COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

PRUEBAS DE CORRELACIÓN MÁXIMA, CORRELACIÓN DE DISTANCIA Y COVARIANZA PARA OPTIMIZACIÓN EN EL PROBLEMA DE SELECCIÓN DE VARIABLE

YAMIL BURGUETE FOURZALI

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO
2016

COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALÍAS COMERCIALES DE PRODUCTOS DE INVESTIGACIÓN

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe, **Yamil Burguete Fourzali** , Alumno de esta Institución, estoy de acuerdo en ser partícipe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor **Gustavo Ramírez Valverde** , por lo que otorgo los derechos de autor de mi tesis **Pruebas de correlación máxima, correlación de distancia y covarianza para optimización en el problema de selección de variable** , y de los productos de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre del colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y el que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta Institución.

Pon lugar del campus, a 03 de octubre de 2016

Yamil Burguete Fourzali

Vo. Bo. Dr. Gustavo Ramírez Valverde

La presente tesis titulada: **Pruebas de correlación máxima, correlación de distancia y covarianza para optimización en el problema de selección de variable**, realizada por el alumno: **Yamil Burguete Fourzali**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO _____

Dr. Gustavo Ramírez Valverde

ASESOR _____

Dr. David Sotres Ramos

ASESOR _____

Dr. Benito Ramírez Valverde

Montecillo, Texcoco, México, Octubre de 2016.

PRUEBAS DE CORRELACIÓN MÁXIMA, CORRELACIÓN DE DISTANCIA Y COVARIANZA PARA OPTIMIZACIÓN EN EL PROBLEMA DE SELECCIÓN DE VARIABLE

YAMIL BURGUETE FOURZALI, Mtro.
Colegio de Postgraduados, 2016

RESUMEN

El problema de selección de variable se refiere a la elección del mejor conjunto de predictores que expliquen a la variable respuesta. Varios grupos de investigación han propuesto métodos de selección de variable que operan con mayor o menor eficiencia dependiendo de las condiciones de las variables predictoras. La presente investigación pretende contrastar los métodos de correlación de distancia, correlación máxima, Lasso y Lasso adaptativo en diferentes condiciones de simulación.

Palabras claves: Selección de variable, correlación de distancia, correlación máxima, lasso, lasso adaptativo, prueba de covarianza, simulación.

MAXIMAL CORRELATION, DISTANCE CORRELATION AND COVARIANCE TESTS FOR OPTIMAL VARIABLE SELECTION

YAMIL BURGUETE FOURZALI, M.S.
Colegio de Postgraduados, 2016

ABSTRACT

Variable selection refers to the problem of picking the best possible subset of predictors that explain the response variable. Several research groups have proposed methods for variable selection that work with certain efficiency depending of the conditions among the predictor variables. The main goal of our research is to contrast the methods of distance correlation, maximal correlation, lasso and adaptive lasso in different simulation conditions.

Key words: Variable selection, distance correlation, maximal correlation, lasso, adaptive lasso, covariance test, simulation.

AGRADECIMIENTOS

Agradezco a todos aquellos que me han apoyado a lo largo de mis años para llegar a este momento de mi vida.
A Annelie,
A Esteban, Laila, Ángeles, Pablo,
A Juan, Pablo, Gela,
A los doctores Gustavo, Benito y David,
A Isabel, Gris, Carmen, Juan,
A todos mis profesores del Colegio, el personal administrativo y CONACyT,
A mis compañeros (especialmente Enrique),
A los coaches Fer Ivan, Rodrigo, Lalo, Karen y Toby,
A 9,
A toda la periferia de amigos, tíos, tías, primos, primas, amistades, gente non-grata,
Finalmente, a la ironía, la comedia, los videojuegos y las figuritas coleccionables.

CONTENIDO

LISTA DE FIGURAS

Capítulo 1

Introducción

Todas las actividades humanas se han visto influenciadas por la recopilación y análisis de datos. El desarrollo tecnológico ha permitido el crecimiento y desarrollo de mayores repositorios de información. Aunque existe evidencia para suponer que esto sucede desde la Segunda Guerra Mundial, es más o menos aceptado que en 1962 nace como tal la Era de la Información (Guyon y Elisseeff, 2003). En los últimos años se ha visto un creciente interés en el manejo de datos masivos para la toma decisiones. Como ejemplo se puede observar el nuevo movimiento de la "Revolución de Datos" por parte de las Naciones Unidas (Melamed, 2014).

Sin embargo, este desarrollo ha generado problemas al incrementar la dificultad de predecir con precisión. Por lo que es necesario hacer una selección de la información relevante para predecir una respuesta. Por lo que diversos grupos de investigación han propuesto técnicas para realizar la selección.

A pesar de que han sido propuestas múltiples aproximaciones al problema, no existe una prueba que sea uniformemente mejor que las demás en todas las condiciones. Por consiguiente, es importante poner a prueba las técnicas con la finalidad de saber cómo se comportan. La presente investigación pretende ampliar la evidencia que se tiene de algunas metodologías de selección de variable.

1.1. Conceptos Relevantes

Esta sección se dividirá en dos aspectos, por una parte, sobre algunos grupos de investigación de este tema en estadística y sus hallazgos; por otra,

para observar las aplicaciones del tema de selección de variable en otras disciplinas. Esto sirve para entender un poco más la necesidad de estudiar esta área del conocimiento y de su actual importancia en la práctica científica.

1.1.1. Selección de variable en estadística

Si se lleva una búsqueda rápida en la base de datos de Project Euclid `https://projecteuclid.org/` utilizando la palabra clave de selección de variable, se obtienen arriba de 150 artículos de los cuales 16 pertenecen a este año. A continuación se presentan algunas de las aproximaciones para tratar de abordar este problema.

Uno de los procedimientos mencionados tanto por Sheather (2009) como por Izenman (2008) es el propuesto por Furnival y Wilson (1974) sobre grandes pasos (leaps and bounds). Este método ayuda a seleccionar un modelo sin la necesidad de probar cada uno de los $2^p - 1$ modelos posibles. La motivación central de los autores era reducir la cantidad de modelos a analizar cuando el número de variables era grande, esto era de gran importancia ya que no se contaba con desarrollos tecnológicos para realizarlo.

Zhang y Zamar (2014) mencionan los métodos que dependen del análisis de algún criterio para evaluar modelos utilizando métodos de selección de variables, como el análisis de regresión por pasos (stepwise regression). Comentan que es un método agresivo, en el caso del procedimiento hacia adelante (forward) una vez que se inserta una variable, no puede salir del modelo. De igual manera, en el procedimiento hacia atrás (backward), una vez eliminada una variable del modelo, no puede entrar al modelo. Con lo anterior, también se propuso los modelos por nivel (stagewise regression) donde se buscan las variables predictoras con mayor influencia y en cada paso se analiza la utilidad de cada variable en el modelo y afuera del mismo. En cuanto a los criterios de comparación, algunos de los comúnmente usados en la literatura son; R^2 ajustada, el criterio de información akaike (AIC), el AIC corregido, el criterio de información bayesiano (BIC) ó validación cruzada (Sheather, 2009).

Uno de los trabajos seminales en selección de variable usando métodos de encogimiento fue presentado por Tibshirani durante la década de los noventa, donde propuso la medida de encogimiento y selección de variable conocida como LASSO (Tibshirani, 1996). Él comenta que lo propuso como respuesta a las limitaciones que presentaba la regresión ridge; esta medida hace el paso de encoger los coeficientes estimados, pero no los selecciona. En un trabajo

posterior de Tibshirani (2011), rememora los conceptos que inspiraron su procedimiento. La idea es resolver la regresión a través de una penalización ℓ_1 tratando de encontrar las $\beta = \{\beta_j\}$ que minimizan

$$\sum_{i=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1.1}$$

Tibshirani (2011) comenta que esto es equivalente minimizar la suma de cuadrados del error con la restricción $\sum |\beta_j| \leq s$, lo cual difiere de la regresión ridge sólo en la restricción. En el caso de la regresión ridge, la restricción es: $\sum_j \beta_j^2 \leq t$. Se menciona que, utilizando la forma generalizada de la penalización $(\sum_{j=1}^{p} = \beta_j^q)^{1/q}$, el modelo emerge cuando $q \to 0$. LASSO utiliza el menor valor posible de q, lo que arroja un problema de convexidad, que es atractivo para propósitos de cómputo. Las influencias mencionadas son las siguientes:

- El *garrotte* no negativo propuesto por Breiman (1995). La idea es minimizar con respecto a $\mathbf{c} = \{c_j\}$

$$\sum_{i=1}^{n} \left(y_i - \sum_j c_j x_{ij} \hat{\beta}_j \right)^2, \qquad \text{sujeto a } c_j \geq 0, \quad \sum_{j=1}^{p} c_j \leq t.$$

  donde $\hat{\beta}_j$ son los estimadores de mínimos cuadrados. El problema es que esto no está definido si $p > n$.

- La regresión puente (bridge regression) de Frank y Friedman (1993), que utilizan la penalización $\lambda \sum |\beta_j|^\gamma$ donde λ y γ se estimaban directamente de la muestra.

- Finalmente, Chen et al. (1996) hicieron la propuesta de búsqueda de base (basis pursuit), que utiliza la penalización ℓ_1 en el contexto del análisis de señal.

A pesar de las ventajas que presenta el procedimiento, no obtuvo popularidad directamente después de ser propuesto. Tibshirani (2011) contempló algunas posibles respuestas para esta falta de interés por parte de la comunidad estadística:

- Las computadoras eran poco potentes, por lo que el cómputo era lento.
- El procedimiento era confuso antes de la aparición del algoritmo LARS.
- Los problemas de alta dimensionalidad $p > n$ no eran populares durante los años noventa.
- Eran raros los problemas con grandes grupos de datos.
- Como R seguía en desarrollo, no había mucha difusión del software.

Gracias al desarrollo en la capacidad de cómputo y el interés en problemas de alta dimensionalidad y de grandes muestras, se ha generado una expansión fuerte de correcciones y variantes para el LASSO. Como son el caso del LASSO adaptativo y sus propiedades (Zou, 2006; Ciuperca, 2012), LASSO adaptativo multivariado (Caner y Fan, 2010), LASSO de datos compartidos (Gross y Tibshirani, 2016), LASSO de componentes (Hussami y Tibshirani, 2013), la prueba de covarianza (Lockhart et al., 2013), entre muchos otros.

El trabajo de Zou y Hastie (2005) es probablemente uno de los más reconocidos (y citados) de los avances de LASSO, ellos proponen la técnica de la red elástica (elastic net). Su comentario es que no ha habido una prueba que domine uniformemente a otras, mencionando específicamente a LASSO, regresión ridge y regresión bridge. La red elástica busca resolver algunos de los problemas que tiene LASSO con sobreparametrización $p > n$ como con correlación entre predictores. Esta medida busca capturar los "peces gordos" al producir un modelo escaso con buena presición de predicción, como LASSO, mientras fomenta los efectos de agrupación de variables.

1.1.2. Selección de variable en otras disciplinas

El tema de selección de variable no es de interés exclusivo en estadística. En otras áreas del conocimiento, la predicción es empleada como herramienta para contestar preguntas de investigación con implicaciones importantes para el avance tecnológico y científico. A continuación se muestran cinco ejemplos de aplicaciones implementadas en diferentes disciplinas.

Los investigadores Ducci et al. (2015), trataron de encontrar la influencia de factores ambientales en la evaluación de la conveniencia del hábitat de cuatro subespecies de murciélagos. En su experimento manejaron 381 variables de diferentes condiciones topográficas y ambientales de un área muy

extensa de tierra en Italia. Su meta era estudiar modelos de distribución de los organismos. Su procedimiento incluyó la depuración de las variables asociadas con modelos univariados, una vez hecha esa preselección con una norma definida, procedieron a clasificar las variables restantes y condujeron un análisis multivariado, tratando de eliminar variables que estuvieran intercorrelacionadas. Entre sus múltiples conclusiones, lograron conservar en su modelo las variables predictoras con mayor influencia a la respuesta. Su modelo realizó predicciones adecuadas para cada una de las cuatro subespecies. Y, finalmente, encontraron que sus hallazgos pueden extenderse a otras especies y escalas especiales (tipos de variables).

El segundo ejemplo se tomó de Gerretzen et al. (2016). Cuando se trabaja con datos quimométricos, como cromatografías y espectrometrías, normalmente sólo se busca un grupo de datos que nombran como la información relevante. Para facilitar su obtención, al realizar los experimentos se eligen medidas de preprocesamiento de la información, para remover artefactos que existen en los resultados. El primer problema que emerge con esto es la elección adecuada de métodos de preprocesamiento, y aún llevando esto a cabo, este proceso no facilita la obtención de las variables con mayor influencia, complicando a su vez la interpretación del modelo. Por lo que en este trabajo mezclan dos conceptos de estadística para optimizar sus resultados. Por una parte, el preprocesamiento elegido está directamente relacionado con el diseño del experimento, con esta decisión disminuyen el riesgo de perder información importante; simultáneamente, mezclan tratamientos específicos de selección de variable para mejorar la interpretación. Sus conclusiones incluyen la mejora de la capacidad predictiva del modelo, al disminuir la influencia de variables con poco poder predictivo que sólo incrementaban la varianza. También hablan de la mejora que se obtiene para hacer la interpretación del modelo sobre otros métodos de reducción de variables, como el uso de los llamados filtros.

La aplicación de selección de variable también se ha tratado de implementar para datos metabolómicos (van Reenen et al., 2016). Estos estudios buscan distinguir un grupo de metabolitos distintivos, obtenidos a través de pruebas como espectroscopía o medidas espectrométricas, entre el grupo control y el grupo experimental. De esta manera se pueden caracterizar diferentes condiciones de enfermedad, drogas de tratamiento, toxicidad, factores ambientales, genéticos o efectos fisiológicos. A los metabolitos distintivos se les conoce como biomarcadores. Es muy difícil llevar a cabo estos estudios por ciertas limitaciones que se pueden tener sobre las condiciones de la muestra,

lo cual dificulta los estudios. Continuamente se dan las condiciones de sobreparametrización $p > n$. Buscaron experimentar con selección de variable para ver si podían utilizar mejores variables predictoras para clasificar correctamente a los sujetos en pacientes con y sin enfermedad. Lograron identificar los marcadores más representativos y lograron discriminar correctamente a los sujetos. Con esta técnica planean avanzar el proceso de elegir, reclasificar o identificar pacientes nuevos o grupos de riesgo. A comparación de las medidas que utilizaban, este procedimiento redujo considerablemente su error de clasificación.

También exiten ejemplos en áreas de comportamiento del consumidor, como es el caso del trabajo de Liébana-Cabanillas et al. (2016). En su investigación muestran una aplicación de los métodos de selección relacionados con el manejo de redes sociales, comercio online y sistemas de pago para las empresas e instituciones financieras que buscan consumidores potenciales. Profundizan en la necesidad que tienen las empresas para poder llevar a cabo campañas publicitarias para alcanzar a su público meta. Distinguen que en el comercio online es muy difícil hacer análisis por el volúmen creciente de información. Destacan que las herramientas que se encuentran disponibles no siempre tienen las características necesarias para separar la información irrelevante de la crucial. Los resultados que obtuvieron sugieren mejoras importantes sobre los obtenidos con los métodos clásicos. Lo cual es una ventaja importante para las instituciones financieras que apliquen esta metodología. Este trabajo es un avance que aplica en varias áreas de desarrollo mercadológico y comportamiento del consumidor.

Finalmente, la investigación realizada por Geng et al. (2015) sobre un análisis profundo de la base de datos de la Sociedad Americana de Cancer (ACS por sus siglas en inglés). El interés era la variable respuesta de bienestar general en sobrevivientes de cancer que fueron tratados entre los 18-45 años. Esta variable se buscaba predecir utilizando covariables de caracter demográfico, social y conductual agrupadas en 8 constructos: personalidad, salud física, salud psicológica, salud espiritual, capacidad de enfrentar situaciones de forma activa y pasiva, soporte social y eficacia propia. Para este trabajo, les interesaba conocer los grupos de variables importantes, as como los constructos individuales relacionados con el bienestar. Los hallazgos sirven para diseñar mejores intervenciones enfocadas para jóvenes supervivientes de cancer desde la perspectiva de un programa para el control del cancer.

1.2. El problema

En muchas ocasiones se buscará predecir el valor de una variable Y con un grupo de variables predictoras $X_1, \ldots, X_p$. Existen múltiples razones para buscar esta relación entre variables, como ejemplo, se puede hablar del costo de medición de la variable respuesta, por lo que sería más conveniente predecirla a través de las variables predictoras. Si no se posee la verdadera forma de la relación entre variables, es muy importante que los datos sean representativos de las condiciones de los predictores (Miller, 2002).

Una pregunta que continuamente aparece sobre selección de variable para regresión es ¿Cuántas variables deben ser elegidas? La respuesta no es una cuestión simple, dado que la relación entre sesgo y varianza juega un papel clave en la toma de decisión. Explicado en el libro de Miller (2002), imagine una variable respuesta y relacionada con k variables predictoras $X_1, \ldots, X_k$, de la siguiente manera:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

Donde ϵ tiene $E(\epsilon) = 0$ y $Var(\epsilon) = \sigma^2$, donde $\beta_0, \ldots, \beta_k$ son parámetros desconocidos. Por lo que se utilizará el cuadrado medio para estimar β_i. Los estimadores de mínimos cuadrados denominados b, en notación matricial son:

$$b = (X'X)^{-1}X'y$$

Donde:

$$b' = (b_0, b_1, \ldots, b_k)$$

X es una matríz de dimensiones $n \times (k+1)$ donde la i-ésima hilera contiene los valores $1, x_1, \ldots, x_k$ en la i-ésima observación, y es un vector de tamaño n con los valores observados de la variable a predecir.

Si se toma el vector $x' = (1, x_1, \ldots, x_k)$ la variable a predecir queda definida de la siguiente manera:

$$\begin{aligned} \hat{y} &= x'b \\ &= b_0 + b_1 x_1 + \ldots + b_k x_k \end{aligned}$$

Utilizando la teoría de mínimos cuadrados, se sabe que

$$\text{var}(x'b) = \sigma^2 x'(X'X)^{-1}x \tag{1.2}$$

Si se hace la factorización de Cholesky de $X'X$

$$(X'X)^{-1} = R^{-1}R^{-T}$$

donde el sobreíndice $^{-T}$ se refiere a la inversa de la transpuesta. Sustituyendo en (1.1),

$$\text{var}(x'b) = \sigma^2(x'R^{-1})(x'R^{-T})' \tag{1.3}$$

Con esto definido, se considera predecir y con las primeras p variables de X, con $p < k$. Entonces sea,

$$X = (X_A, X_B)$$

Donde X_A es la matriz con las primeras $(p+1)$ columnas de X y X_B son las $(k-p)$ columnas restantes. Nuevamente, utilizando la factorización de Cholesky,

$$X'_A X_A = R'_A R_A$$

Considerando lo anterior, R_A tiene $(p+1)$ renglones y columnas de R, entonces la inversa R_A^{-1} es idéntica en renglones y columnas de R^{-1}. Entonces, si x_A son los primeros $(p+1)$ elementos de x y b_A es el vector correspondiente de los coeficientes de regresión estimados con mínimos cuadrados. Estos estimadores para el modelo con sólo p variables:

$$\text{var}(x'_A b_A) = \sigma^2(x'_A R_A^{-1})(x'_A R_A^{-1})' \tag{1.4}$$

Los valores predichos de y son la suma de cuadrados de los primeros $(p+1)$ elementos que fueron sumados para obtener la varianza de $x'b$, entonces

$$\text{var}(x'b) \geq \text{var}(x'_A b_A)$$

Por lo que, la varianza de los valores predichos aumenta monotónicamente con el número de variables utilizadas para predicción. Esto se cumple en el caso de modelos lineales con parámetros ajustados con mínimos cuadrados.

En el caso de que se conozca el modelo real,

$$b_A = (X'_A X_A)^{-1} X'_A y$$

entonces,

$$\begin{aligned} E(b_A) &= (X'_A X_A)^{-1} X'_A X\beta \\ &= (X'_A X_A)^{-1} X'_A (X_A, X_B)\beta \\ &= (X'_A X_A)^{-1} (X'_A X_A, X'_A X_B)\beta \\ &= \beta_A + (X'_A X_A)^{-1} X'_A X_B \beta_B \end{aligned}$$

Donde β_A son los primeros $(p+1)$ valores de β mientras que β_B son los últimos $(k-p)$. Con esto, el sesgo estimando y dado por una x específica es:

$$\begin{aligned} x'\beta - E(X_A\beta_A) =& x'_A\beta_A + x'_B\beta_B - x'_A\beta_A \\ & - x'_A(X'_A X_A)^{-1} X'_A X_B \beta_B \\ =& \{x'_B - x'_A(X'_A X_A)^{-1} X'_A X_B\}\beta_B \end{aligned} \tag{1.5}$$

Por lo tanto, si se aumenta el número de variables en el modelo, aumenta la varianza y disminuye el sesgo. Sin embargo, si la variable que entra en el modelo no tiene valor predictivo, entonces sólo aumenta la varianza. Por lo tanto, si la variable no disminuye mucho el sesgo, el incremento de la varianza excederá el beneficio de la reducción.

En el libro de Izenman(2008) se describen dos fenómenos que se deben considerar al hablar de la selección de las variables para el modelo. En ocasiones, se incluirán muchas variables en el modelo, por lo que la función de regresión tendrá la varianza inflada, a esto se le conoce como sobreajuste (overfitting). También puede ocurrir lo contrario, al meter pocas variables en el modelo, la varianza se verá reducida pero el sesgo aumentará; esto afectará la predicción y se le conoce como subajuste (underfitting).

Tanto Miller como Izenman hablan de la necesidad de buscar un balance entre estos dos extremos, tratando de identificar las variables predictoras importantes. Estas quedan definidas como aquellas variables que al salir del modelo afectan considerablemente la precisión de la predicción.

Cabe destacar, si el número de observaciones en la muestra de entrenamiento aumenta, la varianza de la predicción de (1.4) se reduce. En casos prácticos la varianza de predicción estará en el orden de n^{-1} mientras que la

omisión estará en el orden de 1, en otras palabras, la omisión es independiente de n. Por ende, el número de variables en la mejor predicción aumentará si se aumenta la muestra de calibración (Miller, 2002).

Antes de que se proceda a clasificar los métodos para selección de variable, es crucial mencionar que la motivación central es el deseo de generar la mayor precisión de predicción y, en segundo lugar, parsimonia en el modelo de regresión (Izenman, 2008). Esta distinción se comenta porque el principio de parsimonia no necesariamente genera la mejor respuesta, como lo comenta Domingos (1999) en su crítica al razonamiento de la navaja de Occam. Sin profundizar, el concepto de la navaja de Occam se refiere a la consideración que el modelo más simple siempre será la mejor respuesta. Esto último no necesariamente es cierto. Sin embargo, el ánalisis de este problema no es el enfoque central de este trabajo de investigación.

Harrell en 2015, resume el problema de selección de modelos en la meta de desarrollar modelos concisos, eliminando colinealidad, tratando de eliminar coeficientes "insignificantes" en la regresión. La regresión, a pesar de ser una técnica muy popular, si se observa con detenimiento, es fácil descubrir que viola los principios de estimación estadística y prueba de hipótesis (Harrell, 2015). A continuación se plantean algunas razones de lo anterior:

- Los valores de R^2 tienen sesgo dado que se aumentan.
- La prueba de F ó χ^2 no presentan la distribución atribuida. Los métodos de selección de variable están basados en métodos donde sólo se busca probar hipótesis preespecificadas.
- El método arroja errores estándares de estimación de coeficientes de regresión con sesgo a ser más pequeños. A su vez, los intervalos de confianza para valores predecidos falsamente estrechos.
- Los p-valores son pequeños y su corrección apropiada no es un problema simple.
- Los coeficientes estimados están sesgados y necesitan encogimiento.
- En estudios de observación, la selección de variable determina ajuste de factores de confusión que resulta en residuales con factores de confusión.
- En vez de resolver colinealidad, el proceso de selección se ve afectado por la colinealidad.

En las mismas notas de Harrell (2015) aparecen mencionados los hallazgos de Derksen y Keselman (1992) que estudiaron la selección de variable a través de varios métodos como regresión por pasos (stepwise), selección hacia adelante (forward) y eliminación hacia atrás (backward), con lo que concluyeron:

- La correlación entre predictores afecta la frecuencia en la que se obtiene los predictores "auténticos" del modelo final.
- El número de predictores p afecta el número de predictores falsos que entran en el modelo final.
- El tamaño de la muestra tiene poca importancia práctica para determinar el número de variables "auténticas" al modelo final.
- La determinación de los coeficientes puede ser estimada apropiadamente si se ajusta por el número total de predictores candidatos en vez del número de variables en el modelo final.
- Recomiendan el uso de medidas de preanálisis para determinar si existe relación (lineal o no lineal) entre predictores.

Este último punto es sugerido porque durante el análisis se encuentra que el modelo final, propuesto por las estimaciones, contiene el número real de predictores auténticos menos de la mitad de las veces.

Con todo lo mencionado sobre los problemas de este tema, las ideas que se están trabajando en estadística y su utilidad para otras disciplinas, se puede ver que todavía existen muchas áreas de oportunidad para investigación. Sin embargo, como lo sugiere la literatura en el área, hasta ahora no se ha podido encontrar una respuesta definitiva de una prueba que sea uniformemente mejor para llevar a cabo este procedimiento. Lo que se puede hacer es seguir analizando las medidas que se tienen para observar su funcionamiento.

1.3. Objetivos

El objetivo de la presente investigación es ampliar el conocimiento sobre el uso de métodos de selección de variable. Se pretende obtener evidencia sobre el comportamiento de los métodos de correlación máxima, correlación de distancia, LASSO, LASSO adaptativo y la prueba de covarianza en condiciones

de simulación del modelo lineal, colinealidad entre predictores y colinealidad tipo Toeplitz.

1.4. Hipótesis

Se espera que los métodos basados en penalización tengan un mejor desempeño a comparación de los métodos que trabajan a través de subconjuntos (subsetting). A su vez, se espera que la prueba de covarianza tenga resultados similares a las pruebas de LASSO y LASSO adaptativo.

Capítulo 2

Marco Teórico

En este capítulo se muestra una clasificación de los métodos de selección de variable y se describirán de forma detallada las herramientas a utilizar y las razones para la selección de cada una.

2.1. Clasificación del método de selección de variable

Existen varias formas de clasificar las técnicas de selección de variable, para propósitos de la presente investigación se manejará la clasificación propuesta por Yenigün y Rizzo (2015). Ellos separan las técnicas de selección de variable en métodos de selección a través de subconjuntos y métodos de encogimiento.

2.1.1. Métodos de selección a través de subconjuntos

La información de esta sección fue consultada en el libro de Clarke et al. (2009). La idea central de estos procedimientos es simple, se ajustan todos los modelos posibles y se elige el mejor basado en algún criterio. Algunos de los criterios normalmente utilizados son R^2_{adj}, error cuadrado medio, o la C_k de Mallow. Si se consideran p variables predictoras, se tienen $2^p - 1$ modelos posibles. Por lo que, si p es muy grande, se vuelve excesivamente demandante encontrar una respuesta. Considerando la anterior, múltiples algoritmos han sido propuestos para identificar el mejor subconjunto a través de los llamados

procedimientos voraces (greedy). Estos buscan eliminar múltiples modelos que tienen resultados subóptimos.

Un buen ejemplo de un algoritmo propuesto es el de grandes pasos (leaps and bounds) (Furnival y Wilson, 1974). El cómputo de la respuesta tiene cierta complejidad pero su idea para reducir la cantidad de modelos a comparar se basa en la siguiente idea

$$\varphi_1 \subset \varphi_2 \implies \text{SCE}(\varphi_1) \geq \text{SCE}(\varphi_2)$$

donde φ es la representación del conjunto total de datos, φ_1 y φ_2 son subconjuntos y SCE se refiere a la suma de cuadrados del error definido como $\text{SCE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Se asume en estos procedimientos que las variables son independientes, sin embargo se han utilizado en casos donde no necesariamente lo son.

Este procedimiento pertenece al grupo de técnicas de ramas y pasos (branch and bounds), que buscan eliminar un grupo de modelos candidatos estimando los límites superior e inferior de la SCE. Sin embargo, se pueden utilizar otros criterios como son R^2. Estos procedimientos reducen el cómputo de todos los modelos, sin embargo, cuando se tiene un grupo grande de variables predictoras el tiempo de procesamiento aumenta demasiado (Clarke et al., 2009).

Como consecuencia de lo anterior, se empezaron a proponer métodos para búsqueda secuencial como la regresión por pasos (stepwise), selección hacia adelante (forward), y eliminación hacia atrás (backward). El caso de la selección hacia adelante (forward) se inicia con un modelo sin variables, se va insertando una a una eligiendo la variable que obtenga la máxima disminución de la SCE del modelo actual. Si existen p variables en el modelo actual, el nuevo SCE al agregar una otra variable es

$$\text{SCE}_{p+1}(j) = \text{SCE}_p - \frac{[y^T(I - H_p)\mathbf{x}_j]^2}{\mathbf{x}_j(I - H_p)\mathbf{x}_j}, \tag{2.1}$$

donde $\mathbf{x}_j$ es un vector de datos de la nueva variable y $H_p = X_p(X_p^T X_p)^{-1}X_p^T$ es la matríz proyección del modelo de p variables.

Agregar la variable que obtenga la máxima disminución en SCE es equivalente a elegir la P_{p+1} variable que su correlación parcial es máxima con la respuesta, considerando las variables en el modelo. El método acaba cuando agregar una variable no mejora significativamente el ajuste del modelo bajo

cierto criterio. Uno de los criterios utilizados es llevar a cabo la comparación del ajuste con el valor crítico del estadístico F para probar la hipótesis $H_0 : \beta_{p+1} = 0$ para el modelo con la $(p+1)$ variable. Por lo que, la variable X_{p+1} es agregada al modelo si

$$F_{k+1} = \max \left(\frac{\mathrm{SCE}_p - \mathrm{SCE}_{p+1}}{\mathrm{SCE}_{p+1}/(n-p-1)} \right) > F_{crit} \tag{2.2}$$

donde $F_{crit} = F(\alpha; 1, n-p-1)$, este valor es el que sirve como el umbral de entrada de la variable al modelo.

Por otra parte, la eliminación hacia atrás (backward) funciona considerando el modelo con todas las k variables y a cada paso se remueve la variable que hace la menor contribución. Suponga que hay p variables, $p \leq k$, en el modelo actual la matriz diseo correspondiente es X_p. El nuevo SCE al borrar la j-ésima $(1 \leq j \leq p)$ variable del modelo actual con p variables es

$$\mathrm{SCE}_{p-1}(j) = \mathrm{SCE}_k + \frac{(\hat{\beta}_j)^2}{s^{jj}}, \tag{2.3}$$

donde $\hat{\beta}_p = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ es un vector de los coeficientes actuales y s^{jj} es el j-ésimo elemento de la diagonal de $(X_1^T X_1)^{-1}$. Al igual que en el caso de selección hacia adelante (forward), estos pasos se repiten hasta que borrar una variable empieza a afectar el ajuste del modelo. Y, de igual manera, se puede utilizar el criterio del estadístico F. Se elimina a la variable X_j en el modelo si

$$F_j = \min \left(\frac{\mathrm{SCE}_{p-1} - \mathrm{SCE}_p}{\mathrm{SCE}_p/(n-p)} \right) < F_{crit} \tag{2.4}$$

donde $F_{crit} = F(\alpha; 1, n-p)$, en este caso, este valor sirve como un umbral para permitir la salida de la variable.

Estas dos metodologías para elegir el modelo son muy restrictivas que una vez elegida la entrada o salida de una variable, la decisión no es reversible. Para tratar de disminuir el nivel restrictivo, existe otra propuesta que es un híbrido de la selección hacia adelante y la eliminación hacia atrás, se le conoce como regresión por pasos (stepwise). Inicia sin variables en el modelo y utilizando 2.2 inserta variables al modelo hasta que el criterio se cumpla. De ahí se procede a eliminar las variables mientras se satisface el criterio 2.4. Una vez completada la eliminación de las variables, se vuelven

a examinar las variables hasta satisfacer 2.2 en la entrada. Se sigue llevando este procedimiento hasta que el modelo cumpla ambos criterios.

A pesar de que este método resulta muy positivo dado que reduce mucho el costo computacional para encontrar el modelo, tiene desventajas:

- No garantiza obtener el grupo de variables óptimo.
- Es muy inestable aún cuando los cambios en los datos son pequeños.

Esto específicamente se comenta en el caso de utilizar un criterio de selección basado en los errores al cuadrado. Existen otros criterios que pueden ser utilizados para elegir el mejor modelo. A continuación se presenta una pequeña explicación de algunos criterios comúnmente utilizados.

2.1.2. Criterios para selección de modelo

Existen múltiples criterios para la evaluación de los modelos. Aquí se presentan algunos de los que comúnmente se menciona en la literatura (Sheather, 2009):

El coeficiente de determinación R^2, se define como la proporción de la variabilidad total de la respuesta Y explicada por el modelo de regresión, esto es,

$$R^2 = \frac{\text{SCreg}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}} \tag{2.5}$$

donde

$$\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad \text{Suma de cuadrados totales}$$

$$\text{SCR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad \text{Suma de cuadrados residuales}$$

$$\text{SCreg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad \text{Suma de cuadrados de la regresión}$$

Para compensar esto, se define el coeficiente de determinación ajustado R^2_{adj},

$$R^2_{adj} = 1 - \frac{SCR/(n-p-1)}{SST/(n-1)}, \tag{2.6}$$

donde p es el número de predictores en el modelo actual. Es posible observar que el aumento de variables predictoras en el modelo incrementa la R^2_{adj} si la prueba F excede 1. Elegir el modelo con el mayor valor de R^2_{adj} tiende a sobreparametrizar.

Otro criterio es el Criterio de Información de Akaike (AIC por sus siglas en inglés), que se basa en el balance de la bondad de ajuste y una penalización por la complejidad del modelo. A diferencia del caso anterior, el modelo con menor AIC es el mejor modelo. Se define,

$$\text{AIC} = 2[-\log(L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2|Y)) + K]$$

donde K es una medida de complejidad y se define $K = p + 2$, y $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2$ son estimados en el modelo ajustado. K es necesaria para incrementar la log-verosimilitud al agregar un predictor irrelevante al modelo. Cuando se ocupa R, el AIC se calcula:

$$\text{AIC} = n \log\left(\frac{\text{SCR}}{n}\right) + 2p$$

Otro criterio es el criterio de información bayesiano (BIC), se define

$$\text{BIC} = -2\log(L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2|Y)) + K\log(n)$$

donde $K = p+2$, es el número de parámetros estimados en el modelo. Al igual que con AIC, el modelo con menor valor BIC es el mejor modelo. Se puede observar mucha similaridad con el cálculo del AIC salvo por la modificación del término de penalización con $\log(n)$. Por esta razón, BIC penaliza más que AIC cuando los modelos son complejos, simplificándolos más que AIC.

Finalmente, se puede utilizar un criterio de validación cruzada (James et al., 2013). La motivación detrás de este método es la dificultad de tener un grupo de datos de entrenamiento y otro para la prueba. La validación cruzada sirve estima el error de prueba tomando un subconjunto de los datos de entrenamiento, y tratar de estimarlos con el resto de las observaciones. Existe más de un método, sin embargo, el pertinente para el presente trabajo es la validación cruzada sacando una observación. Simplemente separa los datos en dos partes, se toma una observación (x_1, y_1) que se utilizará para validar y el resto de las observaciones $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ para entrenar. Se ajusta el modelo con $n-1$ observaciones y se predice $\hat{y_1}$ con la observación x_1. Entonces se realiza la estimación del error cuadrado medio $\text{ECM}_1 = (y_1 - \hat{y_1})^2$, que es un estimador aproximadamente insesgado del error de prueba.

A pesar de ser aproximadamente insesgado, el ECM_1 es una estimación pobre por su alta variabilidad, dado que sólo se calcula de una observación. Por lo que se repite el procedimiento n veces hasta que se calcula $\mathrm{ECM}_1, \ldots, \mathrm{ECM}_n$. Con esto, el estimador de validación cruzada sacando una observación es

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{ECM}_i \tag{2.7}$$

Algunas ventajas de este criterio,

- Poco sesgado.
- Como se ajusta el modelo con $n-1$ observaciones, tiende a no sobreestimar el error de prueba.
- Los resultados son consistentes.

A pesar de parecer una medida idónea, tiene algunas desventajas. En primer lugar es muy demandante en su implementación porque el modelo se ajusta n veces. Si resulta ser que n es muy grande, puede consumir demasiado tiempo de procesamiento el cómputo.

2.1.3. Métodos de penalización

Como se ha comentado, los métodos que hacen subconjuntos de predictores producen modelos interpretables, con la meta de tener menor error de predicción a comparación del modelo completo. Sin embargo, como es un proceso discreto tiene altos niveles de varianza, por lo que el error de predicción no necesariamente se reduce a comparación del modelo completo. Por estas razones, como lo muestra Hastie et al. (2009), se empezaron a utilizar los métodos de penalización. Estos son más continuos, por lo que no sufren mucho del problema de alta variabilidad. Aunque existen varios métodos, sólo se hablará específicamente de dos: regresión ridge y LASSO.

La regresión ridge penaliza los coeficientes de regresión, esto los reduce. Por lo que los coeficientes ridge minimizan la suma de cuadrados residual,

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

donde $\lambda \geq 0$ es un parámetro de complejidad que controla la proporción de la penalización. Mientras mayor sea λ, aumenta la penalización. Los coeficientes se reducen hacia cero y entre ellos. Cuando hay varios predictores correlacionados en un modelo de regresión lineal, los coeficientes son mal determinados y exhiben alta varianza. Sin embargo, si se impone límites de tamaño en los coeficientes, el problema se reduce. Una de las críticas principales es que no reduce los coeficientes a cero, sólo los acerca.

Tomando esta limitación, otra medida fue propuesta para tratar de obtener no sólo la penalización de los valores, sino también selección de variable. El LASSO utiliza la penalización ℓ_1 en vez de ℓ_2 que usa ridge. Con este cambio, las soluciones a y_i son no lineales, por lo que no hay una expresión cerrada como en ridge. Sin embargo, existen algoritmos con los que se calcula todo el camino de soluciones variando λ, con el mismo costo computacional de ridge. Por las propiedades de la penalización, ciertos coeficientes se vuelven cero.

A continuación se hablará de las técnicas que se van a utilizar. Las primeras dos, correlación máxima y correlación de distancia, son técnicas que se pueden agrupar en los métodos a través de subconjuntos. Mientras que LASSO, LASSO adaptativo y la prueba de covarianza se clasifican como técnicas de penalización.

2.2. Correlación máxima (MC)

Medida propuesta por Hirschfeld y Gebelein (en Bryc et al., 2002). Aunque existen múltiples formas de representar la MC, se tomó la definición dada por Yenigün y Rizzo (2015), donde se define como:

$$S(X, Y) = \sup_{f,g} \rho(f(X), g(Y)) \tag{2.8}$$

Tanto X como Y tienen varianzas finitas positivas, $f(\cdot)$ y $g(\cdot)$ son funciones medibles en espacios de Borel, además $\rho(U, V)$ se refiere al coeficiente de correlación entre las dos variables. Para poder presentar las propiedades de esta medida de dependencia, es importante recapitular los postulados de Rényi, por lo que en el siguiente apartado se profundiza más sobre ese tema.

2.2.1. Propiedades fundamentales de medidas de dependencia

Rényi (1959) describió unos postulados que son la base para entender la fuerza de la asociación entre dos variables y comparó diversas medidas de dependencia para ver que tanto cumplían con los mismos. Uno de sus comentarios es que, si la prueba cumple conlos postulados, entonces tiene la capacidad de detectar la fuerza de la relación entre dos variables. En la siguiente sección se resumen sus hallazgos.

Sea U y V variables aleatorias en un espacio de probabilidad donde ninguna sea constante con probabilidad 1. Para medir la fuerza de la dependencia entre U y V se necesita alguna forma de caracterizar su relación. Normalmente, se elige un rango $[0, 1]$ donde 1 corresponde cuando existe dependencia y 0 cuando son independientes. Con estas convenciones establecidas, los siguientes postulados son necesarios para examinar si medida de dependencia, representada por $\delta(U, V)$, es apropiada:

A) $\delta(U, V)$ se define como la asociación entre las variables aleatorias U y V, ninguna es constante con probabilidad 1.

B) $\delta(U, V) = \delta(V, U)$.

C) $0 \leq \delta(U, V) \leq 1$.

D) $\delta(U, V) = 0$ si y sólo si U y V son independientes.

E) $\delta(U, V) = 1$ si existen una dependencia estricta, como $U = g(V)$ ó $V = f(U)$, entre U y V. Donde $g(\cdot)$ y $f(\cdot)$ son funciones medibles en espacios de Borel.

F) Si $f(\cdot)$ y $g(\cdot)$ mapean en el eje real uno a uno, entonces $\delta(f(U), g(V)) = \delta(U, V)$.

G) Si la distribución conjunta de U y V es normal, entonces $\delta(U, V) = |\rho(U, V)|$ donde $\rho(U, V)$ es el coeficiente de correlación entre U y V.

Después de definir estos postulados, Rényi analizó algunas medidas de dependencia con estos. Empezando con el coeficiente de correlación, que se define como:

$$\rho(U, V) = \frac{\mathbf{M}(UV) - \mathbf{M}(U)\mathbf{M}(V)}{\mathbf{D}(U)\mathbf{D}(V)} \tag{2.9}$$

donde $\mathbf{M}(\cdot)$ es la media y $\mathbf{D}^2(\cdot)$ es la varianza. Como el rango del estadístico es de $[-1, +1]$, sólo satisface (C) con el valor absoluto. A su vez, con $|\rho(U, V)|$ satiface (B),(C), y (G). Algunas otras conclusiones sobre el coeficiente de correlación son:

- Sólo se encuentra definido si $D(U)$ y $D(V)$ son finitos y positivos.
- Puede desaparecer aún cuando U y V tengan una dependencia funcional entre ellas. Rényi (1959) lo ejemplifica de la siguiente manera: $U \sim \mathcal{U}(-1, +1)$ y $V = 5U^3 - 3U$, se tiene $\rho(U, V) = 0$.
- $|\rho(U, V)|$ es igual a 1 si y sólo si hay una relación lineal entre U y V.

También analizó la MC definida en 2.8. El comentario central es que cumple con todos los postulados, lo cual la hace una buena medida de asociación muy fuerte, al punto de poder detectar relaciones no lineales entre variables. Sin embargo, la desventaja más fuerte es que no siempre existe el supremo de $\rho(f(U), g(V))$. A pesar de las limitaciones que presentaba Rényi, somete a evaluación una forma interesante de resolver el problema. Propone que, si existen algunas funciones $f_0(\cdot)$ y $g_0(\cdot)$ que cumplan la siguiente igualdad:

$$S(U, V) = \rho(f_0(U), g_0(V)) \tag{2.10}$$

Entonces la correlación máxima para U y V puede ser obtenida. Claramente esta aproximación es muy limitante, dado que no siempre se van a encontrar con facilidad las funciones óptimas $f_0(\cdot)$ y $g_0(\cdot)$. Por lo que, tiempo después se buscó una alternativa no paramétrica, misma que se explica a continuación.

2.2.2. Algoritmo ACE

Dado que la MC no siempre existe, se pueden utilizar alternativas para estimar el máximo. Una propuesta por Breiman y Friedman (1985) realiza la estimación al buscar unas funciones $\theta^*(Y)$ y $\phi_1^*(X_1), \ldots, \phi_k^*(X_k)$ que minimicen la varianza no explicada por la regresión. A esta solución se le llama el método de estimación de transformaciones óptimas (ACE por su nombre en inglés). Se muestra el principio de este método, si se define la varianza no explicada por la regresión como:

$$\epsilon^2(\theta, \phi_1, \ldots, \phi_k) = \frac{E\{[\theta(Y) - \sum_{i=1}^{k} \phi_i(X_i)]^2\}}{E\theta^2(Y)} \tag{2.11}$$

Por lo que se definen transformaciones óptimas como:

$$\epsilon^2(\theta^*, \phi_1^*, \ldots, \phi_k^*) = \min_{\theta, \phi_1, \ldots, \phi_k} \epsilon^2(\theta, \phi_1, \ldots, \phi_k) \tag{2.12}$$

Específicamente en el caso bivariado cuando $k = 1$, la transformación óptima de $\theta^*(Y)$ y $\phi_1^*(X)$ satisface

$$\rho^*(X, Y) = \rho(\theta^*, \phi^*) = \max_{\theta, \phi} \rho[\theta(Y), \phi(X)] \tag{2.13}$$

donde ρ es el coeficiente de correlación y ρ^* es la MC entre X y Y. Existen muchas ventajas para utilizar este método:

- Es computacionalmente eficiente
- Como es un procedimiento no paramétrico, elige la transformación óptima con muy pocos supuestos
- Permite la comparación entre diferentes tipos de variables (continuas, ordinales, categóricas) dado que las transformaciones de $\theta(Y), \phi_1(X_1), \ldots, \phi_p(X_p)$ asumen valores en la línea real.
- Puede estimar Y en lugar de $\theta^*(Y)$ porque existe una función inversa θ^{*-1} para θ^*:

$$\hat{\theta}^*(Y) = \sum_{i=1}^{k} \hat{\phi}_i^*(X_i)$$

$$Y = \hat{\theta}^{*-1}\left(\sum_{i=1}^{k} \hat{\phi}_i^*(X_i)\right)$$

Este algoritmo posee las mismas propiedades que la MC. Lo que lo vuelve la alternativa computacional idónea para trabajar.

2.3. Correlación de distancia (DC)

La DC es una medida de asociación entre variables propuesta por Székely, Rizzo y Barikov (2007), en su trabajo proponen medidas análogas para el cálculo de la covarianza y correlación, con las funciones “dcov” y “dcor”

respectivamente. Algunos resultados sugieren que si la estructura de dependencia es no lineal, la prueba de covarianza de distancia es más fuerte que la razón de verosimilitud. Este estadístico también puede detectar estructuras de dependencia no monótonas. Al evaluar esta medida con los postulados de Rényi, se puede observar que los primeros cuatro postulados se cumplen, sin embargo, el resto sólo se cumplen bajo condiciones especiales; (E) sólo sucede si las funciones son lineales, (F) si las transformaciones son ortogonales, y (G) sólo si X y Y son normales bivariadas (Yenigün y Rizzo, 2015).

La DC empírica se define como un número no negativo $R_n(\mathbf{X}, \mathbf{Y})$ entre dos variables aleatorias X y Y con $E(X) < \infty$ y $E(Y) < \infty$ y está dada por la raíz cuadrada de $R_n^2(\mathbf{X}, \mathbf{Y})$:

$$R_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{V_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{V_n^2(\mathbf{X})V_n^2(\mathbf{Y})}}, & V_n^2(\mathbf{X})V_n^2(\mathbf{Y}) > 0, \\ 0, & V_n^2(\mathbf{X})V_n^2(\mathbf{Y}) = 0. \end{cases} \tag{2.14}$$

Algunas propiedades son que si $R = 0$, las variables X y Y son independientes. El rango es $0 \leq R \leq 1$, en el caso bivariado, R es una función de ρ, y $R(X, Y) \leq |\rho(X, Y)|$ siendo iguales cuando $\rho = \pm 1$. Los estadísticos dependientes de distancia se definen a continuación para una muestra aleatoria $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ de la distribución conjunta de vectores aleatorios X y Y en $\Re^p$ y $\Re^q$ respectivamente, entonces se define

$$a_{kl} = |X_k - X_l|_p, \qquad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \qquad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

donde $k, l = 1, \ldots, n$. Se definen de manera similar los estadísticos para $b_{kl} = |Y_k - Y_l|_q$. Con esto, la covarianza de distancia empírica $V_n(X, Y)$ es un número no negativo que proviene de

$$V_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}. \tag{2.15}$$

De manera similar, $V_n(\mathbf{X})$ es un número no negativo definido por

$$V_n^2(\mathbf{X}) = V_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2. \tag{2.16}$$

2.4. LASSO

LASSO es un acrónimo de "Least Absolute Shrinkage and Selection Operator" (operador de encogimiento y selección de mínimos absolutos) que nace como una herramienta para mejorar el resultado de los mínimos cuadrados ordinarios, permitiendo hacer simultáneamente estimación y selección de variable (Tibshirani, 1996). Esta mejora es la reducción de algunos coeficientes a 0, con lo que se aumenta un poco el sesgo pero se reduce la varianza, logrando así mayor precisión predictiva. Por otra parte, también simplifica la explicación al reducir el número de predictores. Los estimadores de LASSO se definen como:

$$\hat{\beta}(lasso) = \arg\min_{\beta} \|y - \sum_{i=1}^{p} x_i \beta_i\|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \tag{2.17}$$

Donde λ es un parámetro de regularización no negativo. El segundo término es la penalización ℓ_1. Esta penalización es la diferencia central entre la regresión LASSO y la regresión ridge, donde se utiliza la penalización ℓ_2. Esta diferencia es muy importante porque, por la penalización que utiliza, la regresión ridge no encoge los coeficientes exactamente a cero, mientras que LASSO encoge los coeficientes, y además los vuelve exactamente 0 si λ es suficientemente grande (Ng, 2013). En la figura 2.1 se muestra la diferencia en la forma en que opera la penalización entre LASSO y ridge, donde es mucho más simple entender porqué la penalización ℓ_1 permite con mayor facilidad que algunos coeficientes sean iguales a cero.

Hay una gran cantidad de investigación realizada en las propiedades de la estimación penalizada de verosimilitud. Fan y Lv (2010) las clasifican en cuatro grandes grupos:

- **Persistencia**: Se refiere a la consistencia del riesgo (pérdida esperada) del modelo estimado.

- **Consistencia y consistencia de selección**: Consistencia del estimador del vector de parámetros ante cierta pérdida, mientras que consistencia de selección se refiere a la consistencia del modelo seleccionado.
- **Propiedad oráculo débil**: Significa que el estimador presenta la misma dispersión que el estimador oráculo con probabilidad asintótica de uno y es consistente.
- **Propiedad oráculo**: Además de lo anterior, el estimador presenta la capacidad limitada por la información de imitar al estimador oráculo.

2.4.1. Consistencia de LASSO

Esta propiedad se mantiene bajo ciertas condiciones de la matríz diseño. Para seleccionar de manera adecuada, el propósito es que pueda estimar $\hat{\beta}$ y que el estimador tenga el mismo soporte que los coeficientes del vector β_0 con probabilidad asimptótitca de 1. Zhao y Yu (2006) caracterizaron la consistencia estudiando una propiedad más conveniente y fuerte a la vez, ls consistencia de signos, donde $P(\text{sgn}(\hat{\beta}) = \text{sgn}(\beta_0)) \to 1$ mientras $n \to \infty$. Para establecer la propiedad débil de oráculo, muestran la condición irrepresentable

$$\|X_2^T X_1 (X_1^T X_1)^{-1} \text{sgn}(\beta_1)\|_\infty < 1 \tag{2.18}$$

esta condición es necesaria para la consistencia de signos del LASSO.Por otra parte, la condición fuerte irrepresentable requere que el lado derecho de 2.18 sea uniformemente limitada por una constante positiva $c < 1$. Donde β_1 es un subvector de β_0 y X_1, X_2 son submatrices de la matriz diseño X. Sin embargo, la condición irrepresentable puede ser restrictiva cuando se trabaja con grandes dimensiones. Es por ello que LASSO fácilmente elige un modelo inconsistente, y tiene a meter falsos positivos en el modelo.

Para establecer la propiedad oráculo débil, además de la escasez caracterizada en los párrafos anteriores, además de la consistencia, se necesita la condición de la matriz diseño

$$\|X_2^T X_1 (X_1^T X_1)^{-1}\|_\infty \leq c \tag{2.19}$$

para alguna constante $c < 1$, esto es más fuerte que la condición fuerte irrepresentable. Además comentan que la penalización ℓ_1 de los coeficientes de regresión de cada variable inactiva ajustada a las p variables activas debe

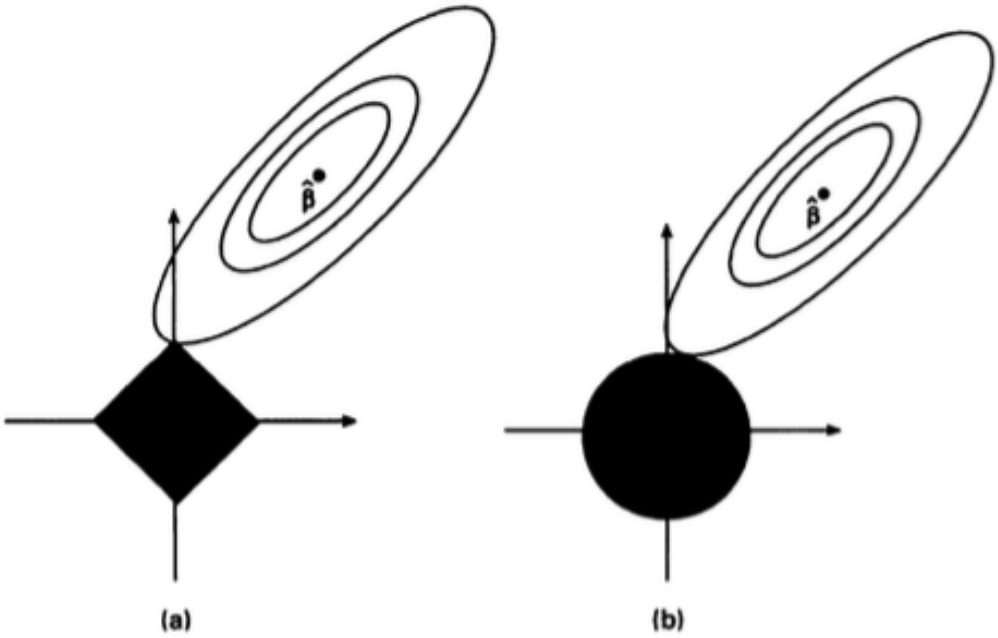

Figura 2.1: Diferencia entre LASSO(a) y Ridge (b) (en Tibshirani, 1996).

ser uniformemente limitada por $c < 1$. Esto muestra que LASSO selecciona un modelo consistente de manera muy limitada, a notar que la penalización ℓ_1 de los coeficientes incrementa cuando se incrementa p.

2.4.2. Propiedad oráculo

En Fan y Li (2001) presentan la verosimilitud penalizada no concava para selección de variables, además introducen el concepto de la propiedad oráculo. Se refiere a cuando un estimador $\hat{\beta}$ tiene escacez(sparcity) de tal manera que $\hat{\beta}_2 = 0$ con una probabilidad de 1 si $n \to \infty$. A su vez, $\hat{\beta}_1$ se comporta como un estimador oráculo limitado por la información, donde $\hat{\beta}_1$ y $\hat{\beta}_2$ son subvectores de $\hat{\beta}$ formados por componentes con soporte en β_0 y β_0^c, respectivamente.

2.5. LASSO adaptativo

El LASSO adaptativo es, al igual que el LASSO, un estimador con penalización ℓ_1 pero el nivel de la penalización se va ajustando (Caner y Fan, 2010). De la misma manera que LASSO, es un método de reducción continua, donde se mejora la precisión de la predicción haciendo un balance entre el aumento de sesgo y la disminución de la varianza. Zou (2006) define el LASSO adaptativo como,

$$\hat{\beta}^{*(n)} = \beta \left\| y - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|. \tag{2.20}$$

Se asume que $\hat{\beta}$ es un estimador raíz n-consistente de β^*, como los mínimos cuadrados. Se elige un valor $\gamma > 0$ y se define el vector $\hat{w} = {}^1/_{|\hat{\beta}|^\gamma}$. Para propósito de la explicación de las propiedades oráculo, sea $\mathcal{A}_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0\}$.

2.5.1. Propiedad oráculo del LASSO adaptativo

La elección apropiada de λ_n hace que el LASSO adaptativo tenga la propiedad oráculo.

Teorema 1. *Suponga que* ${}^{\lambda_n}/_{\sqrt{n}} \to 0$ *y* $\lambda_n n^{(\gamma-1)/2} \to \infty$ *entonces los estimadores de LASSO adaptativo deben satisfacer:*

- *Consistencia en selección de variable:* $\lim_n P(\mathcal{A}_n^* = \mathcal{A}) = 1$
- *Normalidad asintótica:* $\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{*(n)} - \beta_{\mathcal{A}}^*) \xrightarrow{d} N(0, \sigma^2 \times C_{11}^{-1})$

Este teorema muestra que la penalización ℓ_1 es al menos igual de efectiva que otras penalizaciones "oráculo". La revisión y demostración de este teorema se encuentra en Zou (2006). Comenta tres conclusiones importantes:

- $\hat{\beta}$ no requiere ser la n-raíz consistente para el LASSO adaptativo. La condición puede debilitarse. Como ejemplo, suponga una secuencia $\{a_n\}$ siendo $a_n \to \infty$ y $a_n(\hat{\beta} - \beta^*) = O_p(1)$. Entonces las propiedades oráculo se mantienen si $\lambda_n = o(\sqrt{n})$ y ${}^{a_n^\gamma \lambda_n}/_{\sqrt{n}} \to \infty$.

- El parámetro $\hat{w}$ de pesos, dependiente de los datos, es clave. Conforme crezca la muestra, los pesos para estimadores de coeficientes cero se inflan hacia infinito, mientras que los pesos de los estimadores de coeficientes diferentes de cero convergen a una constante finita. Por lo que, se puede estimar simultáneamente y de manera insesgada los coeficientes grandes y los coeficientes cercanos al umbral bajo (small threshold).
- Por definición, la solución de LASSO adaptativo es continua. Esta propiedad no es trivial. Sin continuidad, un procedimiento oráculo puede ser subóptimo. Como ejemplo, esto sucede con la penalización ℓ_γ de la regresión bridge $\lambda \sum |\beta_j|^\gamma$. Si la regresión bridge tiene $0 < \gamma < 1$, la propiedad oráculo funciona (Knight y Fu, 2000). Pero si $\gamma < 1$, la solución de bridge no es continua. Lo que resulta en inestabilidad de predicción del modelo.

Existen resultados interesantes sobre la diferencia del comportamiento del LASSO en comparación con el LASSO adaptativo. Ciuperca (2012) concluyó que el LASSO adaptativo funciona mejor y mantiene sus propiedades oráculo, cuando se tienen que seleccionar variables en un modelo de regresión lineal con múltiples puntos de cambio que no suceden en momentos definidos. Lo único necesario para que funcione es que la muestra sea lo suficientemente grande, independientemente de la ocurrencia del punto de cambio.

2.6. Prueba de covarianza

Esta prueba, propuesta por Lockhart et al. (2013) sirve como una prueba de significancia para las variables predictoras que entran al modelo a través de la solución de LASSO. Muestran que, bajo condiciones como el modelo verdadero siendo lineal, esta prueba tiene una distribución asintótica $Exp(1)$ bajo H_0 que las variables activas se encuentran dentro del modelo elegido. Para definir la prueba de covarianza, se necesita considerar el modelo de regresión lineal, para una variable $y \in \Re^n$ y una matriz de variables predictoras $X \in \Re^{n \times p}$:

$$y = X\beta^* + \epsilon, \epsilon \sim N(0, \sigma^2 I), \tag{2.21}$$

donde $\beta^* \in \Re^p$ son coeficientes por ser estimados. Se utiliza el estimador LASSO que se define como (2.17).

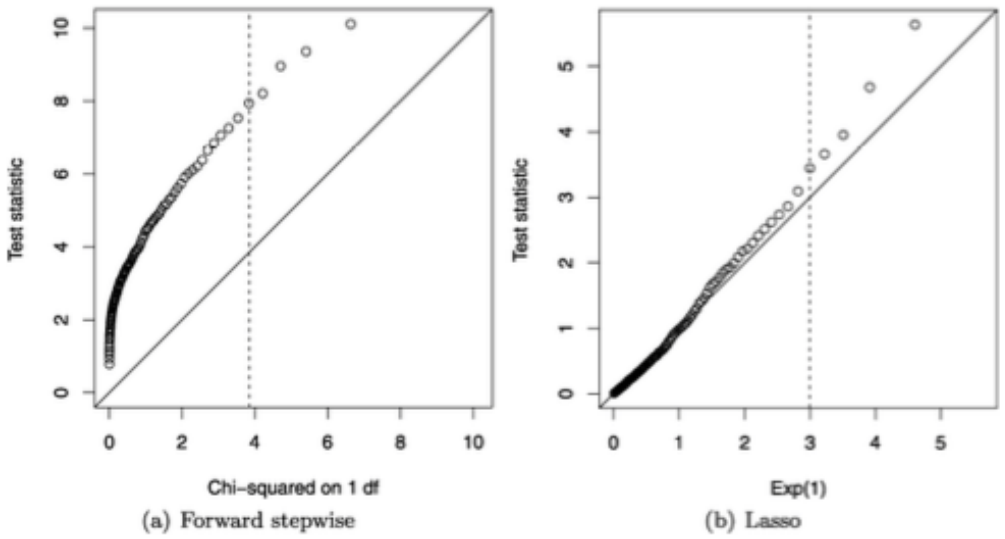

Figura 2.2: Desventaja de χ^2 comparada con T (en Lockhart et al., 2013)

Normalmente, para probar la significancia en el modelo lineal de regresión se opera con dos modelos fijos anidados. Como ejemplo, si M y $M \cup \{j\}$ son subconjuntos fijos de $\{1, \ldots, p\}$, para probar la significancia del j-ésimo predictor en el modelo $M \cup \{j\}$ se utiliza la prueba de la χ^2, que calcula la disminución en la suma de cuadrados residuales entre ambas regresiones,

$$R_j = \frac{(RSS_M - RSS_{M \cup \{j\}})}{\sigma^2} \tag{2.22}$$

y se compara con una χ^2_1 si σ^2 es conocida. En caso de ser desconocida, se reemplaza con la varianza de la muestra, lo que da como resultado a una prueba-F. Sin embargo, esto sólo funciona de manera apropiada cuando M y $M \cup \{j\}$ son fijas. En estos casos, esta condición no aplica, dado que variables entran o salen del modelo a cada paso. Por esta adaptabilidad de los modelos, el error tipo I de la comparación con la χ^2_1 sería mucho mayor al nivel nominal. En la figura 2.6 se observa esta comparación en los cuantiles de la primera entrada en una regresión forward stepwise, se puede observar que al probar con la χ^2_1 en un nivel del 5 %, el error tipo I realmente es del 39 %.

Esto no es un resultado sorprendente, por eso normalmente se utilizan otros criterios para definir cuál es el mejor modelo en vez de una prueba

de hipótesis. Sin embargo, la propuesta por Lockhart et al. (2013) parece ser una solución viable a este problema. Inician con el comentario de que obtienen los resultados de LASSO utilizando el algoritmo LARS (Least Angle Regression) propuesto en Efron et al. (2004), que en uno de sus casos calcula los estimadores LASSO en diferentes valores de λ.

Algunos de las propiedades que se deben considerar del LASSO:

- El paso de los estimadores LASSO $\hat{\beta}(\lambda)$ es continuo y una función linear por parte de λ, con cambios en cada uno de los pasos de lambda que son dependientes de las condiciones de la variable respuesta y y las variables predictoras X.
- En el punto donde $\lambda = \infty$, la solución de $\hat{\beta}(\infty)$ no tiene variables activas en su modelo. Conforme se va reduciendo λ en cada punto entra o sale alguna variable del set activo.
- En cualquier punto del camino de λ, el set activo correspondiente a $A = \sup(\hat{\beta}(\lambda))$ de la solución LASSO son variables predictoras linealmente independientes X_A, que son las columnas de X en A.
- Si una variable de X entra en el modelo activo en un paso de λ, no puede salir del set activo en el paso inmediato siguiente.
- Para matrices X con condiciones de cono positivo, como es el caso de matrices ortogonales, no hay variables que se remuevan del set activo al reducir λ, por ende, el número de pasos es p.

Con esos puntos definidos, la meta de esta prueba es hacer una prueba de significancia de la variable que entra en el set activo. Sea A el set activo justo antes de la entrada de λ_k, se supone que el j-ésimo predictor entra a λ_k. Se define $\hat{\beta}(\lambda k + 1)$ como la solución en el paso λ_{k+1}, usando $A \cup \{j\}$ predictores. A su vez $\tilde{\beta}_A(\lambda_{k+1})$ es la solución sólo utilizando el modelo activo de predictores en X_A, en el paso $\lambda = \lambda_{k+1}$. Esto último es,

$$\tilde{\beta}_A(\lambda_{k+1}) = \underset{\beta_A \in \Re^{|A|}}{\arg\min} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \lambda_{k+1}\|\beta_A\|1. \tag{2.23}$$

Por lo que el estadístico de covarianza se define como

$$T_k = \frac{(\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle)}{\sigma^2} \tag{2.24}$$

Como se observa, el estadístico de covarianza es una función de la diferencia entre $X\hat{\beta}$ y $X_A\tilde{\beta}_A$, de los valores ajustados al incorporar el j-ésimo predictor al set activo, o dejarlo fuera, respectivamente. Los valores ajustados son parametrizados por λ, por lo que es de interés el valor de λ para la evaluación. Si se elige $\lambda = \lambda_k$ como la restricción del set activo, no se puede evaluar la diferencia porque la j-ésima variable tiene un coeficiente de cero a la entrada de λ_k por lo tanto

$$X\hat{\beta}(\lambda_k) = X_A\hat{\beta}_A(\lambda_k) = X_A\tilde{\beta}_A(\lambda_k) \tag{2.25}$$

por ello, se ajusta el valor de $\lambda = \lambda_{k+1}$. Esto permite que el j-ésimo coeficiente tenga su efecto en el ajuste de $X\hat{\lambda}$ antes de la entrada en la siguiente variable en λ_{k+1}.

Además, el estadístico de covarianza utiliza el producto interior de la diferencia con y. El nombre de esta prueba proviene de este producto que se puede pensar como la covarianza.

$$\begin{aligned} T_k =& \langle y - \mu, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y - \mu, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle \\ &+ \langle \mu, X\hat{\beta}(\lambda_{k+1}) - X_A\tilde{\beta}_A(\lambda_{k+1})\rangle \end{aligned} \tag{2.26}$$

Al expander $y = y - \mu + \mu$ con $\mu = X\beta^*$. Cuando X es ortogonal el último término de 2.26 es cero bajo la nula. Mientras más grande sea la covarianza de y con $X\hat{\beta}$ comparada con $X_A\tilde{\beta}_A$, mayor la importancia de la j-ésima variable que entra al modelo. Otra ventaja más, este estadístico admite una distribución nula asimptótica que es simple y exacta. Por lo tanto, bajo la nula que el modelo LASSO contiene los predictores activos, $A \supseteq \text{soporte}\beta^*$,

$$T_k \xrightarrow{d} \text{Exp}(1) \tag{2.27}$$

se deben considerar suposiciones razonables en X y las magnitudes de los coeficientes verdaderos.

Capítulo 3

Método

En este capítulo se describirá la metodología de la investigación. Se describirán los instrumentos y procedimientos que se utilizaron para hacer la comparación entre las herramientas descritas en el capítulo anterior. Se iniciará con la explicación del procedimiento a realizar, una breve descripción del hardware y el software empleados, se expondrá el bosquejo de la operación de las funciones creadas para el programa y finalmente se comentarán las condiciones de simulación.

3.1. Procedimiento

Para comparar el comportamiento de MC, DC, LASSO, LASSO adaptativo (ada LASSO), prueba de covarianza (CovTest), se diseñó un estudio de simulación. Dicho estudio compara la forma de seleccionar variables con las cinco medidas en tres casos diferentes: modelo lineal simple, colinealidad entre los predictores y colinealidad tipo toeplitz. El criterio de selección del mejor modelo se hizo con validación cruzada.

Una vez que se simulan los datos, cada uno de los procedimientos hace la elección del modelo estimado. Este procedimiento se repitió 1000 veces con muestras de tamaño $n = 100$. Los coeficientes estimados se capturan para su posterior análisis.

3.2. Consideraciones generales de hardware y software

Para el presente trabajo se utilizó una computadora MacBook Pro de 15 pulgadas del año 2012 que cuenta con 8 GB 1600 MHz DDR3 de memoria RAM. El istema operativo es OS X El Capitan versión 10.11.6. Para llevar a cabo la simulación y análisis se utilizó el software R (R Core Team, 2016).

Las librerías empleadas para generar el progama de análisis fueron: energy (Rizzo y Székely, 2014), MPV (Braun, 2015), mvtnorm (Genz et al., 2016), Matrix (Maechler, 2016), acepack (Spector et al., 2014), glmnet (Friedman et al., 2016), lars (Efron, 2013), covTest (Lockhart et al., 2013) y Rcmdr (Fox et al., 2016). Cada una de estas librerías se utilizó con sus respectivas dependencias.

3.3. Descripción general del algoritmo

El programa se realizó con los siguientes pasos:

- Cada procedimiento (DC, MC, etc.) se programó como una función.
- Se simula la mátriz diseño X y los coeficientes β.
- Se crean las matrices que almacenarán los resultados de cada procedimiento.
- A partir de este punto, se inicia la parte que se repite 1000 veces del programa:
    - Se calcula el error ϵ y la variable respuesta Y.
    - Se utiliza cada función de cada procedimiento para analizar los datos.
    - Se extraen los valores de los coeficientes elegidos por cada función.
- Se analizan los resultados.

Las funciones se explican a continuación.

3.3.1. MC y DC

La razón para juntar estos dos procedimientos es que, salvo por la forma de calcular la fuerza de la asociación entre variables, el procedimiento para seleccionar variables es exactamente el mismo. Por ende, sólo se describirá uno de los métodos pero aplica para ambos.

Se decribe el método para MC:

A) Se toma cada variable predictora $X_1, \ldots, X_p$ y se calcula su MC con la variable respuesta Y.

B) Se elige la variable X_k que tenga mayor MC con Y, utilizando el algoritmo ACE.

C) Con la X_k seleccionada, se hace una regresión lineal con la variable respuesta $Y \sim X_k$ y cada variable predictora $X_1, \ldots, X_{p-1}$ restantes con X_k.

D) De estas regresiones se extraen los residuales de la variable respuesta a los que se denominará rY, a los residuales con las variables independientes se les denominará como $rX_1, \ldots, rX_{p-1}$.

E) Se vuelve a calcular el MC ahora entre cada $rX_1, \ldots, rX_{p-1}$ y los residuales rY.

F) As se repiten los pasos (B), (C), (D) y (E) hasta que todas las variables predictoras hayan sido elegidas y se encuentren ordenadas en Q.

G) Se calcula paso a paso el error de validación cruzada metiendo en el modelo las variables en el orden establecido en Q.

H) Se elige el modelo que tenga el menor error de validación cruzada.

I) Se registra en una matriz los valores de los coeficientes estimados $\hat{\beta}$.

En el caso de DC, la única diferencia con el otro procedimiento es que se calcula el DC en el paso (B) con el uso del procedimiento "dcor" que es parte del paquete energy (Rizzo y Székely, 2014).

3.3.2. LASSO, LASSO adaptativo y prueba de covarianza

En estos tres procedimiento hay cierta similaridad, los procedimientos se detallan a continuación.

En el caso de LASSO, utilizando el paquete de glmnet (Friedman et al., 2016) fijando los valor de α para que realizara el cálculo del LASSO. Se extraen los coeficientes estimados $\hat{\beta}_{LASSO}$ para posterior análisis.

El procedimiento de LASSO adaptativo fue implementado utilizando el código desarrollado por Stefanski y Boos (2007). Después de verificar los datos, calcula $\hat{w}$, se ajustan los datos con respecto a esta $\hat{w}$ y a partir de ahí se calculan los coeficientes.

En el caso de la prueba de covarianza, primero se calcula el camino de LASSO a través del paquete lars (Efron, 2013). Posterior a esto, se realiza el cálculo del estadístico de prueba a la entrada de cada variable al modelo. La regla de decisión se definió como $T_k < \alpha$, si es así, la variable entra al modelo y se prueba la siguiente variable. En el caso de $T_k > \alpha$ entonces se elige el modelo sin agregar más variables. El valor α se fijó en 0.05. A partir del modelo elegido se calculan los coeficientes.

3.4. Condiciones de simulación

Parte de las condiciones de simulación fueron elegidas con base en el trabajo de Yenigün y Rizzo (2015). Para cada caso, se realizaron 1000 repeticiones con muestras tamaño $n = 100$.

Modelo lineal simple

La primera condición de simulación fue un modelo lineal donde $X_1, \ldots, X_p \sim \text{iid}N(0, 1)$ y $\epsilon \sim N(0, 2)$. El número de variables se fijó como $p = 8$ y $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$. Por lo tanto, y quedó definida como:

$$y = X\beta + \epsilon$$

Colinealidad constante entre predictores

La segunda condición es con colinealidad constante entre predictores con $p = 8$, de una distribución normal multivariada con $X \sim N_p(0, \Sigma)$ donde:

$$\Sigma = \begin{bmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \theta & \cdots & 1 \end{bmatrix}$$

Al igual que en la primera condición, $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ y $\epsilon \sim N(0, 2)$. Este caso se realizó en tres ocasiones, el valor de θ se fijó en 0.6, 0.8 y 0.9 respectivamente. Donde y quedó definida como:

$$y = X\beta + \epsilon$$

Colinealidad tipo Toeplitz entre predictores

La tercera condición fue muy similar a la condición de colinealidad sobre la distribución multivariada de $X \sim N_p(0, \Sigma)$ con la diferencia que Σ se definió como una matriz tipo Toeplitz, de la siguiente manera:

$$\Sigma = \begin{bmatrix} 1 & \theta & \theta^2 & \cdots & \theta^{p-1} \\ \theta & 1 & \theta & \cdots & \theta^{p-2} \\ \theta^2 & \theta & 1 & \cdots & \theta^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta^{p-1} & \theta^{p-2} & \theta^{p-3} & \cdots & 1 \end{bmatrix}$$

El error $\epsilon \sim N(0, 2)$, $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ y $\theta = 0.6$. La variable respuesta quedó definida como

$$y = X\beta + \epsilon$$

Criterio de evaluación de los modelos

El criterio para elegir el modelo fue validación cruzada sacando una observación como lo define James et al. (2013). donde se entrena el modelo con $n - 1$ observaciones y se estima $\hat{y}_j$ con la x_j que se sacó del modelo. Se calcula el error cuadrado medio para esa observación $\text{ECM}_j = (y_j - \hat{y}_j)^2$. Esto se repite con cada observación hasta que se tenga $\text{ECM}_1, \ldots, \text{ECM}_n$, para finalmente obtener el estimador de validación cruzada, definido como

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{ECM}_i$$

Análisis de los resultados

Los resultados se analizaron de la siguiente manera:

- Se contaron las veces que se eligió cada una de las variables por los métodos de selección de variable. Se nombraron 1^{era}, 2^{nda} y 3^{era} respectivamente.
- Se contó cada vez que se eligieron al menos las tres variables, sin importar si las técnicas insertaban otras variables. La condición se nombró "al menos".
- Se realizó un conteo de cada ocasión en la que se eligieron únicamente las tres variables del modelo. Se nombró como "hit rate".
- Se calculó la varianza y el sesgo por cada uno de los estimadores $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ por cada una de las técnicas.
- Finalmente se calculó el error de predicción:

$$\text{Error Pred} = E[(Y - \hat{Y})^2]$$

Capítulo 4

Resultados y Discusión

Los resultados se separan en dos secciones. La primera sección contiene la comparación de tres indicadores del desempeño de las técnicas de selección de variable para todas las condiciones de simulación. La segunda sección muestra los resultados de conteo, varianza y sesgo de cada variable correcta elegida, estos resultados se presentan separados por cada condición de simulación.

4.1. Comparación general del desempeño

Las primeras dos figuras (4.1 y 4.2) muestran el desempeño de las pruebas para los conteos de al menos y "hit rate", mientras que la figura 4.3 compara el error de predicción. En el caso de la figura 4.1 se cuenta la cantidad de veces que al menos se eligieron las tres variables correctas, sin importar si se eligieron más variables al modelo. La figura 4.2 el conteo se realiza sólo cuando se eligieron las tres variables correctas, esta condición se identificó como "hit rate" en la metodología.

4.1.1. Condición de al menos y "hit rate"

En la figura 4.1 se observa el comportamiento de las cinco pruebas para elegir al menos las tres variables correctas. Se puede observar que la prueba con mejor comportamiento en esta condición es LASSO. En el caso de la prueba de covarianza se puede observar que tuvo menor desempeño a comparación del resto de las pruebas, y a lo largo de los resultados se observará este comportamiento. Se puede observar que el comportamiento de todas las

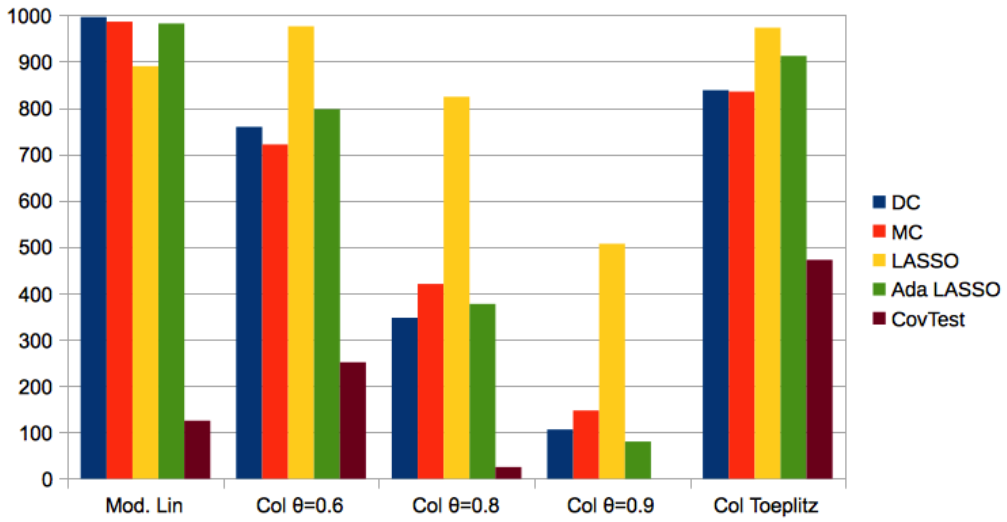

Figura 4.1: Conteo de al menos

técnicas (menos covtest) es bueno en la simulación del modelo lineal. En el caso de colinealidad, mientras θ aumenta las técnicas van fallando cada vez más, la prueba que mejor responde ante esta circunstancia de simulación es LASSO. Finalmente en el caso de la colinealidad tipo toeplitz las pruebas mejoran sus resultados con respecto al caso de colinealidad fijo, específicamente covtest muestra su mejor desempeño en esta condición.

Los resultados anteriores tienen cierta similaridad a los presentados por Yenigün y Rizzo (2015), ellos encontraron resultados parecidos en el caso del modelo lineal, colinealidad con $\theta = 0.6$ y colinealidad toeplitz. Sin embargo, ellos no aumentaron el valor de θ para colinealidad, por lo que no observaron cómo DC y MC empeoraron a comparación de LASSO.

En el caso del conteo de "hit rate" (figura 4.2) se pueden ver diferencias aún más dramáticas entre DC y MC con respecto a LASSO y LASSO adaptativo. Tanto DC como MC insertan más variables en su selección del modelo, y conforme las circunstancias de simulación presentan colinealidad su comportamiento se vuelve peor. En general se observa que LASSO adaptativo elige el modelo correcto la mayor cantidad de veces tanto en el modelo lineal como en los casos de colinealidad fija. En el caso de colinealidad toeplitz LASSO elige el modelo correcto más veces.

Estos resultados son los que muestran bastante diferencia cuando se com-

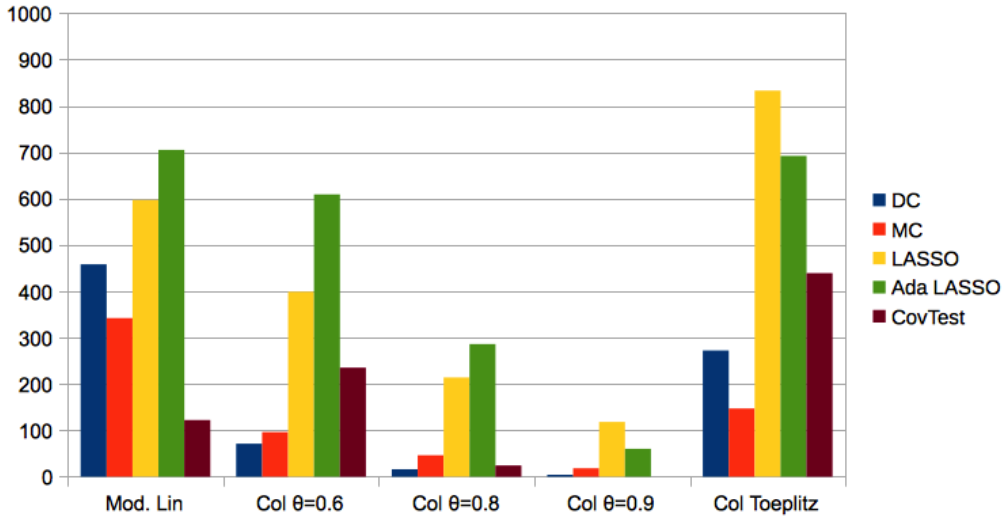

Figura 4.2: Conteo de "Hit rate"

paran con el experimento de Yenigün y Rizzo (2015). Esto sucede porque ellos llaman "hit rate" al conteo que en esta investigación se le llama al menos. En el caso de la prueba de covarianza, a pesar de que tuvo el peor desempeño, se observa un resultado interesante, los resultados son casi los mismos en el conteo de al menos y de "hit rate", lo que significa que si llegaba a elegir las tres variables, normalmente no elige más variables al modelo, a diferencia de lo que pasa con el resto de las pruebas.

4.1.2. Error de predicción

Se muestra en la figura 4.3 el error de predicción de todas las técnicas por cada una de las condiciones de simulación. La prueba de covarianza tuvo el mayor error entre todas las técnicas y condiciones de simulación. Por otra parte, se puede ver que el resto de las técnicas tienen un comportamiento más o menos constante en las diferentes condiciones de simulación. Siendo LASSO la prueba que se comportó peor, seguido de MC y DC. LASSO adaptativo mostró el menor error de predicción en todas las condiciones de simulación. La diferencia en el error de predicción entre LASSO y LASSO adaptativo es comentada por Zou (2006).

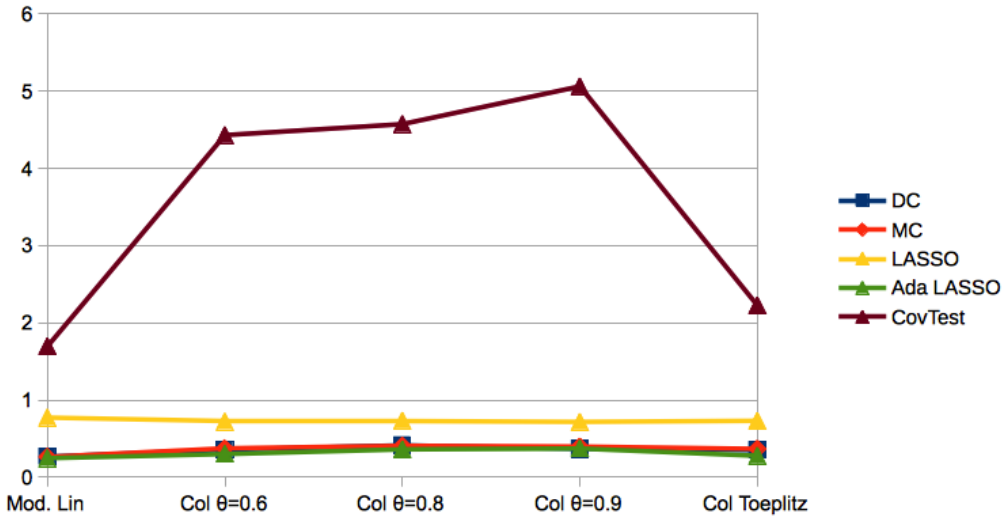

Figura 4.3: Error de predicción

4.2. Resultados por cada condición de simulación

Se presentan a continuación los resultados de las técnicas de selección de variable por cada condición de simulación.

4.2.1. Caso 1. Modelo Lineal

En relación a la selección de variables, en la tabla 4.1 se puede observar que la prueba de covarianza tuvo el peor comportamiento. Esto es, en general eligió menos veces las variables correctas. El resto de las pruebas se comportaron relativamente igual al elegir cada una de las variables individualmente, siendo ligeramente superior los métodos de subconjuntos (DC y MC) en seleccionar un modelo con al menos las variables importantes, sin embargo, los métodos de regresión LASSO fueron superiores en seleccionar exactamente el modelo correcto (”hit rate”), entre ellos, el LASSO adaptativo fue el mejor (como se observó en las figuras 4.1 4.2).

En cuanto a la estimación, en 4.2 se observa nuevamente que la prueba de covarianza tuvo el peor desempeño en lo que se refiere al sesgo. En general, todas las pruebas tuvieron poca varianza en los estimadores. En cuanto al

Método	1^{era}	2^{nda}	3^{era}
MC	995	996	995
DC	1000	999	997
LASSO	943	949	931
Ada LASSO	992	997	993
CovTest	271	265	229

Cuadro 4.1: Tabla con los resultados de la selección de variables en la condición de modelo lineal. Las primeras tres columnas cuentan la entrada de cada una de las tres variables al modelo seleccionado. El valor máximo de conteo es 1000.

Método	Var $\hat{\beta}_1$	Sesgo $\hat{\beta}_1$	Var $\hat{\beta}_2$	Sesgo $\hat{\beta}_2$	Var $\hat{\beta}_3$	Sesgo $\hat{\beta}_3$
MC	0.053	0.002	0.050	0.010	0.053	−0.005
DC	0.049	0.005	0.048	0.012	0.051	−0.004
LASSO	0.059	−0.539	0.056	−0.531	0.059	−0.566
ada LASSO	0.061	−0.100	0.058	−0.094	0.063	−0.114
CovTest	0.076	−0.860	0.079	−0.856	0.066	−0.880

Cuadro 4.2: Varianza y sesgo de los estimadores por método de estimación en el modelo lineal.

sesgo, tanto LASSO como la prueba de covarianza tuvieron un peor desempeño a comparación de las otras medidas, donde MC y DC tuvieron un mejor desempeño.

4.2.2. Caso 2. Colinealidad

En la siguiente sección se muestran los resultados para la condición de colinealidad entre los predictores. Se separan los resultados dependiendo del valor fijado para θ, ya sea 0.6, 0.8 ó 0.9.

Con $\theta = 0.6$

Al existir colinealidad entre predictores (tabla 4.3), las pruebas de MC, DC, LASSO y LASSO adaptativo disminuyen su capacidad de seleccionar las variables del modelo. A pesar de que MC y DC parecen elegir correctamente las variables individualmente, presentan una notoria disminución del "hit rate" (figura 4.2). LASSO, por su parte, se comporta mucho mejor en esta prueba que LASSO adaptativo en cuanto a la selección, sin embargo tiende

Método	1^{era}	2^{nda}	3^{era}
MC	855	910	951
DC	875	923	958
LASSO	989	994	993
Ada LASSO	894	944	958
CovTest	553	646	490

Cuadro 4.3: Tabla de conteos del caso con colinealidad con $\theta = 0.6$.

Método	Var $\hat{\beta}_1$	Sesgo $\hat{\beta}_1$	Var $\hat{\beta}_2$	Sesgo $\hat{\beta}_2$	Var $\hat{\beta}_3$	Sesgo $\hat{\beta}_3$
MC	0.255	-0.034	0.190	-0.000	0.129	0.003
DC	0.234	0.009	0.169	0.022	0.117	0.027
LASSO	0.100	-0.276	0.082	-0.246	0.073	-0.317
ada LASSO	0.216	-0.069	0.166	-0.030	0.126	-0.057
CovTest	0.184	-0.602	0.171	-0.569	0.140	-0.699

Cuadro 4.4: Varianza y sesgo de los estimadores por método de estimación con colinealidad entre predictores $\theta = 0.6$.

a meter más variables al modelo que no son parte del modelo real. Con esto, es notoria la superioridad del LASSO adaptativo en el número de veces que selecciona exactamente el modelo correcto ("hit rate"). Estas diferencias se pueden observar en las figuras 4.1 y 4.2 de la sección anterior.

En la tabla 4.4, un resultado esperado es el incremento en la varianza de los estimadores al compararlo con la condición del modelo lineal. Se observa que LASSO tuvo la menor varianza entre los estimadores, sin embargo, su sesgo es el más elevado, sin considerar a la prueba de covarianza. Aún así, mostró una disminución en el sesgo a comparación de la condición lineal. Tanto DC como MC tuvieron el menor sesgo de los estimadores.

Con $\theta = 0.8$

A medida que aumenta la colinealidad entre las variables predictoras (tabla 4.5), las técnicas disminuyen la selección de las variables correctas. En este caso LASSO es la técnica que idenfica mejor las variables de manera individual, mientras que el LASSO adaptativo encuentra el modelo en más ocasiones que los otros métodos. Es notorio que tanto MC, DC y LASSO disminuyen su rendimiento más como consecuencia del incremento en coli-

Método	1^{era}	2^{nda}	3^{era}
MC	804	763	722
DC	757	760	718
LASSO	956	957	908
Ada LASSO	753	811	724
CovTest	540	381	311

Cuadro 4.5: Tabla de conteos del caso con colinealidad entre predictores con $\theta = 0.8$.

Método	Var $\hat{\beta}_1$	Sesgo $\hat{\beta}_1$	Var $\hat{\beta}_2$	Sesgo $\hat{\beta}_2$	Var $\hat{\beta}_3$	Sesgo $\hat{\beta}_3$
MC	0.391	−0.020	0.375	−0.088	0.443	−0.100
DC	0.418	−0.016	0.383	−0.048	0.462	−0.061
LASSO	0.151	−0.303	0.125	−0.380	0.156	−0.402
ada LASSO	0.428	−0.097	0.319	−0.114	0.410	−0.162
CovTest	0.173	−0.643	0.125	−0.768	0.125	−0.804

Cuadro 4.6: Varianza y sesgo de los estimadores por método de estimación con colinealidad entre predictores $\theta = 0.8$.

nealidad, dado que tienden a meter variables demás al modelo estimado.

La tabla 4.6 muestra un aumento en los sesgos y las varianzas de los predictores con respecto a la condición con $\theta = 0.6$. Nuevamente el error de predicción de la prueba de covarianza es el peor. Otro patrón que empieza a emerger es que los errores de predicción son relativamente similares en todas las condiciones de simulación. LASSO muestra poca varianza pero más sesgo que las otras técnicas.

Con $\theta = 0.9$

En esta condición de simulación todas las pruebas se comportaron de forma pobre (tabla 4.7). Aún así, LASSO tuvo el mejor comportamiento sobre la elección individual de las variables predictoras. Su comportamiento es aún mejor que el de LASSO adaptativo, que se ha comportado mejor en las otras condiciones de simulación. Esta es la condición donde la prueba de covarianza tuvo el peor comportamiento rendimiento. Las técnicas a través de subconjuntos también fallaron mucho, donde la elección de las variables fue mucho menor que en las otras condiciones de simulación.

En la tabla 4.8 se observa un aumento en el sesgo de todas las pruebas

Método	1^{era}	2^{nda}	3^{era}
MC	537	607	575
DC	512	656	594
LASSO	761	862	830
Ada LASSO	504	583	548
CovTest	371	375	280

Cuadro 4.7: Tabla de conteos del caso con colinealidad entre predictores con $\theta = 0.9$.

Método	Var $\hat{\beta}_1$	Sesgo $\hat{\beta}_1$	Var $\hat{\beta}_2$	Sesgo $\hat{\beta}_2$	Var $\hat{\beta}_3$	Sesgo $\hat{\beta}_3$
MC	0.944	-0.145	0.739	-0.092	0.722	-0.161
DC	0.907	-0.138	0.624	-0.024	0.669	-0.115
LASSO	0.252	-0.449	0.236	-0.346	0.225	-0.414
ada LASSO	0.910	-0.181	0.739	-0.131	0.716	-0.213
CovTest	0.200	-0.733	0.198	-0.727	0.168	-0.786

Cuadro 4.8: Varianza y sesgo de los estimadores por método de estimación con colinealidad entre predictores $\theta = 0.9$.

en comparación al resto de las condiciones de colinealidad. A pesar de este aumento, el error de predicción se mantuvo relativamente similar entre colinealidad $\theta = 0.8$ y $\theta = 0.9$. La varianza también tuvo un incremento entre los estimadores, aunque la colinealidad parece haber afectado menos al LASSO con respecto a las otras técnicas de selección de variable.

4.2.3. Caso 3. Colinealidad Toeplitz

En el caso de colinealidad Toeplitz (tabla 4.9), todas las técnicas tuvieron un mejor desempeño que en el caso de colinealidad constante con $\theta = 0.6$. En este caso el LASSO fue la prueba que eligió en más ocasiones a las variables de manera individual, a su vez predijo en más ocasiones el modelo real (como se observa en la figura 4.2). Esta condición fue en la que mejor se comportó la prueba de covarianza. MC y DC predijeron el modelo exacto en menos ocasiones pero eligieron en múltiples ocasiones las variables correctas, metiendo más variables de las que tiene el modelo.

En la tabla 4.10 se observan tendencias similares a las otras condiciones sobre el aumento del sesgo en el caso de LASSO. Estos resultados se asemejan al caso de colinealidad fija con $\theta = 0.6$. LASSO tiene menor varianza en los

Método	1^{era}	2^{nda}	3^{era}
MC	910	952	970
DC	913	956	968
LASSO	987	1000	985
Ada LASSO	957	982	971
CovTest	542	905	515

Cuadro 4.9: Tabla de conteos del caso con colinealidad Toeplitz con $\theta = 0.6$.

Método	Var $\hat{\beta}_1$	Sesgo $\hat{\beta}_1$	Var $\hat{\beta}_2$	Sesgo $\hat{\beta}_2$	Var $\hat{\beta}_3$	Sesgo $\hat{\beta}_3$
MC	0.164	-0.034	0.128	0.007	0.097	-0.011
DC	0.158	-0.032	0.120	0.008	0.096	-0.008
LASSO	0.074	-0.337	0.060	-0.155	0.058	-0.427
ada LASSO	0.128	-0.069	0.108	-0.003	0.091	-0.102
CovTest	0.230	-0.535	0.122	-0.275	0.175	-0.611

Cuadro 4.10: Varianza y sesgo de los estimadores por método de estimación con colinealidad Toeplitz con $\theta = 0.6$.

predictores, el sesgo sigue siendo menor para MC, DC y LASSO adaptativo. Entre todas las condiciones de simulación, en esta condición fue donde la prueba de covarianza se comportó mejor, tanto de agregar las variables en el modelo como de tener menor sesgo.

Las técnicas de selección por subcojuntos DC y MC mostraron evidencia de tener un error de predicción bajo, a pesar de no elegir el modelo correcto seleccionado con la misma frecuencia que LASSO o LASSO adaptativo. En el trabajo de Yenigün y Rizzo (2015) se observaban resultados mucho mejores de DC y MC en cuanto a la selección del modelo. Esto se puede deber a que ellos consideraron la condición "al menos" como la predicción del modelo, además de sólo simular colinealidad con $\theta = 0.6$. En esos casos, se puede ver un comportamiento muy similar entre la selección de LASSO con DC y MC.

El resultado que muestra LASSO en todas las condiciones de simulación es consistente con los resultados de Knight y Fu (2000) donde observaron que LASSO puede elegir los predictores reales si la cantidad de datos es suficiente, sin embargo tiende a elegir modelos con variables extras. También es consistente con los resultados de Zhao y Yu (2006), muestran que la propie-

dad de consistencia no necesariamente se traduce en buen rendimiento. Esto se observa por resultados como el error de predicción. Es interesante que este error se mantuviera similar en todas las condiciones.

Por su parte, el LASSO adaptativo muestra ventajas sobre LASSO en cuanto a la selección de modelos, además de tener un menor error de predicción. Zou (2006) comenta que, al igual que en LASSO, el desempeño del modelo seleccionado no está sujeto a las propiedades oráculo. Sin embargo, la técnica muestra mejores resultados por su manejo de la penalización ponderada ℓ_1. A pesar de haber tenido mejores resultados, cuando la colinealidad es alta, el LASSO adaptativo también tiende a fallar, esto es consistente con resultados obtenidos por Zou y Zhang (2009).

Finalmente, a pesar de que la prueba de covarianza parece ser una mala elección tanto en selección de modelo como en predicción, se debe considerar que no es necesariamente una prueba comparable con las otras. Se puede mejorar la selección de variable y reducir el error de predicción si se ajusta el valor de α. Lockhart et al. (2014) comentan en sus conclusiones que se están realizando pruebas para encontrar los p-valores validos para este procedimiento.

Capítulo 5

Conclusiones

La presente investigación intentó explorar las diferencias que presentan varios métodos de selección de variable en diferentes condiciones de simulación. La motivación era obtener o ampliar un poco la evidencia que existe sobre estos métodos.

Las conclusiones de la presente investigación son:

- Las pruebas basadas en penalización funcionaron mejor para la selección de variables, sin embargo, para la predicción DC y MC se comportaron mejor que LASSO y parecido a LASSO adaptativo.
- Considerando la evidencia, la mejor técnica para selección de variable es el método del LASSO adaptativo. Tanto por su bajo error de predicción como su propiedad oráculo. Salvo en el caso de alta colinealidad entre predictores, donde es mejor utilizar LASSO.
- DC y MC muestran error de predicción bajo pero no son buenas medidas para seleccionar el modelo correcto.
- La prueba de covarianza mostró resultados poco efectivos pero no se debe descartar su importancia, considerando que es una buena propuesta para realizar pruebas de hipótesis para seleccin de modelos.

5.1. Limitaciones y áreas de oportunidad

La presente investigación muestra resultados interesantes en la comparación de las técnicas de selección de variables. Sin embargo, existen varias

limitaciones que pueden ser consideradas en futura investigación, así como áreas de oportunidad:

- Las simulaciones que se llevaron a cabo para este proyecto fueron basadas en el modelo lineal. Considerando que DC y MC pueden detectar asociación entre variables aún en condiciones no lineales, sería interesante comparar las técnicas tanto en modelos no lineales como en modelos lineales generalizados.

- En la presente investigación se trabajó en la condición de la muestra de tener $n > p$, sin embargo, hay una creciente demanda por el estudio de la condición contraria cuando $p > n$. Hay un creciente interés para tratar de buscar técnicas que soporten menos observaciones.

- Sería interesante hacer la comparación de modelos para su selección utilizando otros criterios además de validación cruzada. Utilizando AIC, BIC ó R^2_{adj}.

- En el caso de implementar los otros criterios, también sería recomendado utilizar un set de entrenamiento y un set para la prueba.

Literatura Citada

Braun, W. J. (2015). MPV: Data Sets from Montgomery, Peck and Vining's Book.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384.

Breiman, L. y Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Bryc, W., Dembo, A., y Kagan, A. (2002). On The Maximum Correlation Coefficient. Technical Report 2002-25, Stanford University.

Caner, M. y Fan, Q. (2010). The adaptive lasso method for instrumental variable selection. Technical report, Working Paper, North Carolina State University.

Chen, S. S., Donoho, D. L., y Saunders, M. A. (1996). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

Ciuperca, G. (2012). Model selection by LASSO methods in a change-point model. *Statistical Papers*, 55(2):349–374.

Clarke, B., Fokoue, E., y Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Derksen, S. y Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3(4):409–425.

Ducci, L., Agnelli, P., Di Febbraro, M., Frate, L., Russo, D., Loy, A., Carranza, M. L., Santini, G., y Roscioni, F. (2015). Different bat guilds perceive their habitat in different ways: a multiscale landscape approach for variable selection in species distribution modelling. *Landscape Ecology*, 30(10):2147–2159.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., y others (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Efron, T. H. a. B. (2013). lars: Least Angle Regression, Lasso and Forward Stagewise.

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

Fan, J. y Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*, 20(1):101–148.

Fox, J., Bouchet-Valat, M., Andronic, L., Ash, M., Boye, T., Calza, S., Chang, A., Grosjean, P., Heiberger, R., Pour, K. K., Kerns, G. J., Lancelot, R., Lesnoff, M., Ligges, U., Messad, S., Maechler, M., Muenchen, R., Murdoch, D., Neuwirth, E., Putler, D., Ripley, B., Ristic, M., Wolf, P., y Wright, K. (2016). Rcmdr: R Commander.

Frank, l. E. y Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., Simon, N., y Tibshirani, R. (2016). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

Furnival, G. M. y Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics*, 16(4):499.

Geng, Z., Wang, S., Yu, M., Monahan, P. O., Champion, V., y Wahba, G. (2015). Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study: Group Variable Selection Via LES Penalty. *Biometrics*, 71(1):53–62.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., y Hothorn, T. (2016). mvtnorm: Multivariate Normal and t Distributions.

Gerretzen, J., Szymańska, E., Bart, J., Davies, A. N., van Manen, H.-J., van den Heuvel, E. R., Jansen, J. J., y Buydens, L. M. (2016). Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Analytica Chimica Acta*, 938:44–52.

Gross, S. M. y Tibshirani, R. (2016). Data Shared Lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235.

Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer Series in Statistics. Springer International Publishing, Cham.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Hussami, N. y Tibshirani, R. (2013). A Component Lasso. *arXiv:1311.4472 [cs, stat]*.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer Texts in Statistics. Springer New York, New York, NY.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Knight, K. y Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

Liébana-Cabanillas, F., Herrera, L., y Guillén, A. (2016). Variable selection for payment in social networks: Introducing the Hy-index. *Computers in Human Behavior*, 56:45–55.

Lockhart, R., Taylor, J., Tibshirani, R., y Tibshirani, R. (2013). covTest: Computes covariance test for adaptive linear modelling.

Lockhart, R., Taylor, J., Tibshirani, R. J., y Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

Maechler, D. B. a. M. (2016). Matrix: Sparse and Dense Matrix Classes and Methods.

Melamed, C. (2014). The data revolution is coming and it will unlock the corridors of power. *The Guardian*.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press. Google-Books-ID: 7p59iir822sC.

Ng, S. (2013). Variable selection in predictive regressions. *Handbook of Economic Forecasting*, 2(Part B):752–789.

R Core Team (2016). R: A language and environment for statistical computing.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rizzo, M. L. y Székely, G. J. (2014). energy: E-statistics (energy statistics).

Sheather, S. (2009). *A Modern Approach to Regression with R*. Springer Texts in Statistics. Springer New York, New York, NY.

Spector, P., Friedman, J., Tibshirani, R., y Lumley, T. (2014). acepack: ace() and avas() for selecting regression transformations.

Stefanski, L. y Boos, D. (2007). Adaptative LASSO software for R.

Székely, G. J., Rizzo, M. L., y Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

van Reenen, M., Reinecke, C. J., Westerhuis, J. A., y Venter, J. H. (2016). Variable selection for binary classification using error rate p-values applied to metabolomics data. *BMC Bioinformatics*, 17(1).

Yenigün, C. D. y Rizzo, M. L. (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, 85(8):1692–1705.

Zhang, H. y Zamar, R. H. (2014). Least angle regression for model selection. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(2):116–123.

Zhao, P. y Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

Zou, H. y Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733–1751.

Anexo A

Anexo: Programa

A.1. Funciones

```
#############
#Librerias:
#############
library("energy")
library("MPV")
library("mvtnorm")
library("Matrix")
library("acepack")
library("glmnet")
library("lars")
library("covTest")

############
#Funciones:
############
############
#DC: in Full
############
DC <- function(X,Y){

####Eligiendo la primera variable
  Dce <- c(rep(0,ncol(X)))
```

```
  dece1 <- c(rep(0,ncol(X)))
  for(i in 1:ncol(X)){
    a <- dcor(X[,i],Y)
    dece1[i] <- a
  }
  p <- which.max(dece1)
  Dce[1] <- p
  varnumber <- seq(1,ncol(X),1)
  eq <- as.data.frame(X)
  colnames(eq)<-varnumber
  eq1 <- eq[,-p]
  eq2 <- eq[,p]
  g <- Y
  j <- 1

####eligiendo de la 2nda a la p-2 variable
  while (ncol(as.matrix(eq1)) > 2){
    Xar <- matrix((rep(0)),nrow(eq1),ncol(eq1))
    g <- lm(g~eq2)
    gr <- g$residuals
    for(i in 1:ncol(eq1)){
      Xa <- lm(eq1[,i]~eq2)
      Xar[,i] <- Xa$residuals
    }
    dece1 <- c(rep(0,ncol(Xar)))
    for(i in 1:ncol(Xar)){
      a <- dcor(Xar[,i],gr)
      dece1[i] <- a
    }
    p <- which.max(dece1)
    Dce[j+1] <- as.integer(colnames(eq1)[p])
    eq2 <- eq1[,p]
    eq1 <- eq1[,-p]
    g <- gr
    j <- j+1
  }
####eligiendo la pen\'ultima variable
  Xar <- matrix((rep(0)),nrow= nrow(eq1),ncol= 2)
```

```
  g <- lm(g~eq2)
  gr <- g$residuals
  for(i in 1:ncol(Xar)){
    Xa <- lm(eq1[,i]~eq2)
    Xar[,i] <- Xa$residuals
  }
  dece1 <- c(rep(0,ncol(Xar)))
  for(i in 1:ncol(Xar)){
    a <- dcor(Xar[,i],gr)
    dece1[i] <- a
  }
  p <- which.max(dece1)
  Dce[j+1] <- as.integer(colnames(eq1)[p])
 ####eligiendo la ltima variable (trampita)
  Dce[length(Dce)] <- as.integer(colnames(eq)[-Dce])
 ####reacomodando la matriz de datos
  eq <- eq[,match(Dce,colnames(eq))]

 ####el PRESS
  prvec <- c(rep(0,ncol(eq)))
  gog <- as.matrix(eq)
  for(i in 1:ncol(gog)){
    mod <- lm(Y~gog[,1:i])
    pressmo <- PRESS(mod)
    prvec[i] <- pressmo
  }
  finmod <- lm(Y~gog[,1:which.min(prvec)])
  qe <- finmod$coefficients
  names(qe) <- c("Intercept",Dce[1:which.min(prvec)])
####data from PRESS
  return(list("PRDC"=qe,"DCv"=Dce))
}
###############
#MC: in Full
###############
MC <- function(X,Y){
####Eligiendo la primera variable
  Mce <- c(rep(0,ncol(X)))
```

```
  emece1 <- c(rep(0,ncol(X)))
  for(i in 1:ncol(X)){
    a <- ace(X[,i],Y)
    emece1[i] <- a$rsq
  }
  p <- which.max(emece1)
  Mce[1] <- p
  varnumber <- seq(1,ncol(X),1)
  eq <- as.data.frame(X)
  colnames(eq)<-varnumber
  eq1 <- eq[,-p]
  eq2 <- eq[,p]
  g <- Y
  j <- 1
####eligiendo de la 2nda a la p-2 variable
  while (ncol(as.matrix(eq1)) > 2){
    Xar <- matrix((rep(0)),nrow(eq1),ncol(eq1))
    g <- lm(g~eq2)
    gr <- g$residuals
    for(i in 1:ncol(eq1)){
      Xa <- lm(eq1[,i]~eq2)
      Xar[,i] <- Xa$residuals
    }
    emece1 <- c(rep(0,ncol(Xar)))
    for(i in 1:ncol(Xar)){
      a <- ace(Xar[,i],gr)
      emece1[i] <- a$rsq
    }
    p <- which.max(emece1)
    Mce[j+1] <- as.integer(colnames(eq1)[p])
    eq2 <- eq1[,p]
    eq1 <- eq1[,-p]
    g <- gr
    j <- j+1
  }
####eligiendo la penltima variable
  Xar <- matrix((rep(0)),nrow= nrow(eq1),ncol= 2)
  g <- lm(g~eq2)
```

```
  gr <- g$residuals
  for(i in 1:ncol(Xar)){
    Xa <- lm(eq1[,i]~eq2)
    Xar[,i] <- Xa$residuals
  }
  emece1 <- c(rep(0,ncol(Xar)))
  for(i in 1:ncol(Xar)){
    a <- ace(Xar[,i],gr)
    emece1[i] <- a$rsq
  }
  p <- which.max(emece1)
  Mce[j+1] <- as.integer(colnames(eq1)[p])
 ####eligiendo la ltima variable (trampita)
  Mce[length(Mce)] <- as.integer(colnames(eq)[-Mce])
 ####reacomodando la matriz de datos
  eq <- eq[,match(Mce,colnames(eq))]
 ####el PRESS
  prvec <- c(rep(0,ncol(eq)))
  gog <- as.matrix(eq)
  for(i in 1:ncol(gog)){
    mod <- lm(Y~gog[,1:i])
    pressmo <- PRESS(mod)
    prvec[i] <- pressmo
  }
  finmod <- lm(Y~gog[,1:which.min(prvec)])
  qe <- finmod$coefficients
  names(qe) <- c("Intercept",Mce[1:which.min(prvec)])
  #data from PRESS
  return(list("PRMC"=qe,"MCv"=Mce))
}
####################################
#LASSO (Least Absolute Shrinkage and Selection Operator)
####################################
LAS <- function(X,Y){
lingo <- cv.glmnet(X,Y,alpha=1)
colingo <- coef(lingo)
colin <- colingo[,1]
names(colin) <- c("Intercept",seq(1,ncol(X),1))
```

```
return(list("CoLASSO"= colin))
}

###################
#Adaptative LASSO
#tomado de http://www4.stat.ncsu.edu/~boos/var.select/lasso.adapt.bic2.txt
##################
adaLASSO<-function(x,y){

  ok<-complete.cases(x,y)
  x<-x[ok,]                             # get rid of na's
  y<-y[ok]                              # since regsubsets can't handle na's
  m<-ncol(x)
  n<-nrow(x)
  x<-as.matrix(x)                       # in case x is not a matrix

  #  standardize variables like lars does
  one <- rep(1, n)
  meanx <- drop(one %*% x)/n
  xc <- scale(x, meanx, FALSE)          # first subtracts mean
  normx <- sqrt(drop(one %*% (xc^2)))
  names(normx) <- NULL
  xs <- scale(xc, FALSE, normx)         # now rescales with norm (not sd)

  out.ls=lm(y~xs)                       # ols fit on standardized
  beta.ols=out.ls$coeff[2:(m+1)]        # ols except for intercept
  w=abs(beta.ols)                       # weights for adaptive lasso
  xs=scale(xs,center=FALSE,scale=1/w)  # xs times the weights
  object=lars(xs,y,type="lasso",normalize=FALSE)
  # METER EL CV.GLMNET
  # get min BIC
  # bic=log(n)*object$df+n*log(as.vector(object$RSS)/n)   # rss/n version
  sig2f=summary(out.ls)$sigma^2         # full model mse
  bic2=log(n)*object$df+as.vector(object$RSS)/sig2f       # Cp version
  step.bic2=which.min(bic2)             # step with min BIC

  fit=predict.lars(object,xs,s=step.bic2,type="fit",mode="step")$fit
  coeff=predict.lars(object,xs,
```

```
  s=step.bic2,type="coef",mode="step")$coefficients
  coeff=coeff*w/normx                  # get back in right scale
  st=sum(coeff !=0)                    # number nonzero
  mse=sum((y-fit)^2)/(n-st-1)          # 1 for the intercept

  # this next line just finds the variable id of coeff. not equal 0
  if(st>0) x.ind<-as.vector(which(coeff !=0)) else x.ind<-0
  intercept=as.numeric(mean(y)-meanx%*%coeff)
  #  return(list(fit=fit,st=st,mse=mse,x.ind=x.ind,
  #  coeff=coeff,intercept=intercept,object=object,
  #              bic2=bic2,step.bic2=step.bic2))
  chiprito <- c(intercept,coeff)
  names(chiprito) <- c("Intercept",seq(1,ncol(X),1))
  return(list("LASADA" = chiprito))
}

####################################
#Cov-Test (Covariance Test for the Lasso)
####################################
covT <- function(X,Y){
  achu <- lars(X,Y)
  achupar <- covTest(achu,X,Y)
  scrap <- achupar$results[,3]
  z <- 1
  while (scrap[z] <= 0.05){
    z <- z+1
  }
  gon <- capture.output(coef(achu))
  COVART <-coef(achu)[z,]
#  cat("Iteration",t,"\n","CovTest p-value",scrap,"\n",
# "Coef matrix",gon,"Coef Chosen", z,"\n" ,COVART,"\n",sep="\n" ,
#      file="/Users/Vulcan/Dropbox/tesista Yamil/Programas/
#      Funciones/Outputs/CovTest Data.csv", append= TRUE)
  return(list("covTcoef"= COVART))
}
```

A.2. Simulación

```
#Fixed Part n=100
X <- matrix(rnorm(800,0,1),100,8)
Bta <- c(1,1,1,0,0,0,0,0)
#cat("Modelo Lineal Simple", "\n", sep="\n",
file="/Users/Vulcan/Dropbox/tesista Yamil/Programas/
Funciones/Outputs/CovTest Data.csv", append= TRUE)

linDC <- data.frame(matrix(NA,ncol=1+ncol(X),nrow=1000))
names(linDC) <- c("Intercept",seq(1,ncol(X),1))

linMC <- data.frame(matrix(NA,ncol=1+ncol(X),nrow=1000))
names(linMC) <- c("Intercept",seq(1,ncol(X),1))

linLASSO <- data.frame(matrix(NA,ncol=1+ncol(X),nrow=1000))
names(linLASSO) <- c("Intercept",seq(1,ncol(X),1))

linadaLASSO <- data.frame(matrix(NA,ncol=1+ncol(X),nrow=1000))
names(linadaLASSO) <- c("Intercept",seq(1,ncol(X),1))

lincovT <- data.frame(matrix(NA,ncol=1+ncol(X),nrow=1000))
names(lincovT) <- c("Intercept",seq(1,ncol(X),1))

YhatcovT <- matrix(data=NA,nrow=(100),ncol=(1000))

#Simulaci\'on para Modelo Lineal Simple
#Repeat N=100
#aqu la repeticin
for(t in 1:1000){
Er <- rnorm(100,0,2)
Y <- X%*%Bta+Er


#DC
chale <- DC(X,Y)
```

```
for(p in 1:length(chale$PRDC)){
  linDC[t,names(linDC)==names(chale$PRDC[p])] <- chale$PRDC[p]
}
#MC
chole <- MC(X,Y)
for(p in 1:length(chole$PRMC)){
  linMC[t,names(linMC)==names(chole$PRMC[p])] <- chole$PRMC[p]
}
#LASSO
chile <- LAS(X,Y)
for(p in 1:length(chile$CoLASSO)){
  linLASSO[t,names(linLASSO)==names(chile$CoLASSO)[p]] <- chile$CoLASSO[p]
}

#Adaptative LASSO
chele <- adaLASSO(X,Y)
for(p in 1:length(chele$LASADA)){
  linadaLASSO[t,names(linadaLASSO)==names(chele$LASADA)[p]] <- chele$LASADA[p]
}

#CovTest
chule <- covT(X,Y)
names(chule$covTcoef) <- seq(1,ncol(X),1)
for(p in 1:length(chule$covTcoef)){
  lincovT[t,names(lincovT)==names(chule$covTcoef)[p]] <- chule$covTcoef[p]
}
YhatcovT[,t] <- chule$covTpred
print(t)
}

#Para cambiar los NA por 0 en los data frames de LASSO
linMC[is.na(linMC)] <- 0
linDC[is.na(linDC)] <- 0

#Conteos

Results <- data.frame(matrix(NA,ncol=12,nrow=5))
rownames(Results) <- c("DC","MC","LASSO","Ada LASSO","CovTest")
```

```
colnames(Results) <- c("1st Var","2nd Var","3rd Var","Al menos","Hit Rate",
                       "B1 Var","B1 Ses","B2 Var","B2 Ses","B3 Var","B3 Ses",
                       "Error de Prediccin")

#DC

#Sesgo y MSE B1
SesB1 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  SesB1[t] <- linDC[t,2]-1
}
B1Ses <- mean(SesB1) #Sesgo B1
MSEB1 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  MSEB1[t] <- SesB1[t]^2
}
B1MSE <- mean(MSEB1) #MSE B1
B1Var <- B1MSE - B1Ses^2 #Varianza B1

#Sesgo y MSE B2
SesB2 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  SesB2[t] <- linDC[t,3]-1
}
B2Ses <- mean(SesB2) #Sesgo B2
MSEB2 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  MSEB2[t] <- SesB2[t]^2
}
B2MSE <- mean(MSEB2) #MSE B2
B2Var <- B2MSE - B2Ses^2 #Varianza B2

#Sesgo y MSE B3
SesB3 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  SesB3[t] <- linDC[t,4]-1
}
B3Ses <- mean(SesB3) #Sesgo B3
```

```
MSEB3 <- vector(length=nrow(linDC))
for(t in 1:nrow(linDC)){
  MSEB3[t] <- SesB3[t]^2
}
B3MSE <- mean(MSEB3) #MSE B3
B3Var <- B3MSE - B3Ses^2 #Varianza B3

#Yhat estimate and Yhat - trueY
trueY <- X%*%Bta
trueX <- cbind(1,X) # Para B0

##############################
#Error de prediccion entre trueY y Yhat
##############################
#Como se tiene que hacer con cada Beta estimada
#hacemos un for loop para todo
YhatDClin <- matrix(data=NA, ncol=nrow(linDC),nrow=nrow(Y))
for(i in 1:nrow(linDC)){
Yhat <- trueX%*%as.matrix(t(linDC[i,]))
YhatDClin[,i] <- Yhat
}
# Ahora realizamos la resta de cada columna yhat de truey
erpred <- matrix(data=NA,ncol=ncol(YhatDClin),nrow=nrow(YhatDClin))
for(i in 1:ncol(YhatDClin)){
errado <- (trueY-YhatDClin[,i])^2
erpred[,i] <- errado
}

#########
# finalmente se calcula el ECM
#########
vecerp <- vector(length = ncol(erpred))
for(i in 1:ncol(erpred)){
  gatuza <- sum(erpred[,i])
  gatuzon <- gatuza/100
vecerp[i] <- gatuzon
}
linDCerp <- mean(vecerp)
```

```
###############
#Se extraen los resultados
###############
#DC
Results[1,1] <- sum(linDC[,2] != 0)
Results[1,2] <- sum(linDC[,3] != 0)
Results[1,3] <- sum(linDC[,4] != 0)
Results[1,4] <- sum((linDC[,2] !=0) &
(linDC[,3] !=0) & (linDC[,4] !=0))

Results[1,5] <- sum((linDC[,2] !=0) & (linDC[,3] !=0) &
(linDC[,4] !=0) & (linDC[,5] ==0) & (linDC[,6] ==0) &
(linDC[,7] ==0) & (linDC[,8] ==0) & (linDC[,9] ==0))

Results[1,6] <- B1Var
Results[1,7] <- B1Ses
Results[1,8] <- B2Var
Results[1,9] <- B2Ses
Results[1,10] <- B3Var
Results[1,11] <- B3Ses
Results[1,12] <- linDCerp



#MC
#Sesgo y MSE B1
SesB1 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  SesB1[t] <- linMC[t,2]-1
}
B1Ses <- mean(SesB1) #Sesgo B1
MSEB1 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  MSEB1[t] <- SesB1[t]^2
}
B1MSE <- mean(MSEB1) #MSE B1
B1Var <- B1MSE - B1Ses^2 #Varianza B1
```

```
#Sesgo y MSE B2
SesB2 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  SesB2[t] <- linMC[t,3]-1
}
B2Ses <- mean(SesB2) #Sesgo B2
MSEB2 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  MSEB2[t] <- SesB2[t]^2
}
B2MSE <- mean(MSEB2) #MSE B2
B2Var <- B2MSE - B2Ses^2 #Varianza B2

#Sesgo y MSE B3
SesB3 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  SesB3[t] <- linMC[t,4]-1
}
B3Ses <- mean(SesB3) #Sesgo B3
MSEB3 <- vector(length=nrow(linMC))
for(t in 1:nrow(linMC)){
  MSEB3[t] <- SesB3[t]^2
}
B3MSE <- mean(MSEB3) #MSE B3
B3Var <- B3MSE - B3Ses^2 #Varianza B3

############################
#Error de prediccion entre trueY y Yhat
############################
#Como se tiene que hacer con cada Beta estimada
#hacemos un for loop para todo
YhatMClin <- matrix(data=NA, ncol=nrow(linMC),nrow=nrow(Y))
for(i in 1:nrow(linMC)){
  Yhat <- trueX%*%as.matrix(t(linMC[i,]))
  YhatMClin[,i] <- Yhat
}
# Ahora realizamos la resta de cada columna yhat de truey
```

```
erpred <- matrix(data=NA,ncol=ncol(YhatMClin),nrow=nrow(YhatMClin))
for(i in 1:ncol(YhatMClin)){
  errado <- (trueY-YhatMClin[,i])^2
  erpred[,i] <- errado
}

#########
# finalmente se calcula el ECM
#########
vecerp <- vector(length = ncol(erpred))
for(i in 1:ncol(erpred)){
  gatuza <- sum(erpred[,i])
  gatuzon <- gatuza/100
  vecerp[i] <- gatuzon
}
linMCerp <- mean(vecerp)



Results[2,1] <- sum(linMC[,2] != 0)
Results[2,2] <- sum(linMC[,3] != 0)
Results[2,3] <- sum(linMC[,4] != 0)
Results[2,4] <- sum((linMC[,2] !=0) &
(linMC[,3] !=0) & (linMC[,4] !=0))

Results[2,5] <- sum((linMC[,2] !=0) & (linMC[,3] !=0) &
(linMC[,4] !=0) & (linMC[,5] ==0) & (linMC[,6] ==0) &
(linMC[,7] ==0) & (linMC[,8] ==0) & (linMC[,9] ==0))

Results[2,6] <- B1Var
Results[2,7] <- B1Ses
Results[2,8] <- B2Var
Results[2,9] <- B2Ses
Results[2,10] <- B3Var
Results[2,11] <- B3Ses
Results[2,12] <- mean(linMCerp)

#LASSO
```

```
#Sesgo y Varianza

SesB1 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  SesB1[t] <- linLASSO[t,2]-1
}
B1Ses <- mean(SesB1) #Sesgo B1
MSEB1 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  MSEB1[t] <- SesB1[t]^2
}
B1MSE <- mean(MSEB1) #MSE B1
B1Var <- B1MSE - B1Ses^2 #Varianza B1

#Sesgo y MSE B2
SesB2 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  SesB2[t] <- linLASSO[t,3]-1
}
B2Ses <- mean(SesB2) #Sesgo B2
MSEB2 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  MSEB2[t] <- SesB2[t]^2
}
B2MSE <- mean(MSEB2) #MSE B2
B2Var <- B2MSE - B2Ses^2 #Varianza B2

#Sesgo y MSE B3
SesB3 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  SesB3[t] <- linLASSO[t,4]-1
}
B3Ses <- mean(SesB3) #Sesgo B3
MSEB3 <- vector(length=nrow(linLASSO))
for(t in 1:nrow(linLASSO)){
  MSEB3[t] <- SesB3[t]^2
}
```

```
B3MSE <- mean(MSEB3) #MSE B3
B3Var <- B3MSE - B3Ses^2 #Varianza B3

YhatLASSOlin <- matrix(data=NA, ncol=nrow(linLASSO),nrow=nrow(Y))
for(i in 1:nrow(linLASSO)){
  Yhat <- trueX%*%as.matrix(t(linLASSO[i,]))
  YhatLASSOlin[,i] <- Yhat
}
# Ahora realizamos la resta de cada columna yhat de truey
erpred <- matrix(data=NA,
                 ncol=ncol(YhatLASSOlin),nrow=nrow(YhatLASSOlin))
for(i in 1:ncol(YhatLASSOlin)){
  errado <- (trueY-YhatLASSOlin[,i])^2
  erpred[,i] <- errado
}

#########
# finalmente se calcula el ECM
#########
vecerp <- vector(length = ncol(erpred))
for(i in 1:ncol(erpred)){
  gatuza <- sum(erpred[,i])
  gatuzon <- gatuza/100
  vecerp[i] <- gatuzon
}
linLASSOerp <- mean(vecerp)

Results[3,1] <- sum(linLASSO[,2] != 0)
Results[3,2] <- sum(linLASSO[,3] != 0)
Results[3,3] <- sum(linLASSO[,4] != 0)
Results[3,4] <- sum((linLASSO[,2] !=0) &
(linLASSO[,3] !=0) & (linLASSO[,4] !=0))

Results[3,5] <- sum((linLASSO[,2] !=0) & (linLASSO[,3] !=0) &
(linLASSO[,4] !=0) & (linLASSO[,5] ==0) & (linLASSO[,6] ==0) &
(linLASSO[,7] ==0) & (linLASSO[,8] ==0) & (linLASSO[,9] ==0))

Results[3,6] <- B1Var
```

```
Results[3,7] <- B1Ses
Results[3,8] <- B2Var
Results[3,9] <- B2Ses
Results[3,10] <- B3Var
Results[3,11] <- B3Ses
Results[3,12] <- linLASSOerp

#adaLASSO

#Sesgo y Varianza

SesB1 <- vector(length=nrow(linadaLASSO))
for(t in 1:nrow(linadaLASSO)){
  SesB1[t] <- linadaLASSO[t,2]-1
}
B1Ses <- mean(SesB1) #Sesgo B1
MSEB1 <- vector(length=nrow(linadaLASSO))
for(t in 1:nrow(linadaLASSO)){
  MSEB1[t] <- SesB1[t]^2
}
B1MSE <- mean(MSEB1) #MSE B1
B1Var <- B1MSE - B1Ses^2 #Varianza B1

#Sesgo y MSE B2
SesB2 <- vector(length=nrow(linadaLASSO))
for(t in 1:nrow(linadaLASSO)){
  SesB2[t] <- linadaLASSO[t,3]-1
}
B2Ses <- mean(SesB2) #Sesgo B2
MSEB2 <- vector(length=nrow(linadaLASSO))
for(t in 1:nrow(linadaLASSO)){
  MSEB2[t] <- SesB2[t]^2
}
B2MSE <- mean(MSEB2) #MSE B2
B2Var <- B2MSE - B2Ses^2 #Varianza B2

#Sesgo y MSE B3
SesB3 <- vector(length=nrow(linadaLASSO))
```

```
for(t in 1:nrow(linadaLASSO)){
  SesB3[t] <- linadaLASSO[t,4]-1
}
B3Ses <- mean(SesB3) #Sesgo B3
MSEB3 <- vector(length=nrow(linadaLASSO))
for(t in 1:nrow(linadaLASSO)){
  MSEB3[t] <- SesB3[t]^2
}
B3MSE <- mean(MSEB3) #MSE B3
B3Var <- B3MSE - B3Ses^2 #Varianza B3

YhatadaLASSOlin <- matrix(data=NA,
                                  ncol=nrow(linadaLASSO),nrow=nrow(Y))
for(i in 1:nrow(linadaLASSO)){
  Yhat <- trueX%*%as.matrix(t(linadaLASSO[i,]))
  YhatadaLASSOlin[,i] <- Yhat
}
# Ahora realizamos la resta de cada columna yhat de truey
erpred <- matrix(data=NA,ncol=ncol(YhatadaLASSOlin),
                           nrow=nrow(YhatadaLASSOlin))
for(i in 1:ncol(YhatadaLASSOlin)){
  errado <- (trueY-YhatadaLASSOlin[,i])^2
  erpred[,i] <- errado
}

#########
# finalmente se calcula el ECM
#########
vecerp <- vector(length = ncol(erpred))
for(i in 1:ncol(erpred)){
  gatuza <- sum(erpred[,i])
  gatuzon <- gatuza/100
  vecerp[i] <- gatuzon
}
linadaLASSOerp <- mean(vecerp)

Results[4,1] <- sum(linadaLASSO[,2] != 0)
Results[4,2] <- sum(linadaLASSO[,3] != 0)
```

```
Results[4,3] <- sum(linadaLASSO[,4] != 0)
Results[4,4] <- sum((linadaLASSO[,2] !=0) &
(linadaLASSO[,3] !=0) & (linadaLASSO[,4] !=0))

Results[4,5] <- sum((linadaLASSO[,2] !=0) & (linadaLASSO[,3] !=0) &
(linadaLASSO[,4] !=0) & (linadaLASSO[,5] ==0) & (linadaLASSO[,6] ==0) &
(linadaLASSO[,7] ==0) & (linadaLASSO[,8] ==0) & (linadaLASSO[,9] ==0))

Results[4,6] <- B1Var
Results[4,7] <- B1Ses
Results[4,8] <- B2Var
Results[4,9] <- B2Ses
Results[4,10] <- B3Var
Results[4,11] <- B3Ses
Results[4,12] <- linadaLASSOerp

#covTest

#Sesgo y Varianza

SesB1 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  SesB1[t] <- lincovT[t,2]-1
}
B1Ses <- mean(SesB1) #Sesgo B1
MSEB1 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  MSEB1[t] <- SesB1[t]^2
}
B1MSE <- mean(MSEB1) #MSE B1
B1Var <- B1MSE - B1Ses^2 #Varianza B1

#Sesgo y MSE B2
SesB2 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  SesB2[t] <- lincovT[t,3]-1
}
B2Ses <- mean(SesB2) #Sesgo B2
```

```
MSEB2 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  MSEB2[t] <- SesB2[t]^2
}
B2MSE <- mean(MSEB2) #MSE B2
B2Var <- B2MSE - B2Ses^2 #Varianza B2

#Sesgo y MSE B3
SesB3 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  SesB3[t] <- lincovT[t,4]-1
}
B3Ses <- mean(SesB3) #Sesgo B3
MSEB3 <- vector(length=nrow(lincovT))
for(t in 1:nrow(lincovT)){
  MSEB3[t] <- SesB3[t]^2
}
B3MSE <- mean(MSEB3) #MSE B3
B3Var <- B3MSE - B3Ses^2 #Varianza B3

erpred <- matrix(data=NA,ncol=ncol(YhatcovT),nrow=nrow(YhatcovT))
for(i in 1:ncol(YhatcovT)){
  errado <- (trueY-YhatcovT[,i])^2
  erpred[,i] <- errado
}

#########
# finalmente se calcula el ECM
#########
vecerp <- vector(length = ncol(erpred))
for(i in 1:ncol(erpred)){
  gatuza <- sum(erpred[,i])
  gatuzon <- gatuza/100
  vecerp[i] <- gatuzon
}
lincovTerp <- mean(vecerp)

Results[5,1] <- sum(lincovT[,2] != 0)
```

```
Results[5,2] <- sum(lincovT[,3] != 0)
Results[5,3] <- sum(lincovT[,4] != 0)
Results[5,4] <- sum((lincovT[,2] !=0) &
(lincovT[,3] !=0) & (lincovT[,4] !=0))

Results[5,5] <- sum((lincovT[,2] !=0) & (lincovT[,3] !=0) &
(lincovT[,4] !=0) & (lincovT[,5] ==0) & (lincovT[,6] ==0) &
(lincovT[,7] ==0) & (lincovT[,8] ==0) & (lincovT[,9] ==0))

Results[5,6] <- B1Var
Results[5,7] <- B1Ses
Results[5,8] <- B2Var
Results[5,9] <- B2Ses
Results[5,10] <- B3Var
Results[5,11] <- B3Ses
Results[5,12] <- lincovTerp

######
#Respaldo los resultados
######
write.xlsx(Results,"/Users/Vulcan/Desktop/ModLineal.xlsx")
#Simulacin de colinealidad constante
```

A.3. Variaciones de la simulación

Lo único que cambia para las otras simulaciones que se llevaron a cabo es el cálculo de la matriz diseño. A continuación se muestra sólo las variaciones del código.

```
#Condiciones de simulaci\'on de Colinealidad constante theta .6
bonk <- diag(8)
qu <- .6 ###este valor se fij en 0.6, 0.8 y 0.9
bonk[bonk !=1] <- qu
bonk

X <- rmvnorm(100,mean=c(0,0,0,0,0,0,0,0),sigma=bonk)
```

```
Bta <- c(1,1,1,0,0,0,0,0)

#Condiciones de simulaci\'on de Colinealidad Toeplitz
qu <- 0.6
buga <- c(1,qu,qu^2,qu^3,qu^4,qu^5,qu^6,qu^7)
bunk <- toeplitz(buga)

X <- rmvnorm(100,mean=c(0,0,0,0,0,0,0,0),sigma=bunk)
Bta <- c(1,1,1,0,0,0,0,0)
```