ANÁLISIS DEL PESO DE LOS RECIÉN NACIDOS EN PUERTO RICO USANDO REGRESIÓN A CUANTILES

Por

ANA CRISTINA MORENO CASTILLA

Proyecto sometido en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS ESTADÍSTICA

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

2018

Aprobada por:

Pedro A. Torres Saavedra, Ph.D. Presidente, Comite Graduado	Fecha
Dámaris Santana Morant, Ph.D. Co-Presidente, Comite Graduado	Fecha
Edgardo Lorenzo González, Ph.D. Miembro, Comite Graduado	Fecha
Carlos Hernández Hernández, Ph.D. Representante de Estudios Graduados	Fecha
Olgamary Rivera Marrero, Ph.D. Directora del Departamento	Fecha

Abstract

This project illustrates the use of the grouped *Smoothly Clipped Absolute Deviation* (SCAD), a variable selection method based on regularization, in quantile regression to model the weight of newborns in Puerto Rico in 2009-2011. Quantile regression models allows us to study the effect of the independent variables on the different quantiles of the dependent variable and thus to have a complete idea of the relationship between these variables and the distribution of the response. Therefore, one can conclude which, prenatal care variables, sociodemographic factors, and health conditions are associated, not only to the low weight and overweight newborns, but also to other parts of the distribution of the newborn weights such as the median.

A quantile regression model for newborn weight with several covariates was adjusted for different quanitles of interest. Some of the variables included in the model were age of the mother, weeks of gestation and sex of the newborn. However, due to the large number of possible explanatory variables associated to a particular quantile of the newborn weights, a variable selection method based on regularization, namely SCAD, was implemented to come up with a subset of covariates significantly associated with the quantile of the newborn weights.

Our results suggest that the factors significantly associated to the newborn weights depend on the quanitle being modeled, although there are some explanatory variables that are consistently selected regardless the quantile. For instance, the weight gain of the mother was maintained in all cases.

Resumen

Este proyecto muestra la aplicación del método de selección de variables *Smoothly Clipped Absolute Deviation* (SCAD), un método de selección de variables basado en regularización, en regresión a cuantiles para modelar el peso de los recién nacidos en Puerto Rico en los años 2009-2011. Los modelos de regresión cuantil nos permiten estudiar el efecto de las variables independientes sobre los diferentes cuantiles de la variable dependiente y así tener una idea completa de la relación entre estas variables y la variable dependiente. Esto permite identificar cuáles variables de cuidado prenatal, factores sociodemográficos, y condiciones de salud están asociadas, no sólo con el bajo peso y el sobrepeso de recién nacidos, sino que también con otras parte de la distribución de los pesos de los recién nacidos como por ejemplo, la mediana.

Un modelo de regresión a cuantiles para el peso de los recién nacidos con varias covariables fue ajustado para diferentes cuantiles de interés. Algunas variables incluidas en el modelo fueron edad de la madre, las semanas de gestación y el sexo del recién nacido. Sin embargo, debido al gran número de variables explicativas posibles asociadas a un cuantil particular del peso de los recién nacidos, se implementó un método de selección de variables basado en la regularización, llamado SCAD, para llegar a un subconjunto de covariables significativamente asociadas con el cuantil de los pesos de los recién nacidos.

Nuestros resultados sugieren que los factores significativamente asociados al peso de los recién nacidos dependen del cuantil que se modele, aunque hay algunas variables explicativas que se seleccionan consistentemente independientemente del cuantil. Por ejemplo, el aumento de peso de la madre se mantuvo en todos los casos.

A Dios por su misericordia...

A mi madre Myriam por su gran amor y comprensión...

A mi esposo Cristian, por su apoyo incondicional...

AGRADECIMIENTOS

A Dios por todo lo que ha hecho en mi vida.

Al Dr. Pedro Torres, por su apoyo, orientación, paciencia y disposición en la realización de este trabajo.

A la Dra. Dámaris Santana por su colaboración y al Dr. Edgardo Lorenzo por su apoyo.

A mi padre Raúl por su apoyo, a todos mis demás familiares, en especial a mi abuela.

A ti Cristian, por tu motivación en todo momento. Mil gracias!

A la Universidad de Puerto Rico por darme la oportunidad de realizar mis estudios.

A todos mis amigos y compañeros de estudio. Gracias por su apoyo.

Índice general

Índice de tablas

Índice de figuras

LISTA DE ABREVIATURAS

RCIU	Retardo del Crecimiento Intrauterino.
APGAR	Apariencia, pulso, gesticulación, actividad y respiración.
NCHS	Centro Nacional de Estadísticas de la Salud.
i.i.d.	Independiente e idénticamente distribuidas.
indep	Independiente.
SCE	Suma de cuadrados del error.
LS	(Least Squares) (Mínimos Cuadrados).
ML	(Maximun Likelihood) (Máxima Verosimilitud).
LASSO	(Least Absolute Shrinkage and Selection Operator).
SCAD	(Smoothly Clipped Absolute Deviation).
BIC	(Criterio de información Bayesiana).
GCV	Validación cruzada generalizada.
ANOVA	Análisis multifactor de varianza.
APGAR5	Apariencia, pulso, gesticulación, actividad y respiración a los 5 minutos.
NA	Datos faltantes.

LISTA DE SÍMBOLOS

inf	Ínfimo.
min	Mínimo.
$\prod$	Productoria.
σ^2	Varianza.
α	Nivel de significancia.
$z_{\alpha/2}$	Puntuación z tal que el área bajo la curva normal estándar a la derecha de $z_{\alpha/2}$ es $\alpha/2$.
t	Prueba t de Student.
α	Nivel de significancia.
F	Función de densidad acumulada.
p	Valor p.
$N(0, \sigma^2)$	Distribución Normal estándar.
$E(x)$	Promedio esperado de x.
τ	Cuantil.
F^{-1}	Función inversa de F.
$Var(x)$	Varianza de x
$\mathbf{L}(\theta, x)$	Función de máxima verosimiltud.
$F_n(y)$	Función de distribución empírica de y.
ρ	Función de verificación.
$Q(\tau)$	τ- ésimo Cuantil.
$\hat{\beta}_{i,\tau}$	Coeficiente estimado de regresión a Cuantil.

Introducción

El nacer con bajo peso está asociado con serios problemas de salud tales como discapacidades por el resto de su vida o morir antes de cumplir su primer año de vida. Un recién nacido se clasifica con bajo peso si pesa menos de 5 libras y 8 onzas (2,500 g) y como muy bajo peso si pesa menos de 3 libras y 4 onzas (1,500 g). Esta categoría de bajo peso al nacer incluye a los bebés nacidos antes de término ($<$ 37 semanas de gestación) y a los bebés pequeños para su edad gestacional (Cruz et al., 2009).

Para los años 2009-2011, Puerto Rico enfrentó una alta tasa de nacimientos de bebés prematuros. La tasa de bebés prematuros para el 2009 fue alrededor de 17.6 %, para el 2010 de 16.5 % (Cruz et al., 2010) y para el 2011 de 17.7 % (ENDI, 2011). Además del nacimiento prematuro, otra razón para nacimientos con bajo peso es el retardo del crecimento intrauterino (RCIU) definido como crecimiento fetal menor al potencial debido a factores genéticos o ambientales. La definición de RCIU se basa en la disminución de la velocidad de incremento ponderal que se manifiesta en peso bajo el percentil 10 para la edad gestacional (Rybertt et al., 2016).

Dentro de los factores de riesgo más comunes del bajo peso al nacer se han encontrado el embarazo en la adolescencia, la desnutrición en la madre, el hábito de fumar, la hipertensión arterial durante el embarazo, la sepsis cervicovaginal, anemia y los embarazos gemelares, entre otros (Peraza et al., 2001).

Debido a que en los años 2009-2011 Puerto Rico presentó una alta de bebés prematuros, este trabajo busca determinar qué factores están asociados con el bajo peso al nacer empleando regresión a cuantiles. De la misma manera, este proyecto pretende examinar, los factores asociados al peso de los recién nacidos en PR, con

especial énfasis a diferentes cuantiles a través del soporte de la distribución (por ejemplo, recién nacidos con sobrepeso).

En Puerto Rico se han hecho estudios acerca del bajo peso al nacer (Becerra et al., 1993). Estos autores cuantificaron las contribuciones relativas de la edad materna, la educación, el estado civil, el hospital de nacimiento y el uso de la atención prenatal a la alta incidencia de bajo peso al nacer y mortalidad infantil en Puerto Rico. Estos autores realizaron un análisis de 257,537 nacidos vivos que ocurrieron entre 1986 y 1989 entre residentes de Puerto Rico y 3373 muertes infantiles correspondientes. Usaron modelos de regresión múltiple con respuesta binomial para calcular los riesgos poblacionales atribuibles ajustados para cada variable. Las estimaciones de estos autores indican que aproximadamente 6 de cada 10 muertes infantiles en la isla son potencialmente evitables si se erradica el bajo peso al nacer, independientemente de otros factores asociados. La eliminación de los riesgos asociados con los factores sociodemográficos y socioeconómicos (incluído el hospital de nacimiento) podría reducir la incidencia del bajo peso al nacer en Puerto Rico por un tercio. Específicamente, la eliminación de los riesgos asociados con la desventaja socioeconómica de las mujeres que dan a luz solo en hospitales públicos podría disminuir la incidencia de bajo peso al nacer de Puerto Rico en un 28 %, independientemente de otros factores considerados en este estudio.

Otro estudio sobre el bajo peso de recién nacidos en Puerto Rico fue llevado a cabo por Campos et al. (2008). Estos autores incluyeron 216 neonatos nacidos entre 1999 y 2003. Se parearon en dos grupos: adecuados para la edad gestacional y pequeños para la edad gestacional basado en el género, el año de nacimiento y el peso al nacer. El período de observación fue desde el nacimiento hasta la fecha de alta. Se usaron dos pruebas t para determinar la diferencia en la tasa de crecimiento entre los grupos. La regresión simple se utilizó para establecer el efecto de las morbilidades

en la tasa de ganancia de peso. Cuando todas las variables fueron analizadas usando el modelo de regresión lineal, solo teniendo un bajo puntaje de apariencia, pulso, gesticulación, actividad y respiración (APGAR)($p = 0.02$) y ser pequeño para la edad gestacional ($p = 0.0004$) fueron significativos. Campos et al. (2008) llegaron a la conclusión que los patrones de crecimiento de neonatos de muy bajo peso al nacer son diferentes en función de la adecuación de su peso al nacer, y que la disparidad en la tasa de crecimiento no se explica por las diferencias en la incidencia de morbilidades que afectan el crecimiento. También concluyeron que la ganancia de peso promedio de los neonatos adecuados para la edad gestacional fue menor que la de los pequeños para la edad gestacional.

El análisis del bajo peso al nacer o sobrepeso empleando regresión a cuantiles no se ha llevado a cabo con datos de Puerto Rico. El uso de regresión a cuantiles sí se ha llevado a cabo en Estados Unidos (Abrevaya, 2001). La investigación se realizó con datos del Centro Nacional de Estadísticas de la Salud (NCHS en inglés) en los años 1992 y 1996. Para cada nacimiento, hay datos proporcionados sobre la madre y el resultado del nacimiento. Los datos de la madre consisten en información sobre las características personales (por ejemplo, edad, raza, estado civil, educación, lugar de residencia) y el comportamiento durante el embarazo (por ejemplo, medidas de atención prenatal, tabaquismo durante el embarazo). Los datos de los resultados del nacimiento consisten en información sobre las características del niño (peso, sexo, posibles trastornos, entre otros). El conjunto de datos completo para un año dado constaba de aproximadamente 4 millones de observaciones, lo que hizo que el análisis fuera difícil de llevarse a cabo.

Para reducir el tamaño de la muestra, Abrevaya (2001) usó solo los nacimientos que ocurrieron en el mes de junio (es decir, Junio de 1992 y junio de 1996). La muestra se limitó a nacimientos únicos y de madres blancas o negras, entre 18

y 45 años, y residentes de los Estados Unidos. Las observaciones para las cuales había información faltante sobre cualquier variable relevante fueron descartadas. Desafortunadamente, todos los nacimientos que ocurrieron en California, Indiana, Nueva York y Dakota del Sur tuvieron que ser retirados de la muestra debido a que estos estados no se hizo una pregunta sobre el fumar durante el embarazo o no se solicitó la información en una forma compatible con los estándares del NCHS (*National Center for Health Statistics*). Los tamaños de muestra resultantes fueron 199,108 observaciones para 1992 y 191,748 observaciones para 1996. La conclusión fue que varios factores, entre ellos (la raza, la educación y la atención prenatal), tienen un impacto significativamente mayor en los cuantiles inferiores y un menor impacto en los cuantiles superiores. El factor peso ganado de la madre tiene un efecto mayor en la distribución del peso al nacer en los cuantiles inferiores. El trabajo realizado por Abrevaya (2001) de cierta manera está relacionado con la investigación en curso.

Otro trabajo relacionado con la investigación actual se llevó a cabo por Fallah et al. (2015). En este trabajo el objetivo fue evaluar el peso al nacer de los neonatos utilizando la regresión a cuantil en la provincia de Zanjan en Irán. Este estudio descriptivo se realizó utilizando datos pre-registrados (marzo 2010 - marzo 2012) de neonatos en centros de salud urbanos y rurales de la provincia de Zanyán utilizando muestreo de conglomerados multietápico. Los datos se analizaron mediante regresión lineal múltiple y regresión a cuantiles usando el programa estadístico SAS 9.2. Fallah et al. (2015) obtuvieron una asociación significativa entre el incremento de la edad materna, la edad gestacional, el nivel de educación materna con el aumento de peso de los infantes. Concluyeron, además, que los resultados de la regresión lineal múltiple y la regresión a cuantil no fueron idénticos. Ellos recomiendan el uso de

la regresión a cuantil cuando se dispone de una variable respuesta asimétrica o de datos con valores atípicos.

Este trabajo es relevante dado que el bajo peso al nacer está asociado con varios problemas de salud, por lo tanto, nos interesa conocer los factores relacionados con el bajo peso al nacer. Utilizamos regresión a cuantiles ya que este método es más abarcador que la regresión a la media debido a que nos permite estudiar el efecto de las covariables en diferentes cuantiles de la distribución de la variable respuesta (no se enfatiza en la media solamente), y por lo tanto, proporciona una idea más completa de la relación entre las covariables y la distribución de la variable respuesta. La influencia de los efectos de algunas variables en diferentes cuantiles condicionales de la variable respuesta no se puede examinar con un modelo de media condicional. Otra razón para usar regresión a cuantiles es que en el método no se hacen suposiciones de distribución. Además, las estimaciones de regresión a cuantil son más robustas frente a los valores atípicos en las mediciones de la variable respuesta. De otro lado, en este trabajo se aplica un método reciente de selección de variables en regresión a cuantiles. La propuesta de usar regresión a cuantiles conjuntamente con SCAD agrupada es muy atractiva para llegar a conclusiones importantes en términos de los factores asociados a diferentes niveles del peso de recién nacidos en Puerto Rico.

OBJETIVOS

Ojetivo general

Construir un modelo estadístico que relacione el peso de los recién nacidos en PR con algunas variables explicativas, con particular énfasis en el bajo peso y sobrepeso, usando regresión a cuantiles y el método de selección de variables SCAD agrupada.

Objetivos específicos

1 Determinar qué factores sociodemográficos cuidados prenatales y condiciones de salud influyen en el peso de los recién nacidos, en particular el bajo peso y sobrepeso.

2 Utilizar el método de selección de variables aplicado a regresión a cuantiles para seleccionar modelos razonables que sirvan para explicar el peso en diferentes cuantiles de la variable respuesta.

3 Interpretar los resultados de los modelos seleccionados.

Capítulo 1
MARCO TEÓRICO

1.1. Regresión lineal

En un problema de regresión estamos interesados principalmente en estudiar la relación entre la media de una variable aleatoria Y y un conjunto de variables explicativas. La variable aleatoria Y se denomina variable dependiente o respuesta. Las variables que influencian a Y se denominan variables independientes, explicativas o variables predictoras (Milton, 2007).

1.1.1. Modelo de regresión lineal simple clásico

Considere datos que consisten de pares $\{(x_i, Y_i) \ : \ i = 1, 2, ..., n\}$ basados en una muestra aleatoria. Para cada observación i se tiene una respuesta Y_i y una sola covariable x_i. En el modelo lineal simple clásico se asume que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n. \tag{1.1}$$

donde ε_i satisface, $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, (i.i.d. significa independiente e idénticamente distribuidos) (Hao and Naiman, 2007).

Los supuestos del Modelo 1.1 implican que:

1. $E(Y_i \mid \boldsymbol{x}_i) = \beta_0 + \beta_1 x_i$.

2. $var(Y_i \mid \boldsymbol{x}_i) = \sigma^2$.

3. $Y_i \mid \boldsymbol{x}_i \overset{indep.}{\sim} N(\beta_0 + \beta_1 x_i,\, \sigma^2)$,

donde indep. significa que las $Y_i's$ son independientes.

Para la estimación de parámetros del Modelo 1.1, denotadas por $\hat{\beta}_0$ y $\hat{\beta}_1$, una opción es minimizar la suma de los cuadrados de los residuales. Esta suma con frecuencia se denomina suma de los cuadrados del error respecto de la recta de regresión y se denota como SCE. Este procedimiento de minimización para estimar los parámetros se denomina ***Mínimos Cuadrados*** (*Least Squares*) (Walpole et al., 2012). El problema que tenemos planteado es, hallar los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que la recta que pasa por los puntos $(x_i, \hat{Y}_i)$ se ajuste lo mejor posible a los puntos (x_i, Y_i). Al emplear la suma de cuadrados del error se tiene

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Al derivar SCE con respecto β_0 y β_1 e igualar a cero se obtiene

$$\hat{\beta}_{0_{(LS)}} = \overline{Y} - \hat{\beta}_1 \overline{x},$$

y

$$\hat{\beta}_{1_{(LS)}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Estas estimaciones de los parámetros dan la ecuación estimada de regresión

$$\hat{y}_i = \hat{\beta}_{0(LS)} + \hat{\beta}_{1(LS)} x_i.$$

Un estimador insesgado de la (varianza residual) $\boldsymbol{\sigma}^2$ en el Modelo 1.1 esta dado por:

$$\hat{\sigma}^2_{LS} = \frac{SCE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}.$$

De la media $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$ podemos ver que β_0 representa la estimación del promedio de la variable respuesta cuando $x_i = 0$. La interpretación para β_1 es el cambio esperado en Y_i dado un aumento de una unidad en x_i. Es muy importante saber en qué unidades se miden las covariables $x_{i's}$, de modo que el significado de un aumento de una unidad se pueda expresar claramente (Seltman, 2012).

Los parámetros β_0 , β_1 y σ^2 en el Modelo 1.1 también se pueden estimar usando otos métodos tales como máxima verosimilitud (Andrzej Galecki, 2013). Más detalles de este método se discuten en la Sección 1.2.

1.2. Modelo de regresión lineal múltiple

Suponga que tiene n observaciones independientes. Para cada observación i se observa una variable respuesta Y_i y d covariables en un vector columna $\boldsymbol{x}_i$, es decir, $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})^t$, de dimensión $d < n$. En el modelo lineal clásico se asume que (Andrzej Galecki, 2013)

$$Y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, .., n, \tag{1.2}$$

donde ε_i satisface que, $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ y $\boldsymbol{\beta}$ es un vector columna con d parámetros, $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^t$.

Los supuestos del modelo anterior implican que :

1. $E(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}$.

2. $var(Y_i \mid \boldsymbol{x}_i) = \sigma^2$.

3. $Y_i \mid \boldsymbol{x}_i \overset{indep.}{\sim} N(\boldsymbol{x}_i^t \boldsymbol{\beta}, \sigma^2)$.

El modelo anterior puede ser expresado por medio de matrices. Defina la siguiente matriz del modelo de dimensión $n \times d$

$$\boldsymbol{X}_{n\times d} = \begin{pmatrix} x_{1,1} & x_{1,2} & . & . & . & x_{1,d} \\ x_{2,1} & x_{2,2} & . & . & . & x_{2,d} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_{n,1} & x_{n,2} & . & . & . & x_{n,d} \end{pmatrix}.$$

Cada fila en $\boldsymbol{X}$ representa el vector de covariables de cada uno de los n individuos. Es decir,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^t \\ \boldsymbol{x}_2^t \\ . \\ . \\ . \\ \boldsymbol{x}_n^t \end{pmatrix}.$$

Por lo tanto, el modelo se puede escribir como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.3}$$

donde $\boldsymbol{Y} = (Y_1, ..., Y_n)^t$ y $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^t$ es un vector columna con los n errores del modelo. También se asume que el error, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, $\boldsymbol{I}_n$ es la matriz identidad de orden n. Los supuestos para el modelo anterior implican que:

1. $E(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$.

2. $Var(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$.

3. $\boldsymbol{Y} \mid \boldsymbol{X} = N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

Como en regresión simple, un método de estimación de $\boldsymbol{\beta}$ para el Modelo 1.3 es el de ***Mínimos Cuadrados***. El método consiste en minimizar la suma de cuadrados de errores dada por

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Al derivar la expresión anterior se obtienen las ecuaciones normales

$$(\boldsymbol{X}^t \boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^t \boldsymbol{Y}.$$

La matriz $\boldsymbol{X}$ del Modelo 1.3 se asume de rango completo. Por lo tanto, $(\boldsymbol{X}^t \boldsymbol{X})$ es no singular y la solución mínimos cuadrados para los coeficientes del modelo esta dada por

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y}.$$

Un estimador insesgado para $\boldsymbol{\sigma}^2$ en el Modelo 1.3 está dado por

$$\hat{\boldsymbol{\sigma}}^2_{LS} = \frac{1}{n-d} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}_{LS})^2 = \frac{1}{n-d} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})^t (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}).$$

El modelo lineal simple clásico definido en la Ecuación 1.2 implica que las observaciones Y_i, son independientes y normalmente distribuidas. En consecuencia, la función de verosimilitud para este modelo se define de la siguiente manera:

$$\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \boldsymbol{Y}) = (2\pi\boldsymbol{\sigma}^2)^{-n/2} \boldsymbol{\Pi}_{i=1}^{n} e^{\left[-\frac{(Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2}{2\boldsymbol{\sigma}^2}\right]}.$$

Los estimadores de máxima verosimilitud de los parámetros del modelo están dados por:

$$\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y},$$

y

$$\hat{\boldsymbol{\sigma}}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{ML})^2.$$

Note que el estimador ML de $\boldsymbol{\beta}$ es igual que el estimador LS pero el estimador ML de $\boldsymbol{\sigma}^2$ tiene un sesgo negativo (anteriormente vimos que $\hat{\boldsymbol{\sigma}}^2_{LS}$ es un estimador insesgado para $\boldsymbol{\sigma}^2$). Este sesgo se debe a que la estimación de $\boldsymbol{\beta}$ no se tiene en cuenta al calcular $\hat{\boldsymbol{\sigma}}^2_{ML}$ (no ajusta el denominador o los grados de libertad).

1.3. Regresión a cuantiles

La regresión a cuantiles complementa el enfoque de la regresión de mínimos cuadrados ordinarios de la media condicional. La regresión a cuantiles ofrece una estrategia sistemática para examinar cómo las covariables influyen en la ubicación, la escala y la forma de toda la distribución para distintos cuantiles de la variable dependiente (Koenker, 2005).

Definición 1.3.1. ***Cuantil***

Dado un $\tau \in (0,1)$ y una variable aleatoria Y (continua o discreta), el τ-ésimo cuantil es definido como: $Q(\tau) = F^{-1}(\tau) = inf\{y \in Y : F(y) \geq \tau\}$, que es la función inversa de F, donde F es la función de distribución acumulada de Y y inf denota el ínfimo.

En la anterior definición usamos el ínfimo para definir la inversa de una función de distribución acumulada arbitraria.

En la Figura 1–1 la función de distribución acumulada $F(y)$ es continua y estrictamente creciente. Para $\tau \in (0,1)$, $F^{-1}(\tau)$ se define como en la figura, es decir $F^{-1}(\tau)$ es el número único ξ tal que $F(\xi) = \tau$. Por un resultado de análisis, $F^{-1}(\tau)$ es continua y estrictamente creciente en τ.

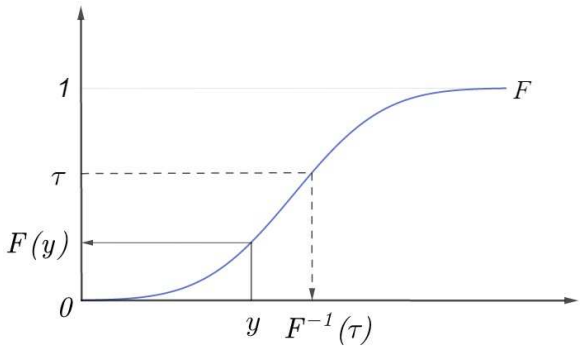


Figura 1–1: Función de densidad acumulada y su inversa.

La anterior definición sirve para funciones continuas y estrictamente crecientes. El caso general donde F es una función de distribución de probabilidad arbitraria se puede observar en la Figura 1–2, que F puede tener saltos y puntos planos. Es decir, no hay una verdadera inversa a F en el sentido habitual. Se puede definir un tipo de inversa F^{-1} a F, de tal manera que $F^{-1}(\tau)$ sea el valor y más pequeño tal que $F(y) \geq \tau$. Esta definición funciona para $\tau = \tau_2$, τ_3 y τ_1.

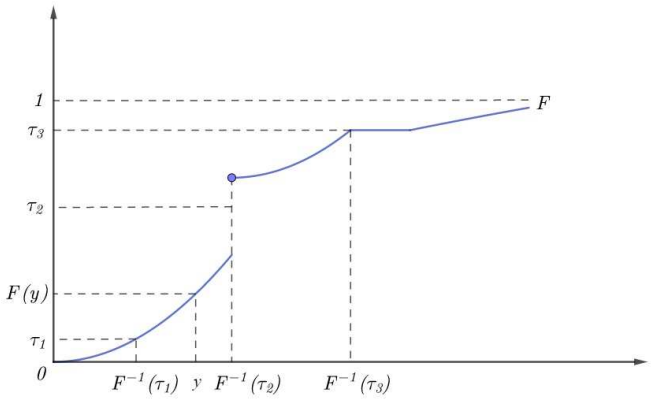


Figura 1–2: Función de distribución acumulada F y función cuantil.

Por otro lado, sea $S_n = \{Y_1, ..., Y_n\}$ una muestra de n observaciones i.i.d., entonces $F(y)$ se puede estimar usando (Thas, 2010)

$$\hat{F}_n(y) = \frac{1}{n}\#\{Y_i \in S_n : Y_i \leq y\} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y). \tag{1.4}$$

Es decir, $\hat{F}_n(y)$ es igual al número de observaciones de muestra no mayores que y, dividido por el tamaño de muestra n. Este estimador de la función de distribución de y se conoce como la función distribución empírica.

Análogamente, es posible definir una estimación de los cuantiles por medio de la distribución empírica en 1.4, de la siguiente manera (López and Mora, 2007):

$$\hat{Q}(\tau) = inf\{y : \hat{F}_n(y) \geq \tau\} \tag{1.5}$$

El problema en 1.5 es equivalente a solucionar el siguiente problema de optimización:

$$\hat{Q}(\tau) = \underset{\varepsilon_\tau \in \mathbb{R}}{min} \left[\sum_{Y_i \geq \varepsilon_\tau} \mid Y_i - \varepsilon_\tau \mid + \sum_{Y_i < \varepsilon_\tau} (1-\tau) \mid Y_i - \varepsilon_\tau \mid \right], \tag{1.6}$$

donde Y_i son los distintos valores que toman las observaciones de la muestra para la variable Y y ε_τ es el valor que minimiza la expresión anterior. Este valor ε_τ es el valor de la observación que deja una proporción τ de la muestra por debajo y $1-\tau$ por encima , donde $\tau \in (0, 1)$, correspondiente al cuantil que se quiere estimar.

Otra forma habitual de presentar el problema de la Expresión 1.6 es minimizando una suma asimétricamente ponderada de residuales absolutos (simplemente dando diferentes pesos a los residuales positivos y negativos). La solución al siguiente problema nos conduciría a las estimación del cuantil τ:

$$\hat{Q}(\tau) = \underset{\varepsilon_\tau \in \mathbb{R}}{min} [\textstyle\sum_{i=1}^{n} \rho_\tau(Y_i - \varepsilon_\tau)],$$

donde $\rho_\tau(u) = u(\tau - I(u < 0))$ se conoce como función de chequeo con $\tau \in (0, 1)$ y $I(\cdot)$ es la función indicadora definida de la siguiente manera:

$$I(u < 0) = \begin{cases} 1, & si\, u < 0; \\ 0, & si\, u \geq 0. \end{cases}$$

Como se puede observar en la Figura 1–3 la función de verificación ρ tiene pendiente (peso ponderado) τ respecto del eje x sobre la derecha y tiene pendiente (peso ponderado) $\tau - 1$ respecto del eje x sobre la izquierda. Si $\tau = 0.5$, la función de chequeo sería simétrica, lo cual implica que la minimización de la suma de residuales absolutos deba igualar el número de residuales positivos y negativos. Esto asegura que u tenga el mismo número de observaciones por debajo y por encima de la mediana.

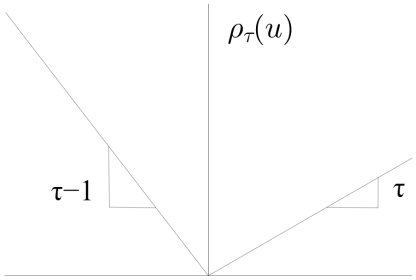


Figura 1–3: Representación gráfica de la función de chequeo ρ

1.4. Regresión a cuantil Simple

El modelo de regresión a cuantil simple se define como (Hao and Naiman, 2007):

$$Y_i = \beta_{0,\tau} + \beta_{1,\tau} x_i + \varepsilon_{i,\tau}, \tag{1.7}$$

donde $0 < \tau < 1$ indica la proporción de la población con puntuaciones por debajo del cuantil τ. El modelo de regresión a cuantil se puede formular de manera equivalente al modelo de regresión lineal simple clásico definido en el Modelo 1.1 con

una declaración sobre los términos de error ε_i; se requiere que el τ-ésimo cuantil del error condicionado al valor que tome la variable explicativa x_i sea cero. Ya que $\beta_{0,\tau} + \beta_{1,\tau}x_i$ es constante, tenemos que el $\tau-$ésimo cuantil condicional de y dado x está dado por:

$$Q_\tau(Y_i \mid x_i) = \beta_{0,\tau} + \beta_{1,\tau}x_i + Q_\tau(\varepsilon_{i,\tau} \mid x_i) = \beta_{0,\tau} + \beta_{1,\tau}x_i.$$

Esto asegura la identificabilidad del intercepto $\beta_{0,\tau}$ en la la función $Q(Y_i \mid x_i)$, conocida como la **función cuantil condicional** de Y_i dado x_i.

La interpretación de los coeficientes en el modelo de regresión a cuantiles es análoga a la interpretación de los coeficientes en el modelo de regresión lineal. Estos coeficientes representan la derivada parcial del cuantil condicional de la variable dependiente con respecto a una variable explicativa (Spatz, 2006):

$$\hat{\beta}_{i,\tau} = \frac{\partial Q_\tau(Y_i|x_i)}{\partial x_i}, \ (i = 0, 1).$$

Esto es, cada coeficiente $\hat{\beta}_{i,\tau}$ puede interpretarse como la tasa de cambio en el $\tau-$ésimo cuantil de la distribución de la variable dependiente y por cada cambio unitario en el valor del regresor x.

1.4.1. Estimación de los parámetros

Se busca minimizar una suma ponderada de residuales $Y_i - \hat{Y}_i$ donde los residuales positivos reciben un peso de τ y los residuales negativos reciben un peso de $1 - \tau$. Formalmente, los estimadores del $\tau-$ésimo cuantil $\hat{\beta}_{0,\tau}$ y $\hat{\beta}_{1,\tau}$ se escogen minimizando (Koenker and Bassett, 1978)

$$\hat{\beta}_\tau = \min_{\beta \in \mathbb{R}} \left[\sum_{i=1}^{n} \rho_\tau (Y_i - b) \right] = \min_{b \in \mathbb{R}} \left[\tau \sum_{Y_i \geq b} \mid Y_i - b \mid + (1-\tau) \sum_{Y_i < b} \mid Y_i - b \mid \right], \quad (1.8)$$

donde $b = \hat{\beta}_{0,\tau} + \hat{\beta}_{1,\tau} x_i$, $\hat{\beta}_\tau = (\hat{\beta}_{0,\tau}, \hat{\beta}_{1,\tau})$ y $\rho_\tau(.)$ es la función de chequeo introducida anteriormente. La estimación de la los parámetros $\hat{\beta}_{0,\tau}$ y $\hat{\beta}_{1,\tau}$ en regresión a cuantiles es usualmente llevada a cabo resolviendo un problema de programación lineal debido a que la función objetivo en la Ecuación 1.8 no es diferenciable. Dicho problema puede ser resuelto mediante diferentes algoritmos entre ellos se encuentran: método Simplex (Koenker and d'Orey, 1987), el método del punto interior Frisch-Newton y el enfoque de Frisch-Newton con preprocesamiento (Portnoy and Koenker, 1997). Se puede encontrar más detalles sobre estos algoritmos para calcular los coeficientes en regresión a cuantiles en Koenker (2005).

1.4.2. Comparación regresión lineal mínimos cuadrados versus regresión a cuantiles

Para hacer la comparación entre regresión lineal simple mínimos cuadrados y regresión a cuantiles, simulamos una base de datos con una variable predictora x (100 datos uniformemente distribuidos entre 0 y 1) y una variable respuesta y de la forma $y = 7 + 0.08x + \varepsilon$, donde ε es el error aleatorio distribuido normalmente con media 0 y varianza no constante de la forma $(0.004x)^2$, dependiente de x. (Ver código en Apéndice A). Un modelo de regresión lineal simple puede tener varianza no constante. Ahora, si es el modelo lineal simple clásico la varianza es constante. Aquí generamos datos de un modelo con varianza no constante para ilustrar la idea que la regresión a la media ignora el cambio de la distribución condicional de Y dado x. Es decir, la relación de Y con x solo se enfoca en la media. Esto es independiente del ajuste de un modelo lineal con varianza constante o no constante. En nuestro

caso, decidimos ilustrar el ajuste asumiendo un modelo con varianza constante, que resulta en estimaciones menos eficientes que si asumimos el verdadero modelo del cual se generaron los datos. A continuación los resultados de los ajustes:

$Coeficientes$	LS	$\tau = 0.10$	$\tau = 0.20$	$\tau = 0.30$	$\tau = 0.40$	$\tau = 0.50$	$\tau = 0.60$	$\tau = 0.70$	$\tau = 0.80$	$\tau = 0.90$
β_0	6.994	6.995	6.998	6.998	6.998	6.999	6.998	6.999	6.998	6.999
β_1	0.096	0.044	0.058	0.066	0.075	0.083	0.091	0.110	0.128	0.133

Tabla 1–1: Comparación regresión lineal (mínimos cuadrados LS) vs regresión a cuantil para cuantiles $\tau = 0.1, 0.2, ..., 0.9$.

En la Figura 1–4 se observa que la estimación del intercepto no cambia mucho pero la pendiente de la recta de cada cuantil es distinta. Esto significa que el predictor x influye de forma distinta en cada cuantil de la variable respuesta. Por ejemplo, 0.128 es el coeficiente estimado de $\beta_{1,0.80}$, que representa la tasa de cambio en el cuantil 0.80 de la distribución de la variable respuesta y por cada cambio unitario de la variable predictora x. Note lo siguiente, el coeficiente estimado de $\beta_{1,0.90}$ es mayor que $\beta_{1,0.10}$, esto indica que para cuantiles superiores, la variable predictora x tiene más influencia en la variable respuesta y. También se puede notar que la línea de regresión mínimos cuadrados tiene más inclinación que la línea de regresión a cuantil para $\tau = 0.50$. No obstante teóricamente las dos líneas deben coincidir ya que los errores en el modelo simulado son simétricos (en teoría la media y la mediana de los errores es igual a 0).

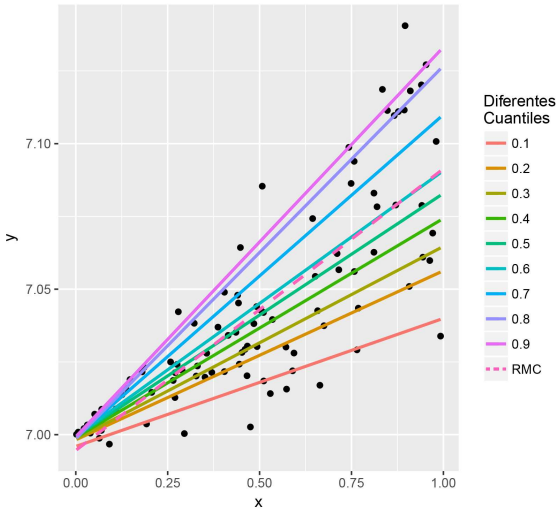


Figura 1–4: Líneas de regresión ajustadas con varios cuantiles $\tau = 0.1, ..., 0.9$. Línea de regresión con mínimos cuadrados (RMC).

1.4.3. Inferencia en regresión a cuantiles

Para hacer inferencias y construir intervalos de confianza para los coeficientes, tenemos que bajo el supuesto de que los errores sean i.i.d., los estimadores $\hat{\beta}_{i,\tau}$ tendrán la propiedad de que asintóticamente, $(\hat{\beta}_{i,\tau} - \beta_{i,\tau})/s_{\hat{\beta}_{i,\tau}}$, donde $s_{\hat{\beta}_{i,\tau}}$ es la desviación estándar de $\hat{\beta}_{i,\tau}$, tiene una distribución normal estándar aproximada (Hao and Naiman, 2007).

Estos procedimientos asintóticos basados en suposiciones paramétricas fuertes pueden ser inapropiados para estimar los intervalos de confianza de los coeficientes, ya que los sesgos y valores atípicos observados hacen que la distribución de los errores no sean i.i.d. (Koenker, 1994). Como alternativa, se recomienda métodos basados en *Bootstrap* para la estimación del error estándar de los coeficientes. Estos métodos se aplican independientemente de la función de densidad de probabilidad de la variable respuesta y del error (Hao and Naiman, 2007).

En el enfoque basado en *Bootstrap*, introducido por (Efron, 1979), para la estimación del error estándar de los coeficientes, se extraen M muestras aleatorias con reemplazo de tamaño n. El error estándar s_{boot} de un coeficiente $\beta_{i,\tau}$ se estima calculando la desviación estándar de los coeficientes estimados $\hat{\beta}_{i,\tau}$ en las muestras con reemplazo. Las estimaciones de *Bootstrap* pueden usarse para calcular un intervalo de confianza para $\hat{\beta}_{i,\tau}$ haciendo uso de la estimación del error estándar y la aproximación normal: $\hat{\beta}_{i,\tau} \pm z_{\alpha/2} \cdot s_{boot}$, donde $z_{\alpha/2}$ es la puntuación z tal que el área bajo la curva normal estándar a la derecha de $z_{\alpha/2}$ es $\alpha/2$, y α es el nivel de significancia. Alternativamente, podemos calcular un intervalo de confianza basado en la distribución empírica de los cuantiles calculados en cada muestra bootstrap (Hao and Naiman, 2007).

En la Figura 1–5 cada punto representa el coeficiente de regresión estimado de un cuantil $\hat{\beta}_{1,\tau}$ con su intervalo de confianza del 95 % (región gris). La línea horizontal continua es el coeficiente de regresión estimado para el predictor utilizando mínimos cuadrados ordinarios y las líneas horizontales discontinuas sus límites de confianza del 95 %. Se puede ver que los cuantiles inferiores y superiores no coinciden con la estimación de mínimos cuadrados. Para el caso específico de $\tau = 0.9$ el coeficiente estimado $\hat{\beta}_1$ es igual a 0.154, es decir, en un incremento de una unidad en la variable predictora x, la variable respuesta y puede aumentar 0.154 unidades en el cuantil 0.9. En cambio la estimación con mínimos cuadrados de $\hat{\beta}_1$ es menor (0.097), lo cual significa que el incremento en una unidad de la variable x, se asocia con un incremento promedio de 0.097 unidades en la variable respuesta y.

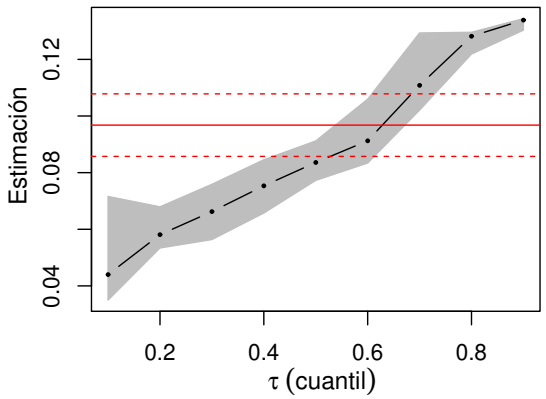


Figura 1–5: Resultados de la regresión a cuantiles. Estimación puntual e intervalos de confianza (95 %) para $\tau = 0.1, 0.2, ..., 0.9$.

1.5. Regresión a cuantil múltiple

En la sección anterior se trabajó con regresión a cuantil simple, la cual es apropiada en muchas situaciones cuando solamente se dispone de una variable explicativa. Sin embargo, muchos problemas involucran varias variables explicativas.

El modelo de regresión a cuantil múltiple se puede expresar de la siguiente manera (Buchinsky, 1998)

$$Y_i = \boldsymbol{x}_i^t \boldsymbol{\beta}_\tau + \boldsymbol{\varepsilon}_{i,\tau}, \quad 0 < \tau < 1. \tag{1.9}$$

En el modelo anterior se supone que $Q_\tau(\varepsilon_{i,\tau} \mid \boldsymbol{x}_i) = 0$, donde $\varepsilon_{i,\tau}$ es el término residual del modelo de regresión en el τ-ésimo cuantil.

El τ-ésimo cuantil condicionado a la covariable $\boldsymbol{x}_i$ se puede expresar de la siguiente manera (Costanzo and Desimoni, 2017)

$$Q_\tau(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}_\tau, \tag{1.10}$$

donde para cada observación i se observa una variable respuesta Y_i y d covariables $\boldsymbol{x}_i$, donde $\boldsymbol{x}_i$, es un vector columna de variables explicativas $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})^t$, $\boldsymbol{\beta_\tau} = (\beta_1, ..., \beta_d)^t$ y τ es el cuantil de interés.

1.5.1. Estimación de los parámetros del modelo

Los métodos de regresión cuantil se basan en minimizar los residuos absolutos ponderados asimétricamente. El estimador de regresión a cuantil denotado por $\hat{\boldsymbol{\beta}}_\tau$ se encuentra mediante la solución del problema de minimización (Buchinsky, 1998)

$$\hat{\boldsymbol{\beta}}_\tau = \frac{1}{n} \underset{\boldsymbol{\beta}}{min} \left[\sum_{Y_i \geq \boldsymbol{x}_i^t \boldsymbol{\beta}} \tau \mid Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta} \mid + \sum_{Y_i < \boldsymbol{x}_i^t \boldsymbol{\beta}} \mid Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta} \mid \right] = \frac{1}{n} \underset{\beta}{min} \left[\sum_{i=1}^{n} \rho_\tau (Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}) \right], \tag{1.11}$$

donde $\hat{\boldsymbol{\beta}}_\tau$ son los estimadores del τ-ésimo cuantil, $\hat{\boldsymbol{\beta}}_\tau = (\hat{\beta}_{1,\tau}, ..., \hat{\beta}_{d,\tau})^t$.

La estimación de los parámetros en un modelo de regresión a cuantil se puede obtener usando el procedimiento `PROC QUANTREG` de SAS y la librería `quantreg` de R . Ver pasos en Apéndice A.

1.6. Método de selección de variables

En las aplicaciones de regresión para fines de predicción e interpretación, a menudo hay un gran número de variables independientes disponibles. Usualmente existe incertidumbre sobre cuáles de estas variables independientes deberían incluirse en el modelo final, ya que una predicción adecuada puede ser posible utilizando solo un subconjunto de las variables disponibles (Thompson, 1978). Es decir, seleccionar un modelo adecuado podría redundar en predicciones con poco sesgo y menor variabilidad.

Existen varios métodos de selección de variables tanto clásicos como modernos.

En este capítulo se discuten dos enfoques para seleccionar variables: métodos de regresión paso a paso y métodos de regularización. En los métodos de regresión paso a paso los subconjuntos de variables del modelo se identifican secuencialmente agregando o eliminando, dependiendo del método, la variable que tiene el mayor impacto en la suma de cuadrados residuales. Entre los métodos de selección de variables regresión paso a paso se encuentran: selección hacia adelante, eliminación hacia atrás y selección paso a paso (John O. Rawlings, 1998). De otro lado, existen los métodos de penalización o regularización. Estos métodos, en el contexto de regresión lineal, minimizan la suma de cuadrados del error con una penalidad que tiene como objetivo seleccionar variables y estimar al mismo tiempo. Estos métodos de regularización han tenido un auge reciente, particularmente para problemas con un número grande de variables comparado al número de individuos.

En la siguiente sección nos enfocaremos a describir los métodos de regularización o penalización, el énfasis en la aplicación de este trabajo.

1.7. Métodos de regularización o penalización

Este enfoque implica ajustar un modelo de regresión que involucre todas las variables predictoras. Sin embargo, los coeficientes estimados de variables que no son significativas (coeficientes relativamente pequeños) se reducen a cero en relación con las estimaciones de mínimos cuadrados. Esta contracción (también conocida como regularización o encogimiento *(shrinkage)* tiene el efecto de reducir la varianza, a pesar de que se obtienen estimaciones sesgadas. Dependiendo de qué tipo de penalización se realice, algunos de los coeficientes pueden estimarse exactamente cero (Gareth James, 2015). Ha habido una enorme cantidad de actividad de investigación dedicada a los métodos de regularización entre ellas: i) LASSO, ii) LASSO agrupado (Yuan and Lin, 2007) donde las variables se incluyen o excluyen en grupos, iii) el

selector de Dantzig (Candes and Tao, 2007), una versión ligeramente modificada del LASSO, iv) la red elástica (Zou and Hastie, 2005) para las variables correlacionadas que usa una penalización que combina dos normas, v) métodos que utilizan penalizaciones no cóncavas, como SCAD (Fan and Li, 2001), vi) el LASSO gráfico (Friedman J, 2008) para la estimación de covarianza dispersa y gráficos no dirigidos, entre otros. Todos estos métodos tienen como característica que estiman coeficientes y seleccionan términos en el modelo simultáneamente.

1.8. Método de penalización con mínimos cuadrados

El método de mínimos cuadrados ordinarios no funciona cuando el número de variables p es superior al número de observaciones n. Una alternativa es usar regresión penalizada que permite crear un modelo de regresión lineal con menos variables en el modelo al agregar una penalidad a la función objetivo (Bruce and Bruce, 2017; Gareth et al., 2014). Esto también se conoce como métodos de contracción o regularización. La consecuencia de imponer esta penalización es reducir los valores de los coeficientes de variables no significativas hacia cero.

Esta contracción (también conocida como regularización) tiene el efecto de reducir la varianza, aunque el sesgo podría aumentar. Este método consiste en estimar el vector de parámetros β, minimizando la expresión

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} p_\lambda(\mid \beta_j \mid),$$

donde $p_\lambda(\cdot)$ es la función de penalización y λ es el parámetro de penalización (también conocido como *tuning parameter*) (Racine et al., 2014). La elección de la función de penalización determina el tipo de estimador. En la literatura reciente una elección popular es $p_\lambda(\mid u \mid) = \lambda \mid u \mid$ que produce el estimador LASSO propuesto por Tibshirani (1996) y el método SCAD propuesto por (Fan and Li, 2001). Este último

se usó en este trabajo debido a su buen desempeño en modelos de regresión lineal (Fan and Li, 2001).

1.9. Regresión penalizada con SCAD en un modelo lineal

Fan and Li (2001) propusieron la función de penalización SCAD (Smoothly Clipped Absolute Deviation, en inglés). Estos autores describieron las siguientes condiciones de una buena función de penalización:

a) Insesgamiento (*Unbiasedness*): el estimador resultante es casi insesgado cuando el verdadero parámetro desconocido es grande para evitar un sesgo de modelado innecesario.

b) Esparcimiento (*Sparsity*): el estimador resultante es una regla de umbral, que establece automáticamente pequeños coeficientes a cero para reducir la complejidad del modelo.

c) Continuidad (*Continuity*): el estimador resultante es continuo en los datos para evitar la inestabilidad en la predicción del modelo.

El Estimador de Mínimos Cuadrados con la función de penalización SCAD minimiza la función de criterio (Jung, 2014)

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_\lambda(\mid \beta_j \mid),$$

donde

$$p_\lambda(\mid \beta \mid) = \begin{cases} \lambda \mid \beta \mid, & \text{si } \ 0 \leq \beta < \lambda, \\ \frac{(a^2-1)\lambda^2-(\mid\beta\mid-a\lambda)^2}{2(a-1)}, & \text{si } \lambda \leq \mid \beta \mid < a\lambda, \\ \frac{1}{2}(a+1)^2\lambda^2, & \text{si } \ \mid \beta \mid > a\lambda. \end{cases} \tag{1.12}$$

Su derivada se convierte en

$$p'_\lambda(\mid \beta \mid) = \begin{cases} \lambda, & \text{si } \ 0 \leq \beta < \lambda, \\ \frac{a\lambda-\mid\beta\mid}{a-1}, & \text{si } \lambda \leq \mid \beta \mid < a\lambda, \\ 0, & \text{si } \ \mid \beta \mid > a\lambda, \end{cases}$$

donde a es uno de los parámetros de ajuste que puede elegirse mediante validación cruzada o validación cruzada generalizada tal implementación puede ser computacionalmente costosa. Jung (2014) y Fan and Li (2001) recomiendan utilizar $a = 3.7$, estos autores calcularon el riesgo de Bayes a través de la integración numérica los riesgos de Bayes alcanzan sus mínimos en $a \approx 3.7$. El parámetro de ajuste $\lambda > 0$ controla la compensación entre el ajuste del modelo y el esparcimiento del modelo (Jung, 2014).

Fan and Li (2001) encontraron los valores de los parámetros de ajuste a través de la validación cruzada y la validación cruzada generalizada. Wang et al. (2007a) utilizaron el parámetro de ajuste al minimizar una función de criterio de tipo BIC y Jung (2014) propone hallar el parámetro de ajuste por validación cruzada generalizada (GCV).

La regresión penalizada con SCAD no solo selecciona covariables importantes de manera consistente, sino que también produce estimadores de parámetros tan eficientes como si se conociera el modelo verdadero, es decir, tiene la Propiedad del Oráculo (*Oracle Property*, en inglés) (Jung, 2014) .

La Propiedad del Oráculo establece que cuando los parámetros verdaderos tienen algunos componentes iguales a cero, se estiman como 0 con una probabilidad que tiende a 1, y los componentes distintos de cero se estiman tan bien como cuando se conoce el submodelo correcto (Fan and Li, 2001).

En la Figura 1–6 se puede observar que la función de penalización SCAD es singular en el origen, una condición necesaria para esparcimiento en la selección de variables. Además, la penalización SCAD es no convexa sobre $(0, +\infty)$ para reducir el sesgo de la estimación.

Note también que la penalización de SCAD reduce a cero los coeficientes pequeños , mientras que se estabiliza cuando los coeficientes son grandes; esta función de penalización no penaliza excesivamente valores grandes de β.

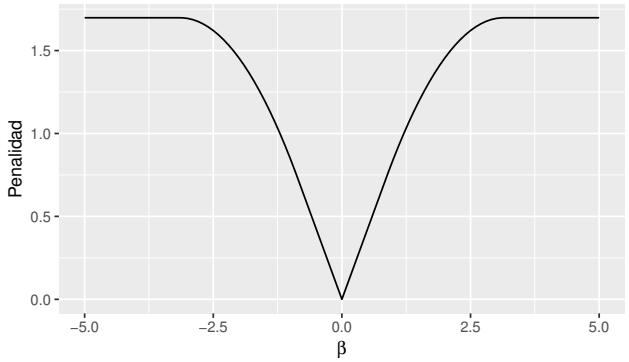


Figura 1–6: Función de penalidad SCAD para un coeficiente β.

El razonamiento detrás de las penalizaciones se puede entender considerando las soluciones univariadas. Considere la regresión lineal simple $\boldsymbol{Y}$ sobre $\boldsymbol{x}$ con la solución de minímos cuadrados denotada por z (Breheny and Huang, 2015). Para este problema de regresión lineal simple la estimación SCAD tienen la siguiente forma (Fan, 1997; Fan and Li, 2001).

$$\hat{\beta} = \begin{cases} sgn(z)(\mid z \mid -\lambda)_+, & si \mid z \mid < 2\lambda, \\ \{(a-1)z - sgn(z)a\lambda\}/(a-2), & si\ 2\lambda < \mid z \mid \leq a\lambda, \\ z, & si \mid z \mid > a\lambda, \end{cases}$$

donde la función $sgn(x)$ es la función signo, definida por:

$$sgn(x) := \begin{cases} -1, & si\, x < 0, \\ 0, & si\ x = 0, \\ 1, & si\, x > 0, \end{cases}$$

y la función $(x)_+$ es la función parte positiva, definida por:

$$(x)_+ := \begin{cases} x, & si\, x \geq 0, \\ 0, & si\ x < 0. \end{cases}$$

En la Figura 1–7, cuando $\mid z \mid > a\lambda$ (cuando los coeficientes son coeficientes grandes), la solución se amplía completamente a la solución de mínimos cuadrados no penalizados (línea punteada). Cuando $-\lambda < z < \lambda$, los coeficientes estimados con la penalización SCAD se reducen a cero (note la continuidad en las soluciones).

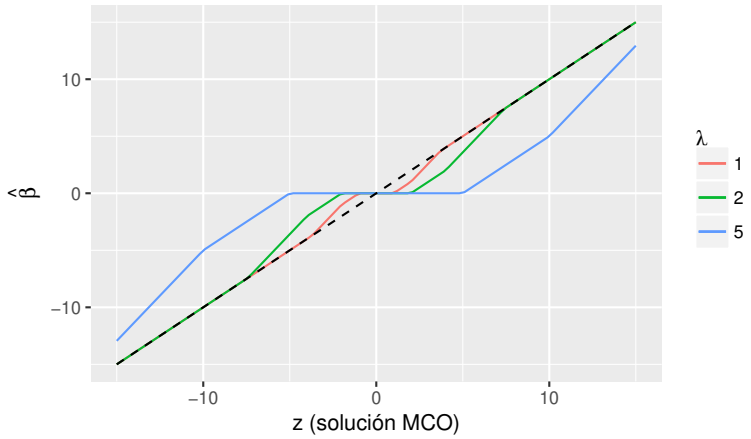


Figura 1–7: Función de umbral SCAD con $\lambda = 1, 2, 5$ y $a = 3.7$. $\hat{\beta}$ es la solución de mínimos cuadrados penalizada por SCAD y z es la solución de mínimos cuadrados.

1.10. Regresión lineal penalizada con SCAD agrupada

En muchos problemas de regresión nos interesa encontrar factores explicativos importantes para predecir la variable respuesta donde cada factor explicativo puede estar representado por un grupo de variables. El ejemplo más común es el problema del análisis multifactor de varianza **(ANOVA)**, en el cual cada factor puede tener varios niveles y se puede expresar a través de un grupo de variables indicadoras. El objetivo de **ANOVA** es a menudo seleccionar importantes efectos e interacciones principales para una predicción precisa. Otro ejemplo es el modelo aditivo con componentes polinomiales o no paramétricos. En ambas situaciones, cada componente en el modelo aditivo puede expresarse como una combinación lineal de varias funciones básicas de la variable medida original. En tales casos, la selección de variables importantes corresponde a la selección de grupos de funciones básicas. En ambos ejemplos, la selección de variables generalmente equivale a la selección de factores importantes (grupos de variables) en lugar de variables individuales, ya que cada factor corresponde a una variable medida y está directamente relacionada con el valor de la medición (Yuan and Lin, 2006b) .

Considere el problema de regresión general con J factores como (Yuan and Lin, 2006b) :

$$\boldsymbol{Y} = \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j + \varepsilon, \tag{1.13}$$

donde J es el número de grupos, $\boldsymbol{Y}$ es el vector de las variables respuesta $\boldsymbol{Y} = (Y_1, ..., Y_n)^t$, $\boldsymbol{X}_j$ es una matriz $n \times p_j$ correspondiente al j-ésimo factor y β_j es un vector de coeficiente de tamaño p_j, $j = 1, ..., J$ y el error, $\varepsilon \sim N_n(0, \sigma^2 \boldsymbol{I})$. Asumiendo que cada $\boldsymbol{X}_j$ está ortonormalizado, es decir $\boldsymbol{X}_j^t \boldsymbol{X}_j = \boldsymbol{I}_{p_j}$, $j = 1, ..., J$. Denote $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_J)$ y $\boldsymbol{\beta} = (\beta_1^t, ..., \beta_J)^t$. La Ecuación 1.13 puede ser escrita como $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$.

Cada uno de los factores en la Ecuación 1.13 puede ser categórico o continuo. El modelo ANOVA tradicional es el caso especial en el que todos los factores son categóricos y el modelo aditivo es un caso especial en el que todos los factores son continuos. Es claramente posible incluir factores categóricos y continuos en la Ecuación 1.13.

El objetivo es seleccionar factores importantes (grupos de variables) para una estimación precisa en la Ecuación 1.13. Varios métodos se han introducido para este problema en los últimos años, uno de los más conocido es LASSO agrupado (Yuan and Lin, 2006b) y SCAD agrupado (Wang et al., 2007b).

El LASSO agrupado tiene muchas propiedades atractivas, pero no posee la consistencia de selección de nivel grupal. De hecho, tiende a seleccionar más grupos (variables) de los necesarios. Se han propuesto métodos penalizados no convexos que satisfacen la propiedad del oráculo para la selección de variables grupales. Un

estimador penalizado de grupo no convexo se define como el minimizador (Lee et al., 2016)

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{X}_j\boldsymbol{\beta}_j\|_2^2 + \sum_{j=1}^{J} K_{\lambda_j}(\|\beta_j\|_2), \tag{1.14}$$

donde $K_{\lambda_j}(\cdot)$ son las funciones de penalización no convexa definidas en la Ecuación 1.12 y λ_j son los parámetros de regularización (Fan and Li, 2001). Se ha establecido $\lambda_j = \lambda\sqrt{p}_j$ para algún $\lambda > 0$, para p_j al igual número de covariables en el $j-$ésimo grupo. Si $K_\lambda(\cdot)$ es la penalización en Fan and Li (2001), el estimador se convierte en agrupado SCAD (Wang et al., 2007b).

Breheny and Huang (2012) y Wei and Zhu (2012) ampliaron el algoritmo de Yuan and Lin (2006a) para encontrar el estimador agrupado SCAD. Obtuvieron la forma explícita de la solución para cada iteración bajo el supuesto de ortogonalidad dentro de cada grupo. Las soluciones de forma cerrada del grupo SCAD en el grupo j-ésimo se dan de la siguiente manera (Lee et al., 2016):

$$\hat{\boldsymbol{\beta}}_j^{gSCAD} = \begin{cases} S(\boldsymbol{s}_j, \lambda_j) & si \;\; \|\boldsymbol{s}_j\|_2 < 2\lambda_j, \\ \frac{a-1}{a-2} S(\boldsymbol{s}_j, \frac{a\lambda_j}{a-1}) & si \;\; 2\lambda_j < \|\boldsymbol{s}_j\|_2 \leq a\lambda_j. \\ \boldsymbol{s}_j & si \;\; \|\boldsymbol{s}_j\|_2 > a\lambda_j, \end{cases} \tag{1.15}$$

para algún $a > 2$, donde $S(\mathbf{z}, \lambda) = (1 - \frac{\lambda}{\|\boldsymbol{z}\|_2})_+\boldsymbol{z}$ es el operador de umbral suave *(soft-threshding)*, correspondiente a la solución LASSO bajo un diseño ortonormal (Huang et al., 2012).

En cualquier enfoque de penalización para la selección de variables, una pregunta difícil es cómo determinar los parámetros de penalización. Esta pregunta es aún

más difícil en los métodos de selección grupal. Los criterios ampliamente utilizados, incluidos el AIC (Akaike, 1973) y el BIC (Schwarz, 1978), requieren la estimación de la varianza del error y los grados de libertad (Huang et al., 2012).

1.11. Regresión a cuantil con penalidad SCAD

Considere el modelo de regresión a cuantil dado anteriormente

$$Y_i = \boldsymbol{x}_i^t \boldsymbol{\beta}_\tau + \boldsymbol{\varepsilon}_{i,\tau} \ ,$$

donde

$$Q_\tau(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}_\tau.$$

Para cada observación i se observa una variable respuesta Y_i y d covariables $\boldsymbol{x}_i$, donde $\boldsymbol{x}_i$, es un vector columna de variables explicativas $\boldsymbol{x}_i = (x_{1i}, ..., x_{di})^t$, de dimensión $d < n$, $\boldsymbol{\beta}$ (es el vector de coeficientes), es un vector columna de dimensión d, es decir $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^t$, τ es el cuantil condicional de interés y se supone que $Q_\tau(\varepsilon_{i,\tau} \mid \boldsymbol{x}_i) = 0$ y $\varepsilon_{i,\tau}$ es el término residual del modelo de regresión en el τ-ésimo cuantil.

El estimador de regresión a cuantil denotado por $\hat{\boldsymbol{\beta}}_\tau$ se encuentra mediante la solución del problema de minimización

$$\hat{\beta}_\tau = \frac{1}{n} \min_{\beta} \left[\sum_{i=1}^n (Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})\right].$$

El estimador de regresión a cuantil con la función de penalización SCAD minimiza la función objetivo (Wang et al., 2012)

$$Q(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}) + \sum_{j=1}^d p_\lambda(\mid \beta_j \mid),$$

donde $\rho_\tau(u) = u\{\tau - I(u < 0)\}$, es la función de pérdida de cuantiles (o función de chequeo), y $p_\lambda(.)$ es una función de penalización con un parámetro de ajuste λ. El parámetro de ajuste λ controla la complejidad del modelo. La función de penalización SCAD cuantil $p_\lambda(.)$, es la de Fan and Li (2001) definida en la Ecuación 1.12.

1.12. Regresión a cuantil con penalidad SCAD agrupada

Para estimar β_τ, el estimador de regresión a cuantil con la función de penalización SCAD agrupada minimiza la función objetivo

$$\beta_\tau = \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}) + \sum_{j=1}^{J} K_{\lambda_j}(\| \beta_j \|_2),$$

donde $K_{\lambda_j}(\cdot)$ son las funciones de penalización no convexa de Fan and Li (2001) definida en la Ecuación 1.12 pero reemplazando el valor absoluto (norma L_1) por $\| \cdot \|_2$, definida como,

$$L_2 = \| x \|_2 = (x_1^2 + ... + x_n^2)^{1/2} = [\sum_{i=1}^{n} | x_i |^2]^{1/2}.$$

Los λ_j son los parámetros de regularización, donde se ha establecido $\lambda_j = \lambda\sqrt{p}_j$ para algún $\lambda > 0$ y p_j es el número de covariables en el $j-$ésimo grupo.

Este método de selección de variables en regresión a cuantiles se encuentra implementado en la función `rq` de la librería `quantreg` de R. Debido a sus buenas propiedades estadísticas mencionadas en la literatura, en este trabajo se usará este método de penalización para determinar los factores asociados al peso de los recién nacidos en PR en los años 2009-2011.

Capítulo 2
DESCRIPCIÓN DE LOS DATOS

2.1. Descripción de la base de datos

Los datos que se van a utilizar para este estudio son los nacimientos registrados en Puerto Rico en los años 2009- 2011, los cuales incluye información sociodemográfica sobre el recién nacido y los padres, al igual que aspectos de salud que pueden afectar el nacimiento. Esta información está disponible en la página web, http://www.estadisticas.gobierno.pr/iepr/Estadisticas/Basesdedatos

La Tabla 2–1 muestra información acerca de las bases de datos en cada año que se incluyó en este trabajo.

Año	Número de registros	Cantidad de variables
2009	44,833	314
2010	42,247	314
2011	41,130	314

Tabla 2–1: Registros originales

En un primer paso de este trabajo se procedió a hacer una limpieza y edición de las bases de datos originales como se discute a continuación. Del total de registros entre 2009 y 2011, se escogieron registros de nacimientos únicos (*single*) de cada año, raza blanca y negra de los padres de cada año, se eliminaron registros cuya información en la variable es desconocida o incompleta. Luego se utilizó la función

`imagmiss` de la librería `dprep` que calcula el porcentaje de datos faltantes (NA). Se encontró que en las tres bases de datos las variables que tenían datos faltantes eran las mismas (el porcentaje de datos faltantes varía de 94.27 % a 100 %) excepto la base de datos de 2011 que tiene 7 columnas más con datos faltantes. (Tabla 2–2).

Año	# de columnas con datos faltantes
2009	106
2010	106
2011	113

Tabla 2–2: Número de columnas con datos faltantes por año

En la base de datos hay dos columnas con datos constantes que no nos interesa en el estudio las cuales son: REVISION: (A repetida), OTERR: TERRITORIO DE OCURRENCIA (PR repetido). Por lo tanto, la cantidad de columnas que se van a eliminar en cada base de datos fue 115 columnas.

Al hacer lo anteriormente mencionado nos quedan las bases de datos con la siguiente información Tabla 2–3.

Año	Número de registros	Cantidad de variables
2009	40,836	199
2010	38,134	199
2011	37,252	199

Tabla 2–3: Registros sin datos faltantes

Luego se hizo un muestreo estratificado por año de 2000 registros. Estas tres bases de datos se unieron para obtener una nueva base de datos con 6000 registros y 199 variables. La motivación para llevar a cabo el muestreo es disminuir la complejidad al tratar de ajustar los modelos debido a que usualmente requiere bastantes

recursos computacionales. En la práctica estos modelos se pueden ajustar si se cuenta con buenos recursos computacionales, por ejemplo, un servidor o "clusters" con mayor capacidad. La mayor limitación proviene de la demanda de las rutinas de optimización y de la estimación de varios cuantiles. Vale la pena recordar que el número de observaciones en nuestra aplicación es igual a 116222, un número que trae problemas en el ajuste de los modelos usando una computadora portatil típica.

Después se analizó cada una de las 199 variables de la base de datos con 6000 registros y se eliminaron aquellas variables bajo los siguientes criterios: algunas variables relacionadas al momento del parto, por ejemplo si usaron fórceps para sacar al bebé, día de la semana que nació el bebé, entre otras. Se eliminaron variables que tenían dos categorías, con muy pocas respuestas por ejemplo, cigarrillos en el primer trimestre 5999 ninguno y 2 cigarrillos diarios sólo una respuesta alternativa; algunas variables medidas después del parto, por ejemplo, ventilación asistida. También se sacaron variables que que se repiten con distintas categorías (quedando las mejor representadas). Al final, la base de datos quedó con 6000 registros y 28 variables, incluida la variable respuesta (DBWT: Peso al nacer). La Tabla 2–4 lista las variables con su respectivo tipo.

$VARIABLE$	$TIPO$ $CUANTITATIVA/CATEGORICA$
$MAGER$ ($Edad\,de\,la\,madre$)	Cuantitativa
$FAGECOMB$ ($Edad\,del\,padre$)	Cuantitativa
$DBWT$ ($Peso\,al\,nacer$)	Cuantitativa
$APGAR5$ ($Puntaje\,de\,5\,minutos$)	Categórica
$DWGT$ ($Peso\,de\,la\,madre\,en\,la\,entrega$)	Cuantitativa
DOB_YY ($Año\,de\,nacimiento$)	Categórica
$DMETH\_REC$ ($Método\,de\,nacimiento$)	Categórica
LBO_REC ($Orden\,de\,nacimiento$)	Categórica
RF_DIAB ($Diabetes\,en\,pre-embrazo$)	Categórica
RF_GEST ($Diabetes\,gestacional$)	Categórica
RF_PHYP ($Hipertensión\,pre-gestacional$)	Categórica
RF_GHYP ($Hipertensión\,gestacional$)	Categórica
RF_PPTERM ($Parto\,pretérmino\,anterior$)	Categórica
RF_PPOUTC ($Resultados\,de\,embarazo\,precario$)	Categórica
URF_DIAB ($Diabetes$)	Categórica
URF_CHYPER ($Hipertensión\,crónica$)	Categórica
URF_PHYPER ($Hipertensión\,asociada\,al\,embarazo$)	Categórica
SEX ($Sexo\,del\,bebé$)	Categórica
$MBRACE$ ($Raza\,de\,la\,madre$)	Categórica
MAR ($Estado\,cívil\,de\,la\,madre$)	Categórica
$FBRACE$ ($Raza\,del\,padre$)	Categórica
$PRECARE\_REC$ ($Comienzo\,del\,cuidado\,prenatal$)	Categórica
$WTGAINC$ ($Aumento\,de\,peso\,de\,la\,madre$)	Cuantitativa
$MEDUC$ ($Educación\,de\,la\,madre$)	Categórica
$PREVIS\_REC$ ($Número\,de\,vistitas\,prenatales$)	Categórica
$GESTREC10$ ($Gestación\,semanas$)	Categórica
$MRCNTY$ ($Pueblo\,residencia\,de\,la\,madre$)	Categórica
$OCNTY$ ($Pueblo\,de\,ocurrencia$)	Categórica

Tabla 2–4: Variables seleccionadas para el análisis

2.2. Análisis descriptivo

En las variables seleccionadas para el análisis se tiene 23 variables cualitativas con información acerca de condiciones médicas de la madre, estado civil de la madre, educación, condado de ocurrencia, raza de los padre, sexo del recién nacido, año de nacimiento, entre otras. Este análisis descriptivo nos permitirá tener una mejor idea de las variables explicativas que serán usadas en los modelos y también nos permitirá entender mejor los resultados de los modelos.

En las Tabla 2–5 a 2–11 se encuentran registros con información acerca de las condiciones médicas de la madre, evidenciándose que menos del 4 % en la muestra de las madres presentaron una condición médica. Las condiciones más frecuentes fueron: hipertensión gestacional (3.38 %), hipertensión asociada al embarazo (3.38 %), diabetes(3 %) y diabetes gestacional (2.33 %), mientras que las menos frecuentes fueron: hipertensión pre gestacional (1.27 %), hipertensión crónica (1.27 %) y diabetes en pre embarazo (0.67 %).

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_DIAB	*SI*	40	0.67 %
(*Diabetes en pre embarazo*)	*NO*	5960	99.33 %

Tabla 2–5: Diabetes en pre-embarazo

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_GEST	*SI*	140	2.33 %
(*Diabetes gestacional*)	*NO*	5860	97.67 %

Tabla 2–6: Diabetes gestacional

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_PHYP	SI	76	1.27 %
$(Hipertensión\ pre-gestacional)$	NO	5924	98.73 %

Tabla 2–7: Hipertensión pre-gestacional

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_GHYP	SI	203	3.38 %
$(Hipertensión\ gestacional)$	NO	5797	96.62 %

Tabla 2–8: Hipertensión gestacional

VARIABLE	RESPUESTA	FRECUENCIA	%
URF_DIAB	SI	180	3.00 %
$(Diabetes)$	NO	5820	97.00 %

Tabla 2–9: Diabetes

VARIABLE	RESPUESTA	FRECUENCIA	%
URF_CHYPER	SI	76	1.27 %
$(Hipertensión\ crónica)$	NO	5924	98.73 %

Tabla 2–10: Hipertensión crónica

VARIABLE	RESPUESTA	FRECUENCIA	%
URF_PHYPER	SI	203	3.38 %
$(Hipertensión\ asociada\ al\ embarazo)$	NO	5797	96.62 %

Tabla 2–11: Hipertensión asociada al embarazo

Los registros con información acerca del estado civil la madre, revelan que más del 50 % de las madres en la muestra no están casadas (Tabla 2–12).

VARIABLE	RESPUESTA	FRECUENCIA	%
MAR (*Estado civil de la madre*)	*CASADA*	2173	36.22 %
	NO CASADA	3827	63.78 %

Tabla 2–12: Estado civil de la madre

En la Tabla 2–13 se encuentra los registros con información acerca del número de nacimientos, se destaca que los partos primerizos son más frecuentes en la muestra.

VARIABLE	RESPUESTA	FRECUENCIA	%
LBO_REC (*Número de nacimientos*)	1*er nacimiento*	2778	46.3 %
	2*do nacimiento*	2131	35.52 %
	3*er nacimiento ó más*	1091	18.18 %

Tabla 2–13: Número de nacimientos

En cuanto a la educación de la madre (Tabla 2–14), según los datos las graduadas de secundaria son más frecuentes en la muestra (41.03 %) y las de maestría o doctorado las menos frecuentes (4.87 %).

VARIABLE	RESPUESTA	FRECUENCIA	%
MEDUC (*Educación de la madre*)	*noveno grado ó menos*	984	16.4 %
	graduado de secundaria	2462	41.03 %
	grado asociado	1059	17.65 %
	licenciatura	1203	20.05 %
	maestría/doctorado	292	4.87 %

Tabla 2–14: Educación de la madre

En relación a la información del pueblo de residencia de la madre, se puede observar que más del 50 % de las madres de la muestra residen en un pueblo con menos de 100, 000 habitantes, siguiéndole el pueblo de San Juan con un 9.3 % de las residentes de la muestra (Tabla 2–15).

VARIABLE	RESPUESTA	FRECUENCIA	%
MRCNTY (*Pueblo de residencia madre*)	*Bayamon*	324	5.4 %
	Caguas	241	4.02 %
	Carolina	261	4.35 %
	Mayaguez	122	2.03 %
	Ponce	356	5.93 %
	San Juan	558	9.3 %
	Pob menor de 100, 000 *hab*	4138	68.97 %

Tabla 2–15: Pueblo de residencia de la madre

En la Tabla 2–16 se encuentra la información acerca del método de nacimiento, mostrando que más del 50 % de los partos de la muestra fueron de forma vaginal. En la información concerniente a los partos pretérmino anterior, se evidencia que solo un 1.35 % de los partos de la muestra fueron pretérmino (Tabla 2–17). La Tabla 2–18 es correspondiente al comienzo del cuidado prenatal, la cual revela con mayor frecuencia las madres en la muestra que iniciaron el cuidado prenatal desde el primer hasta el cuarto mes de embarazo (76.75 %).

VARIABLE	RESPUESTA	FRECUENCIA	%
DMETH_REC (*Método de nacimiento*)	*VAGINAL*	3182	53.03 %
	CESAREA	2818	46.97 %

Tabla 2–16: Método de nacimiento

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_PPTERM (*Parto pretérmino anterior*)	*SI*	81	1.35 %
	NO	5919	98.65 %

Tabla 2–17: Parto pretérmino anterior

VARIABLE	RESPUESTA	FRECUENCIA	%
PRECARE_REC (*Comienzo cuidado prenatal*)	*del* 1*er al* 4*to mes*	4605	76.75 %
	del 4*to al sexto mes*	1256	20.93 %
	séptimo al último ó no lo hizo	139	2.32 %

Tabla 2–18: Comienzo del cuidado prenatal

Los datos concernientes acerca del número de visitas prenatales, muestran con mayor frecuencia aquellas madres en la muestra que hicieron de 11 a 12 visitas prenatales (Tabla 2–19). Con referencia a las semanas de gestación, se evidencia que más del 60 % son aquellas madres en la muestra que tuvieron de 37 a 39 semanas de gestación (Tabla 2–20). La raza de los padres la Tabla 2–21 y la Tabla 2–22 revelan que la raza blanca es predominante en la muestra con más del 80 %.

VARIABLE	RESPUESTA	FRECUENCIA	%
$PREVIS\_REC$ (*Número de visitas prenatales*)	10 *visitas ó menos*	1533	25.55 %
	11 *a* 12 *visitas*	1833	30.55 %
	13 *ó más visitas*	2634	43.9 %

Tabla 2–19: Número de vistas prenatales

VARIABLE	RESPUESTA	FRECUENCIA	%
$GESTREC10$ (*Gestación semanas*)	20 *a* 36 *semanas*	938	15.63 %
	37 *a* 39 *semanas*	4032	67.2 %
	40 *o más semanas*	1030	17.17 %

Tabla 2–20: Semanas de gestación

VARIABLE	RESPUESTA	FRECUENCIA	%
$MBRACE$ (*Raza de la madre*)	$BLANCA$	5364	89.4 %
	$NEGRA$	636	10.6 %

Tabla 2–21: Raza de la madre

VARIABLE	RESPUESTA	FRECUENCIA	%
$FBRACE$ (*Raza del padre*)	$BLANCO$	5252	87.53 %
	$NEGRO$	748	12.47 %

Tabla 2–22: Raza del padre

En la Tabla 2–23 se encuentra presente la información sobre el sexo del recién nacido en la muestra, revelando que hubo un poco más de nacimientos del sexo masculino (51.8 %).

En cuanto a la puntuación de (Apariencia, pulso, gesticulación, actividad y respiración al primer y quinto minuto de recién nacido) APGAR5 en la Tabla 2–24, se observa que más del 90 % de los recién nacidos en la muestra tuvieron un buen puntaje $(7-10)$.

VARIABLE	RESPUESTA	FRECUENCIA	%
SEX (*Sexo del recién nacido*)	*MASCULINO*	3108	51.8 %
	FEMENINO	2892	48.2 %

Tabla 2–23: Sexo del recién nacido

VARIABLE	RESPUESTA	FRECUENCIA	%
*APGAR*5 *Apariencia, pulso, gesticulación, actividad y respiración* 5 *minutos*	*puntuación* $1-3$	6	0.1 %
	puntuación $4-6$	28	0.47 %
	puntuación $7-10$	5966	99.43 %

Tabla 2–24: Puntuación APGAR 5 minutos

La información registrada acerca del pueblo de ocurrencia indica que más del 30 % de los nacimientos en la muestra ocurrieron en un pueblo con menos de $100,000$ habitantes, siguiéndole el pueblo de San Juan con un 27.1 % de nacimientos ocurridos en la muestra (Tabla 2–25).

VARIABLE	RESPUESTA	FRECUENCIA	%
$OCNTY$ ($Pueblo\,de\,ocurrencia$)	$Bayamon$	672	11.2 %
	$Caguas$	416	6.93 %
	$Carolina$	152	2.53 %
	$Mayaguez$	433	7.22 %
	$Ponce$	839	13.98 %
	$San\,Juan$	1626	27.1 %
	$pob\,menor\,de\,100,000\,hab$	1862	31.03 %

Tabla 2–25: Pueblo de ocurrencia

En la Tabla 2–26 está suministrada la información referente a los resultados de embarazo precario, se evidencia que solamente un 0.42 % de los nacimientos en la muestra, es decir 25 bebés no sobrevivieron.

VARIABLE	RESPUESTA	FRECUENCIA	%
RF_PPOUTC ($Resultados\,de\,embarazo\,precario$)	SI	25	0.42 %
	NO	5975	99.58 %

Tabla 2–26: Resultados de embarazo precario

A continuación se tiene una tabla sobre la información de las variables cuantitativas.

VARIABLE	Mínimo	Q_1	Mediana	Media	Q_3	Máximo	SD
$MAGER$ ($Edad\,de\,la\,madre(años)$)	13	21	25	25.4	30	45	5.97
$FAGECOMB$ ($Edad\,del\,padre(años)$)	14	23	27	28.31	32	68	7.18
$DWGT$ ($Peso\,de\,la\,madre\,en\,la\,entrega(libras)$)	99	146	165	171.9	191	370	36.40
$DBWT$ ($Peso\,del\,recién\,nacido\,(gramos)$)	425	2807	3090	3086	3402	5245	501.83
$WTGAINC$ ($Aumento\,de\,peso\,madre(libras)$)	0	18	26	26.25	34	98	12.92

Tabla 2–27: Estadísticas descriptivas de variables cuantitativas. Nota: la variable DWGT: Peso de la madre en la entrega está truncada en 99 y 400.

La edad de los padres es mayor en promedio (mediana) que la edad de las madres. Se evidencia también que hay más padres con edades más altas que madres (Figura 2–1).

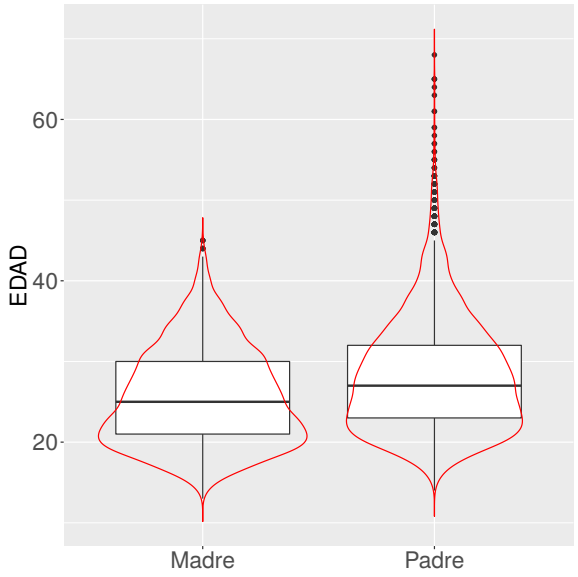


Figura 2–1: Edad de los padres (años)

En la Figura 2–2 se puede observar que la distribución de los pesos de la madre al momento de la entrega tiene un sesgo positivo y tiene presencia de valores atípicos. El peso promedio (mediana) es de 165 libras y aproximadamente el 25 % de los pesos de las madres al momento de la entrega es superior a 190 libras.

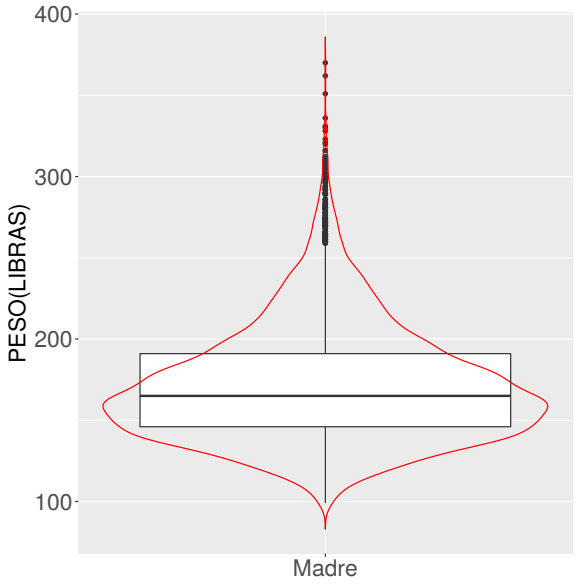


Figura 2–2: Peso (libras) de la madre al momento del nacimiento

De acuerdo con la Figura 2–3, el peso de los niños es mayor en promedio (mediana) que el de las niñas. La desviación estándar del peso de los niños es 516.7 gramos mientras que para niñas es 480.84 gramos. Note que en la Tabla 2–27 se muestra un aparente dato atípico correspondiente al peso de una bebe. No podemos hacer nada hasta que contemos con información extra para descartarlo como dato atípico. Esta es una de las ventajas de regresión a quantiles; ante presencia de datos atípicos, la regresión a quantiles es más robusta que el método de mínimos cuadrados.

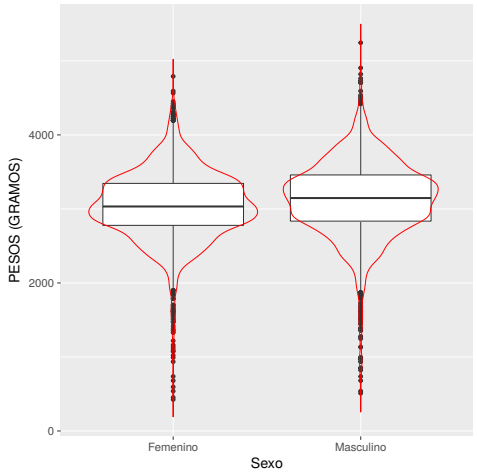


Figura 2–3: Peso (gramos) del bebé

La distribución del aumento de peso de la madre tiene un sesgo positivo (hay valores grandes menos frecuentes en comparación con los aumentos de peso menores a 25 libras). El aumento promedio (mediana) es de 26 libras. (Figura 2–4).

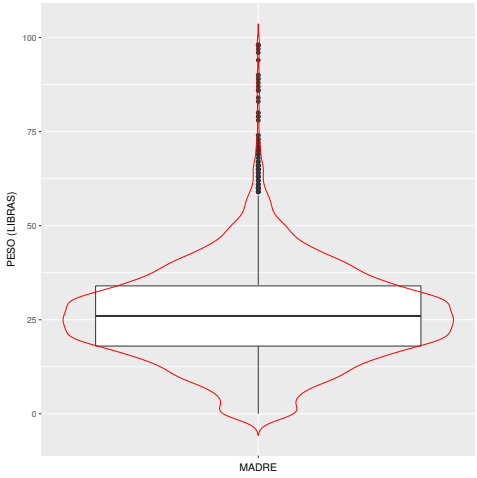


Figura 2–4: Aumento de peso (libras) de la madre

Capítulo 3
APLICACIÓN DE REGRESIÓN A CUANTILES

En este capítulo se quiere analizar la relación de las covariables seleccionadas preliminarmente en el Capítulo 2 con la distribución de los pesos de los recién nacidos usando regresión a cuantiles y selección de variables vía penalización SCAD.

Debido a que existen variables explicativas categóricas con muy baja frecuencia en una de las categorías (ejemplo: hipertensión crónica, SI: 1.27 %, NO: 98.73 %), al emplear el método de selección de variables SCAD cuantil agrupada discutida en la Sección 1.12 con todas las variables consideradas en el estudio se generó un error debido a la singularidad en la matriz de diseño particularmente para cuantiles en los extremos de la distribución. El método de regresión a cuantil estima el cuantil condicional dado un valor de variable explicativa. Así, cuando la variable explicativa es categórica con una frecuencia pequeña en un nivel esto creó problemas de singularidad. En regresión a la media este fenómeno conlleva estimaciones con errores estándar muy grandes para los coeficientes.

Por lo tanto, se eliminaron 13 variables con estas características, la variable con mayor % en una categoría que se eliminó de la lista fue (resultados de embarazo precario, SI: 0.42 %, NO: 99.58 %). La Tabla 3–1 lista las variables consideradas para ajustar los modelos de regresión a cuantiles junto con el método de penalización SCAD.

VARIABLE	SIGNIFICADO
(MBRACE)	raza de madre
(MAR)	estado civil madre
(MEDUC)	educación madre
(FBRACE)	raza padre
(LBO_REC)	número de nacimientos
(PREVIS_REC)	número de visitas prenatales
(WTGAINC)	aumento de peso de la madre
(DOB_YY)	año de nacimiento
(SEX)	sexo del bebé
(GESTREC10)	semanas de gestación
(OCNTY)	pueblo de ocurrencia
(DMETH_REC)	método de nacimiento
(MAGER)	edad madre
(DWGT)	peso de la madre en la entrega
(FAGECOMB)	edad del padre

Tabla 3–1: Variables para ajustar los modelos

Se adoptó el método de selección de variables SCAD agrupado debido a que se tiene varias variables categóricas que se pueden agrupar de forma natural (cada variable categórica puede tener varios niveles y se puede expresar a través de un grupo de variables indicadoras o *dummy*) y a sus propiedades mencionadas en el empleó Capítulo 1. Se usó el programa estadístico `R` para la selección de variables usando la función `cv.rq.group.pen` de la librería `rqPen` (ver Sección 1.12). Esta función usa el método de selección de variables SCAD agrupada en regresión a cuantiles. Se usa el criterio BIC para seleccionar el parámetro de regularización λ. La función cv.rq.group.pen automáticamente genera una secuencia de $\lambda's$ y selecciona el λ con menor valor BIC. El ajuste de cada modelo con la respectiva selección de variable se llevó cabo para cada cuantil de interés.

3.1. Selección de variables y ajuste de modelos de regresión a cuantil para ciertos valores de τ

Para cada $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$ se hace la selección de variables y se analiza que variables quedan elegidas con el método SCAD agrupado. Estos valores

de τ fueron seleccionados arbitrariamente para cubrir distintos puntos de la distribución del peso de los recién nacidos. También se hace la selección de variables con el método SCAD agrupado con mínimos cuadrados para tener un punto de comparación (esto también nos ayuda a constratar los resultados de la regresión a la media y regresión a la mediana para determinar, por ejemplo, asimetría en la distribución del peso de los recién nacidos). Es decir, se ajusta un modelo lineal múltiple tradicional con SCAD agrupado y las regresiones a cuantiles con SCAD agrupado para contrastar los resultados de los métodos. En regresión a cuantiles la estimación de λ óptimo se hace para cada τ. Este proceso arroja un conjunto de variables significativas asociadas al cuantil respectivo de manera independiente.

La Tabla 3–2 muestra los valores óptimos de λ junto con las variables significativas seleccionadas por el método SCAD agrupado marcadas con $*$ en la respectiva celda. Note que la variable edad de la madre, peso de la madre en la entrega, aumento de peso madre y semanas de gestación salieron seleccionadas en los distintos τ y por el método de mínimos Cuadrados. Las variable raza de madre, la raza del padre y educación de la madre no fueron seleccionadas en los distintos τ e inclusive usando el método SCAD agrupado con mínimos cuadrados. En el Apéndice A se pueden encontrar los comandos de R usados para ajustar los modelos. Todas las variables cuantitativas fueron centradas en sus medianas, edad de la madre (mediana 25 años), peso de entrega de la madre (mediana 165 libras), aumento de peso de la madre (mediana 26 libras) y edad del padre (mediana 27 años).

Variable	$\tau = 0.05$ $\lambda = 0.013$	$\tau = 0.25$ $\lambda = 0.007$	$\tau = 0.50$ $\lambda = 0.024$	$\tau = 0.75$ $\lambda = 0.003$	$\tau = 0.95$ $\lambda = 0.008$	MC $\lambda = 10.152$
MAGER(edad madre)	*	*	*	*	*	*
DWGT(peso de la madre en la entrega)	*	*	*	*	*	*
FAGECOMB(edad del padre)		*		*	*	
MBRACE(raza de madre)						
MAR(estado civil madre)						*
MEDUC(educación madre)						
FBRACE(raza padre)						
LBO_REC(número de nacimientos)						*
PREVIS_REC(número visitas prenatales)						*
WTGAINC(aumento de peso madre)	*	*	*	*	*	*
DOB_YY(año de ocurrencia)						
SEX(sexo del bebé)		*	*	*		*
GESTREC10(semanas de gestación)	*	*	*	*	*	*
OCNTY(pueblo de ocurrencia)						*
DMETH_REC(método de nacimiento)		*				

Tabla 3–2: Variables seleccionadas usando el método de regresión a cuantiles para un cuantil dado τ y el método de mínimos cuadrados con SCAD (MC)(*). El valor óptimo de λ fue seleccionado usando el criterio BIC.

La función de R que implementa regresión a cuantil y SCAD presenta funcionalidades limitadas en términos de la obtención de los errores estándar y los valores p de los coeficientes. Este problema no es exclusivo de regresión a cuantil. Otros métodos de regularización, áun en el contexto de mínimos cuadrados, todavía sufren de esta limitación en su implementación, debido especialmente a la dificultad de incorporar la incertidumbre al seleccionar λ en las estimaciones finales de los coeficientes. Por lo tanto, una vez se seleccionaron las variables para los diferentes cuantiles se procedió a ajustar un modelo de regresión a cuantiles para cada τ usando la función de R convencional `rq()`, la cual si proporciona los errores estándar y los p-valores. Los métodos para estimar los errores estándar fueron revisados en la Subsección 1.4.3 . En el Apéndice A se tienen los códigos de la información condensada en la Tabla 3–3.

A continuación se presenta algunas interpretaciones de los coeficientes cuantiles estimados en la Tabla 3–3. En los demás casos, las interpretaciones siguen una línea similar.

El peso estimado al nacer de un bebé nacido con bajo peso (en el cuantil 0.05) entre 20 a 36 semanas de gestación, cuya madre tiene 25 años, tuvo un aumento de peso de 26 libras con peso de entrega de 165 libras, es de 1510.43 gramos; en contraste, con el peso estimado al nacer de un bebé nacido con alto peso (en el cuantil 0.95) entre las mismas semanas de gestación mencionadas, cuya madre tiene 25 años, tuvo un aumento de peso de 26 libras con peso de entrega al nacimiento de 165 libras, es de 3506.51 gramos.

De acuerdo al modelo de regresión mínimos cuadrados, por cada año adicional, manteniendo las demás variables constantes, el peso de los bebés recién nacidos aumenta en promedio 1.88 gramos. Manteniendo las demás variables constantes,

los resultados de la regresión cuantil indican que el efecto de la edad de la madre tiene un impacto positivo para el peso del recién nacido a partir del primer cuartil. Para el tercer cuartil ($\tau = 0.75$) el peso del recién nacido aumenta 9.26 gramos por cada año adicional de la madre, manteniendo las demás variables fijas. Note que el coeficiente estimado asociado al aumento del peso de la madre en la regresión mínimos cuadrados y la regresión cuantil de la mediana, es aproximadamente el mismo (5.3). Esto significa que un cambio de una unidad en el aumento del peso de la madre está asociado con un aumento promedio (tanto media como mediana) en el peso del recién nacido de 5.3 gramos. El coeficiente para el peso de la madre en la entrega en el modelo de regresión mínimos cuadrados es 2.54 gramos, el cual es más bajo para dicho coeficiente en el modelo de regresión cuantil para $\tau = 0.95$. Esto sugiere que, si bien un aumento de una libra en el peso de entrega de la madre da lugar a un aumento promedio de 2.54 gramos, el aumento es más sustancial para la mayoría de recién nacidos con alto peso al nacer.

Los bebés recién nacidos que tuvieron entre (37 a 39 semanas de gestación) pesan aproximadamente 965 gramos más en comparación con los bebés prematuros (de 20 a 36 semanas de gestación) en el cuantil 0.05. Sin embargo, pesan cerca de 318 gramos más en el cuantil 0.75. Los niños son obviamente más grandes que las niñas en cerca de 94 gramos de acuerdo con la estimación de mínimos cuadrados, pero como se desprende de los resultados de la regresión cuantil, la disparidad es menor en el cuantil 0.25 (82.08) gramos y un poco mayor en el cuantil 0.50 (102.64) gramos.

VARIABLE	$\tau = 0.05$		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$		$\tau = 0.95$		$\boldsymbol{MC}(SCAD)$	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercepto	**1510.44**	**75.61**	**2335.14**	**25.79**	**2654.80**	**21.71**	**2996.6**1	**25.45**	**3506.51**	**33.45**	**2516.31**	**27.03**
Edad madre	−1.64	2.28	**6.39**	**1.51**	**7.09**	**1.63**	**9.26**	**1.77**	**7.43**	**2.65**	1.88	1.12
Peso de la madre en la entrega	**1.61**	**0.42**	**2.29**	**0.19**	**2.58**	**0.19**	**3.01**	**0.23**	**4.39**	**0.34**	**2.54**	**0.16**
Aumento de peso de la madre	**3.34**	**1.04**	**4.29**	**0.51**	**5.26**	**0.53**	**4.99**	**0.62**	**5.97**	**0.94**	**5.34**	**0.45**
Semanas de gestación (20 a 36 semanas)(referencia)												
Semanas de gestación (37 a 39 semanas)	**965.46**	**76.64**	**501.63**	**24.99**	**396.18**	**22.14**	**318.09**	**25.82**	**233.96**	**35.68**	**446.12**	**16.00**
Semanas de gestación (40 o más semanas)	**1054.44**	**81.55**	**635.22**	**30.08**	**553.00**	**25.38**	**501.29**	**29.82**	**461.11**	**44.32**	**611.11**	**20.13**
Sexo(Femenino)(referencia)												
Sexo(Masculino)			**82.08**	**12.63**	**102.64**	**13.69**	**95.45**	**14.97**			**94.44**	**11.34**
Método de nacimiento (Vaginal)(referencia)												
Método de nacimiento (Cesárea)			**−33.11**	**12.75**								
Edad del padre			−0.65	1.21			−1.89	1.46	0.79	2.34		
Estado civil (Casada)(referencia)											−23.42	12.76
Estado civil (No casada)											−23.42	12.76
Orden de nacimiento (1er nacimiento)(referencia)												
Orden de nacimiento (2do nacimiento)											**72.29**	**13.33**
Orden de nacimiento (3 er nacimiento)											**79.18**	**16.88**
Número de visitas prenatales (10 visitas o menos)(referencia)												
Número de visitas prenatales (11 a 12 visitas)											12.59	15.26
Número de visitas prenatales (13 o más visitas)											**58.98**	**14.46**
Pueblo de ocurrencia (Bayamón)(referencia)												
Pueblo de ocurrencia (Caguas)											−27.32	27.42
Pueblo de ocurrencia (Carolina)											66.93	39.69
Pueblo de ocurrencia (Mayaguez)											−2.98	27.16
Pueblo de ocurrencia (Ponce)											**64.45**	**22.74**
Pueblo de ocurrencia (San Juan)											**72.87**	**20.32**
Pueblo de ocurrencia (Pob menor a 100,000 hab)											**42.37**	**19.83**

Tabla 3–3: Ajustes modelos (en negrilla variables significativas con un nivel de significancia de 5 %). Estimaciones (SE= Error estándar) de los coeficientes cuantiles para $\tau = 0.05$, 0.25, 0.50, 0.75 *y* 0.95, y la regresión lineal con mínimos cuadrados (MC).

3.2. Ajuste de modelo con la unión de variables seleccionadas en los distintos τ

Una vez se hizo la selección de variables y ajuste de las regresiones a cuantiles con diferentes valores de τ, se ajustó un modelo con regresión a cuantil en R, con aquellas variables que fueron seleccionadas en al menos un τ de la Tabla 3–2. Las variables son las siguientes:

Variable	Significado
MAGERC	Edad de la madre
DWGTC	Peso de la madre en la entrega
WTGAINCC	Aumento de peso de la madre
GESTREC10	Semanas de gestación
DMETH_REC	Método de nacimiento
SEX	Sexo del bebé
FAGECOMB	Edad del padre

Tabla 3–4: Variables que fueron seleccionadas en al menos un τ

Este modelo incluye cuatro variables continuas todas ellas centradas en su mediana: edad de la madre (mediana 25 años), peso de entrega de la madre (mediana 165 libras), aumento de peso de la madre (mediana 26 libras) y edad del padre (mediana 27 años). Se incluyen tres variables categóricas: sexo del bebé, semanas de gestación y si el parto ocurrió por cesárea o no. La semanas de gestación fue categorizada como (20 semanas a 36 semanas de gestación) (categoría de referencia), (37 semanas a 39 semanas de gestación) y (40 o más semanas de gestación).

En la Tabla 3–5 se resume la información de los ajuste de los modelos de regresión a cuantil para $\tau = 0.05, 0.25.0.50, 0.75, 0.95$ realizada en apéndice A.

VARIABLE	$\tau = 0.05$		$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$		$\tau = 0.95$		*RMC*	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercepto	**1483.99**	**112.19**	**2352.83**	**42.62**	**2679.79**	**43.51**	**3048.21**	**47.76**	**3429.13**	**75.56**	**2645.34**	**34.86**
Edad madre	−1.73	3.11	**6.39**	**1.51**	**7.49**	**1.59**	**9.26**	**1.73**	**8.20**	**2.91**	**6.41**	**1.33**
Peso de la madre en la entrega	**1.58**	**0.43**	**2.29**	**0.19**	**2.59**	**0.19**	**3.01**	**0.22**	**4.14**	**0.34**	**2.64**	**0.17**
Aumento de peso de la madre	**3.65**	**1.03**	**4.29**	**0.51**	**5.22**	**0.54**	**5.01**	**0.61**	**5.79**	**0.84**	**4.97**	**0.45**
Semanas de gestación (20 a 36 semanas) (referencia)												
Semanas de gestación (37 a 39 semanas)	**943.18**	**87.13**	**501.63**	**24.99**	**390.84**	**22.23**	**317.50**	**25.64**	**207.25**	**32.53**	**450.31**	**16.05**
Semanas de gestación (40 o más semanas)	**1041.58**	**91.75**	**635.22**	**30.09**	**548.74**	**25.72**	**500.48**	**29.52**	**447.59**	**39.93**	**613.01**	**20.04**
Método de nacimiento (Vaginal) (referencia)												
Método de nacimiento (Cesárea)	**−74.00**	**26.15**	−33.11	12.75	−9.91	13.86	−0.49	15.18	7.50	21.98	−22.03	11.62
Sexo (Femenino) (referencia)												
Sexo (Masculino)	**51.50**	**25.32**	**82.08**	**12.59**	**104.06**	**13.24**	**95.20**	**14.75**	**114.59**	**21.41**	**90.34**	**11.41**
Edad del padre	1.57	2.53	−0.65	1.21	−0.55	1.35	−1.87	1.43	1.38	2.50	−0.91	1.09

Tabla 3–5: Ajustes modelos variables de la Tabla 3–4 (en negrilla variables significativas con un nivel de significancia de 5 %)

En la Tabla 3–5 se ajustaron modelos para valores de $\tau = 0.05, 0.25, 0.50, 0.75$ y 0.95 y en la Figura 3–3 están las estimaciones de los coeficientes de regresión a cuantiles con intervalos de confianza del 95 % para $\tau = 0.05, 0.10, 0.15, 0.20, ..., 0.95$ esto es para poder apreciar mejor estas estimaciones.

Los coeficientes de una regresión a cuantil para una covariable particular revelan el efecto de un cambio de unidad en la covariable sobre los cuantiles de la distribución de la variable respuesta. Se destaca este efecto de cambio mediante una vista gráfica en la Figura 3–3 que examina los coeficientes para distintos valores de τ. Para una covariable particular se gráfica los coeficientes y sus intervalos de confianza del 95 % (región gris), donde el efecto de la variable predictora $\hat{\beta}_\tau$ está en el eje y, y el valor del cuantil τ en el eje x, adicionalmente se añade la estimación de mínimos cuadrados (línea continua) y las líneas discontinuas son los límites de confinza del 95 %.

En el primer panel de la Figura 3–3, el (INTERCEPTO) puede interpretarse como el cuantil condicional estimado de la distribución del peso al nacer de una bebé cuya madre y padre tienen 25 y 27 años respectivamente, donde la madre tiene de 20 a 36 semanas de gestación, la madre pesa 165 libras, la madre tuvo un aumento de peso de 26 libras y su parto fue vaginal. Aproximadamente el 10 % de estos bebés pesan menos de 2000 gramos en comparación con los pesos de los bebés en el cuantil 0.90 que pesan aproximadamente 3300 gramos.

La Tabla 3–5 indica que, en el cuantil $\tau = 0.05$, la edad de la madre no es un factor muy determinante en la distribución del peso de los recién nacidos, ya que no es significativo con $\alpha = 5\,\%$. No obstante, en el panel (EDAD DE LA MADRE) de la Figura 3–3 a partir del cuantil $\tau = 0.10$ hasta aproximadamente el cuantil $\tau = 0.85$, la edad de la madre tiene un efecto positivo sobre el peso al nacer. El efecto es menor para los cuantiles superiores a $\tau = 0.85$ aproximadamente. En el

siguiente panel, (PESO DE LA MADRE EN LA ENTREGA) tiene un efecto positivo sobre el peso al nacer especialmente en los cuantiles superiores de la distribución del peso al nacer. En el panel (AUMENTO DE PESO DE LA MADRE) se observa un efecto positivo sobre el peso al nacer especialmente en los cuantiles centrales de la distribución del peso al nacer.

De acuerdo a la Figura 3–3, la diferencia entre el peso de los recién nacidos entre 37 a 39 semanas de gestación y los recién nacidos entre 20 a 36 semanas de gestación es bastante grande, especialmente en la cola inferior de la distribución del peso al nacer. La diferencia en el peso al nacer entre un bebé de 37 a 39 semanas de gestación y los recién nacidos entre 20 a 36 semanas de gestación sobre el cuantil $\tau = 0.05$ es aproximadamente 900 gramos como se evidencia en el panel (GESTACIÓN EN SEMANAS 37 A 39 SEMANAS). Un comportamiento similar pero más marcado se presenta al comparar los pesos de los recién nacidos entre 40 o más semanas de gestación y los recién nacidos entre 20 a 36 semanas de gestación; ahora, esta diferencia en el cuantil $\tau = 0.05$ aproximadamente 1000 gramos.

La diferencia entre el peso de los recién nacidos entre 37 a 39 semanas de gestación y los recién nacidos entre 20 a 36 semanas de gestación en el cuantil $\tau = 0.25$ es 501.63, podemos ver esta diferencia gráficamente en los dos primeros diagramas de caja de la Figura 3–1. Similarmente, la diferencia en el cuantil 25 % del peso de los recién nacidos entre las semanas 37 a 39 y 40 o más de gestación es estimado por el modelo en la Tabla 3–5, un número que también se puede apreciar en los últimos dos diagramas de caja.

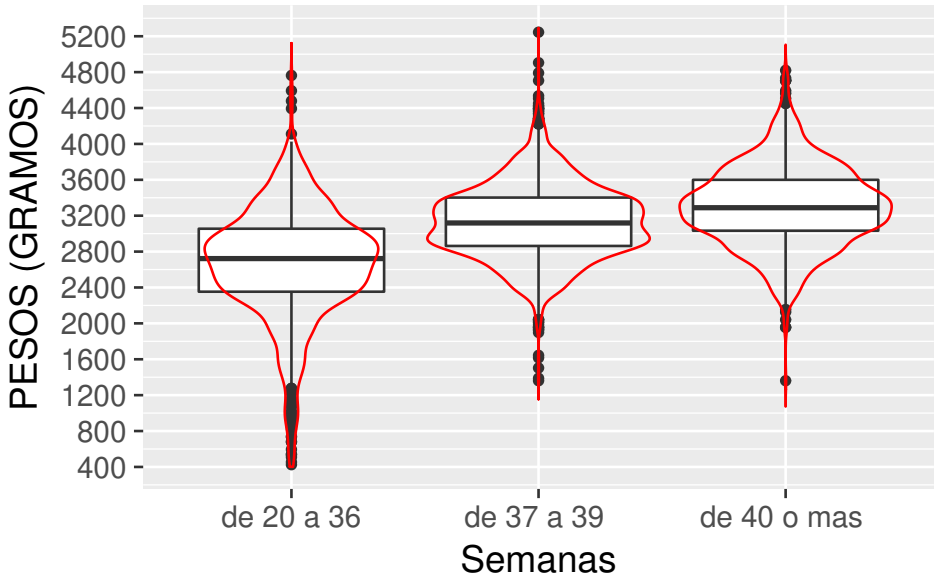


Figura 3–1: Peso (gramos) de los bebes en las distintas semanas de gestación. Note la tendencia, entre más semanas de gestación el bebé haya tenido, su peso es mayor

Para las madres que tuvieron un parto vaginal sus bebés tienen un peso mayor en cuantiles bajos que los bebés nacidos por cesárea. A partir del cuantil 0.40 el efecto de un parto vaginal no es significativo a un nivel $\alpha = 5\,\%$, ver panel (MÉTODO DE NACIMIENTO CESÁREA). En los diagramas de caja de la Figura 3–2 se puede observar para el cuantil 0.25 que los bebés que nacen por parto vaginal pesan un poco más en promedio que los que nacen por parto cesárea y en el cuantil 0.75 los que nacen por cesárea pesan en promedio un poco más que los que nacieron por parto vaginal.

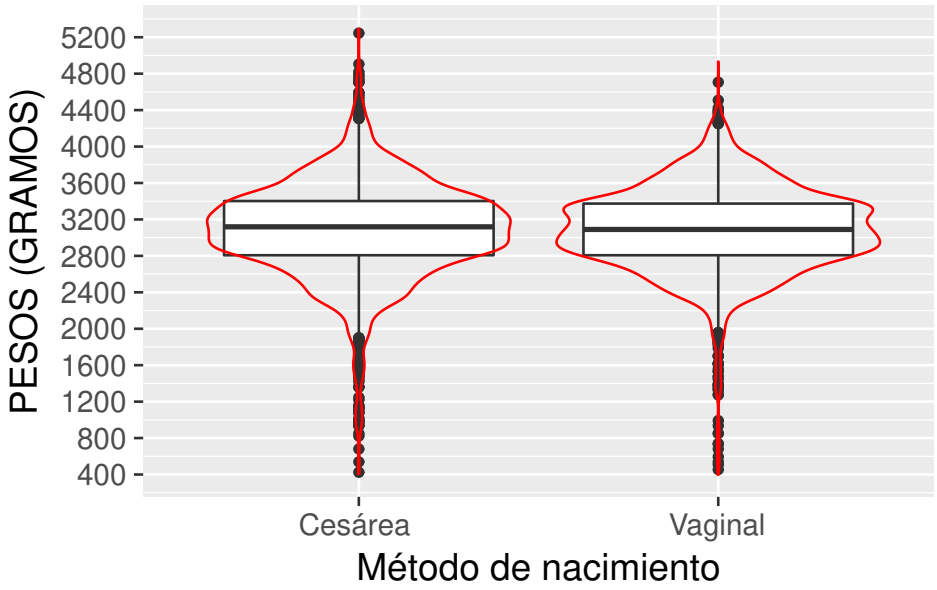


Figura 3–2: Peso (gramos) de los bebes mediante el método de nacimiento.

Como se puede observar en la Figura 3–3 los niños pesan en promedio 90 gramos más que las niñas, de acuerdo a la estimación de regresión de mínimos cuadrados. La regresión cuantil nos brinda más información. En los cuantiles inferiores ($\tau < 0.15$), la diferencia es mucho menor alrededor de 45 gramos. Para el cuantil $\tau = 0.60$ la diferencia es mayor que la estimada por mínimos cuadrados, aproximadamente 110 gramos. Es decir, la diferencia de peso entre niños y niñas cambia dependiendo si están bajos de peso o en sobrepeso.

Como se observa en el panel (EDAD DEL PADRE), la cual no es significativa con un nivel de significancia $\alpha = 0.05$, tanto en el ajuste de mínimos cuadrados como en el ajuste de regresión a cuantil para todos los cuantiles.

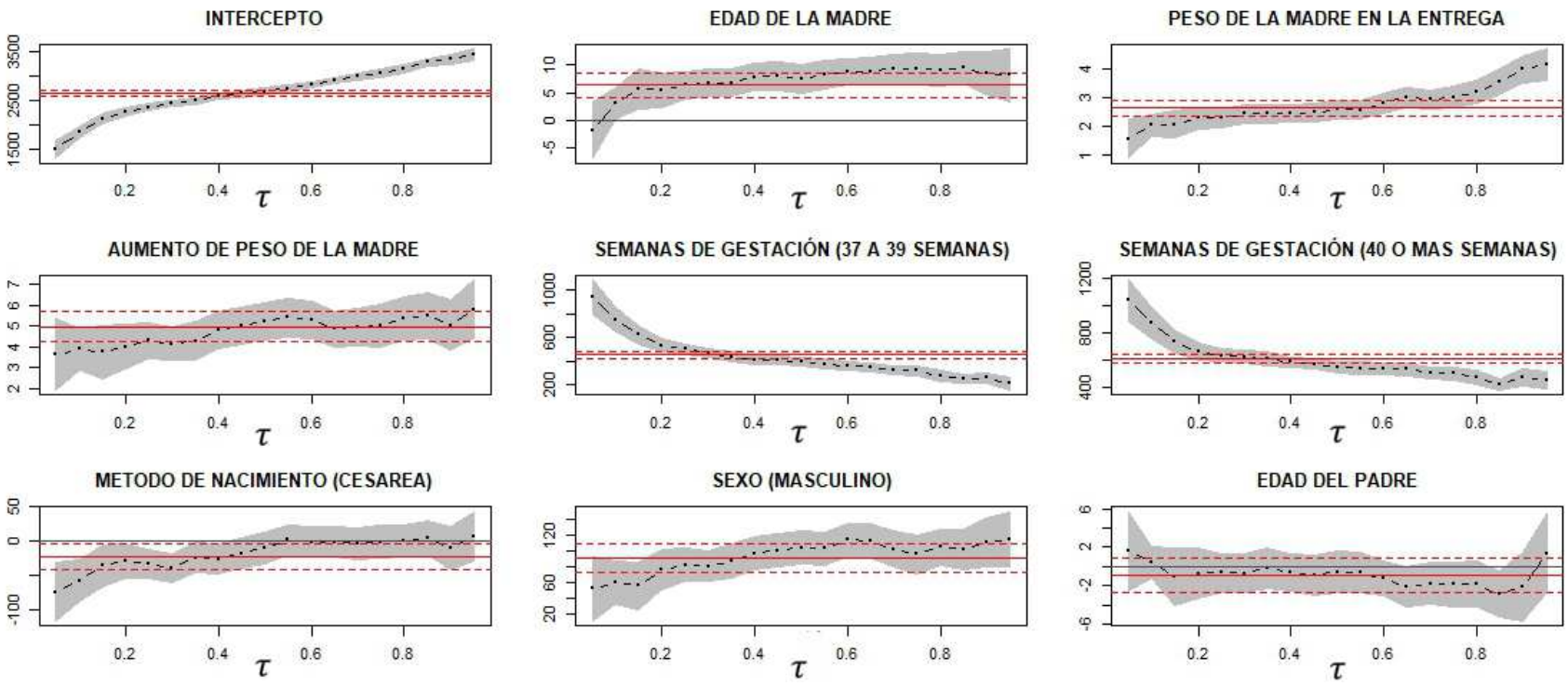


Figura 3–3: Estimaciones de los coeficientes de regresión a cuantiles con intervalos de confianza del 95 % para $\tau = 0.05, 0.10, 0.15, ..., 0.95$.

Capítulo 4
CONCLUSIONES Y TRABAJOS FUTUROS

4.1. Conclusiones generales

Las estimaciones por regresión a cuantiles permiten analizar el comportamiento del peso al nacer, particularmente en las colas de la distribución. Cabe destacar lo siguiente: la edad del padre no resultó ser relevante en el estudio, en cambio la edad de la madre fue relevante excepto en los cuantiles muy inferiores de la distribución del peso al nacer, esta conclusión es similar a la de Fallah et al. (2015). Un factor muy influyente son las semanas de gestación. Este resultado se dio también en el estudio de Fallah et al. (2015), debido a que entre más semanas de gestación se tenga, el bebé va a tener un mayor peso y la regresión a cuantil en este estudio revela que la diferencia es mucho mayor en los cuantiles inferiores y va decreciendo gradualmente a medida que aumentan los cuantiles. Con referencia al aumento de peso de la madre se obtuvo un efecto positivo sobre el peso al nacer especialmente en los cuantiles centrales de la distribución del peso al nacer, pero en el trabajo de Abrevaya (2001) tuvo un efecto mayor en los cuantiles inferiores. Se confirmó mediante este estudio que en promedio los niños pesan más que las niñas y esto es más evidente alrededor del cuantil 0.60: aproximadamente 110 gramos más a favor de los niños, siendo más evidente para los recién nacidos alrededor del cuantil 0.60. Este resultado no se hubiera podido obtener por medio de la regresión mínimos cuadrados, ya que solo se estimaría la diferencia promedio. Cabe notar que la diferencia en el peso de los recién nacidos cambia dependiendo del cuantil de referencia. Entre los bebés que

nacen con bajo peso, los de parto por cesárea presentan menos peso que aquellos que nacieron por parto vaginal.

El uso de regresión a cuantiles conjuntamente con SCAD agrupada sugiere las ventajas en términos de interpretación con respecto a regresión a cuantiles es un método que ofrece ventajas en términos de interpretación con respecto a la regresión tradicional, tal como se ilustró con la aplicación en este trabajo. Cuando regresión a cuantiles se combina con SCAD, esto permite el uso de regresión a cuantiles en una serie de problemas donde existen bastantes variables predictoras. Si bien es cierto que en este trabajo se redujo la cantidad de variables predictoras a un subconjunto relativamente pequeño, la metodología ilustrada en este proyecto sirve para escenarios más complejos. Además, SCAD funciona bien cuando existen variables explicativas agrupadas como es el caso de variables categóricas. En nuestra aplicación, la variable semanas de gestación tiene tres niveles. En los métodos de selección tradicionales o no agrupados suele pasar que estos tres niveles se seleccionan por separado (dos variables dummy), así que una de ellas puede salir seleccionada. En SCAD selecciona el grupo completo con los tres niveles o no se selecciona la variable.

4.2. Trabajos futuros

Existen varias formas en las que este trabajo se puede extender. Primero, usar otros métodos de selección de variables como por ejemplo LASSO agrupado (Yuan and Lin, 2007) y compararlo con SCAD. A pesar de que en este trabajo se implementó un muestreo por razones computacionales, la comparación de los métodos se podría llevar a cabo con más datos y mejores recursos de cómputo.

Segundo, considerar la inclusión de otras variables que no se tuvieron en cuenta en este estudio, como por ejemplo, nacimientos múltiples, año de nacimiento y otras razas aparte de blanca y negra.

Por último, como se puede apreciar en la Figura 3–3, los coeficientes cuantiles $\beta(\tau)$ parecen tener formas funcionales conocidas (polinómicas, cuadráticas, logarítmicas, exponencial entre otras). Por ejemplo, las estimaciones de los coeficientes cuantiles beta(tau) del peso de la madre en el nacimiento (entrega) en función de tau se pueden resumir con una línea recta. De manera similar, para el aumento del peso de la madre.

El proceso de ajustar regresiones cuantiles no paramétricas para cada valor de τ para obtener cada gráfica en esta figura puede ser tedioso para una escala fina de τ. Se puede asumir que $\beta(\tau)$ sigue una forma funcional específica para hacer el proceso de estimación más eficiente sin sacrificar las propiedades estadísticas de los estimadores (Frumento and Bottai, 2016). Por ejemplo, la representación en el panel EDAD DE LA MADRE de la Figura 3–3 se asemeja a la forma de la gráfica de la función logarítmica y también hay una similitud con la gráfica de la función exponencial en el panel SEMANAS DE GESTACIÓN. Vale la pena recordar que las gráficas en la Figura 3–3 se crean estimando el modelo para cada cuantil τ. Así, si se decide estimar el modelo para una secuencia fina de $\tau's$, y quizás bajo un escenario complejo con bastantes variables y observaciones, el ajuste del modelo puede ser computacionalmente pesado. Así, algunos autores están optando por asumir formas funcionales para estos coeficientes, simplificando así la estimación para cada cuantil.

APÉNDICES

Apéndice A
PROGRAMAS DE R PARA AJUSTAR

comparación de regresion cuantil y lineal

```
set.seed(20)
x <- runif(100)
# Intercepto:
b_0 <- 7
# Pendiente:
b_1 <- 0.08
# error aleatorio normal con varianza no constante:
set.seed (20)
e <- rnorm(100,mean = 0, sd = 0.04*x)
# variable respuesta:
y <- b_0 + b_1*x + e
data_set <- data.frame(x,y)
# ajuste modelo minimos cuadrados:
fit.ls <- lm(y ~ x, data = data_set)
summary(fit.ls)

##
## Call:
## lm(formula = y ~ x, data = data_set)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.056762 -0.009902  0.000103  0.007880  0.059239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.994597   0.003732 1874.29   <2e-16 ***
## x           0.096781   0.006713   14.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02016 on 98 degrees of freedom
## Multiple R-squared:  0.6796,Adjusted R-squared:  0.6763
## F-statistic: 207.9 on 1 and 98 DF,  p-value: < 2.2e-16

# ajuste modelo quantil:
library(quantreg)

fit.qr <- rq(y ~ x, data=data_set, tau = 1:9/10)
summary(fit.qr,se="ker")

##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.1
##
## Coefficients:
```

```
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99596    0.00363 1927.93018    0.00000
## x              0.04400    0.01112    3.95841    0.00014
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.2
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99823    0.00355 1970.84564    0.00000
## x              0.05810    0.01021    5.68885    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.3
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99846    0.00369 1898.43483    0.00000
## x              0.06624    0.01034    6.40496    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.4
##
## Coefficients:
```

```
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99897    0.00359 1948.58193    0.00000
## x              0.07534    0.01036    7.27413    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99931    0.00393 1778.76285    0.00000
## x              0.08359    0.01142    7.31910    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.6
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99948    0.00392 1787.70760    0.00000
## x              0.09124    0.01179    7.73829    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.7
##
## Coefficients:
```

```
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99915    0.00440 1590.06832    0.00000
## x              0.11082    0.01164    9.51677    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.8
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99855    0.00467 1499.46515    0.00000
## x              0.12816    0.01143   11.21606    0.00000
##
## Call: rq(formula = y ~ x, tau = 1:9/10, data = data_set)
##
## tau: [1] 0.9
##
## Coefficients:
##             Value      Std. Error t value    Pr(>|t|)
## (Intercept)    6.99930    0.00383 1826.28354    0.00000
## x              0.13387    0.00947   14.13022    0.00000

library(ggplot2)


#lineas de ajuste
ggplot(data_set, aes(x,y)) + geom_point() + geom_quantile(quantiles = 1:9/10,
   aes(colour = as.factor(..quantile..)),size=1)+
```

```
  geom_smooth(method="lm",se = FALSE,aes(x = x,colour = as.factor("RMC")),
size=1)+labs( colour="Diferentes\nCuantiles")
```

```
#otra grafica
plot(summary(fit.qr), parm="x",main = "")
```

Gráfica de penalidad SCAD y umbral

penalidad SCAD

```
p1 <- ggplot(data = data.frame(x = 0), mapping = aes(x = x))


penalidadSCAD <- function(b){
  a=3.7
  lambda=0.85
  ifelse((0 <= abs(b)) & (abs(b)<=lambda),lambda*abs(b),
         ifelse((lambda <= abs(b)) & (abs(b)<= a*lambda),
                -(b^2-2*a*lambda*abs(b)+lambda^2)/(2*(a-1)),
                (a+1)*lambda^2 / 2) )


}
lambda=0.85
p1 + stat_function(fun = penalidadSCAD ) +xlim(-5,5)+
  labs(title="",x=expression(beta), y = "Penalidad")
```

gráfica del umbral

```
library(ggplot2)
p1 <- ggplot(data = data.frame(x = 0), mapping = aes(x = x))
```

```
Penalised_least_square_estimatorsSCAD <- function(z){
  a=3.7
  lambda=2
  ifelse( (abs(z)<=2*lambda), ifelse(abs(z)<lambda,0,sign(z)*(abs(z)-lambda)),
          ifelse((2*lambda < abs(z)) & (abs(z)<= a*lambda),
                 ((a-1)*z-sign(z)*a*lambda )/(a-2),
                 z ))

}
lambda=2
LS <- function(x) x
p1 + stat_function(fun = Penalised_least_square_estimatorsSCAD)+
stat_function(fun = LS,geom = "line",linetype = 2 )+xlim(-10,10)+
labs(title="",x="z", y = expression(hat(beta)))
```

Gráficas de variables cuantitativas

```
library(readr)
BASEFINAL6000_28VAR <- read.csv("~/Proyecto/Base de datos/BASEFINAL6000_28VAR")
#EDAD PADRES
PADRES=c(rep("Madre",6000),rep("Padre",6000))
EDAD=c(BASEFINAL6000_28VAR$MAGER,BASEFINAL6000_28VAR$FAGECOMB)
padresedad=data.frame(PADRES,EDAD)

library(ggplot2)
GRAF <- ggplot(padresedad, aes(x = PADRES, y =EDAD))+geom_boxplot()+
geom_violin(trim = FALSE,color="red",
            alpha=0)+theme(axis.title.x = element_blank())
```

```
GRAF
```

```
#PESO MADRE EN LIBRAS
Madre=c(rep("Madre",6000))
PESOLIBRAS=c(BASEFINAL6000_28VAR$DWGT)
pesomadre=data.frame(Madre,PESOLIBRAS)


library(ggplot2)
GRAF1 <- ggplot(pesomadre,aes(x = Madre, y = PESOLIBRAS))+geom_boxplot()+
geom_violin(trim = FALSE,color="red",alpha=0)+ylab("PESO (LIBRAS)")
GRAF1
```

```
#peso gramos del bebe
Peso.F<-(as.vector(BASEFINAL6000_28VAR[BASEFINAL6000_28VAR$SEX=="F",
                 which (colnames(BASEFINAL6000_28VAR)=="DBWT")]))
Peso.M<-(as.vector(BASEFINAL6000_28VAR[BASEFINAL6000_28VAR$SEX=="M",
                 which (colnames(BASEFINAL6000_28VAR)=="DBWT")]))
Pesos <- c(Peso.F, Peso.M)
Sexo <- c(rep("Femenino",length(Peso.F)),rep("Masculino",length(Peso.M)))


data=data.frame(Sexo,Pesos)


library(ggplot2)
ggplot(data, aes(x = Sexo, y = Pesos))+
geom_boxplot()+geom_violin(trim = FALSE,color="red",alpha=0)+
ylab("PESOS (GRAMOS)")+ theme_grey(base_size = 15)
```

```
#aumento de peso en libras de la madres
Madre=c(rep("Madre",6000))
PESOL=c(BASEFINAL6000_28VAR$WTGAINC)
pesomadre=data.frame(Madre,PESOL)


library(ggplot2)
GRAF2 <- ggplot(pesomadre,aes(x = Madre, y = PESOL))+geom_boxplot()+
geom_violin(trim = FALSE,color="red",alpha=0)+ylab("PESO (LIBRAS)")+
  theme( axis.text.x=element_blank())+xlab("MADRE")
GRAF2
```

Estimación en R

- paquete `quantreg` (contribuido por Koenker)
- La sintaxis de la función rq ()

  `library(quantreg)`

  `rq(formula, tau=.5, data, method=\ br", ...)`

Estimación in SAS

- paquete: SAS/STAT PROC QUANTREG
- Sintaxis básica

  `PROC QUANTREG DATA` =sas-data-set $<$ *options* $>$;

  `BY` variables;

  `CLASS` variables;

  `MODEL` response = independents $<$ /*options* $>$;

  `RUN`;
- Para especificar el nivel de cuantil:

Use la opción QUANTILE en la declaración MODELO

`MODELY =X / QUANTILE = <number list | ALL>;`

- Para especificar el algoritmo:

  Use la opción `ALGORITHM` en el `PROC`

  declaración `QUANTREG`

```
library(readr)
BASEFINAL6000_28VAR <- read.csv("~/Proyecto/Base de datos/BASEFINAL6000_28VAR")
library(readr)
v1=as.factor(BASEFINAL6000_28VAR$MRCNTY)
v2=as.factor(BASEFINAL6000_28VAR$MBRACE)
v3=as.factor(BASEFINAL6000_28VAR$MAR)
v4=as.factor(BASEFINAL6000_28VAR$MEDUC)
v5=as.factor(BASEFINAL6000_28VAR$FBRACE)
v6=as.factor(BASEFINAL6000_28VAR$LBO_REC)
v7=as.factor(BASEFINAL6000_28VAR$PRECARE_REC)
v8=as.factor(BASEFINAL6000_28VAR$PREVIS_REC)
v9=as.factor(BASEFINAL6000_28VAR$WTGAINC)
v10=as.factor(BASEFINAL6000_28VAR$RF_DIAB)
v11=as.factor(BASEFINAL6000_28VAR$RF_GEST)
v12=as.factor(BASEFINAL6000_28VAR$RF_PHYP)
v13=as.factor(BASEFINAL6000_28VAR$RF_GHYP)
v14=as.factor(BASEFINAL6000_28VAR$RF_PPOUTC)
v15=as.factor(BASEFINAL6000_28VAR$URF_DIAB)
v16=as.factor(BASEFINAL6000_28VAR$URF_CHYPER)
v17=as.factor(BASEFINAL6000_28VAR$URF_PHYPER)
v18=as.factor(BASEFINAL6000_28VAR$DOB_YY)
```

```
v19=as.factor(BASEFINAL6000_28VAR$SEX)
v20=as.factor(BASEFINAL6000_28VAR$GESTREC10)
v21=as.factor(BASEFINAL6000_28VAR$OCNTY)
v22=as.factor(BASEFINAL6000_28VAR$RF_PPTERM)
v23=as.factor(BASEFINAL6000_28VAR$DMETH_REC)
v24=as.factor(BASEFINAL6000_28VAR$APGAR5)
v25=as.numeric(BASEFINAL6000_28VAR$MAGER)
V26=as.numeric(BASEFINAL6000_28VAR$FAGECOMB)
V27=as.numeric(BASEFINAL6000_28VAR$DWGT)
y=as.numeric(BASEFINAL6000_28VAR$DBWT)

xfactors<- model.matrix(y~v2+v3+v4+v5+v6+v8+v18+v19+v20+v21+v23)[,-1]
x <- as.matrix(data.frame((BASEFINAL6000_28VAR$MAGER)-25,
as.numeric(BASEFINAL6000_28VAR$DWGT)-165,(BASEFINAL6000_28VAR$FAGECOMB)-27,
                          (BASEFINAL6000_28VAR$WTGAINC)-26,xfactors))
my.groups <- c (1,2,3,4,rep(5,length( names(table(v2)))-1),
rep(6,length( names(table(v3)))-1),rep(7,length( names(table(v4)))-1),
rep(8,length( names(table(v5)))-1),rep(9,length( names(table(v6)))-1),
rep(10,length( names(table(v8)))-1),rep(11,length( names(table(v18)))-1),
rep(12,length( names(table(v19)))-1), rep(13,length( names(table(v20)))-1),
rep(14,length( names(table(v21)))-1), rep(15,length( names(table(v23)))-1))

v=c("MAGER.C","DWGT.C","FAGECOMB.C","WTGAINC.C","v2","v3","v4","v5","v6","v8",
    "v18","v19","v20","v21","v23")

#####################################################################
#t 0.05
```

```r
library(rqPen)

cv_model.05 <- cv.rq.group.pen(x, y, groups=my.groups, tau = 0.05,
                    penalty = "SCAD",criteria = "BIC",eps = 0.001)

cv_model.05

##
## Coefficients:
## (Intercept)          x1          x2          x3          x4          x5
## 1467.037057   -1.792201    1.611917    0.000000    3.237406    0.000000
##          x6          x7          x8          x9         x10         x11
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x12         x13         x14         x15         x16         x17
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x18         x19         x20         x21         x22         x23
##    0.000000 1009.853867 1107.587002    0.000000    0.000000    0.000000
##         x24         x25         x26         x27
##    0.000000    0.000000    0.000000    0.000000
##
## Cross Validation (or BIC) Results
##          lambda       BIC
## 1   1.000000000  14.69707
## 2   0.932603347  14.69707
## 3   0.869749003  14.69707
## 4   0.811130831  14.69707
## 5   0.756463328  14.69707
## 6   0.705480231  14.69707
```

```
## 7   0.657933225  14.69707
## 8   0.613590727  14.69707
## 9   0.572236766  14.69707
## 10  0.533669923  14.69707
## 11  0.497702356  14.69707
## 12  0.464158883  14.69707
## 13  0.432876128  14.69707
## 14  0.403701726  14.69779
## 15  0.376493581  14.69779
## 16  0.351119173  14.70570
## 17  0.327454916  14.70581
## 18  0.305385551  14.70261
## 19  0.284803587  14.70284
## 20  0.265608778  14.70258
## 21  0.247707636  14.70260
## 22  0.231012970  14.70280
## 23  0.215443469  14.72427
## 24  0.200923300  14.72481
## 25  0.187381742  14.72446
## 26  0.174752840  14.72443
## 27  0.162975083  14.72445
## 28  0.151991108  14.72416
## 29  0.141747416  14.72473
## 30  0.132194115  14.72465
## 31  0.123284674  14.72410
## 32  0.114975700  14.72470
## 33  0.107226722  14.72476
```

```
## 34  0.100000000  14.72407
## 35  0.093260335  14.72423
## 36  0.086974900  14.72437
## 37  0.081113083  14.72468
## 38  0.075646333  14.72415
## 39  0.070548023  14.72480
## 40  0.065793322  14.72444
## 41  0.061359073  14.72408
## 42  0.057223677  14.72421
## 43  0.053366992  14.72406
## 44  0.049770236  14.72407
## 45  0.046415888  14.72413
## 46  0.043287613  14.72406
## 47  0.040370173  14.72411
## 48  0.037649358  14.72466
## 49  0.035111917  14.72439
## 50  0.032745492  14.72424
## 51  0.030538555  14.72387
## 52  0.028480359  14.72411
## 53  0.026560878  14.72099
## 54  0.024770764  14.72133
## 55  0.023101297  14.72353
## 56  0.021544347  14.72111
## 57  0.020092330  14.64685
## 58  0.018738174  14.64597
## 59  0.017475284  14.64674
## 60  0.016297508  14.64511
```

```
## 61  0.015199111  14.64558
## 62  0.014174742  14.64690
## 63  0.013219411  14.64508
## 64  0.012328467  14.64512
## 65  0.011497570  20.06976
## 66  0.010722672  20.03157
## 67  0.010000000  19.92353
## 68  0.009326033  19.96334
## 69  0.008697490  20.07895
## 70  0.008111308  19.98836
## 71  0.007564633  19.58646
## 72  0.007054802  19.95438
## 73  0.006579332  19.51029
## 74  0.006135907  19.22382
## 75  0.005722368  69.73460
## 76  0.005336699  19.20292
## 77  0.004977024  98.93960
## 78  0.004641589  88.20352
## 79  0.004328761 105.65546
## 80  0.004037017 103.46748
## 81  0.003764936 110.72431
## 82  0.003511192 113.49827
## 83  0.003274549 108.62169
## 84  0.003053856 115.61365
## 85  0.002848036 124.25060
## 86  0.002656088 116.72196
## 87  0.002477076 110.38124
```

```
## 88  0.002310130 111.76962
## 89  0.002154435 112.72128
## 90  0.002009233 129.02790
## 91  0.001873817 122.99912
## 92  0.001747528 119.34618
## 93  0.001629751 119.82910
## 94  0.001519911 114.15582
## 95  0.001417474 119.28644
## 96  0.001321941 111.52495
## 97  0.001232847 118.19350
## 98  0.001149757 109.20911
## 99  0.001072267 113.71186
## 100 0.001000000 128.95086
```

```
cbind(names(coefficients(cv_model.05)[-1]),
      as.numeric(coefficients(cv_model.05))[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))
```

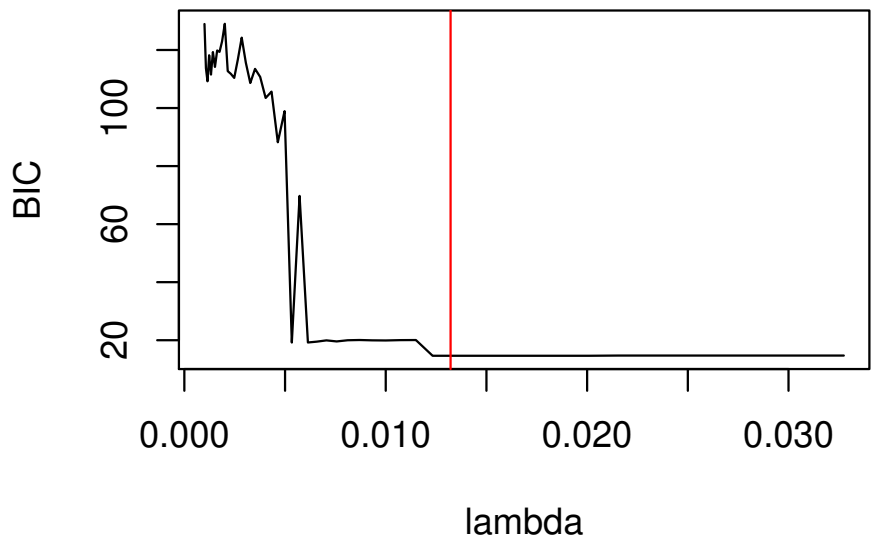
```
##                                 my.groups Variable
##  [1,] "x1"  "-1.79220093491442" "1"       "MAGER.C"
##  [2,] "x2"  "1.61191716207497"  "2"       "DWGT.C"
##  [3,] "x3"  "0"                 "3"       "FAGECOMB.C"
##  [4,] "x4"  "3.23740552467634"  "4"       "WTGAINC.C"
##  [5,] "x5"  "0"                 "5"       "v2"
##  [6,] "x6"  "0"                 "6"       "v3"
##  [7,] "x7"  "0"                 "7"       "v4"
##  [8,] "x8"  "0"                 "7"       "v4"
##  [9,] "x9"  "0"                 "7"       "v4"
```

```
## [10,] "x10" "0"                  "7"       "v4"
## [11,] "x11" "0"                  "8"       "v5"
## [12,] "x12" "0"                  "9"       "v6"
## [13,] "x13" "0"                  "9"       "v6"
## [14,] "x14" "0"                  "10"      "v8"
## [15,] "x15" "0"                  "10"      "v8"
## [16,] "x16" "0"                  "11"      "v18"
## [17,] "x17" "0"                  "11"      "v18"
## [18,] "x18" "0"                  "12"      "v19"
## [19,] "x19" "1009.8538669227"    "13"      "v20"
## [20,] "x20" "1107.58700226686"   "13"      "v20"
## [21,] "x21" "0"                  "14"      "v21"
## [22,] "x22" "0"                  "14"      "v21"
## [23,] "x23" "0"                  "14"      "v21"
## [24,] "x24" "0"                  "14"      "v21"
## [25,] "x25" "0"                  "14"      "v21"
## [26,] "x26" "0"                  "14"      "v21"
## [27,] "x27" "0"                  "15"      "v23"
```

```
#LAMBDA OPTIMO
cv_model.05$lambda.min
```

```
## [1] 0.01321941
```

```
#GRAFICA LAMBDA VS BIC
m05<- as.matrix(cv_model.05$cv)
plot(m05[50:100,1],m05[50:100,2],type="l",xlab = "lambda",ylab = "BIC")
abline(v=cv_model.05$lambda.min,col="red")
```




```
#t .25
library(rqPen)

#tao 0.25
cv_model.25 <- cv.rq.group.pen(x, y, groups=my.groups, tau = 0.25,penalty = "SCAD"
                               criteria = "BIC",eps = 0.001)
cv_model.25

##
## Coefficients:
## (Intercept)          x1          x2          x3          x4          x5
## 2336.236202    6.513249    2.296545   -0.692545    4.295066    0.000000
##          x6          x7          x8          x9         x10         x11
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x12         x13         x14         x15         x16         x17
```

```
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x18         x19         x20         x21         x22         x23
##   83.055564  500.213573  634.082173    0.000000    0.000000    0.000000
##         x24         x25         x26         x27
##    0.000000    0.000000    0.000000  -34.084319
##
## Cross Validation (or BIC) Results
##          lambda      BIC
## 1   6.000000000 14.12128
## 2   5.595620081 14.12128
## 3   5.218494016 14.12128
## 4   4.866784985 14.12128
## 5   4.538779965 14.12128
## 6   4.232881386 14.12128
## 7   3.947599348 14.12128
## 8   3.681544364 14.12128
## 9   3.433420596 14.12128
## 10  3.202019539 14.12128
## 11  2.986214139 14.12128
## 12  2.784953300 14.12128
## 13  2.597256769 14.12128
## 14  2.422210355 14.12200
## 15  2.258961484 14.11255
## 16  2.106715041 14.11121
## 17  1.964729498 14.10805
## 18  1.832313305 14.10696
## 19  1.708821521 14.10346
```

```
## 20  1.593652670 14.10181
## 21  1.486245814 14.10168
## 22  1.386077820 14.10036
## 23  1.292660814 14.09723
## 24  1.205539802 14.09750
## 25  1.124290454 14.09760
## 26  1.048517040 14.09675
## 27  0.977850501 14.09559
## 28  0.911946650 14.09603
## 29  0.850484498 14.09557
## 30  0.793164689 14.09523
## 31  0.739708044 14.09541
## 32  0.689854197 14.09541
## 33  0.643360333 14.09541
## 34  0.600000000 14.09541
## 35  0.559562008 14.09541
## 36  0.521849402 14.08099
## 37  0.486678498 14.08110
## 38  0.453877997 14.08099
## 39  0.423288139 14.08119
## 40  0.394759935 14.08114
## 41  0.368154436 14.08118
## 42  0.343342060 14.08112
## 43  0.320201954 14.08115
## 44  0.298621414 14.08111
## 45  0.278495330 14.08109
## 46  0.259725677 14.08113
```

```
## 47  0.242221036 14.08070
## 48  0.225896148 14.08120
## 49  0.210671504 14.08070
## 50  0.196472950 14.08096
## 51  0.183231331 14.08087
## 52  0.170882152 14.08097
## 53  0.159365267 14.08117
## 54  0.148624581 14.08118
## 55  0.138607782 14.08103
## 56  0.129266081 14.08110
## 57  0.120553980 14.08087
## 58  0.112429045 14.08097
## 59  0.104851704 14.08112
## 60  0.097785050 14.08105
## 61  0.091194665 14.07903
## 62  0.085048450 14.07951
## 63  0.079316469 14.07912
## 64  0.073970804 14.07922
## 65  0.068985420 14.07923
## 66  0.064336033 14.07923
## 67  0.060000000 14.07912
## 68  0.055956201 14.07932
## 69  0.052184940 14.07901
## 70  0.048667850 14.07947
## 71  0.045387800 14.07905
## 72  0.042328814 14.07918
## 73  0.039475993 14.07918
```

```
## 74  0.036815444 14.07904
## 75  0.034334206 14.07911
## 76  0.032020195 14.07912
## 77  0.029862141 14.07900
## 78  0.027849533 14.01848
## 79  0.025972568 14.01865
## 80  0.024222104 14.01834
## 81  0.022589615 14.01847
## 82  0.021067150 14.01831
## 83  0.019647295 14.01841
## 84  0.018323133 14.01838
## 85  0.017088215 14.01534
## 86  0.015936527 14.01531
## 87  0.014862458 14.01683
## 88  0.013860778 14.01669
## 89  0.012926608 14.01658
## 90  0.012055398 14.01661
## 91  0.011242905 14.01654
## 92  0.010485170 14.01651
## 93  0.009778505 14.01648
## 94  0.009119466 14.01639
## 95  0.008504845 14.01666
## 96  0.007931647 14.01205
## 97  0.007397080 14.01153
## 98  0.006898542 14.01164
## 99  0.006433603 14.01657
## 100 0.006000000 14.01981
```

```
#resumen variables
cbind(names(coefficients(cv_model.25)[-1]),
      as.numeric(coefficients(cv_model.25))[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))
```

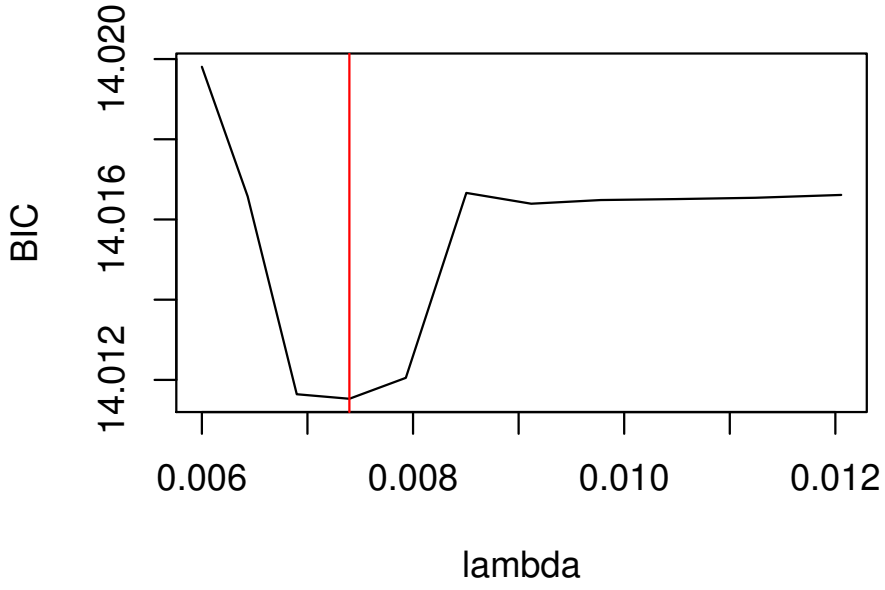
```
##                                  my.groups Variable
##  [1,] "x1"  "6.5132488090314"    "1"       "MAGER.C"
##  [2,] "x2"  "2.29654541975658"   "2"       "DWGT.C"
##  [3,] "x3"  "-0.692545035090024" "3"       "FAGECOMB.C"
##  [4,] "x4"  "4.29506572256253"   "4"       "WTGAINC.C"
##  [5,] "x5"  "0"                  "5"       "v2"
##  [6,] "x6"  "0"                  "6"       "v3"
##  [7,] "x7"  "0"                  "7"       "v4"
##  [8,] "x8"  "0"                  "7"       "v4"
##  [9,] "x9"  "0"                  "7"       "v4"
## [10,] "x10" "0"                  "7"       "v4"
## [11,] "x11" "0"                  "8"       "v5"
## [12,] "x12" "0"                  "9"       "v6"
## [13,] "x13" "0"                  "9"       "v6"
## [14,] "x14" "0"                  "10"      "v8"
## [15,] "x15" "0"                  "10"      "v8"
## [16,] "x16" "0"                  "11"      "v18"
## [17,] "x17" "0"                  "11"      "v18"
## [18,] "x18" "83.0555637693792"   "12"      "v19"
## [19,] "x19" "500.213573034494"   "13"      "v20"
## [20,] "x20" "634.082172601737"   "13"      "v20"
## [21,] "x21" "0"                  "14"      "v21"
```

```
## [22,] "x22" "0"                    "14"      "v21"
## [23,] "x23" "0"                    "14"      "v21"
## [24,] "x24" "0"                    "14"      "v21"
## [25,] "x25" "0"                    "14"      "v21"
## [26,] "x26" "0"                    "14"      "v21"
## [27,] "x27" "-34.0843185117762"    "15"      "v23"
```

```
#lambda optimo
cv_model.25$lambda.min
```

```
## [1] 0.00739708
```

```
#GRAFICA LAMBDA VS BIC
m25<- as.matrix(cv_model.25$cv)
plot(m25[90:100,1],m25[90:100,2],type="l",xlab = "lambda" ,ylab = "BIC")
abline(v=cv_model.25$lambda.min,col="red")
```




```
library(rqPen)

#tao 0.50
cv_model.50 <- cv.rq.group.pen(x, y, groups=my.groups, tau = 0.5,
                               penalty = "SCAD",criteria = "BIC" ,eps = 0.000001)
cv_model.50

##
## Coefficients:
## (Intercept)          x1          x2          x3          x4          x5
## 2654.631971    7.144019    2.589022    0.000000    5.253154    0.000000
##          x6          x7          x8          x9         x10         x11
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x12         x13         x14         x15         x16         x17
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
```

```
##         x18         x19         x20         x21         x22         x23
##  103.247770  396.195217  553.333577    0.000000    0.000000    0.000000
##         x24         x25         x26         x27
##    0.000000    0.000000    0.000000    0.000000
##
## Cross Validation (or BIC) Results
##           lambda      BIC
## 1   1.587096e+04 13.94670
## 2   1.380375e+04 13.94670
## 3   1.200580e+04 13.94670
## 4   1.044203e+04 13.94670
## 5   9.081945e+03 13.94670
## 6   7.899013e+03 13.94670
## 7   6.870159e+03 13.94670
## 8   5.975314e+03 13.94670
## 9   5.197023e+03 13.94670
## 10  4.520106e+03 13.94670
## 11  3.931357e+03 13.94670
## 12  3.419294e+03 13.94670
## 13  2.973928e+03 13.94670
## 14  2.586571e+03 13.94670
## 15  2.249667e+03 13.94670
## 16  1.956646e+03 13.94670
## 17  1.701791e+03 13.94670
## 18  1.480131e+03 13.94670
## 19  1.287342e+03 13.94670
## 20  1.119665e+03 13.94670
```

```
## 21  9.738272e+02 13.94670
## 22  8.469853e+02 13.94670
## 23  7.366646e+02 13.94670
## 24  6.407133e+02 13.94670
## 25  5.572597e+02 13.94670
## 26  4.846761e+02 13.94670
## 27  4.215466e+02 13.94670
## 28  3.666397e+02 13.94670
## 29  3.188845e+02 13.94670
## 30  2.773495e+02 13.94670
## 31  2.412244e+02 13.94670
## 32  2.098047e+02 13.94670
## 33  1.824774e+02 13.94670
## 34  1.587096e+02 13.94670
## 35  1.380375e+02 13.94670
## 36  1.200580e+02 13.94670
## 37  1.044203e+02 13.94670
## 38  9.081945e+01 13.94670
## 39  7.899013e+01 13.94670
## 40  6.870159e+01 13.94670
## 41  5.975314e+01 13.94670
## 42  5.197023e+01 13.94670
## 43  4.520106e+01 13.94670
## 44  3.931357e+01 13.94670
## 45  3.419294e+01 13.94670
## 46  2.973928e+01 13.94670
## 47  2.586571e+01 13.94670
```

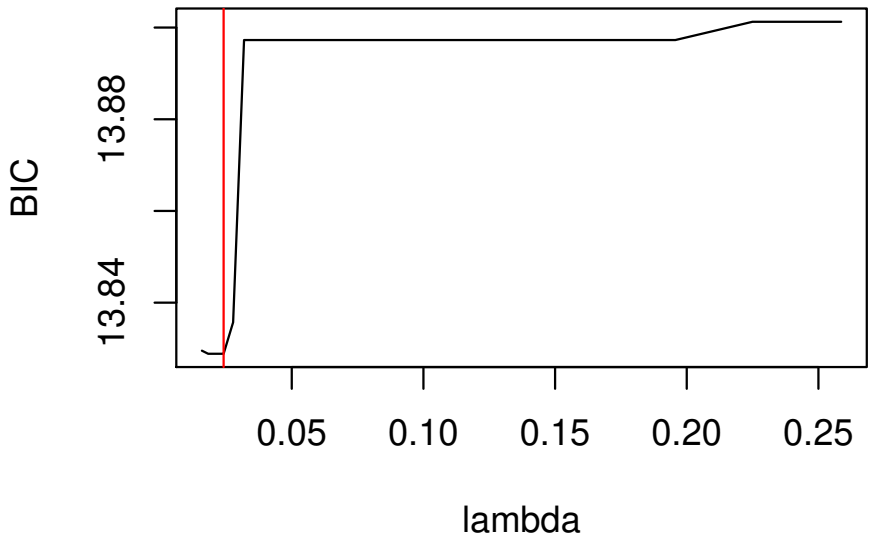
```
## 48  2.249667e+01 13.94670
## 49  1.956646e+01 13.94670
## 50  1.701791e+01 13.94670
## 51  1.480131e+01 13.94670
## 52  1.287342e+01 13.94670
## 53  1.119665e+01 13.94670
## 54  9.738272e+00 13.94670
## 55  8.469853e+00 13.94670
## 56  7.366646e+00 13.94670
## 57  6.407133e+00 13.94670
## 58  5.572597e+00 13.94670
## 59  4.846761e+00 13.94670
## 60  4.215466e+00 13.94670
## 61  3.666397e+00 13.94670
## 62  3.188845e+00 13.94225
## 63  2.773495e+00 13.93417
## 64  2.412244e+00 13.93013
## 65  2.098047e+00 13.92590
## 66  1.824774e+00 13.92360
## 67  1.587096e+00 13.92036
## 68  1.380375e+00 13.91662
## 69  1.200580e+00 13.91491
## 70  1.044203e+00 13.91438
## 71  9.081945e-01 13.91430
## 72  7.899013e-01 13.91391
## 73  6.870159e-01 13.90126
## 74  5.975314e-01 13.90126
```

```
## 75  5.197023e-01 13.90126
## 76  4.520106e-01 13.90126
## 77  3.931357e-01 13.90126
## 78  3.419294e-01 13.90126
## 79  2.973928e-01 13.90126
## 80  2.586571e-01 13.90126
## 81  2.249667e-01 13.90126
## 82  1.956646e-01 13.89727
## 83  1.701791e-01 13.89727
## 84  1.480131e-01 13.89727
## 85  1.287342e-01 13.89727
## 86  1.119665e-01 13.89727
## 87  9.738272e-02 13.89727
## 88  8.469853e-02 13.89727
## 89  7.366646e-02 13.89727
## 90  6.407133e-02 13.89727
## 91  5.572597e-02 13.89727
## 92  4.846761e-02 13.89727
## 93  4.215466e-02 13.89727
## 94  3.666397e-02 13.89727
## 95  3.188845e-02 13.89727
## 96  2.773495e-02 13.83571
## 97  2.412244e-02 13.82885
## 98  2.098047e-02 13.82885
## 99  1.824774e-02 13.82885
## 100 1.587096e-02 13.82953
```

```r
#resumen variables
cbind(names(coefficients(cv_model.50)[-1]),
      as.numeric(coefficients(cv_model.50))[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))
```

```
##                                my.groups Variable
##  [1,] "x1"  "7.14401919800299" "1"       "MAGER.C"
##  [2,] "x2"  "2.58902223107692" "2"       "DWGT.C"
##  [3,] "x3"  "0"                "3"       "FAGECOMB.C"
##  [4,] "x4"  "5.25315379225056" "4"       "WTGAINC.C"
##  [5,] "x5"  "0"                "5"       "v2"
##  [6,] "x6"  "0"                "6"       "v3"
##  [7,] "x7"  "0"                "7"       "v4"
##  [8,] "x8"  "0"                "7"       "v4"
##  [9,] "x9"  "0"                "7"       "v4"
## [10,] "x10" "0"                "7"       "v4"
## [11,] "x11" "0"                "8"       "v5"
## [12,] "x12" "0"                "9"       "v6"
## [13,] "x13" "0"                "9"       "v6"
## [14,] "x14" "0"                "10"      "v8"
## [15,] "x15" "0"                "10"      "v8"
## [16,] "x16" "0"                "11"      "v18"
## [17,] "x17" "0"                "11"      "v18"
## [18,] "x18" "103.247769878614" "12"      "v19"
## [19,] "x19" "396.195216646594" "13"      "v20"
## [20,] "x20" "553.333576756342" "13"      "v20"
## [21,] "x21" "0"                "14"      "v21"
```

```
## [22,] "x22" "0"                "14"      "v21"
## [23,] "x23" "0"                "14"      "v21"
## [24,] "x24" "0"                "14"      "v21"
## [25,] "x25" "0"                "14"      "v21"
## [26,] "x26" "0"                "14"      "v21"
## [27,] "x27" "0"                "15"      "v23"

#lambda optimo
cv_model.50$lambda.min

## [1] 0.02412244

#GRAFICA LAMBDA VS BIC
m<- as.matrix(cv_model.50$cv)
plot(m[80:100,1],m[80:100,2],type="l",xlab = "lambda" ,ylab = "BIC")
abline(v=cv_model.50$lambda.min,col="red")
```




```
library(rqPen)

#tao 0.75
cv_model.75 <- cv.rq.group.pen(x, y, groups=my.groups, tau = 0.75,
                               penalty = "SCAD",criteria = "BIC" ,eps = 0.00001)

cv_model.75

##
## Coefficients:
## (Intercept)          x1          x2          x3          x4          x5
## 3001.765011    9.384676    3.009614   -2.147639    5.049761    0.000000
##          x6          x7          x8          x9         x10         x11
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
##         x12         x13         x14         x15         x16         x17
##    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
```

```
##         x18         x19         x20         x21         x22         x23
##   96.878489  311.232464  495.084673    0.000000    0.000000    0.000000
##         x24         x25         x26         x27
##    0.000000    0.000000    0.000000    0.000000
##
## Cross Validation (or BIC) Results
##           lambda      BIC
## 1   6.000000e+00 14.13069
## 2   5.341291e+00 14.13069
## 3   4.754897e+00 14.13069
## 4   4.232881e+00 14.13069
## 5   3.768175e+00 14.13069
## 6   3.354486e+00 14.13069
## 7   2.986214e+00 14.13069
## 8   2.658373e+00 14.11573
## 9   2.366524e+00 14.11121
## 10  2.106715e+00 14.10800
## 11  1.875430e+00 14.10565
## 12  1.669536e+00 14.10349
## 13  1.486246e+00 14.09974
## 14  1.323078e+00 14.09891
## 15  1.177824e+00 14.09767
## 16  1.048517e+00 14.09638
## 17  9.334057e-01 14.09692
## 18  8.309318e-01 14.09692
## 19  7.397080e-01 14.09692
## 20  6.584993e-01 14.09729
```

```
## 21  5.862060e-01 14.08371
## 22  5.218494e-01 14.08385
## 23  4.645582e-01 14.08404
## 24  4.135567e-01 14.08386
## 25  3.681544e-01 14.08420
## 26  3.277366e-01 14.08409
## 27  2.917561e-01 14.08402
## 28  2.597257e-01 14.08349
## 29  2.312117e-01 14.08411
## 30  2.058282e-01 14.07828
## 31  1.832313e-01 14.07830
## 32  1.631153e-01 14.07831
## 33  1.452077e-01 14.07831
## 34  1.292661e-01 14.07827
## 35  1.150746e-01 14.07829
## 36  1.024412e-01 14.07828
## 37  9.119466e-02 14.07828
## 38  8.118287e-02 14.07828
## 39  7.227021e-02 14.07829
## 40  6.433603e-02 14.07829
## 41  5.727291e-02 14.07830
## 42  5.098521e-02 14.07828
## 43  4.538780e-02 14.07898
## 44  4.040490e-02 14.07954
## 45  3.596906e-02 14.07967
## 46  3.202020e-02 14.07959
## 47  2.850486e-02 14.07964
```

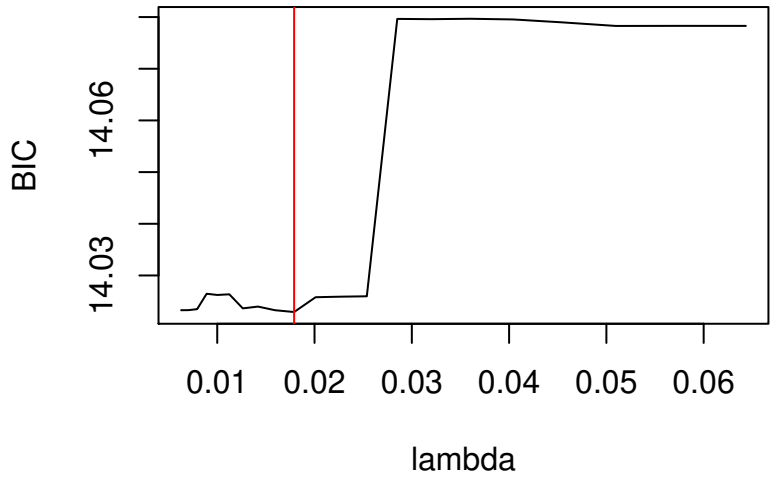
```
## 48  2.537546e-02 14.02597
## 49  2.258961e-02 14.02589
## 50  2.010962e-02 14.02579
## 51  1.790188e-02 14.02293
## 52  1.593653e-02 14.02328
## 53  1.418694e-02 14.02398
## 54  1.262942e-02 14.02364
## 55  1.124290e-02 14.02635
## 56  1.000860e-02 14.02623
## 57  8.909810e-03 14.02646
## 58  7.931647e-03 14.02348
## 59  7.060872e-03 14.02330
## 60  6.285695e-03 14.02328
## 61  5.595620e-03 14.02346
## 62  4.981305e-03 14.02350
## 63  4.434433e-03 14.02362
## 64  3.947599e-03 14.02302
## 65  3.514212e-03 14.02496
## 66  3.128405e-03 14.02553
## 67  2.784953e-03 14.02530
## 68  2.479207e-03 14.02494
## 69  2.207028e-03 14.02840
## 70  1.964729e-03 14.02777
## 71  1.749032e-03 14.02829
## 72  1.557015e-03 14.02857
## 73  1.386078e-03 14.02843
## 74  1.233907e-03 14.02841
```

```
## 75  1.098443e-03 14.02852
## 76  9.778505e-04 14.02888
## 77  8.704973e-04 14.02908
## 78  7.749298e-04 14.02819
## 79  6.898542e-04 14.02885
## 80  6.141186e-04 14.02899
## 81  5.466977e-04 14.02876
## 82  4.866785e-04 14.02981
## 83  4.332485e-04 14.02892
## 84  3.856844e-04 14.02879
## 85  3.433421e-04 14.02994
## 86  3.056483e-04 14.02982
## 87  2.720927e-04 14.02975
## 88  2.422210e-04 14.02974
## 89  2.156288e-04 14.02946
## 90  1.919560e-04 14.02870
## 91  1.708822e-04 14.02872
## 92  1.521219e-04 14.02836
## 93  1.354212e-04 14.02908
## 94  1.205540e-04 14.03003
## 95  1.073190e-04 14.03015
## 96  9.553697e-05 14.03012
## 97  8.504845e-05 14.03015
## 98  7.571141e-05 14.03014
## 99  6.739944e-05 14.03015
## 100 6.000000e-05 14.03027
```

```r
#resumen variables
cbind(names(coefficients(cv_model.75)[-1]),
      as.numeric(coefficients(cv_model.75))[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))
```

```
##                                my.groups Variable
##  [1,] "x1"  "9.3846756303186"   "1"       "MAGER.C"
##  [2,] "x2"  "3.00961393437427"  "2"       "DWGT.C"
##  [3,] "x3"  "-2.14763893308636" "3"       "FAGECOMB.C"
##  [4,] "x4"  "5.04976069990881"  "4"       "WTGAINC.C"
##  [5,] "x5"  "0"                 "5"       "v2"
##  [6,] "x6"  "0"                 "6"       "v3"
##  [7,] "x7"  "0"                 "7"       "v4"
##  [8,] "x8"  "0"                 "7"       "v4"
##  [9,] "x9"  "0"                 "7"       "v4"
## [10,] "x10" "0"                 "7"       "v4"
## [11,] "x11" "0"                 "8"       "v5"
## [12,] "x12" "0"                 "9"       "v6"
## [13,] "x13" "0"                 "9"       "v6"
## [14,] "x14" "0"                 "10"      "v8"
## [15,] "x15" "0"                 "10"      "v8"
## [16,] "x16" "0"                 "11"      "v18"
## [17,] "x17" "0"                 "11"      "v18"
## [18,] "x18" "96.8784888391023"  "12"      "v19"
## [19,] "x19" "311.232463573977"  "13"      "v20"
## [20,] "x20" "495.084673263797"  "13"      "v20"
## [21,] "x21" "0"                 "14"      "v21"
```

```
## [22,] "x22" "0"                    "14"      "v21"
## [23,] "x23" "0"                    "14"      "v21"
## [24,] "x24" "0"                    "14"      "v21"
## [25,] "x25" "0"                    "14"      "v21"
## [26,] "x26" "0"                    "14"      "v21"
## [27,] "x27" "0"                    "15"      "v23"

#lambda optimo
cv_model.75$lambda.min

## [1] 0.01790188

#GRAFICA LAMBDA VS BIC
m75<- as.matrix(cv_model.75$cv)
plot(m75[40:60,1],m75[40:60,2],type="l",xlab = "lambda" ,ylab = "BIC")
abline(v=cv_model.75$lambda.min,col="red")
```



```
library(rqPen)

#tao 0.95
library(rqPen)
cv_model.95 <- cv.rq.group.pen(x, y, groups=my.groups, tau = 0.95,penalty = "SCAD"
                               criteria = "BIC" ,eps = 0.0001)

cv_model.95

##
## Coefficients:
##  (Intercept)           x1           x2           x3           x4
## 3523.6461164    7.4386746    4.2695747    0.8843724    5.9447561
##           x5           x6           x7           x8           x9
##    0.0000000    0.0000000    0.0000000    0.0000000    0.0000000
##          x10          x11          x12          x13          x14
##    0.0000000    0.0000000    0.0000000    0.0000000    0.0000000
##          x15          x16          x17          x18          x19
##    0.0000000    0.0000000    0.0000000    0.0000000  215.5672783
##          x20          x21          x22          x23          x24
##  437.5076412    0.0000000    0.0000000    0.0000000    0.0000000
##          x25          x26          x27
##    0.0000000    0.0000000    0.0000000
##
## Cross Validation (or BIC) Results
##           lambda       BIC
## 1   2.0000000000  14.68287
## 2   1.8223255122  14.68287
```

```
## 3   1.6604351363  14.68287
## 4   1.5129266551  14.68287
## 5   1.3785224209  14.68287
## 6   1.2560582884  14.68031
## 7   1.1444735319  14.66697
## 8   1.0428016576  14.66090
## 9   0.9501620324  14.64268
## 10  0.8657522562  14.64269
## 11  0.7888412119  14.64273
## 12  0.7187627328  14.64283
## 13  0.6549098326  14.64164
## 14  0.5967294481  14.64261
## 15  0.5437176485  14.64274
## 16  0.4954152712  14.64283
## 17  0.4514039439  14.64211
## 18  0.4113024617  14.64137
## 19  0.3747634846  14.64238
## 20  0.3414705295  14.64226
## 21  0.3111352288  14.64281
## 22  0.2834948326  14.64285
## 23  0.2583099330  14.64265
## 24  0.2353623905  14.64195
## 25  0.2144534444  14.62414
## 26  0.1954019915  14.62422
## 27  0.1780430171  14.62423
## 28  0.1622261662  14.62412
## 29  0.1478144407  14.62412
```

```
## 30  0.1346830132  14.62420
## 31  0.1227181455  14.62401
## 32  0.1118162037  14.62411
## 33  0.1018827603  14.62421
## 34  0.0928317767  14.62420
## 35  0.0845848575  14.62421
## 36  0.0770705719  14.62352
## 37  0.0702238347  14.62363
## 38  0.0639853428  14.62352
## 39  0.0583010613  14.62340
## 40  0.0531217557  14.62111
## 41  0.0484025653  14.62073
## 42  0.0441026148  14.62091
## 43  0.0401846601  14.62117
## 44  0.0366147656  14.62182
## 45  0.0333620107  14.62032
## 46  0.0303982217  14.62094
## 47  0.0276977274  14.62100
## 48  0.0252371377  14.62030
## 49  0.0229951399  14.62085
## 50  0.0209523151  14.62090
## 51  0.0190909691  14.62042
## 52  0.0173949801  14.62052
## 53  0.0158496580  14.62084
## 54  0.0144416180  14.62079
## 55  0.0131586645  14.62069
## 56  0.0119896850  14.62096
```

```
## 57  0.0109245544  14.62058
## 58  0.0099540471  14.57983
## 59  0.0090697570  14.57900
## 60  0.0082640248  14.57861
## 61  0.0075298716  14.57951
## 62  0.0068609386  14.58385
## 63  0.0062514317  14.58314
## 64  0.0056960717  14.58291
## 65  0.0051900484  14.58233
## 66  0.0047289788  14.58257
## 67  0.0043088694  99.52908
## 68  0.0039260813 113.77995
## 69  0.0035772991 118.63439
## 70  0.0032595017 106.95072
## 71  0.0029699365 111.78549
## 72  0.0027060955 101.92217
## 73  0.0024656935 117.32864
## 74  0.0022466481 104.80734
## 75  0.0020470620 117.21284
## 76  0.0018652067 122.54852
## 77  0.0016995069 114.47991
## 78  0.0015485274 108.35929
## 79  0.0014109605 124.00821
## 80  0.0012856146 127.46474
## 81  0.0011714042 117.62713
## 82  0.0010673398 117.53174
## 83  0.0009725203 120.30645
```

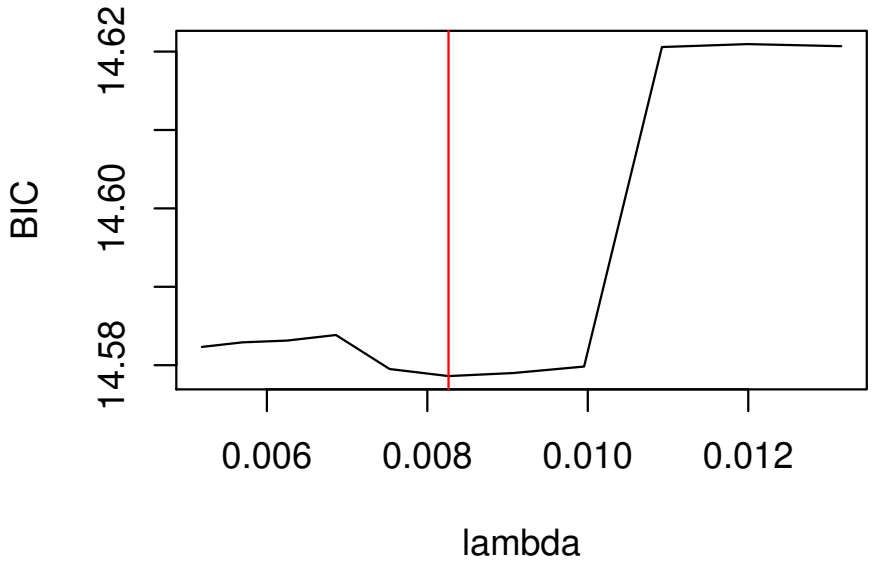
```
## 84  0.0008861243 120.19782
## 85  0.0008074035 117.30054
## 86  0.0007356760 114.76186
## 87  0.0006703205 120.48643
## 88  0.0006107711 118.60998
## 89  0.0005565119 119.40111
## 90  0.0005070729 129.92979
## 91  0.0004620259 122.01807
## 92  0.0004209808 124.59583
## 93  0.0003835821 115.66328
## 94  0.0003495057 116.87706
## 95  0.0003184566 122.10893
## 96  0.0002901658 114.67180
## 97  0.0002643882 116.62976
## 98  0.0002409007 119.44705
## 99  0.0002194998 118.41950
## 100 0.0002000000 117.97247

#resumen variables
cbind(names(coefficients(cv_model.95)[-1]),
      as.numeric(coefficients(cv_model.95))[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))

##                                my.groups Variable
##  [1,] "x1"  "7.43867459687005" "1"       "MAGER.C"
##  [2,] "x2"  "4.26957472283701" "2"       "DWGT.C"
##  [3,] "x3"  "0.88437236397668" "3"       "FAGECOMB.C"
##  [4,] "x4"  "5.94475613211683" "4"       "WTGAINC.C"
```

```
##  [5,] "x5"  "0"                 "5"       "v2"
##  [6,] "x6"  "0"                 "6"       "v3"
##  [7,] "x7"  "0"                 "7"       "v4"
##  [8,] "x8"  "0"                 "7"       "v4"
##  [9,] "x9"  "0"                 "7"       "v4"
## [10,] "x10" "0"                 "7"       "v4"
## [11,] "x11" "0"                 "8"       "v5"
## [12,] "x12" "0"                 "9"       "v6"
## [13,] "x13" "0"                 "9"       "v6"
## [14,] "x14" "0"                 "10"      "v8"
## [15,] "x15" "0"                 "10"      "v8"
## [16,] "x16" "0"                 "11"      "v18"
## [17,] "x17" "0"                 "11"      "v18"
## [18,] "x18" "0"                 "12"      "v19"
## [19,] "x19" "215.567278302899" "13"      "v20"
## [20,] "x20" "437.507641183013" "13"      "v20"
## [21,] "x21" "0"                 "14"      "v21"
## [22,] "x22" "0"                 "14"      "v21"
## [23,] "x23" "0"                 "14"      "v21"
## [24,] "x24" "0"                 "14"      "v21"
## [25,] "x25" "0"                 "14"      "v21"
## [26,] "x26" "0"                 "14"      "v21"
## [27,] "x27" "0"                 "15"      "v23"
```

```
#lambda optimo
cv_model.95$lambda.min
```

```
## [1] 0.008264025
```

```
#GRAFICA LAMBDA VS BIC
m95<- as.matrix(cv_model.95$cv)
plot(m95[55:65,1],m95[55:65,2],type="l",xlab = "lambda" ,ylab = "BIC")
abline(v=cv_model.95$lambda.min,col="red")
```



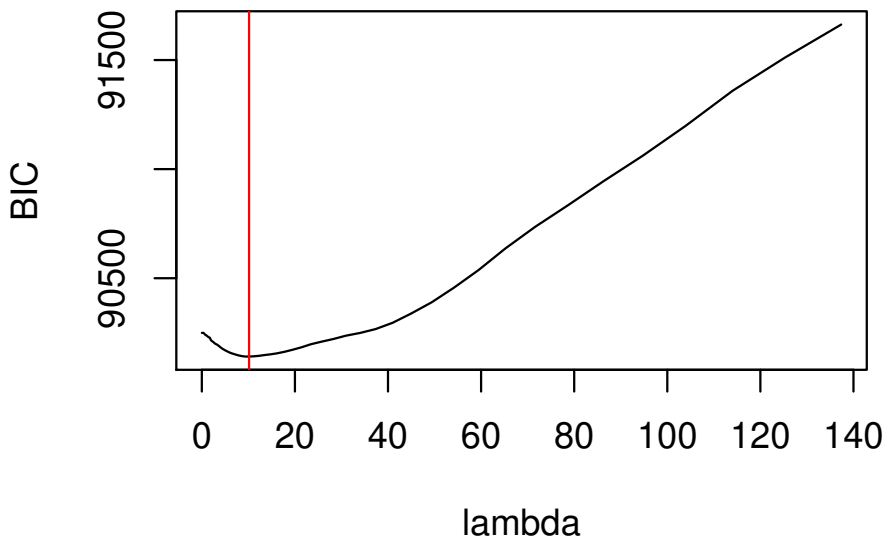
SCAD GROUP MINIMOS CUADRADOS

```
# minimos cuadtrados con scad group
library(grpreg)

fit <- grpreg(x, y, group=my.groups, penalty = c("grSCAD"))
mod = select(fit, "BIC")
cbind(names(mod$beta)[-1],as.numeric(mod$beta)[-1],
      my.groups, Variable=rep(v,as.numeric(table(my.groups))))

##
```

```
##  [1,] "X.BASEFINAL6000_28VAR.MAGER....25"          "1.63540776524025"
##  [2,] "as.numeric.BASEFINAL6000_28VAR.DWGT....165" "2.6364368985257"
##  [3,] "X.BASEFINAL6000_28VAR.FAGECOMB....27"       "0"
##  [4,] "X.BASEFINAL6000_28VAR.WTGAINC....26"        "5.21063117144757"
##  [5,] "v22"                                        "0"
##  [6,] "v32"                                        "-7.25301800635254"
##  [7,] "v42"                                        "0"
##  [8,] "v43"                                        "0"
##  [9,] "v44"                                        "0"
## [10,] "v45"                                        "0"
## [11,] "v52"                                        "0"
## [12,] "v62"                                        "53.8215130568317"
## [13,] "v63"                                        "59.3070148190952"
## [14,] "v82"                                        "5.82624597914011"
## [15,] "v83"                                        "29.1642141603667"
## [16,] "v182010"                                    "0"
## [17,] "v182011"                                    "0"
## [18,] "v19M"                                       "92.3578862067172"
## [19,] "v202"                                       "449.734498007991"
## [20,] "v203"                                       "617.385746544241"
## [21,] "v2125"                                      "-7.14104921423724"
## [22,] "v2131"                                      "14.5382487173867"
## [23,] "v2197"                                      "-0.124434582841369"
## [24,] "v21113"                                     "15.8910828589116"
## [25,] "v21127"                                     "18.5815069092945"
## [26,] "v21999"                                     "9.70990640950468"
## [27,] "v232"                                       "0"
```

```
##       my.groups Variable
##  [1,] "1"       "MAGER.C"
##  [2,] "2"       "DWGT.C"
##  [3,] "3"       "FAGECOMB.C"
##  [4,] "4"       "WTGAINC.C"
##  [5,] "5"       "v2"
##  [6,] "6"       "v3"
##  [7,] "7"       "v4"
##  [8,] "7"       "v4"
##  [9,] "7"       "v4"
## [10,] "7"       "v4"
## [11,] "8"       "v5"
## [12,] "9"       "v6"
## [13,] "9"       "v6"
## [14,] "10"      "v8"
## [15,] "10"      "v8"
## [16,] "11"      "v18"
## [17,] "11"      "v18"
## [18,] "12"      "v19"
## [19,] "13"      "v20"
## [20,] "13"      "v20"
## [21,] "14"      "v21"
## [22,] "14"      "v21"
## [23,] "14"      "v21"
## [24,] "14"      "v21"
## [25,] "14"      "v21"
## [26,] "14"      "v21"
```

```
## [27,] "15"      "v23"

#lambda optimo

mod$lambda

## [1] 10.15225

#GRAFICA LAMBDA VS BIC

plot( fit$lambda,mod$IC,type="l",xlab = "lambda" ,ylab = "BIC")

abline(v=mod$lambda,col="red")
```




```
#ajuste de modelo con regresion a quantile para tao=0.05

library(quantreg)

MAGERC=(MAGER)-25

DWGTC=as.numeric(DWGT)-165
```

```
WTGAINCC=(WTGAINC)-26
FAGECOMBC=(FAGECOMB)-27
quantreg005<-rq(y ~ MAGERC +DWGTC+WTGAINCC+factor(GESTREC10),tau=0.05)
summary(quantreg005)

##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10),
##     tau = 0.05)
##
## tau: [1] 0.05
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        1510.43930   75.61499   19.97539    0.00000
## MAGERC               -1.64303    2.28675   -0.71850    0.47248
## DWGTC                 1.61189    0.42126    3.82639    0.00013
## WTGAINCC              3.33557    1.04363    3.19613    0.00140
## factor(GESTREC10)2  965.46114   76.63636   12.59795    0.00000
## factor(GESTREC10)3 1054.43747   81.54761   12.93033    0.00000

#ajuste de modelo con regresion a quantile para tao=0.25
library(quantreg)
quantreg025<-rq(y ~ MAGERC+DWGTC+FAGECOMBC+WTGAINCC
                +SEX+factor(GESTREC10)+factor(DMETH_REC),tau=0.25)
summary(quantreg025)

##
## Call: rq(formula = y ~ MAGERC + DWGTC + FAGECOMBC + WTGAINCC + SEX +
```

```
##     factor(GESTREC10) + factor(DMETH_REC), tau = 0.25)
##
## tau: [1] 0.25
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        2335.14161   25.78523   90.56121    0.00000
## MAGERC                6.39824    1.50769    4.24375    0.00002
## DWGTC                 2.29369    0.19173   11.96308    0.00000
## FAGECOMBC            -0.65494    1.21392   -0.53953    0.58954
## WTGAINCC              4.29751    0.50664    8.48245    0.00000
## SEXM                 82.08421   12.58618    6.52177    0.00000
## factor(GESTREC10)2  501.62903   24.99226   20.07137    0.00000
## factor(GESTREC10)3  635.22310   30.08582   21.11371    0.00000
## factor(DMETH_REC)2  -33.10919   12.75446   -2.59589    0.00946
```

```
#ajuste de modelo con regresion a quantile para tao=0.50
library(quantreg)
quantreg050<-rq(y ~ MAGERC+DWGTC+WTGAINCC
                +SEX+factor(GESTREC10),tau=0.50)

summary(quantreg050)
```

```
##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + SEX + factor(GESTREC10),
##     tau = 0.5)
##
## tau: [1] 0.5
```

```
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        2654.80293   21.70777  122.29734    0.00000
## MAGERC                7.09574    1.16339    6.09920    0.00000
## DWGTC                 2.58455    0.19571   13.20609    0.00000
## WTGAINCC              5.26312    0.53803    9.78226    0.00000
## SEXM                102.63871   13.16861    7.79420    0.00000
## factor(GESTREC10)2  396.18528   22.13613   17.89767    0.00000
## factor(GESTREC10)3  553.00450   25.38189   21.78737    0.00000

#ajuste de modelo con regresion a quantile para tao=0.75
library(quantreg)
quantreg075<-rq(y ~ MAGERC+DWGTC+FAGECOMBC
                +WTGAINCC+SEX+factor(GESTREC10),tau=0.75)
summary(quantreg075)

##
## Call: rq(formula = y ~ MAGERC + DWGTC + FAGECOMBC + WTGAINCC + SEX +
##     factor(GESTREC10), tau = 0.75)
##
## tau: [1] 0.75
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        2996.61087   25.45480  117.72282    0.00000
## MAGERC                9.26765    1.76819    5.24131    0.00000
## DWGTC                 3.01235    0.23096   13.04260    0.00000
```

```
## FAGECOMBC            -1.89368    1.46277   -1.29458    0.19551
## WTGAINCC              4.99615    0.62461    7.99881    0.00000
## SEXM                 95.45499   14.97002    6.37641    0.00000
## factor(GESTREC10)2  318.09700   25.81654   12.32144    0.00000
## factor(GESTREC10)3  501.29716   29.81661   16.81268    0.00000

#ajuste de modelo con regresion a quantile para tao=0.95
library(quantreg)
quantreg095<-rq(y ~ MAGERC+DWGTC+FAGECOMBC+WTGAINCC
                +factor(GESTREC10),tau=0.95)
summary(quantreg095)

##
## Call: rq(formula = y ~ MAGERC + DWGTC + FAGECOMBC + WTGAINCC +
##     factor(GESTREC10),tau = 0.95)
##
## tau: [1] 0.95
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        3506.50810   33.45000  104.82835    0.00000
## MAGERC                7.43387    2.64919    2.80609    0.00503
## DWGTC                 4.38640    0.34352   12.76889    0.00000
## FAGECOMBC             0.79025    2.33974    0.33775    0.73556
## WTGAINCC              5.97400    0.94130    6.34654    0.00000
## factor(GESTREC10)2  233.96182   35.67927    6.55736    0.00000
## factor(GESTREC10)3  461.11191   44.32011   10.40412    0.00000
```

```
#ajuste de modelo con SCAD GROUP MC
mc=lm(y~ MAGERC+DWGTC+WTGAINCC+as.factor(MAR)
      +as.factor(LBO_REC)+as.factor(PREVIS_REC)
      +as.factor(SEX)+as.factor(GESTREC10)
      +as.factor(OCNTY),data = BASEFINAL6000_28VAR)
summary(mc)

##
## Call:
## lm(formula = y ~ MAGERC + DWGTC + WTGAINCC + as.factor(MAR) +
##     as.factor(LBO_REC) + as.factor(PREVIS_REC) + as.factor(SEX) +
##     as.factor(GESTREC10) + as.factor(OCNTY), data = BASEFINAL6000_28VAR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2305.51  -260.14     5.29   274.91  2060.84
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2516.3089    27.0273  93.103  < 2e-16 ***
## MAGERC                    1.8808     1.1246   1.672 0.094504 .
## DWGTC                     2.5455     0.1623  15.682  < 2e-16 ***
## WTGAINCC                  5.3454     0.4547  11.756  < 2e-16 ***
## as.factor(MAR)2         -23.4164    12.7587  -1.835 0.066507 .
## as.factor(LBO_REC)2      72.2927    13.3342   5.422 6.14e-08 ***
## as.factor(LBO_REC)3      79.1846    16.8845   4.690 2.80e-06 ***
## as.factor(PREVIS_REC)2   12.5895    15.2574   0.825 0.409325
```

```
## as.factor(PREVIS_REC)3   58.9799    14.4578   4.079 4.57e-05 ***
## as.factor(SEX)M          94.4408    11.3452   8.324  < 2e-16 ***
## as.factor(GESTREC10)2   446.1216    16.0082  27.868  < 2e-16 ***
## as.factor(GESTREC10)3   611.1074    20.1258  30.364  < 2e-16 ***
## as.factor(OCNTY)25      -27.3156    27.4151  -0.996 0.319111
## as.factor(OCNTY)31       66.9317    39.6986   1.686 0.091849 .
## as.factor(OCNTY)97       -2.9776    27.1591  -0.110 0.912702
## as.factor(OCNTY)113      64.4472    22.7411   2.834 0.004613 **
## as.factor(OCNTY)127      72.8701    20.3152   3.587 0.000337 ***
## as.factor(OCNTY)999      42.3657    19.8346   2.136 0.032724 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 438.5 on 5982 degrees of freedom
## Multiple R-squared:  0.2386,Adjusted R-squared:  0.2365
## F-statistic: 110.3 on 17 and 5982 DF,  p-value: < 2.2e-16
```

ajuste del modelo general

```
library(quantreg)
#modelo general cuantiles
quantreg<-rq(y ~ MAGERC+DWGTC+WTGAINCC+factor(GESTREC10)+
factor(DMETH_REC)+SEX+FAGECOMB,tau=c(0.05,0.25,0.50,0.75,0.95))
summary(quantreg)

##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10) +
##     factor(DMETH_REC) + SEX + FAGECOMB, tau = c(0.05, 0.25, 0.5,
```

```
##     0.75, 0.95))
##
## tau: [1] 0.05
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        1483.99939  112.19679   13.22675    0.00000
## MAGERC               -1.72569    3.11212   -0.55451    0.57925
## DWGTC                 1.57766    0.42547    3.70806    0.00021
## WTGAINCC              3.65367    1.03391    3.53384    0.00041
## factor(GESTREC10)2  943.18220   87.12841   10.82520    0.00000
## factor(GESTREC10)3 1041.57701   91.75142   11.35216    0.00000
## factor(DMETH_REC)2  -74.00380   26.14809   -2.83018    0.00467
## SEXM                 51.50435   25.31694    2.03438    0.04196
## FAGECOMB              1.57091    2.52629    0.62182    0.53408
##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10) +
##     factor(DMETH_REC) + SEX + FAGECOMB, tau = c(0.05, 0.25, 0.5,
##     0.75, 0.95))
##
## tau: [1] 0.25
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        2352.82506   42.61617   55.20968    0.00000
## MAGERC                6.39824    1.50769    4.24375    0.00002
## DWGTC                 2.29369    0.19173   11.96308    0.00000
```

```
## WTGAINCC              4.29751    0.50664    8.48245    0.00000
## factor(GESTREC10)2  501.62903   24.99226   20.07137    0.00000
## factor(GESTREC10)3  635.22310   30.08582   21.11371    0.00000
## factor(DMETH_REC)2  -33.10919   12.75446   -2.59589    0.00946
## SEXM                 82.08421   12.58618    6.52177    0.00000
## FAGECOMB             -0.65494    1.21392   -0.53953    0.58954
##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10) +
##     factor(DMETH_REC) + SEX + FAGECOMB, tau = c(0.05, 0.25, 0.5,
##     0.75, 0.95))
##
## tau: [1] 0.5
##
## Coefficients:
##                    Value      Std. Error t value    Pr(>|t|)
## (Intercept)        2679.79793   43.50628   61.59565    0.00000
## MAGERC                7.49562    1.59741    4.69237    0.00000
## DWGTC                 2.59379    0.19956   12.99751    0.00000
## WTGAINCC              5.22228    0.54083    9.65602    0.00000
## factor(GESTREC10)2  390.84575   22.23227   17.58011    0.00000
## factor(GESTREC10)3  548.74382   25.72265   21.33310    0.00000
## factor(DMETH_REC)2   -9.91878   13.86326   -0.71547    0.47435
## SEXM                104.06892   13.24077    7.85973    0.00000
## FAGECOMB             -0.55568    1.35050   -0.41146    0.68075
##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10) +
##     factor(DMETH_REC) + SEX + FAGECOMB, tau = c(0.05, 0.25, 0.5,
```

```
##     0.75, 0.95))
##
## tau: [1] 0.75
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        3048.20531   47.76030   63.82299    0.00000
## MAGERC                9.26824    1.73446    5.34358    0.00000
## DWGTC                 3.01518    0.22753   13.25198    0.00000
## WTGAINCC              5.01176    0.61526    8.14582    0.00000
## factor(GESTREC10)2  317.50437   25.64589   12.38032    0.00000
## factor(GESTREC10)3  500.48110   29.52004   16.95395    0.00000
## factor(DMETH_REC)2   -0.49109   15.18727   -0.03234    0.97421
## SEXM                 95.20190   14.75857    6.45062    0.00000
## FAGECOMB             -1.87722    1.43607   -1.30720    0.19120
##
## Call: rq(formula = y ~ MAGERC + DWGTC + WTGAINCC + factor(GESTREC10) +
##     factor(DMETH_REC) + SEX + FAGECOMB, tau = c(0.05, 0.25, 0.5,
##     0.75, 0.95))
##
## tau: [1] 0.95
##
## Coefficients:
##                     Value      Std. Error t value    Pr(>|t|)
## (Intercept)        3429.13462   75.56272   45.38130    0.00000
## MAGERC                8.20438    2.91114    2.81827    0.00484
## DWGTC                 4.14871    0.34476   12.03353    0.00000
```

```
## WTGAINCC               5.79051    0.84863    6.82337    0.00000
## factor(GESTREC10)2   207.25218   32.53336    6.37045    0.00000
## factor(GESTREC10)3   447.59701   39.93490   11.20817    0.00000
## factor(DMETH_REC)2     7.50468   21.98387    0.34137    0.73284
## SEXM                 114.59050   21.41813    5.35016    0.00000
## FAGECOMB               1.38856    2.50754    0.55375    0.57977
```

codigo grafica de paneles

```
# grafica de paneles
quantreg.all<-rq(y ~MAGERC+DWGTC+WTGAINCC+factor(GESTREC10)+
factor(DMETH_REC)+SEX+FAGECOMB,tau=seq(0.05,0.95,by=0.05))

quantreg.plot<-summary(quantreg.all)

plot(quantreg.plot,main = c("INTERCEPTO", "EDAD DE LA MADRE",
 "PESO DE LA MADRE EN LA ENTREGA", "AUMENTO DE PESO DE LA MADRE",
 "GESTACION EN SEMANAS (37 A 39 SEMANAS)",
 "GESTACION EN SEMANAS (40 O MAS SEMANAS)",
 "METODO DE ENTREGA (CESAREA)", "SEXO (MASCULINO)", "EDAD DEL PADRE"))
```

Bibliografía

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes,United states. *Empirical Economics*, 26:247–257.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. second international symposium on information theory (tsahkadsor, 1971). *Akadémiai Kiadó, Budapest*, page 267–281.

Andrzej Galecki, T. B. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Texts in Statistics. Springer, 2013 edition.

Becerra, J. et al. (1993). Low birthweight and infant mortality in puerto rico,United states. *American Journal of Public Health*, 83(11):1572– 1576.

Breheny, P. and Huang, J. (2012). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 1:1–15.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25(2):173–187.

Bruce, P. and Bruce, A. (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly Media, 1 (early release) edition.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*, 33(1):88–126.

Campos, M. et al. (2008). Rate of weight gain in very-low birth weight puerto rican neonates. *Puerto Rico health sciences journal*, 27(2):141–145.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Costanzo, A. and Desimoni, M. (2017). Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using invalsi survey data. *Large-scale Assess Education*, 5(1):1–25.

Cruz, M. et al. (2009). Índice integral de la salud materna e infantil por municipios, Puerto Rico. *Departamento de Salud de Puerto Rico*, 80:4–17.

Cruz, M. et al. (2010). Índice integral de la salud materna e infantil por municipios, Puerto Rico. *Departamento de Salud de Puerto Rico*, 1:4–16.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

ENDI (2011). Alta tasa de nacimientos prematuros en Puerto Rico. elnuevodia.com. Recuperado de: https://www.elnuevodia.com/noticias/locales/notaaltatasadenacimientosprematurosenpuertorico-1110611/.

Fallah, R. et al. (2015). Birthweight related factors in northwestern iran: Using quantile regression method,United states. *Global Journal of health Science*, 8(7):116–125.

Fan, J. (1997). "comments on 'wavelets in statistics: A review' by a. antoniadis,". *Journal of the Italian Statistical Association*, 6:131–138.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456)), pages = 1348–1360,.

Friedman J, Hastie T, T. R. (2008). Sparse inverse covariance estimation with thegraphical lasso. *Biostatistics*, 9), pages = 432–441,.

Frumento, P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics*, pages 74–84.

Gareth, J. et al. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company.

Gareth James, Daniela Witten, T. H. R. T. (2015). *An Introduction to Statistical Learning with Applications in R*. Springer.

Hao, L. and Naiman, D. Q. (2007). *Quantile Regression (Quantitative Applications in the Social Sciences)*, volume 149),.

Huang, J. et al. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499.

John O. Rawlings, Sastry G. Pantula, D. A. D. (1998). *Applied Regression Analysis: A Research Tool, Second Edition*. Springer.

Jung, K.-M. (2014). Robust estimator with the scad function in penalized linear regression. *The SIJ Transactions on Computer Science Engineering and its Applications (CSEA)*, 2(4):156–160.

Koenker, R. (1994). Confidence intervals for regression quantiles. *n Proceedings of the 5th Prague symposium on asymptotic statistics*, page 349–359.

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

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

Koenker, R. and d'Orey, V. (1987). Computing regression quantiles. *Applied Statistics*, 36:383–393.

Lee, S. et al. (2016). Sparse optimization for nonconvex group penalized estimation. *Journal of Statistical Computation and Simulation*, 86(3):597–610.

López, H. A. and Mora, H. M. (2007). Calculus of the estimators of linear quantile regression by the method accpm. *Revista Colombiana de Estadística*, 30(1):53–68.

Milton, J. S. (2007). *Estadística para biología y ciencias de la salud*. McGRAW-HILL/INTERAMERICANA, 3ra ampliada edition.

Peraza, G. et al. (2001). Factores asociados al bajo peso al nacer,Cuba. *Revista Cubana de Medicina General Integra*, 17(5):490– 510.

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

Racine, J. et al. (2014). *The Oxford Handbook of Applied Nonparametric and Semi-parametric Econometrics and Statistics*. Oxford University Press.

Rybertt, T. et al. (2016). Retardo de crecimiento intrauterino: Consecuencias a largo plazo. *Revista Médica Clínica Las Condes*, pages 509–513.

Schwarz, G. (1978). Estimating the dimension of a mode. *Ann. Statist*, 6(2):461–464.

Seltman, H. (2012). *Experimental Design and Analysis*. Carnegie Mellon University.

Spatz, J. (2006). *Poverty and Inequality in the Era of Structural Reforms: The Case of Bolivia*. Kieler Studien - Kiel Studies 336. Springer-Verlag Berlin Heidelberg.

Thas, O. (2010). *Comparing Distributions*. Springer Series in Statistics. Springer-Verlag New York.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part i. a review and evaluation. *International Statistical review*, 46:1–19.

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

Walpole, R. E. et al. (2012). *Probabilidad y estadística para ingeniería y ciencias*. Pearson, 9 edition.

Wang, H. et al. (2007a). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics*, 25:347–355.

Wang, L. et al. (2007b). Group scad regression analysis for microarray time course gene expression data. *Journal of Statistical Computation and Simulation*, 23(12):1486–1494.

Wang, L. et al. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107:214–222.

Wei, F. and Zhu, H. (2012). Group coordinate descent algorithms for nonconvex penalized regression. *Computational Statistics and Data Analysis*, 56:316–326.

Yuan, M. and Lin, Y. (2006a). Model selection and estimation in regression with groped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67.

Yuan, M. and Lin, Y. (2006b). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B,*, 68(1):49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320.