biswas, T., Koshelev, A. S., Mazumdar, A., and Vernov, S. Y. (2012). Stable bounce and inflation in non-local higher derivative cosmology. *Journal of Cosmology and Astroparticle Physics*, 2012(8). Cited By :166.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pages 88–126.

Buen-Abad, M. A., Essig, R., McKeen, D., and Zhong, Y. . (2022). Cosmological constraints on dark matter interactions with ordinary matter. *Physics Reports*, 961:1–35. Cited By :2.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Galarza, C., Castro, L. M., Louzada, F., and Lachos, V. H. (2020). Quantile regression for nonlinear mixed effects models: a likelihood based perspective. *Statistical Papers*, 61:1281–1307.

Galarza, C., Lachos, V., Barbosa Cabral, C., and Castro Cepero, L. (2017). Robust quantile regression using a generalized class of skewed distributions. *Stat*, 6(1):113–130.

Gaztanaga, E. (2022). How the big bang ends up inside a black hole. *Universe*, 8(5).

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, 11(1):105–121.

He, S., Sun, Y., Zhao, L., and Zhang, Y. . (2022). The universality of islands outside the horizon. *Journal of High Energy Physics*, 2022(5).

He, X. and Shi, P. (1996). Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58(2):162–181.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590.

Huber, P. (1981). Robust statistics. new york: John wiley and sons. *HuberRobust statistics1981*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2019). *An introduction to statistical learning*. Springer.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. W. and d'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Applied statistics*, pages 383–393.

Lachos, V. H., Chen, M.-H., Abanto-Valle, C. A., and Azevedo, C. L. (2015). Quantile regression for censored mixed-effects models with applications to hiv studies. *Statistics and its Interface*, 8(2):203.

Léna, P., Lebrun, F., and Mignard, F. (2013). *Observational astrophysics*. Springer Science & Business Media.

Marshall, M. (2020). Simulated infrared images from webb and hubble. https://webbtelescope.org/contents/media/images/2020/51/4754-Image?keyword=redshift.

Meshcheryakov, A. V., Glazkova, V. V., Gerasimov, S. V., and Mashechkin, I. V. (2018). Measuring the probabilistic photometric redshifts of x-ray quasars based on the quantile

regression of ensembles of decision trees. *Astronomy Letters*, 44(12):735–753. Cited By :11.

Morlini, I. (2006). On multicollinearity and concurvity in some nonlinear multivariate models. *Statistical Methods and Applications*, 15(1):3–26.

NASA and Feild, A. (2018). Redshifted light from distant galaxies. https://webbtelescope.org/contents/media/images/4195-Image?keyword=redshift.

Penrose, R. (2005). *The road to reality: A complete guide to the laws of the universe*. Random house.

Penzias, A. A. and Wilson, R. W. (1965). A measurement of excess antenna temperature at 4080 mc/s. *The Astrophysical Journal*, 142:419–421.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Rauch, M. (1998). The lyman alpha forest in the spectra of quasistellar objects. *Annual Review of Astronomy and Astrophysics*, 36(1):267–316.

Schneider, D. P., Hall, P. B., Richards, G. T., Berk, D. E. V., Anderson, S. F., Fan, X., Jester, S., Stoughton, C., Strauss, M. A., SubbaRao, M., et al. (2005). The sloan digital sky survey quasar catalog. iii. third data release. *The Astronomical Journal*, 130(2):367.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Sou, H., Wang, J. ., Xie, Z. ., Kang, W. ., and Cai, Z. . (2022). The relation between x-ray and ultraviolet variability of quasars. *Monthly Notices of the Royal Astronomical Society*, 512(4):5511–5519.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wang, H. J., Zhu, Z., and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B):3841–3866.

Wichitaksorn, N., Choy, S., and Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, 42(4):579–596.

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579.