USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Modelos de Regresión con Penalizacións

Sergio Mejuto Vázquez

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Modelos de Regresión con Penalizacións

Sergio Mejuto Vázquez

07/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Modelos de Regresión con Penalizacións
Breve descrición do contido
A irrupción nos últimos anos de problemas de regresión onde o número de covariables pode ser moi grande ou mesmo maior que o número de casos, fixo que se desenvolveran modelos de regresión que o mesmo tempo que estiman a relación entre as covariables e a resposta seleccionen as mellores covariables para formar parte do modelo de regresión. Estas técnicas soen incluir un termo de penalización que axuda a esta selección. Entre estas técnicas atópanse a regresión Ridge, o método best subset selection ou o LASSO e as súas variantes. O obxectivo do traballo é presentar estas metodoloxías e comparar o seu funcionamento en estudios de simulación e na aplicación a exemplos clásicos como os que se poden atopar na web UCI Machine Learning.
Recomendacións
Soltura co paquete estatístico R.
Outras observacións

Índice general

Resumen

Los modelos de regresión se basan en los estimadores de mínimos cuadrados, el cual veremos que es fácil que no sea capaz de estimar los coeficientes con fiabilidad en el caso de variables dependientes. Para corregir este problema, se crearon distintos métodos que veremos en este trabajo. La Regresión Ridge y el método LASSO utilizan el estimador de mínimos cuadrados con una penalización, la cual restringe los posibles valores que pueden tomar los coeficientes, y así logran controlar sus valores. Por otro lado, LAR nos da una ordenación de las variables según su importancia, con la cual podemos crear modelos de regresión cogiendo sólo variables importantes. Con estos métodos conseguimos evitar el problema de las variables dependientes.

Abstract

Regression methods are based on the least square estimator, which may not provide a good estimation of the coefficients in the case we have dependant variables. To avoid this problem, different methods were created. Ridge Regression and LASSO use the least square estimator with a penalty that shrinks the values of the coefficients, controling their possible values. On the other hand, LAR order the variables by their importance in the model, and we can use it to create a model by choosing only the most important variables. With these methods we can avoid the problem we had in the case of dependant variables.

Introducción

Para empezar, definamos qué es la regresión. La regresión consiste en intentar averiguar la relación entre una variable respuesta Y y una o más variables explicativas $X_1, ..., X_p$. La principal herramienta para buscar esta relación son los Modelos de regresión.

La gran mayoría de los modelos de regresión utilizan el concepto del estimador por mínimos cuadrados, el cual fue desarrollado por Carl Friedrich Gauss (1777-1855) al intentar predecir la órbita del planeta enano Ceres a partir de los datos que Giuseppe Piazzi, astrónomo italiano descubridor de este planeta, tenía sobre la posición del planeta durante 40 días. Esto sucedió en 1801, aunque la teoría del método no se publicó hasta 1809. Este método también fue desarrollado de manera independiente por el matemático francés Adrien-Marie Legendre (1752-1833) en 1805. En 1829 Gauss demostró que este método es tan exacto y útil porque es óptimo en muchos sentidos, lo cual se resume en el Teorema de Gauss-Márkov.

El nombre de modelos de regresión se remonta a Francis Galton (1822-1911), un prolífico científico con conocimiento en muchas materias, entre ellas las Matemáticas. Además, Galton era primo de Darwin y cuando éste sacó su teoría de la evolución, Galton quiso comprobar su veracidad por lo que decidió estudiar la evolución centrándose en los humanos. Para ello decidió mirar primero las alturas de padres e hijos y llegó a la conclusión de que los padres altos tenían hijos altos, pero en general no tan altos como sus padres y los padres bajos tenían hijos bajos, pero en general no tan bajos como sus padres. Galton dedujo, por tanto, que la altura de los hijos tendía a regresar al promedio, y de ahí se acuñó el término 'regresión'.

Los modelos de regresión se basan en la idea de conocer cómo es la dependencia de Y respecto de X (lineal, polinómica, logarítmica, etc.) y calcular unos coeficientes utilizando el estimador de mínimos cuadrados. Una vez que tenemos esta relación calculada, podemos utilizarla para intentar predecir la variable Y conociendo la variable X.

Los modelos de regresión se han estado desarrollando desde hace tiempo, pues se vio que eran muy útiles a la hora de predecir y establecer si hay correlación entre variables, pues no siempre existen fórmulas exactas que nos relacionen las variables, como podrían

ser las fórmulas de las Leyes de Kepler, que se denominan 'modelos deterministas', sino que muchas veces se debe intentar predecir el valor de Y considerando un error que no podemos predecir, por lo que estos modelos se conocen como 'modelos estocásticos'.

Al principio se utilizaron modelos de regresión lineal, ya que nos sirven para predecir qué valor va a tomar una variable, aunque más adelante se avanzó para poder utilizar estos modelos para predecir, por ejemplo, si un paciente va a morir o no, las cuales no son variables continuas sino que toman valores aislados. Estos modelos se denominan modelos logísticos y posteriormente se generalizó el concepto de modelos de regresión para incluir estos modelos. Hasta finales del siglo pasado solamente se utilizaban modelos lineales, pues si consideramos otro tipo de relación entre las variables (polinómica, logarítmica, etc.), el coste computacional era demasiado grande, pero con la llegada de los softwares estadísticos se ha podido profundizar en otro tipo de relaciones, las cuales se estudian hoy en día para mejorar este campo. En este trabajo vamos a utilizar R, un software estadístico gratuito que nos permitirá realizar los ejemplos y gráficas.

Los modelos de regresión se han utilizado mucho desde que se descubrieron y se siguen utilizando en la actualidad en varios ámbitos como pueden ser la Medicina, pues nos permiten identificar factores de riesgo de una enfermedad, en Economía, para calcular el crecimiento del PIB y son muy utilizados en una rama conocida como Econometría que se encarga de dar validez a las leyes económicas a través de diferentes métodos, y en Biología se pueden utilizar para observar qué factores medioambientales favorecen a ciertas especies o cómo las características fisiológicas de éstas se ven alteradas por cambios en el medio.

Los modelos de regresión con penalizaciones son modelos que tienen la particularidad de que los posibles valores que pueden tomar los coeficientes son limitados, ya sea por limitaciones físicas o limitaciones que nosotros imponemos, y por tanto debemos considerar estas restricciones a la hora de crear el modelo. Por ejemplo, podemos tener un modelo en el que sepamos que los coeficientes deben ser positivos, o podemos imponerle al modelo que la suma de los coeficientes esté cerca de 0 para hacer que sólo los coeficientes que son verdaderamente distintos de 0 sean grandes y disminuir los otros.

En este trabajo vamos a profundizar en los modelos de regresión con penalizaciones, apoyándonos en Friedman et al. (2009), 'The Elements of Statistical Learning', un libro popular en cuanto a conceptos estadísticos y James et al. (2014), 'An Introduction to Statistical Learning', el cual abarca los mismos temas que el otro libro pero de una manera menos técnica y con más ejemplos. (James et al., 2014, véase aquí).

Este trabajo consta de 5 capítulos:

En el primero veremos el estimador de mínimos cuadrados y conceptos generales de los modelos de regresión que luego aplicaremos en los métodos que veremos más adelante y

que nos servirán para entender en qué consisten y compararlos. Todo ésto irá acompañado de ejemplos para hacer más sencilla la comprensión.

En el segundo hablaremos sobre la Regresión Ridge, en qué consiste y características que tienen los modelos obtenidos a partir de este método.

En el tercero hablaremos sobre el método Lasso, veremos sus características generales, su función como seleccionador de variables y lo compararemos con la Regresión Ridge.

En el cuarto explicaremos el método de Least Angle Regression, el cual no es un método para obtener un modelo de regresión, sino una manera de ordenar las variables en función de su importancia.

Por último, en el quinto capítulo tendremos un ejemplo real al cual aplicaremos todos los modelos vistos en este trabajo y hablaremos sobre las salidas que obtenemos.

Capítulo 1

Modelos de regresión

Para entender en qué consisten los modelos de regresión con penalizaciones, primero veamos los modelos de regresión y conceptos relacionados con éstos que luego aplicaremos a los distintos modelos de regresión con penalizaciones que veremos en los siguientes capítulos.

1.1. Idea general de los Modelos de regresión

Para poder predecir los valores que tomará una variable Y (variable respuesta) a partir de una serie de variables X_1, X_2, ..., X_p (variables explicativas) utilizamos los Modelos de regresión. Dichos Modelos de regresión también nos permiten ver la relación entre las dos variables, es decir, si la variable explicativa X_i aumenta entonces la variable respuesta Y aumenta o disminuye y en qué medida lo hace. Esta relación existente entre las dos variables la reflejaremos en un valor β_i que será positivo o negativo dependiendo de si al aumentar la variable X_i, Y aumenta o disminuye respectivamente y el valor numérico que tendrá será la proporción entre las dos variables, es decir, si la variable X_i aumenta en 1, la variable Y variará en esa proporción.

A la hora de tomar medidas de cualquiera de las variables puede haber errores de medición, errores humanos o un margen de error intrínseco a la variable que estamos intentando predecir, por lo que el modelo que obtendremos no predecirá exactamente la variable respuesta, sino que lo hará con un cierto error que no podremos calcular pero que debemos intentar que sea el menor posible. Para verlo mejor utilizaremos un conjunto de datos proveniente de la librería Faraway (2016), donde tenemos un modelo que intenta predecir el número de muertes en un grupo de 30 insectos aplicando distintas concentraciones de insecticida, con lo que obtenemos la gráfica 1.1 en la que vemos que el modelo no es capaz

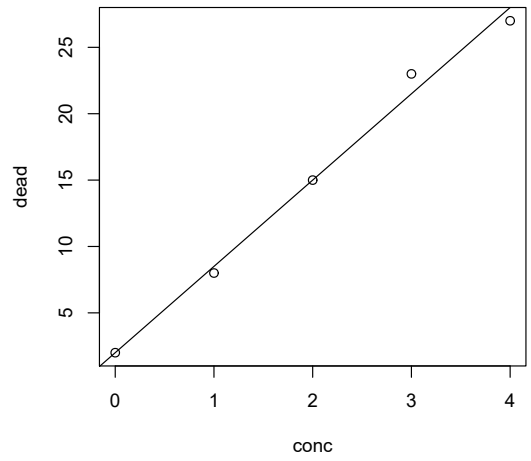

Figura 1.1: *Ejemplo de insecticida. En él vemos que a medida que aumentamos la concentración de insecticida (conc) aumenta el número de muertes (dead) y esto se ve reflejado en la recta de regresión. Vemos, además, que la recta predice con bastante fiabilidad el número de muertes con un cierto margen de error.*

de predecir exactamente el número de muertes pero se aproxima bastante.

Para construir un modelo de regresión debemos estimar el valor de $p+1$ variables, los $\beta_i, i \in \{1, ..., p\}$ que son los coeficientes de las p variables explicativas y el término independiente, que denotaremos por β_0, que representa el valor que tendría la variable respuesta si todas las variables explicativas toman el valor 0. Por lo tanto, el modelo que nos queda tiene la forma

$$Y = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p + \epsilon$$

donde ϵ denota el error que no conocemos.

Para ser capaces de intentar calcular estos $\beta_i, i \in \{1, ..., p\}$, conocemos n observaciones previas de las variables explicativas $X_i, i \in \{1, ..., p\}$ y de la variable respuesta Y. A través de éstas estimaremos el valor que toman los coeficientes intentando encontrar unos β_i que sean capaces de calcular lo más aproximadamente posible el valor de la observación Y_j al sustituir en la fórmula anterior cada una de las variables X_i por la observación j-ésima de dicha variable, $X_{j,i}$, para todas las observaciones $j \in \{1, ..., n\}$, dicho de otra manera, intentar que los errores ϵ_j sean lo más pequeños posibles. Otra condición que cumplen los coeficientes $\beta_i, i \in \{0, ..., p\}$ es que

$$\overline{Y} = \beta_0 + \beta_1 \cdot \overline{X_1} + ... + \beta_n \cdot \overline{X_n}$$

donde $\overline{X_i}$ denota la media de las observaciones de la variable X_i, es decir, $\overline{X_i} = \frac{1}{n}\sum_{j=1}^{n} X_{j,i}$. Por lo tanto, el modelo de regresión siempre pasa por el punto formado por todas las medias de las variables, o lo que es lo mismo, cuando las variables explicativas toman como valor su media se considera que el modelo toma el valor $\overline{Y}$ y el error ϵ es 0.

De ahora en adelante utilizaremos la notación matricial:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Que denotaremos como

$$Y = X\beta + \epsilon$$

En dicha expresión, el vector Y son las n observaciones que tenemos de la variable con el mismo nombre, la matriz X son las n observaciones de cada una de las p variables explicativas y una columna de 1's debido a que el coeficiente β_0 no multiplica a ninguna variable explicativa, β es el vector de coeficientes que tenemos que estimar y por último ϵ es el vector de errores que no podemos conocer.

1.2. Estimación de los coeficientes

Para estimar los $\beta_i, i \in \{0, ..., p\}$ lo haremos utilizando el método de mínimos cuadrados, el cual se aplica escogiendo como estimador el vector $\widehat{\beta}$ que cumple:

$$\widehat{\beta} = \min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=0}^{p} x_{ij}\beta_j)^2$$

Si lo pasamos a notación matricial, la expresión anterior es equivalente a

$$\widehat{\beta} = \min_{\beta} (Y - X\beta)^\intercal (Y - X\beta)$$

Para obtener la solución de esta expresión derivamos e igualamos a 0

$$-X^\intercal(Y - X\beta) - ((Y - X\beta)^\intercal X)^\intercal = 0;$$

$$-X^\intercal(Y - X\beta) - X^\intercal(Y - X\beta) = 0;$$

$$-2X^\intercal(Y - X\beta) = 0;$$

$$2X^\intercal X\beta = 2X^\intercal Y;$$

$$X^\intercal X\beta = X^\intercal Y$$

con lo que obtenemos que el mínimo se alcanza en un β que cumple:

$$X^\intercal X\beta = X^\intercal Y$$

Por lo tanto, el estimador de mínimos cuadrados en notación matricial es

$$\widehat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$$

con lo cual, para que $\widehat{\beta}$ esté bien definida, $X^\intercal X$ no puede ser singular. Se puede considerar otra matriz denominada matriz 'Hat' que se construye como

$$H = X(X^\intercal X)^{-1} X^\intercal$$

y cumple

$$\widehat{Y} = X\widehat{\beta} = HY$$

donde $\widehat{Y}$ denota la estimación que hacemos de Y a través del vector estimado $\widehat{\beta}$.

Ejemplo 1.1. Si consideramos el siguiente conjunto de datos donde tenemos dos variables explicativas $X_1 = (1, 2, 3, 4, 5, 6)$ y $X_2 = (-1, 1, -2, 2, -3, 3)$ y una variable respuesta $Y = 10 + 6 \cdot X_1 - 2 \cdot X_2 + \epsilon$, donde ϵ es el vector de errores perteneciente a una normal $N(0, 1)$. Con lo que hemos visto hasta ahora, podemos crear un modelo de regresión y ver si los coeficientes que obtenemos se parecen o no a los que realmente tenemos. Si calculamos el vector de estimadores $\widehat{\beta}$ obtenemos:

```
beta

>         [,1]
>     9.361049
> x1  6.310315
> x2 -1.991095
```

que serían los estimadores de β_0 y de los coeficientes respectivos a las variables. Además, podemos calcular la matriz H y con ella averiguar cual es el vector $\widehat{Y}$ para ver si los valores que obtenemos se acercan o no a los que realmente tenemos:

```
>      [,1]               [,2]
> [1,] "Y"                "Y estimado"
> [2,] "17.4395243534478" "17.6624592338644"
> [3,] "19.7698225105167" "19.9905844833884"
> [4,] "33.5587083141491" "32.2741851798991"
> [5,] "30.0705083914246" "30.6201202506284"
> [6,] "46.1292877351609" "46.8859111259337"
> [7,] "41.7150649868833" "41.2496560178684"
```

La primera columna son los valores de Y y la segunda, la estimación que obtenemos que podemos ver que se aproxima bastante.

Un coeficiente mayor no implica que la variable explicativa asociada esté más relacionada con la variable respuesta. Para ver si dos variables están relacionadas o no, se utiliza lo que se conoce como el coeficiente de correlación que definimos como:

$$r(X,Y) = \frac{S_{XY}}{S_X S_Y}$$

donde S_{XY} denota la covarianza entre X e Y:

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

con $\bar{X}$ la media de la variable X e $\bar{Y}$ la media de la variable Y. S_X y S_Y son las desviaciones típicas de estas variables:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

y S_Y se calcula de manera análoga.

Este coeficiente tiene valores entre -1 y 1 y cuanto mayor sea en valor absoluto, mayor es la correlación entre las dos variables.

1.3. Hipótesis del modelo

Para que un modelo esté bien construido, éste debe cumplir cuatro hipótesis:

-Linealidad: Suponemos que las variables explicativas y la variable respuesta se relacionan de manera lineal, es decir, siguiendo una ecuación del tipo:

$$Y = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p + \epsilon$$

Esta hipótesis se puede relajar pues como ya comentamos en la Introducción, también se pueden considerar modelos con otros tipos de relaciones como logarítmicos o polinómicos.

-Homocedasticidad: Suponemos que la varianza de todos los errores ϵ_i, $i \in \{1, ..., n\}$ es constante e igual para todos:

$$Var(\epsilon_i) = \sigma^2$$

-Normalidad del error: Suponemos que los errores ϵ_i, $i \in \{1, ..., n\}$ siguen una distribución normal de media 0 y varianza σ^2. Esta hipótesis, junto con la de Homocedasticidad, nos permiten hacer un mejor estudio de las características del modelo.

-Independencia del error: Suponemos que todos los errores son independientes unos de otros.

A la hora de comprobar estas hipótesis tenemos el problema de que el vector de errores ϵ no lo podemos conocer pero sí que podemos conocer una estimación de éstos conocida como 'residuos' cuya definición es

$$\widehat{\epsilon_i} = Y_i - \widehat{\beta_0} - \sum_{j=1}^{p} \widehat{\beta_j} \cdot x_{ij}, \text{ con } i \in \{1, ..., n\}$$

los cuales se utilizarán para comprobar estas hipótesis. Utilizando estos residuos podemos estimar la varianza que tiene el error como

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \widehat{\epsilon_i}^2$$

En notación matricial, esto se puede escribir de la siguiente forma:

$$\widehat{\epsilon} = Y - \widehat{Y} = Y - X \cdot \widehat{\beta}$$

y

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \widehat{\epsilon}^{\intercal} \widehat{\epsilon}$$

Ejemplo 1.2. Si reutilizamos el ejemplo anterior, podemos ver cuáles serían los residuos y estimar la varianza del error $\widehat{\sigma}^2$ y ver si se acerca a 1. Los residuos que tenemos son:

```
res

>             [,1]
> [1,] -0.2229349
> [2,] -0.2207620
> [3,]  1.2845231
> [4,] -0.5496119
> [5,] -0.7566234
> [6,]  0.4654090
```

Y si los utilizamos para obtener una estimación de la varianza del error obtenemos:

```
sig2est

>           [,1]
> [1,] 0.9465311
```

La cual vemos que se acerca bastante al valor que realmente tenemos.

1.4. El t-test. Aplicación en la selección de variables

Otra herramienta muy útil en los modelos de regresión es el t-test, el cual nos permite contrastar la hipótesis nula de que un β_i es igual a una constante, usualmente 0 ya que al hacer esto indirectamente estamos contrastando si X_i realmente tiene algún efecto en Y. Este test se basa en la distribución t de Student. Para aplicar este test, escogemos un nivel de significación (los más usuales son 5 %, 1 % y 0.1 %). Estos niveles de significación nos indican en qué porcentaje de veces se equivoca el test cuando nos asegura que la hipótesis nula es falsa, en otras palabras, al disminuir el nivel de significación estamos haciendo que al test le cueste más rechazar la hipótesis nula.

Cuando aplicamos el t-test a nuestro conjunto de datos, podemos obtener lo que se conoce como p-valor, que representa el porcentaje de veces que, suponiendo la hipótesis nula cierta, obtenemos un valor peor. Al obtener este valor simplemente debemos compararlo con el nivel de significación que tenemos. Si el p-valor es mayor que nuestro nivel de significación, aceptaremos la hipótesis nula y la rechazaremos en caso contrario.

Para construir modelos de regresión existen varias maneras dependiendo de la distribución de los datos o de la cantidad de variables explicativas. Cuando tenemos varias variables explicativas seguramente debamos prescindir de algunas de ellas en favor de otras, pues puede ser que dichas variables explicativas realmente no tengan ninguna relación con la variable respuesta o que estén relacionadas con otras variables explicativas y que si las incluimos, esconderán el efecto real que tienen las otras variables. Veámoslo con un ejemplo.

Ejemplo 1.3. Si consideramos el primer ejemplo relativo al insecticida, existe el comando `lm` en R que nos construye automáticamente el modelo, y el comando `summary` que nos devuelve la siguiente salida:

```
summary(mod)

>
> Call:
> lm(formula = dead ~ conc, data = blis)
>
> Residuals:
>          1          2          3          4          5
> -1.221e-15 -5.000e-01  5.524e-16  1.500e+00 -1.000e+00
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)   2.0000     0.8367    2.39 0.096702 .
> conc          6.5000     0.3416   19.03 0.000317 ***
```

```
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 1.08 on 3 degrees of freedom
> Multiple R-squared:  0.9918,Adjusted R-squared:  0.989
> F-statistic: 362.1 on 1 and 3 DF,  p-value: 0.0003168
```

En el apartado de `Coefficients` vemos las estimaciones de los coeficientes y del intercepto (que es otra manera de nombrar a β_0). Además vemos al final de esta tabla unos puntos en la fila correspondiente a `conc`. Estos puntos se corresponden al t-test que definimos anteriormente. R nos devuelve una serie de puntos al lado del p-valor de cada coeficiente indicando para que niveles de significación podemos aceptar la hipótesis nula. Pero si introducimos en el modelo una nueva variable que denotaremos por `conc2`, que es simplemente el resultado de elevar al cuadrado las concentraciones, obtendremos lo siguiente:

```
summary(mod1)

>
> Call:
> lm(formula = dead ~ conc + conc2, data = bliss)
>
> Residuals:
>       1       2       3       4       5
>  0.4286 -0.7143 -0.4286  1.2857 -0.5714
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)   1.5714     1.1249   1.397   0.2972
> conc          7.3571     1.3325   5.521   0.0313 *
> conc2        -0.2143     0.3194  -0.671   0.5714
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 1.195 on 2 degrees of freedom
> Multiple R-squared:  0.9933,Adjusted R-squared:  0.9866
> F-statistic: 148.1 on 2 and 2 DF,  p-value: 0.006707
```

En este caso vemos que el t-test no es capaz de asegurarnos que `conc` influye en el modelo con demasiada seguridad y rechaza completamente que tanto β_0 como el coeficiente de `conc2` sean distintos de 0. Ésto nos permite ver que la presencia de dos variables relacionadas en el modelo impide ver claramente si las variables influyen o no, pues parte de la influencia que tiene `conc` se la está llevando `conc2`.

1.5. Selección de variables

Cuando queremos hacer un modelo de regresión a partir de una serie de variables, debemos intentar buscar el mejor modelo posible según el criterio que hayamos elegido y que, al mismo tiempo, sea lo más sencillo posible. Para lograr esto, la manera de proceder sería considerar todos los posibles subconjuntos del conjunto de variables explicativas, hacer los modelos con estos subconjuntos y ver cuál es mejor. Sin embargo, esta opción es inviable cuando el número de variables que estamos considerando es grande, pues tenemos que crear y comparar 2^p modelos y esto conlleva un coste computacional elevado.

Para evitar tener que crear todos estos modelos se suele proceder de dos maneras alternativas y una manera intermedia:

- Métodos Forward: Cogemos un modelo muy sencillo, por ejemplo sólo considerando el intercepto β_0, y vamos añadiendo variables según un criterio hasta que dicho criterio nos impida añadir más.

- Métodos Backward: Partimos de un modelo muy complejo con todas las variables que creemos que pueden influir en la variable respuesta, y se van eliminando variables del modelo siguiendo algún criterio hasta que no proceda eliminar ninguna más.

- Métodos Stepwise: Estos métodos parten de cualquier modelo y lo comparan con todos los posibles modelos resultantes de eliminar o añadir variables utilizando el mismo criterio que en los otros dos casos. Realizamos este proceso hasta que el modelo no mejore ni añadiendo ni eliminando variables.

El criterio que hemos mencionado en estos métodos, podemos considerarlo como un criterio de significación o un criterio global. Si consideramos un criterio de significación, en el caso de los métodos forward incluiríamos las variables que al añadirlas, su coeficiente es significativamente distinto de 0 (utilizando el t-test) y pararíamos cuando todas las variables que no están incluidas tienen coeficientes no significativos. En el caso backward, eliminaríamos variables menos significativas hasta que todas las variables que nos quedan en el modelo sean significativas.

Los criterios globales, en lugar de considerar la significación de cada coeficiente, construyen una medida global de cada modelo para decir si un modelo es mejor que otro, por lo que buscaremos el mejor modelo basándonos en esta medida. En nuestro caso, sólo vamos a ver el Criterio de Información de Akaike (AIC) que consiste en:

$$AIC = 2p - 2 \cdot ln(\text{verosimilitud})$$

donde 'verosimilitud' es la función de verosimilitud, que en estadística se refiere a una función de los coeficientes que nos permite hacer inferencia acerca de su valor real a partir

de una serie de observaciones, para lo cual debemos intentar maximizar esta función. Si intentamos minimizar el valor de AIC estaríamos maximizando la verosimilitud y minimizando el número de variables en el modelo p, ambos objetivos suelen ser opuestos. Por lo tanto, al hacer esto estamos intentando encontrar un modelo que incluya exclusivamente variables que realmente aumenten la verosimilitud, y por tanto, influyan realmente en la variable respuesta Y.

Ejemplo 1.4. Consideremos otro conjunto de datos de la librería Faraway (2016) llamado `savings`. En él tenemos las siguientes variables:

- `sr`: Ahorro personal entre renta disponible.
- `pop15`: Porcentaje de población menor de 15 años.
- `pop75`: Porcentaje de población mayor de 75 años.
- `dpi`: Renta disponible per cápita en dólares.
- `ddpi`: Tasa de crecimiento porcentual de dpi.

Ahora construyamos un modelo para intentar predecir `sr` en función de las otras variables. Primero utilicemos un método backward con un criterio de significación. Para ello, hagamos el modelo con todas las variables y utilicemos el comando `summary`:

```
>
> Call:
> lm(formula = sr ~ ., data = datos)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -8.2422 -2.6857 -0.2488  2.4280  9.7509
>
> Coefficients:
>              Estimate Std. Error t value Pr(>|t|)
> (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
> pop15       -0.4611931  0.1446422  -3.189 0.002603 **
> pop75       -1.6914977  1.0835989  -1.561 0.125530
> dpi         -0.0003369  0.0009311  -0.362 0.719173
> ddpi         0.4096949  0.1961971   2.088 0.042471 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 3.803 on 45 degrees of freedom
> Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
> F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Para decidir qué variable eliminaremos miramos cuál tiene el mayor valor en la última columna, y por tanto una peor significación. En este caso es `dpi`. Si la eliminamos del modelo y volvemos a ver la significación obtenemos:

```
>
> Call:
> lm(formula = sr ~ pop15 + pop75 + ddpi, data = datos)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -8.2539 -2.6159 -0.3913  2.3344  9.7070
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  28.1247     7.1838   3.915 0.000297 ***
> pop15        -0.4518     0.1409  -3.206 0.002452 **
> pop75        -1.8354     0.9984  -1.838 0.072473 .
> ddpi          0.4278     0.1879   2.277 0.027478 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 3.767 on 46 degrees of freedom
> Multiple R-squared:  0.3365,Adjusted R-squared:  0.2933
> F-statistic: 7.778 on 3 and 46 DF,  p-value: 0.0002646
```

La variable `pop75` que antes no era significativa, ahora sí lo es (podemos decir que `dpi` ocultaba esta significación) y con ésto, todas las variables del modelo son significativas para alguno de los niveles de confianza usuales y por tanto ya habríamos llegado al modelo óptimo.

Si hacemos ahora un método forward utilizando como criterio el Criterio de Información de Akaike, creamos un modelo con sólo β_0 y calculamos su AIC:

```
mod.null = lm(sr ~ 1, data = datos)
AIC(mod.null)

> [1] 294.8551
```

Ahora hacemos los modelos con cada una de las variables, calculamos su AIC y vemos cuál minimiza más dicho valor:

```
>    pop15    pop75      dpi     ddpi
> 285.2260 291.5768 294.3663 291.9802
```

Vemos que todos mejoran el AIC del modelo sin variables, pero el que más lo hace es `pop15` por lo que la consideramos en el modelo, y procederíamos de manera análoga añadiendo el resto de variables al modelo y viendo si alguna mejora el AIC. Si hacemos ésto, obtendríamos el mismo modelo que en el caso del método backward (lo cual no se cumple en general) y si calculamos su AIC tendríamos:

```
AIC(mods)

> [1] 280.3414
```

1.6. Predicción

Una última cosa que podemos calcular utilizando los modelos de regresión son los intervalos de confianza. Cuando ya hemos obtenido un modelo de regresión que cumple las hipótesis y tenemos calculados sus coeficientes podemos intentar predecir el valor que tomará Y cuando las variables explicativas toman unos nuevos valores que denominaremos X_0. Esta predicción se obtiene simplemente sustituyendo en la fórmula general que teníamos obteniendo

$$\tilde{Y_0} = X_0 \cdot \widehat{\beta}$$

Esta predicción tiene una cierta precisión asociada, es decir, existe un margen de error que expresamos en forma de un intervalo cuyo punto medio es la predicción que habíamos obtenido:

$$\left(\tilde{Y_0} - z, \tilde{Y_0} + z\right) = \left(\tilde{Y_0} \pm z\right)$$

donde z es una medida que tenemos de la precisión de la predicción. Para calcular esta precisión, primero debemos fijarnos en qué estamos prediciendo. Existen dos predicciones cuyo valor numérico es el que hemos calculado pero la precisión es distinta: la estimación de la media condicionada y la predicción de una nueva observación. La predicción de una nueva observación consiste en intentar predecir el valor que obtendremos para $\tilde{Y_0}$ al aplicar las condiciones X_0 mientras que la estimación de la media condicionada consiste en intentar averiguar cuanto valdrá la media de la variable $\tilde{Y_0}$, es decir, si hacemos el mismo experimento con las condiciones X_0 varias veces, cual será la media de todas las $\tilde{Y_0}$ que obtendremos. Por esto, es evidente que en ambos casos el punto medio del intervalo es el mismo, pero en el segundo caso el intervalo será mayor que en el primero. Para calcular esta z primero calculemos la varianza de $\widehat{\beta}$:

$$Var(\widehat{\beta}) = Var((X^\intercal X)^{-1} X^\intercal Y) =$$

$$= (X^\intercal X)^{-1} X^\intercal Var(Y)((X^\intercal X)^{-1} X^\intercal)^\intercal =$$

$$= (X^\intercal X)^{-1} X^\intercal Var(Y) X (X^\intercal X)^{-1}$$

pues $(X^\intercal X)^{-1}$ es simétrica. Para calcular $Var(Y)$, es fácil ver que se cumple $Var(Y) = Var(\epsilon)$ pues $Y = X\beta + \epsilon$ y X y β son datos que tenemos, por lo que la aleatoriedad de la variable Y se la aportan los errores que sí que son variables estadísticas. Además, sabemos como estimar la varianza de cada uno de dichos errores $Var(\epsilon_i) = \widehat{\sigma}^2$ por lo que la varianza de todo el vector será

$$Var(\epsilon) = \widehat{\sigma}^2 I$$

siendo I la matriz identidad (n x n). Por tanto, volviendo a la varianza de $\widehat{\beta}$:

$$Var(\widehat{\beta}) = (X^\intercal X)^{-1} X^\intercal Var(Y) X (X^\intercal X)^{-1} =$$

$$= (X^\intercal X)^{-1} X^\intercal \widehat{\sigma}^2 I X (X^\intercal X)^{-1} =$$

$$= \widehat{\sigma}^2 (X^\intercal X)^{-1} X^\intercal X (X^\intercal X)^{-1} =$$

$$= \widehat{\sigma}^2 (X^\intercal X)^{-1}$$

Además, estos intervalos se construyen para un nivel de significación usualmente denotado como $(1-\alpha)$, donde α denota la probabilidad de que un valor quede fuera del intervalo. Finalmente, el intervalo de confianza para la estimación de la media condicionada es

$$\left[\tilde{Y} \pm t_{\frac{\alpha}{2},(n-(p+1))} \sqrt{X_0 Var(\widehat{\beta}) X_0^\intercal}\right]$$

donde $t_{\frac{\alpha}{2},(n-(p+1))}$ denota el cuantil $\frac{\alpha}{2}$ de la distribución t de Student con $(n-(p+1))$ grados de libertad $T_{(n-(p+1))}$. Para el caso de la predicción de una nueva observación, el intervalo de confianza es

$$\left[\tilde{Y} \pm t_{\frac{\alpha}{2},(n-(p+1))} \sqrt{\widehat{\sigma}^2 + X_0 Var(\widehat{\beta}) X_0^\intercal}\right]$$

Ejemplo 1.5. Para calcular los intervalos de confianza, R tiene un comando que nos permite hacerlo directamente, por lo que no tenemos que hacer los cálculos. Si utilizamos los datos del ejemplo anterior y cogemos una nueva concentración $conc = 1'5$, el intervalo de confianza para la estimación de la media condicionada es:

```
>     fit      lwr      upr
> 1 11.75 10.11948 13.38052
```

Y el intervalo de confianza de una nueva observación es

```
>     fit      lwr      upr
> 1 11.75 7.945457 15.55454
```

Donde vemos que el centro del intervalo, el elemento `fit`, es el mismo en ambos casos y después tenemos los extremos de los intervalos. Vemos que efectivamente, en el segundo caso el intervalo de confianza que obtenemos es mayor que en el primero.

Capítulo 2

Regresión Ridge

Este trabajo se va a centrar en los modelos de regresión con penalizaciones, los cuales son modelos de regresión como los del capítulo anterior pero que además incluyen una penalización en los coeficientes β_i, estableciendo una cota superior en la suma de éstos o en otro tipo de operación que los incluya. Esta penalización puede deberse a limitaciones físicas de las variables o puede ser intencionado. Si volvemos al estimador de los coeficientes, tenemos $\widehat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$, por lo que si nuestra matriz $X^\intercal X$ es singular o próxima a serlo, nuestro estimador tendría mucha varianza. Para restringir este comportamiento, podemos aplicar estas restricciones intencionadas que nos ayudan a mejorar el condicionamiento de dicha matriz y así controlar mejor la estimación que vamos a hacer.

Primero veremos la Regresión Ridge. En español se hizo un intento por traducir el nombre y así utilizar términos como 'Regresión Riscal' o 'Regresión de Arista', aunque no llegaron a ser demasiado conocidos y por tanto en este trabajo utilizaremos el término inglés 'Ridge'.

2.1. Variables centradas

Para empezar, consideremos hacer el siguiente cambio en las variables explicativas, considerando las nuevas variables explicativas $X_i' = X_i - \bar{X}_i, i \in \{1, ..., p\}$. Estas nuevas variables son iguales que las anteriores pero centradas, es decir, con media 0. Como vimos en el capítulo anterior, en cualquier modelo de regresión se tiene que $\beta_0 + \sum_{i=1}^{p} \beta_i \cdot \bar{X}_i = \bar{Y}$ y por tanto, si consideramos las variables explicativas centradas, es fácil ver que $\beta_0 = \bar{Y} = 1/n \sum_{i=1}^{n} Y_i$. A partir de ahora consideraremos que hemos centrado las variables explicativas y por tanto no necesitamos estimar β_0, con lo que podemos considerar en la notación matricial

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

por lo que la fórmula general de los modelos de regresión resulta

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Al definir estas nuevas variables, tenemos que la matriz que queremos evitar que sea singular, $X^\intercal X$, es además lo que se conoce como matriz de covarianzas pues:

$$X^\intercal X = \begin{pmatrix} \sum_{i=1}^n x_{1,i}^2 & \sum_{i=1}^n x_{1,i}x_{2,i} & \cdots & \sum_{i=1}^n x_{1,i}x_{p,i} \\ \sum_{i=1}^n x_{2,i}x_{1,i} & \sum_{i=1}^n x_{2,i}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^n x_{n,i}x_{1,i} & \cdots & \cdots & \sum_{i=1}^n x_{p,i}^2 \end{pmatrix}$$

y utilizando que las variables han sido centradas tenemos:

$$X^\intercal X = \begin{pmatrix} \sum_{i=1}^n (x_{1,i} - \bar{X}_1)^2 & \sum_{i=1}^n (x_{1,i} - \bar{X}_1)(x_{2,i} - \bar{X}_2) & \cdots & \sum_{i=1}^n (x_{1,i} - \bar{X}_1)(x_{p,i} - \bar{X}_p) \\ \sum_{i=1}^n (x_{2,i} - \bar{X}_2)(x_{1,i} - \bar{X}_1) & \sum_{i=1}^n (x_{2,i} - \bar{X}_2)^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^n (x_{n,i} - \bar{X}_n)(x_{1,i} - \bar{X}_1) & \cdots & \cdots & \sum_{i=1}^n (x_{p,i} - \bar{X}_p)^2 \end{pmatrix}$$

donde la componente (i, i) de la diagonal de la matriz es n veces la varianza de la variable X_i, que es una medida de la dispersión de la variable respecto a su media, y en el caso del resto de la matriz, la componente (i, j) es n veces la covarianza entre X_i y X_j, que es también una medida de la varianza conjunta de estas variables.

2.2. Fórmula general. Estimadores

Recordamos la fórmula del estimador de mínimos cuadrados que utilizábamos para obtener el estimador de los coeficientes:

$$\widehat{\beta} = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

La Regresión Ridge consiste en añadirle a esta fórmula una penalización en los coeficientes, obteniendo la siguiente fórmula para la estimación de los coeficientes de Ridge:

$$\widehat{\beta}^{ridge} = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 \leq t \text{ con } t \geq 0$$

Como vemos, estamos restringiendo los valores que toman los coeficientes impidiendo que puedan tomar valores demasiado altos en valor absoluto. El cambio que hicimos previamente al centrar las variables hace que esta restricción funcione mejor, pues si tenemos una variable respuesta y una variable explicativa muy alejadas, el coeficiente será bastante grande ya que debe 'acercar' la variable explicativa a la respuesta. Si aplicamos la Regresión Ridge a un modelo con todas las variables explicativas cerca de la variable respuesta y una alejada, aunque esta última sea la que verdaderamente explica la variable respuesta, su coeficiente sería el más afectado, por lo que al centrar todas las variables estamos evitando que ocurra este problema.

Esta restricción se puede representar de otra manera introduciéndola en la fórmula de mínimos cuadrados:

$$\widehat{\beta}^{ridge} = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=0}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Aquí vemos que aparece un nuevo elemento $\lambda \geq 0$ que se conoce como parámetro de ajuste, el cual está relacionado con el t que teníamos en la restricción anterior. Como podemos ver, la primera parte coincide exactamente con el método de mínimos cuadrados y hemos añadido un elemento $\lambda \sum_{j=1}^{p} \beta_j^2$ que restringe los valores que pueden tomar los coeficientes en función de λ. Si $\lambda = 0$, tenemos que los coeficientes de Ridge coinciden con los que obtenemos por el método de mínimos cuadrados. En otro caso, $\lambda > 0$ nos permite controlar cuanto queremos contraer los coeficientes y se puede ver fácilmente que cuando $\lambda \to \infty$, los coeficientes tienden a 0. A diferencia del método de mínimos cuadrados que nos producía una solución para cada problema, este nuevo método nos produce una solución dependiendo del λ que queramos, por lo que es muy importante saber escoger bien el parámetro para obtener el mejor resultado.

Cuando las variables explicativas que estamos considerando tienen demasiada correlación, los coeficientes se vuelven difíciles de estimar, y puede ocurrir que una variable tenga un coeficiente positivo muy grande que se esté compensando con un coeficiente negativo muy grande de otra variable, por lo que al aplicar esta penalización a los coeficientes al cuadrado estamos impidiendo que ocurra ésto y ajustando mejor los $\beta_i, i \in \{1, ..., p\}$. Observemos que no estamos aplicando esta restricción sobre β_0, pues lo que queremos restringir es el efecto estimado que tienen las variables explicativas sobre la variable respuesta Y, y β_0 es simplemente la media de la variable Y cuando todas las variables explicativas están centradas.

Cuando calculamos el estimador de mínimos cuadrados en notación matricial partíamos de que $RSS = (Y - X\beta)^\intercal (Y - X\beta)$. En el caso de la Regresión Ridge la fórmula que tenemos

es un poco distinta:

$$RSS(\lambda) = (Y - X\beta)^\intercal (Y - X\beta) + \lambda\beta^\intercal\beta$$

Si derivamos e igualamos a 0 al igual que hicimos en el caso de mínimos cuadrados, obtenemos que el estimador de β en la Regresión Ridge es

$$\widehat{\beta}^{ridge} = (X^\intercal X + \lambda I)^{-1} X^\intercal Y$$

con I la matriz identidad $p \times p$.

Si $X^\intercal X$ es singular o está cerca de estarlo, en el caso de mínimos cuadrados teníamos que hacer su inversa lo que provoca que la estimación que obtenemos tenga mucha varianza, sin embargo podemos ver que en este caso le estamos sumando una matriz diagonal constante, lo que nos permite corregir este problema y así evitar hacer la inversa de una matriz singular o mal condicionada.

Ejemplo 2.1. Veamos esto con un ejemplo. Si consideramos tres vectores X_1, X_2, X_3 formados por 100 elementos escogidos de una normal $N(0,1)$ que serán nuestras variables explicativas y nuestra variable respuesta $Y = 10 + 5 \cdot X_1 + 7 \cdot X_2 - 4 \cdot X_3 + \epsilon$, con ϵ un vector de errores con distribución normal. Si centramos las variables y hacemos el modelo de regresión correspondiente a estos datos obtenemos la siguiente salida:

```
summary(mod1)

>
> Call:
> lm(formula = Y ~ x1 + x2 + x3)
>
> Residuals:
>      Min       1Q   Median       3Q      Max
> -1.24569 -0.32696  0.02832  0.33516  1.26605
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  9.98189    0.05258  189.86   <2e-16 ***
> x1           4.97227    0.05844   85.09   <2e-16 ***
> x2           7.02311    0.05473  128.33   <2e-16 ***
> x3          -4.02869    0.05612  -71.79   <2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.5258 on 96 degrees of freedom
> Multiple R-squared:  0.9967,Adjusted R-squared:  0.9966
> F-statistic:  9695 on 3 and 96 DF,  p-value: < 2.2e-16
```

Donde vemos que podemos estimar bastante bien todos los coeficientes. Pero si añadimos una nueva variable $X_4 = X_1 + X_2 + X_3$, obtenemos la siguiente salida:

```
summary(mod2)

>
> Call:
> lm(formula = Y ~ x1 + x2 + x3 + x4)
>
> Residuals:
>      Min       1Q   Median       3Q      Max
> -1.24569 -0.32696  0.02832  0.33516  1.26605
>
> Coefficients: (1 not defined because of singularities)
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  9.98189    0.05258  189.86   <2e-16 ***
> x1           4.97227    0.05844   85.09   <2e-16 ***
> x2           7.02311    0.05473  128.33   <2e-16 ***
> x3          -4.02869    0.05612  -71.79   <2e-16 ***
> x4                NA         NA      NA       NA
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.5258 on 96 degrees of freedom
> Multiple R-squared:  0.9967,Adjusted R-squared:  0.9966
> F-statistic:  9695 on 3 and 96 DF,  p-value: < 2.2e-16
```

Aquí, R detecta que una de las variables es linealmente dependiente de las otras y por tanto hace el modelo sin considerarla, pero en un ejemplo real podemos tener una variable linealmente dependiente a otras, que por errores de medición o por características propias a la variable, sus mediciones no lo muestren exactamente por lo que la matriz X no sería singular y por lo tanto R no haría esta corrección, pero sí que estaría muy próxima a serlo, por lo que nos interesa algún método que nos permita calcular sus coeficientes de manera fiable. Para ver esto, podemos considerar una nueva variable $X_5 = X_1 + X_2 + X_3 + \delta$, la cual es prácticamente X_4 pero con un error δ muy pequeño, pero suficiente para que X no sea exactamente singular. En este caso obtenemos la siguiente salida:

```
summary(mod2)

>
> Call:
> lm(formula = Y ~ x1 + x2 + x3 + x5)
>
```

```
> Residuals:
>      Min       1Q   Median       3Q      Max
> -1.24195 -0.33273  0.02213  0.33088  1.25937
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  9.98333    0.05316 187.803   <2e-16 ***
> x1           6.32687    5.52102   1.146    0.255
> x2           8.37880    5.52543   1.516    0.133
> x3          -2.67174    5.53062  -0.483    0.630
> x5          -1.35762    5.53305  -0.245    0.807
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.5283 on 95 degrees of freedom
> Multiple R-squared:  0.9967,Adjusted R-squared:  0.9966
> F-statistic:  7200 on 4 and 95 DF,  p-value: < 2.2e-16
```

Donde vemos que ya no somos capaces de asegurar que ninguna de las variables influye en el modelo. Esto significa que aunque las estimaciones que obtenemos pueden ser acertadas, en el caso de que no conociéramos los valores reales de estos coeficientes, no podríamos estar seguros ni de que estos valores no sean 0 y realmente ninguna de las variables influya.

Aquí es donde entra la Regresión Ridge, que nos permite obtener una estimación de los coeficientes a pesar de este problema y con mucha más seguridad de los coeficientes que obtenemos. Si aplicamos la Regresión Ridge con un parámetro $\lambda = 0{,}5$, obtenemos las siguientes estimaciones de los coeficientes $\beta_i, i \in \{1, 2, 3, 4\}$:

```
>         [,1]
> x1  2.464133
> x2  4.498541
> x3 -6.493962
> x5  2.502227
```

y β_0:

```
> [1] 9.981889
```

Los coeficientes que obtenemos no son los que teníamos al principio, pero en realidad son bastante acertados pues si redondeamos con una cifra decimal obtenemos que el modelo de regresión que nos da el modelo es $Y = 10 + 2{,}5 \cdot X_1 + 4{,}5 \cdot X_2 - 6{,}5 \cdot X_3 + 2{,}5 \cdot X_5$ y si tenemos en cuenta que $X_5 = X_1 + X_2 + X_3 + \delta$, ya obtendríamos la relación que teníamos al principio con la suma de un error muy pequeño.

2.3. Predicción

Aunque no aparezca en los libros de la bibliografía, podemos calcular los intervalos de confianza para la Regresión Ridge de forma análoga a mínimos cuadrados, pero teniendo en cuenta que el estimador $\widehat{\beta}$ es distinto. En este caso, la varianza de $\widehat{\beta}$ es:

$$Var(\widehat{\beta}) = Var((X^\intercal X + \lambda I)^{-1} X^\intercal Y) =$$

$$= (X^\intercal X + \lambda I)^{-1} X^\intercal Var(Y)((X^\intercal X + \lambda I)^{-1} X^\intercal)^\intercal$$

$$= (X^\intercal X + \lambda I)^{-1} X^\intercal Var(Y) X (X^\intercal X + \lambda I)^{-1}$$

La varianza de Y es la misma, por lo que obtenemos:

$$Var(\widehat{\beta}) = (X^\intercal X + \lambda I)^{-1} X^\intercal \widehat{\sigma}^2 I X (X^\intercal X + \lambda I)^{-1} =$$

$$= \widehat{\sigma}^2 (X^\intercal X + \lambda I)^{-1} X^\intercal X (X^\intercal X + \lambda I)^{-1}$$

y en este caso no podemos reducirlo más. Finalmente, las fórmulas para calcular los intervalos son las mismas, sólo debemos tener en cuenta que hemos centrado las variables. Para la estimación de la media condicionada tenemos:

$$\left[\tilde{Y} \pm t_{\frac{\alpha}{2},(n-p)} \sqrt{X_0 Var(\widehat{\beta}) X_0^\intercal}\right]$$

Y para el caso de la predicción de una nueva observación:

$$\left[\tilde{Y} \pm t_{\frac{\alpha}{2},(n-p)} \sqrt{\widehat{\sigma}^2 + X_0 Var(\widehat{\beta}) X_0^\intercal}\right]$$

Ejemplo 2.2. Considerando el ejemplo anterior podemos intentar calcular $Var(\widehat{\beta})$:

```
Var

>               x1            x2            x3            x5
> x1  0.0027269904 -0.0003151771 -0.0001511995 -0.0003018535
> x2 -0.0003151771  0.0027617421 -0.0004150006 -0.0005304480
> x3 -0.0001511995 -0.0004150006  0.0027031877 -0.0004247816
> x5 -0.0003018535 -0.0005304480 -0.0004247816  0.0013071364
```

Si calculamos el intervalo de confianza para la estimación de la media condicionada con un nivel de significación de $(1 - 0'05)$ para la nueva observación $X_1 = 1$, $X_2 = 1$ y $X_3 = 1$, teniendo en cuenta que entonces $X_5 = 1 + 1 + 1 + \delta$, con $\delta \in N(0, 0'01)$ y obtenemos:

```
>      lwr      upr
> 17.73767 18.14677
```

y en el caso de la predicción de una nueva variable:

```
>      lwr      upr
> 16.87282 19.01163
```

En vista de estos intervalos, podemos calcular el valor real que tendría Y en este caso, pues $Y = 10 + 5 \cdot X_1 + 7 \cdot X_2 - 4 \cdot X_3 + \epsilon$:

```
> [1] 17.50315
```

que como podemos ver está fuera del intervalo de confianza para la estimación de la media, pero dentro del intervalo de confianza para una nueva observación. Ésto se debe a que el intervalo de confianza para la estimación de la media es un intervalo donde estará la media de las observaciones, por lo que si seguimos haciendo observaciones con los mismos datos, la media de todos ellos debería estar dentro de dicho intervalo.

2.4. Descomposición de valores singulares. Grados efectivos de libertad

La descomposición de valores singulares (*singular value decomposition*) de una matriz A $(m \times n)$ es

$$A = UDV^\intercal$$

donde U$(m \times n)$ y V$(n \times n)$ son matrices ortogonales y D es una matriz diagonal $n \times n$ cuyos valores de la diagonal cumplen $d_1 \geq d_2 \geq ... \geq d_n \geq 0$ y son los autovalores de la matriz A.

Esta descomposición la podemos aplicar en nuestro caso a la matriz X. Si utilizamos esta descomposición en el estimador de β por mínimos cuadrados, obtenemos

$$X\widehat{\beta}^{mc} = X(X^\intercal X)^{-1}XY = UU^\intercal Y$$

y en el estimador usando la Regresión Ridge:

$$X\widehat{\beta}^{ridge} = X(X^\intercal X + \lambda I)^{-1}X^\intercal Y = UD(D^2 + \lambda I)^{-1}DU^\intercal Y = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^\intercal Y$$

donde u_j son las columnas de la matriz U.

Como $\lambda \geq 0$, $\frac{d_j^2}{d_j^2+\lambda} \leq 1$. Además, vemos que si $\lambda = 0$, tendríamos la misma fórmula que para mínimos cuadrados. Por lo tanto, podemos decir que la solución de la Regresión Ridge es la solución por mínimos cuadrados contraída por el factor $\frac{d_j^2}{d_j^2+\lambda}$, por lo que cuanto menor sea d_j^2, mayor será la contracción. Ahora estudiemos cuando d_j^2 es pequeño.

d_j es otra manera de expresar las componentes principales de las variables en X. La matriz de covarianzas viene dada por

$$S = X^\intercal X/n$$

donde si aplicamos la descomposición que vimos previamente tenemos

$$S = X^\intercal X/n = VD^2V^\intercal/n$$

($VD^2V^\intercal$ se denomina autodescomposición de $X^\intercal X$).

Las columnas de V, v_j, son las direcciones de las componentes principales de X. La dirección de la primera componente principal, v_1, cumple que $z_1 = X \cdot v_1$ tiene la mayor varianza de todas las combinaciones lineales normalizadas de las columnas de X, la cual vale

$$Var(z_1) = Var(X \cdot v_1) = \frac{d_1^2}{n}$$

además, tenemos que $z_1 = X \cdot v_1 = u_1 \cdot d_1$. z_1 se denomina primera componente principal de X (aquí estamos cometiendo un abuso de notación pues z_1 en realidad es componente principal de $X^\intercal X$ pero se le denota así para abreviar pues se utiliza más que la verdadera componente principal de X) y u_1 primera componente principal normalizada. Esto se puede aplicar la j-ésima componente principal $z_j = X \cdot v_j$ cuya varianza será $Var(z_j) = \frac{d_j^2}{n}$ y por tanto, la última componente principal z_p tiene la menor varianza de todas.

Ejemplo 2.3. Si utilizamos el ejemplo anterior, podemos calcular su descomposición de valores singulares y ver cuales son los valores de la matriz D, es decir, cuales son los autovalores de la matriz X y con ellos calcular los de la matriz de covarianzas S. Los autovalores de X son:

```
S = svd(X)
S$d

> [1] 17.88559950  9.82704225  9.29535541  0.04808657
```

y por tanto los de S son:

```
S$d^2

> [1] 3.198947e+02 9.657076e+01 8.640363e+01 2.312318e-03
```

Aquí vemos que uno de ellos es muy pequeño y es el que provoca que al calcular $(X^\intercal X)^{-1}$ nos produzca esas malas estimaciones. Para corregir este problema, la Regresión Ridge suma a esta diagonal un valor λ que evita que exista algún autovalor tan pequeño. La mejor manera de escoger el λ para utilizar en la Regresión Ridge es uno que sea lo suficientemente grande para que los autovalores que queremos corregir dejen de ser tan pequeños y que al resto de autovalores apenas afecte. En el ejemplo anterior utilizamos $\lambda = 0{,}5$ pues es lo suficientemente grande para corregir el problema del último autovalor y al segundo autovalor más pequeño, 86,40363, apenas le influye sumarle esa cantidad.

Esta descomposición nos permite ver también las direcciones de mayor dispersión de los datos, puesto que V son los autovectores de la matriz de covarianzas S.

Al igual que en los modelos de regresión usuales podemos definir la matriz `hat` de manera análoga:

$$H_\lambda = X(X^\intercal X + \lambda I)^{-1} X^\intercal$$

y utilizando esta matriz, podemos definir la siguiente función:

$$df(\lambda) = tr[X(X^\intercal X + \lambda I)^{-1} X^\intercal] = tr(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Ésta es una función monótona decreciente de λ que denominamos como los grados efectivos de libertad del ajuste de la Regresión Ridge. En el caso de mínimos cuadrados (cuando $\lambda = 0$) podemos ver que $df(\lambda) = p$ y si consideramos la Regresión Ridge, tenemos que cuando $\lambda \to \infty$, $df(\lambda) \to 0$. Estos grados de libertad son una manera de estimar cuantas variables del modelo verdaderamente influyen en la variable respuesta. En el caso de los modelos de regresión usuales, el grado de libertad siempre es p ya que en este caso asumimos que todas las variables influyen en el modelo, pero en la Regresión Ridge esta función depende del λ que escojamos, por lo que debemos elegir bien este parámetro para que refleje bien el número de variables que influyen en el modelo. Tenemos que considerar que en realidad tenemos un grado de libertad más correspondiente al intercepto β_0 que eliminamos antes del vector β y su correspondiente columna en X.

Ejemplo 2.4. En el ejemplo anterior podemos calcular esta función, que para el caso que estamos considerando con $\lambda = 0{,}5$ obtenemos

```
> [1] 2.992138
```

Vemos que tenemos prácticamente 3, que es el número de variables que influyen en Y puesto que X_5 no influye directamente.

2.5. Elección de un buen parámetro λ

Para poder decidir un buen parámetro λ para realizar la Regresión Ridge, se suele utilizar el error cuadrático medio e intentar reducirlo. R tiene una librería que nos permite buscar un λ óptimo llamada `glmnet`. Utilizando el mismo ejemplo de antes obtenemos la gráfica **??** donde vemos a la izquierda el intervalo óptimo para coger el λ cuyos extremos son las líneas verticales discontinuas. R también nos permite obtener el mínimo que sería:

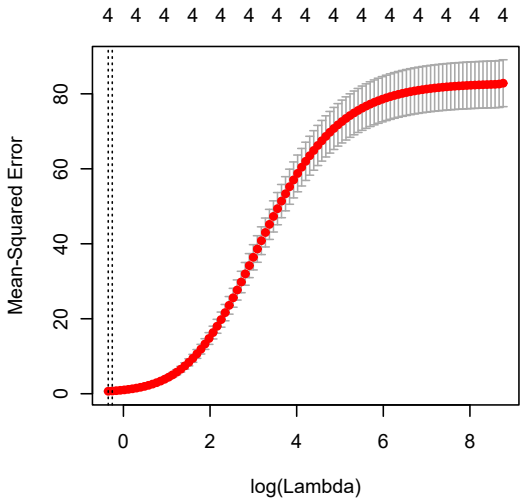

Figura 2.1: *Estamos representando en el eje x el logaritmo de λ y en el eje y el error cuadrático medio. También obtenemos un intervalo donde se encuentran los mejores λ delimitado por las líneas discontinuas verticales.*

```
l = mod3$lambda.min
l

> [1] 0.7042749
```

Esta es la mejor opción para utilizar como λ. Si volviéramos a hacer el ejemplo que hemos hecho durante el capítulo con este nuevo λ obtendríamos como coeficientes:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                     s0
> (Intercept)  9.980481
> x1           3.407972
> x2           5.262055
> x3          -4.981423
> x5           1.329332
```

que vemos que restringen más el valor del coeficiente X_5, que es la variable dependiente de las otras, como buscábamos. También podemos calcular los intervalos de confianza para la misma nueva observación. El intervalo para la estimación de la media condicionada sería

```
>      lwr      upr
> 17.73279 18.14293
```

y el intervalo para la predicción de una nueva observación:

```
>      lwr      upr
> 16.86501 19.01070
```

Aunque sea poco, podemos ver que los intervalos son mejores que considerando $\lambda = 0{,}5$. No cambian demasiado debido a que estábamos considerando un λ bastante próximo al óptimo.

Un problema que presenta la Regresión Ridge es que no elimina variables explicativas del modelo, sino que sólo restringe la suma de los coeficientes, o dicho de otra manera, nunca vamos a tener que $\beta_j = 0$ para algún $j \in \{1, ..., p\}$ a menos que consideremos $\lambda = \infty$, con lo que no nos sirve para descartar variables explicativas que puedan no influir en el modelo.

Capítulo 3

Lasso

El nombre de Lasso proviene de las siglas de 'Least Absolute Shrinkage and Selection Operator'. Este método pretende corregir el mismo problema que la Regresión Ridge, pero además pudiendo hacer que ciertas variables se anulen, cosa que en Ridge nunca llegaba a ocurrir. Esto podría no ser un gran problema, ya que tenemos un modelo con sus coeficientes estimados y la precisión con la que estimamos no se vería comprometida, pero hace muy complicada la interpretación de un modelo, pues da igual las variables que escojamos que siempre obtendremos un coeficiente que las relaciona. Por ello, nos interesa buscar un método que anule coeficientes en caso de ser necesario.

En este capítulo también supondremos que hemos centrado las variables.

3.1. Fórmula general

De manera análoga a la Regresión Ridge, Lasso coge el estimador de mínimos cuadrados y le impone una restricción. El estimador de Lasso es:

$$\widehat{\beta}^{lasso} = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \text{ sujeto a } \sum_{j=1}^{p} |\beta_j| \leq t \text{ con } t \geq 0$$

Si no tuviéramos la penalización, el estimador que obtendríamos sería el de mínimos cuadrados $\widehat{\beta}^{mc}$. Si consideramos $t_0 = \sum_{j=1}^{p} |\widehat{\beta}_j^{mc}|$, entonces para cualquier $t \geq t_0$ se cumple que $\widehat{\beta}^{lasso} = \widehat{\beta}^{mc}$. En cambio, si consideramos un $t = \frac{t_0}{2}$, estaríamos contrayendo el estimador de mínimos cuadrados en un 50 % de media. El estimador de Lasso se puede escribir en la forma Lagrangiana como

$$\widehat{\beta}^{lasso} = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \text{ con } \lambda \geq 0$$

Al igual que el estimador de la Regresión Ridge, si tenemos un $\lambda = 0$, estaríamos obteniendo el estimador de mínimos cuadrados y cuanto mayor sea el λ, mayor será la contracción y más cercanos a cero serán los coeficientes.

Si queremos escribir la fórmula anterior utilizando notación matricial, se puede escribir como:

$$\widehat{\beta}^{lasso} = \min_{\beta}(Y - X\beta)^{\intercal}(Y - X\beta) + \lambda||\beta||_1$$

El problema de Lasso es que no tenemos una manera sencilla de calcular el estimador $\widehat{\beta}^{lasso}$, puesto que no podemos escribir su fórmula como suma o producto de las matrices X o Y. Para poder calcular los coeficientes de Lasso utilizaremos el paquete de R llamado `glmnet` (véase Hastie (2019)) que vimos en el capítulo anterior, el cual incluye comandos de R que nos permiten trabajar con estos métodos sin necesidad de calcular directamente los coeficientes.

La gran diferencia de Lasso respecto a la Regresión Ridge es que Lasso actúa como un método 'Subset Selection', es decir, selecciona variables que no influyen en el modelo y considera sus coeficientes 0. Ahondaremos más en estas diferencias más adelante.

Ejemplo 3.1. Para ver que Lasso selecciona variables al contrario que la Regresión Ridge, podemos utilizar el ejemplo del capítulo anterior. En él considerábamos tres variables explicativas con 100 observaciones pertenecientes a una distribución normal de media 0 y varianza 1:

$$x_{i,j} \in N(0,1), \text{ con } i \in \{1, ..., 100\}, j \in \{1, 2, 3\}$$

la variable respuesta que estamos intentando predecir es $Y = 10 + 5 \cdot X_1 + 7 \cdot X_2 - 4 \cdot X_3 + \epsilon$. Además, en el modelo introducimos una variable $X_5 = X_1 + X_2 + X_3 + \delta$, que es una variable dependiente de las otras con un error muy pequeño para evitar que las filas de X sean linealmente dependientes.

Para buscar el λ óptimo para Lasso podemos hacer de manera análoga a la Regresión Ridge, la gráfica 3.1, que nos delimita con líneas verticales discontinuas el intervalo de λ's óptimos nos permite encontrar el λ óptimo.

Si aplicamos Lasso con dicho λ, obtenemos los siguientes coeficientes:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                    s0
> (Intercept)  9.981889
> x1           4.932654
> x2           6.980159
> x3          -3.991571
> x5           .
```

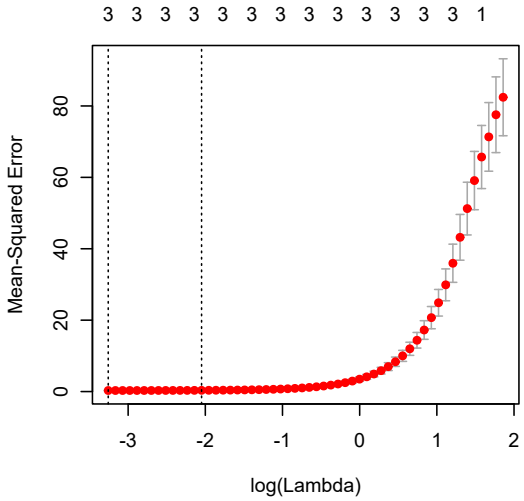

Figura 3.1: *Representación del log(λ) frente al error cuadrático medio de manera análoga a la Regresión Ridge. Los números que tenemos encima de la gráfica son las variables que Lasso considera que influyen en el modelo, que resulta al aplicar el λ correspondiente de la parte inferior.*

Vemos que Lasso aproxima con bastante exactitud los coeficientes y le da el valor 0 al coeficiente de la variable X_5.

3.2. Predicción

Al no tener una fórmula matricial para $\widehat{\beta}^{lasso}$, no podemos calcular $Var(\widehat{\beta})$, y sin este valor, no podemos obtener el intervalo para la estimación de la media condicionada. Sin embargo, sí que podemos obtener el intervalo para la predicción de una nueva observación sin conocer $Var(\widehat{\beta})$.

Como sabemos, cuando aplicamos un modelo de regresión estamos suponiendo que se cumple que

$$Y = X\beta + \epsilon$$

Si aplicamos Lasso y obtenemos una estimación de los coeficientes, estamos suponiendo que se cumple:

$$Y = X\widehat{\beta}^{lasso} + \epsilon$$

Por lo tanto, para una nueva observación obtendríamos:

$$Y_0 = X_0\widehat{\beta}^{lasso} + \epsilon$$

Por lo que la varianza de la nueva observación viene dada por ϵ. Este ϵ cumple dos características que también hemos supuesto ciertas. Una es que se cumple

$$\epsilon_i \in N(0, \sigma^2)$$

y la otra es que dicho σ^2 es el mismo para todos los elementos de ϵ. En el primer capítulo hemos visto que podemos estimar el valor de σ^2 como:

$$\widehat{\sigma}^2 = \frac{1}{n-p}\widehat{\epsilon}^{\intercal}\widehat{\epsilon}$$

Por lo que podemos suponer que

$$Y_0 \in N(X\beta, \widehat{\sigma}^2)$$

y utilizar esto para obtener el intervalo para la predicción de una nueva observación.

Ejemplo 3.2. Si consideramos el mismo ejemplo que utilizamos en el capítulo anterior, podemos ver los intervalos para la predicción de una nueva observación y compararlos con los de la Regresión Ridge. Utilizando los coeficientes estimados de Lasso que obtuvimos anteriormente, podemos obtener los residuos y con ellos calcular una estimación de σ^2:

```
sig2est

> [1] 0.2838807
```

Y utilizando este resultado, podemos obtener que el intervalo para una nueva observación, asumiendo que las variables X_1, X_2 y X_3 valen 1, con un nivel de confianza del 95 % es

```
>      lwr      upr
> 16.85885 18.94741
```

Vemos que obtenemos un intervalo mejor para la predicción de una nueva observación debido a que Lasso elimina del modelo la última variable, que es la que más variabilidad aportaba debido a su dependencia.

3.3. Comparación de Lasso y Regresión Ridge

Podemos ver claramente que la diferencia entre la Regresión Ridge y Lasso es el término $\sum_{j=1}^{p}(\beta_j)^2$ y $\sum_{j=1}^{p}|\beta_j|$. Sin embargo, el término $\sum_{j=1}^{p}|\beta_j|$ de Lasso tiene la propiedad de que si establecemos un λ suficientemente pequeño, esta restricción hará que alguno de los β_j se anule y por tanto, además de corregir el problema de las variables dependientes, nos sirve como un método de selección de variables. Para poder comparar mejor los modelos podemos hacerlo con un ejemplo.

Ejemplo 3.3. Consideramos 3 variables X_1, X_2 y X_3 con 100 muestras de una normal $N(0,1)$, una variable dependiente de éstas $X_4 = X_1 + X_2 + X_3 + \delta$, y una variable $Y = 6 + 3 \cdot X_1 - 7 \cdot X_2 + 4 \cdot X_3 + 2 \cdot X_4 + \epsilon$, con $\epsilon_i \in N(0,0{,}5), \forall i = 1...n$. Entonces, si aplicamos la Regresión Ridge obtendríamos como coeficientes:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                     s0
> (Intercept)  6.004039
> x1           3.636422
> x2          -5.646225
> x3           4.509088
> x4           1.102722
```

Vemos que el coeficiente más contraído es precisamente el de la variable dependiente de las otras, pero aún así obtenemos unos coeficientes que se acercan a los valores reales. Por otro lado, si aplicamos Lasso obtenemos:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                     s0
> (Intercept)  6.071901
> x1           4.995229
> x2          -4.901355
> x3           5.912735
> x4           .
```

Vemos que Lasso descarta que la variable X_4 influya en el modelo, cuando en realidad sí que lo hace. En este caso, la Regresión Ridge nos aporta una mejor estimación pues Lasso tiende a eliminar variables y en este caso todas las variables influyen, por lo que no deberíamos eliminar ninguna.

Sin embargo, si eliminamos de la variable respuesta algunas variables explicativas y cogemos $Y2 = 6 + 3 \cdot X_1 - 7 \cdot X_2 + \epsilon$, obtendremos para la Regresión Ridge:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                      s0
> (Intercept)  5.8657204
> x1           3.3058558
> x2          -5.8469007
> x3           0.4923175
> x4          -0.5742114
```

y para Lasso:

```
> 5 x 1 sparse Matrix of class "dgCMatrix"
>                    s0
> (Intercept)  5.984730
> x1           3.042537
> x2          -6.966655
> x3           .
> x4           .
```

En este caso, Lasso nos aporta una mejor estimación de los coeficientes pues efectivamente, hay variables que no influyen en el modelo. En cambio, la Regresión Ridge nos da estimaciones de estos parámetros a pesar de que sean 0.

Así concluimos que dependiendo del caso, es mejor utilizar un método u otro. Si estamos ante un modelo donde todas las variables tienen coeficientes distintos de cero deberemos utilizar la Regresión Ridge, pues aunque el coeficiente de alguna de las variables sea muy cercano a 0, Ridge nunca lo va a anular. Por otro lado, en casos donde algunas de las variables tienen coeficientes grandes y el resto tienen coeficientes cercanos o iguales a cero, Lasso efectúa una Selección de Variables. El problema es que en un caso real, nosotros no conocemos el valor real de estos coeficientes por lo que no sabemos cuál de los métodos debemos aplicar.

Este efecto que tiene Lasso de seleccionar variables también lo podemos ver gráficamente para el caso de dos dimensiones. En el caso de la Regresión Ridge, la penalización que tenemos es $\sum_{j=1}^{p}(\beta_j)^2 \leq t$, la cual representa una esfera. Si representamos dicha esfera y el punto donde se encuentra el vector de coeficientes estimado por mínimos cuadrados, con una serie de elipses que nos indican las curvas de nivel de $\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$, es decir, regiones en las cuales todos los puntos de la frontera tienen el mismo valor para dicha función. Si representamos ésto obtenemos la gráfica 3.2, en la cual vemos que las curvas de nivel tocan a la circunferencia roja en un punto que sería el estimador de la Regresión Ridge que obtendríamos. En este caso, al ser una circunferencia es bastante improbable que las elipses intersequen exactamente en alguno de los puntos de los ejes, es por eso que la Regresión Ridge no nos sirve como seleccionador de variables.

Para Lasso, la restricción que tenemos es $\sum_{j=1}^{p}|\beta_j| \leq t$, que en este caso es un rombo. Si representamos lo mismo que antes obtenemos la gráfica 3.3. Aquí, el punto, y por tanto el estimador utilizando Lasso, sí es uno de los vértices, en el cual el coeficiente para la primera variable explicativa sería 0 y por consiguiente se eliminaría del modelo. Para el caso de Lasso, la región donde se cumple la penalización tiene vértices así que es más fácil que el punto donde se intersequen las curvas de nivel con la región sea un vértice, y por

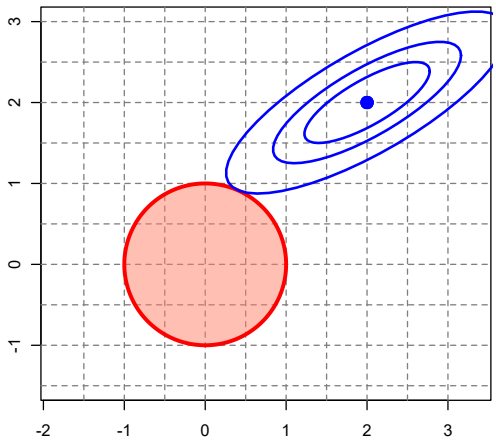

Figura 3.2: *En Regresión Ridge, la circunferencia representa la región donde la penalización que establecemos se cumple y las curvas de nivel de la función que queremos minimizar.*

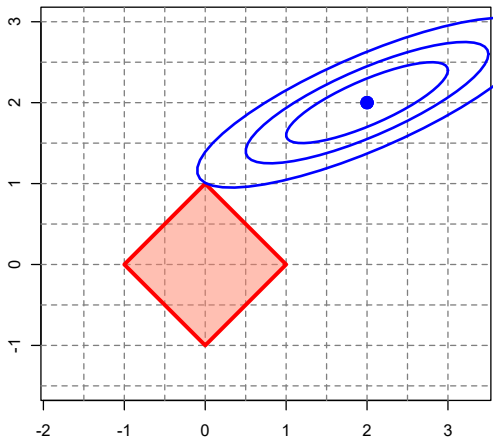

Figura 3.3: *En LASSO, el rombo representa la región donde la penalización que establecemos se cumple y las curvas de nivel de la función que queremos minimizar.*

tanto se anule alguno de los coeficientes de las variables. Otra cosa que podemos apreciar en estas figuras es que si la cota que ponemos a la penalización es suficientemente grande, el estimador por mínimos cuadrados estaría en las regiones rojas y por tanto, obtendríamos que el estimador utilizando la Regresión Ridge o Lasso sería el mismo que para mínimos cuadrados, como vimos en este capítulo.

Para arrojar un poco de luz sobre esta duda, en el último capítulo utilizaremos un ejemplo real y compararemos estos dos métodos para ver cuál predice mejor el valor de una nueva observación.

Debido a que el Lasso nos permite seleccionar variables y la Regresión Ridge tiene un menor coste computacional y es mejor a la hora de restringir los coeficientes de las variables dependientes, existe una penalización intermedia conocida como *elastic net* que se define como:

$$\sum_{j=1}^{p}(\alpha(\beta_j)^2 + (1-\alpha)|\beta_j|) \leq t$$

Esta penalización depende de un $\alpha \in [0,1]$ que determinará el peso que le damos a cada uno de los métodos. Si representamos la región donde se cumple esta penalización obtendríamos un rombo con los lados abombados en mayor o menor medida dependiendo del α que escojamos, si α es cercano a 0, la figura será prácticamente un rombo y si se acerca a 1, será prácticamente un círculo.

Capítulo 4

Least Angle Regression

Se suele abreviar como LAR y consiste en utilizar un procedimiento parecido al de los métodos forward.

4.1. Algoritmo

Un método forward con un criterio de significación consiste en identificar la variable más correlacionada e incluirla en el modelo, ajustar el modelo que estamos considerando y volver a buscar la variable con mayor correlación considerando el nuevo modelo, y así sucesivamente hasta que hayamos incluido en el modelo todas las variables que estén relacionadas.

LAR consiste en escoger la variable con mayor correlación, y en lugar de meterla directamente en el modelo con el coeficiente calculado por mínimos cuadrados, consideramos su coeficiente como 0 y vamos moviendo este coeficiente progresivamente hacia el estimador de mínimos cuadrados. A medida que el coeficiente se acerca al estimador por mínimos cuadrados, la correlación de esta variable decrece. Cuando alguna de las otras variables alcanza la misma correlación que tiene la variable que estamos considerando, entra también en el modelo con coeficiente 0 y se empiezan a mover los dos coeficientes hacia sus respectivas estimaciones por mínimos cuadrados y repetimos el proceso antes descrito. Este proceso se continúa hasta que todas las variables se hayan incluido en el modelo. Para entender mejor el proceso aquí tenemos el algoritmo a seguir:

1. Estandarizar las variables, es decir, las centramos y dividimos sus valores por la desviación típica. Al hacer esto estamos haciendo que las variables tengan media 0 y desviación típica 1. Calcular el residuo $\widehat{\epsilon} = Y - \widehat{Y}$, asumiendo aquí que $\beta_i = 0, \forall i \in \{1, ..., p\}$.

2. Encontrar la variable X_j más correlacionada con los residuos $\widehat{\epsilon}$.

3. Asumir $\beta_j = 0$ y moverlo hacia $\widehat{\beta}_j^{mc}$ hasta que otra variable X_k tenga tanta correlación con el residuo actual como tiene X_j.

4. Asumir $\beta_k = 0$ y mover β_k y β_j hacia $\widehat{\beta_k}^{mc}$ y $\widehat{\beta}_j^{mc}$ respectivamente hasta que otra variable X_l tenga la misma correlación que X_k y X_j.

5. Llevar a cabo este proceso hasta que las p variables explicativas hayan entrado en el modelo y hayamos alcanzado la solución de mínimos cuadrados.

Se le puede añadir un paso más a LAR para que también nos funcione como un método de Selección de Variables. Esta modificación se conoce como Least Angle Regression:Lasso Modification y consiste en:

4a. En caso de que alguna variable tenga un coeficiente distinto de 0 y éste alcance el valor 0, desechamos esta variable del modelo y dejamos de considerarla como candidata a entrar en el modelo.

Este nuevo método es muy eficiente, pues requiere un orden de computación igual al de un ajuste de mínimos cuadrados con p variables explicativas.

Los coeficientes que obtenemos son los mismos que con mínimos cuadrados, lo que nos interesa en este caso es el orden en el que entran las variables, pues nos sirve como criterio para ordenar las variables de más influyente a menos. Con ésto, podemos intentar buscar un modelo mejor eliminando las últimas variables en entrar.

Ejemplo 4.1. Si consideramos el mismo ejemplo que utilizamos cuando vimos los métodos de Selección de variables forward y backward y utilizamos un paquete de R conocido como `selectiveInference` (véase Tibshirani et al. (2017)), el cual tiene una función llamada `lar` que realiza el algoritmo de Least Angle Regresion sin la modificación. Si utilizamos esta función en los datos que teníamos obtenemos la siguiente salida:

```
>
> Call:
> lar(x = X, y = Y)
>
> Sequence of LAR moves:
>  Step Var Sign
>     1   1   -1
>     2   4    1
>     3   2   -1
>     4   3   -1
```

Si nos fijamos en la columna llamada `Var`, tenemos las variables ordenadas por orden de importancia. Cada número está asociado a una variable en función del orden en el que están en la matriz X, que en este caso es:

```
> [1] "pop15" "pop75" "dpi"   "ddpi"
```

Vemos que el orden por importancia sería `pop15`, `ddpi`, `pop75` y `dpi`. Este orden se parece a lo que obteníamos al aplicar el método forward, pues éste nos devolvía un modelo en el que entraban las variables `pop15`, `ddpi`, y `pop75` en este orden y por tanto `dpi` era la menos importante, hasta el punto de que no merecía la pena incluirla. Esto no tiene por qué cumplirse pues los criterios de selección son distintos.

También podemos ver esto en el gráfico 4.1. En él vemos que las dos primeras variables en entrar lo hacen rápidamente, y más tarde entran las otras dos. Las líneas de colores representan los coeficientes de las variables y vemos cómo se mueven hacia sus estimadores por mínimos cuadrados. Además, se observa que el coeficiente de la variable 3 apenas se mueve de 0, lo cual concuerda con la idea de ser poco influyente en el modelo.

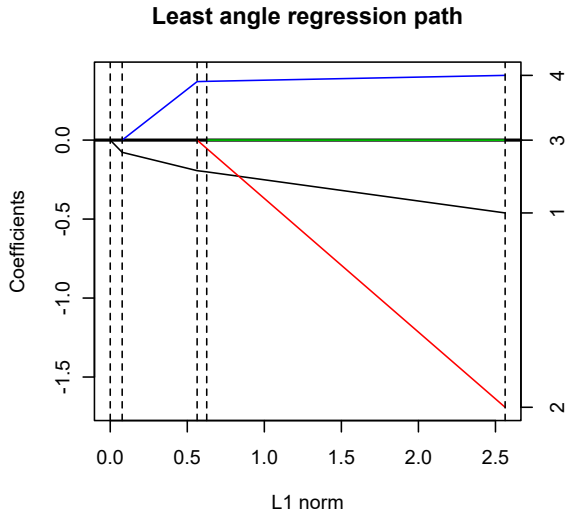

Figura 4.1: *Gráfico que nos muestra los valores que van tomando los coeficientes y los puntos en los que alguna nueva variable entra al modelo.*

4.2. Grados de libertad para LAR y Lasso

Al igual que con la Regresión Ridge podemos definir una fórmula para los grados de libertad de Lasso y de LAR, que es una manera de indicar el número de variables linealmente independientes que hay en el modelo. En el caso de la Regresión Ridge se hacía directamente en función del parámetro λ escogido, pero aquí lo calcularemos a partir del vector de predicciones $\widehat{Y}$. Este vector se calcula aplicando LAR o Lasso (habiendo seleccionado previamente un λ), obtenemos los coeficientes estimados $\widehat{\beta}^{LAR}$ y $\widehat{\beta}^{Lasso}$, que multiplicándolos por la matriz X, obtenemos una estimación de los valores de Y denotada como $\widehat{Y} = (\widehat{Y_1}, ..., \widehat{Y_n})$. Utilizando este vector, definimos los grados de libertad de Lasso y LAR como:

$$df(\widehat{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov(\widehat{Y}_i, Y_i)$$

En esta fórmula, $Cov(\widehat{Y}_i, Y_i)$ indica la covarianza entre el vector de predicciones $\widehat{Y}$ y la variable respuesta Y.

Esta fórmula también cumple que si consideramos que estamos aplicando mínimos cuadrados, tenemos $df(\widehat{Y}) = p$. En el caso de la Regresión Ridge, al aplicar esta fórmula obtenemos $df(\widehat{Y}) = tr(H_\lambda)$ como habíamos visto en el capítulo 2. Por lo tanto, esta fórmula de los grados de libertad es una generalización de la fórmula que habíamos visto para la Regresión Ridge.

Capítulo 5

Ejemplo Real

En este capítulo vamos a considerar un ejemplo real obtenido de una base de datos que la Universidad de California tiene en Internet, elaborada por Dua and Graff (2019), y poniendo en práctica todo lo que hemos visto a lo largo de este trabajo, aplicando todos los métodos, viendo lo que obtenemos y las conclusiones que podemos sacar. El conjunto de datos que consideraremos está formado por los datos de 1994 ciudades de Estados Unidos de las cuales tenemos 128 datos, que prescindiremos de los primeros 5 datos, pues son códigos del Estado, condado y comunidad en los que se encuentra la ciudad y no verdaderos datos numéricos de éstas. Estos datos fueron tomados en 1990. Por lo tanto, consideraremos las 123 variables restantes, las cuales contienen los siguientes datos:

- `Population`: Población.
- `HouseholdSize`: Número medio de domiciliados por hogar.
- `RacePctBlack`: Porcentaje de población afroamericana.
- `RacePctWhite`: Porcentaje de población caucásica.
- `RacePctAsian`: Porcentaje de población de origen asiático.
- `RacePctHisp`: Porcentaje de población de origen hispánico.
- `AgePct12t21`: Porcentaje de población con edades comprendidas entre los 12 y los 21.
- `AgePct12t29`: Porcentaje de población con edades comprendidas entre los 12 y los 29.
- `AgePct16t24`: Porcentaje de población con edades comprendidas entre los 16 y los 24.

- `AgePct65up`: Porcentaje de población mayor de 65.
- `NumbUrban`: Número de personas viviendo en áreas urbanas.
- `PctUrban`: Porcentaje de población que vive en área urbanas.
- `medIncome`: Ingreso medio por hogar.
- `PctWWage`: Porcentaje de hogares con ingresos salariales en 1989.
- `PctWFarmSelf`: Porcentaje de hogares con ingresos provenientes de autónomos o labores agrícolas.
- `PctWInvInc`: Porcentaje de hogares con ingresos provenientes de inversiones o rentas.
- `PctWSocSec`: Porcentaje de hogares con ingresos provenientes de la seguridad social.
- `PctWPubAsst`: Porcentaje de hogares con ingresos provenientes de ayudas sociales.
- `PctWRetire`: Porcentaje de hogares con ingresos provenientes de jubilaciones.
- `medFamInc`: Ingreso medio por familia. Se diferencia del 'ingreso medio por hogar' por los hogares no formados por familias.
- `PerCapInc`: Ingreso per capita.
- `WhitePerCap`: Ingreso per capita de la población caucásica.
- `BlackPerCap`: Ingreso per capita de la población afroamericana.
- `IndianPerCap`: Ingreso per capita de la población nativos americanos.
- `AsianPerCap`: Ingreso per capita de la población de origen asiático.
- `OtherPerCap`: Ingreso per capita de la población de otros orígenes.
- `HispPerCap`: Ingreso per capita de la población de origen hispánico.
- `NumUnderPov`: Población por debajo del umbral de pobreza.
- `PctPopUnderPov`: Porcentaje de población por debajo del umbral de pobreza.
- `PctLess9thGrade`: Porcentaje de población mayor de 25 años con estudios menores o iguales al equivalente estadounidense de 3º de la ESO.
- `PctNotHSGrad`: Porcentaje de población mayor de 25 años sin estudios de bachillerato.

- `PctBSorMore`: Porcentaje de población mayor de 25 años con estudios de bachillerato o superiores.
- `PctUnemployed`: Porcentaje de población mayor de 16 años perteneciente a la población activa y desempleada.
- `PctEmploy`: Porcentaje de población mayor de 16 años con empleo.
- `PctEmplManu`: Porcentaje de población mayor de 16 años empleada en el sector industrial.
- `PctEmplProfServ`: Porcentaje de población mayor de 16 años empleada en el sector servicios.
- `PctOccupManu`: Porcentaje de población mayor de 16 años trabajando en el sector industrial. Se diferencia de 'PctEmplManu' en que en este caso también incluimos a los dueños de las empresas o fábricas que se dedican al sector industrial.
- `PctOccupMgmtProf`: Porcentaje de población mayor de 16 años trabajando en administración o puestos especializados.
- `MalePctDivorce`: Porcentaje de hombres divorciados.
- `MalePctNevMarr`: Porcentaje de hombres que nunca se han casado.
- `FemalePctDiv`: Porcentaje de mujeres divorciadas.
- `TotalPctDiv`: Porcentaje de población que se ha divorciado.
- `PersPerFam`: Número medio de personas por familia.
- `PctFam2Par`: Porcentaje de familias con hijos con dos padres.
- `PctKids2Par`: Porcentaje de niños en familias con 2 padres.
- `PctYoungKids2Par`: Porcentaje de niños menores de 4 años en familias con 2 padres.
- `PctTeen2Par`: Porcentaje de niños entre 12 y 17 años en familias con 2 padres
- `PctWorkMomYoungKids`: Porcentaje de madres de niños menores de 6 años y que pertenecen a la población activa.
- `PctWorkMom`: Porcentaje de madres de niños menores de 18 años y que pertenecen a la población activa.

- `NumIlleg`: Número de niños ilegítimos.
- `PctIlleg`: Porcentaje de niños ilegítimos.
- `NumImmig`: Número total de inmigrantes.
- `PctImmigRecent`: Porcentaje de inmigrantes que inmigraron en los últimos 3 años.
- `PctImmigRec5`: Porcentaje de inmigrantes que inmigraron en los últimos 5 años.
- `PctImmigRec8`: Porcentaje de inmigrantes que inmigraron en los últimos 8 años.
- `PctImmigRec10`: Porcentaje de inmigrantes que inmigraron en los últimos 10 años.
- `PctRecentImmig`: Porcentaje de población que ha inmigrado en los últimos 3 años.
- `PctRecImmig5`: Porcentaje de población que ha inmigrado en los últimos 5 años.
- `PctRecImmig8`: Porcentaje de población que ha inmigrado en los últimos 8 años.
- `PctRecImmig10`: Porcentaje de población que ha inmigrado en los últimos 10 años.
- `PctSpeakEnglOnly`: Porcentaje de gente que sólo habla Inglés.
- `PctNotSpeakEnglWell`: Porcentaje de población que no habla Inglés de manera fluida.
- `PctLargHouseFam`: Porcentaje de hogares habitados por familias que tienen más de 6 personas.
- `PctLargHouseOccup`: Porcentaje de todos los hogares que tienen más de 6 personas.
- `PersPerOccupHous`: Número medio de habitantes por hogares habitados.
- `PersPerOwnOccHous`: Número medio de personas en hogares habitados por su dueño.
- `PersPerRentOccHous`: Número medio de personas por viviendas de alquiler.
- `PctPersOwnOccup`: Porcentaje de gente que vive en hogares habitados por su dueño.
- `PctPersDenseHous`: Porcentaje de población en viviendas 'densas', es decir, con más de una persona por habitación.
- `PctHousLess3BR`: Porcentaje de viviendas con menos de 3 habitaciones.
- `MedNumBR`: Número medio de habitaciones.
- `HousVacant`: Número de viviendas vacías.

- `PctHousOccup`: Porcentaje de viviendas ocupadas.
- `PctHousOwnOcc`: Porcentaje de viviendas habitadas por su dueño.
- `PctVacantBoarded`: Porcentajes de viviendas vacías que están tapiadas.
- `PctVacMore6Mos`: Porcentaje de viviendas vacías que llevan así más de 6 meses.
- `MedYrHousBuilt`: Número medio de viviendas construidas en un año.
- `PctHousNoPhone`: Porcentaje de viviendas ocupadas con teléfono.
- `PctWOFullPlumb`: Porcentaje de viviendas habitadas sin una instalación completa de fontanería.
- `OwnOccLowQuart`: Viviendas ocupadas por su dueño: Primer cuartil.
- `OwnOccMedVal`: Viviendas ocupadas por su dueño: Mediana.
- `OwnOccHiQuart`: Viviendas ocupadas por su dueño: Tercer cuartil.
- `RentLowQ`: Viviendas de alquiler: Primer cuartil del alquiler
- `RentMedian`: Viviendas de alquiler: Mediana del alquiler
- `RentHighQ`: Viviendas de alquiler: Tercer cuartil del alquiler
- `MedRent`: Mediana de la renta bruta.
- `MedRentPctHousInc`: Mediana de la renta bruta como un porcentaje de los ingresos del hogar.
- `MedOwnCostPctInc`: Mediana de los gastos del dueño como un porcentaje de los ingresos del hogar para hogares con hipoteca.
- `MedOwnCostPctIncNoMtg`: Mediana de los gastos del dueño como un porcentaje de los ingresos del hogar para hogares sin hipoteca.
- `NumInShelters`: Número de personas en los refugios para indigentes.
- `NumStreet`: Número de indigentes en la calle.
- `PctForeignBorn`: Porcentaje de población nacida en el extranjero.
- `PctBornSameState`: Porcentaje de población nacida en el mismo estado en el que vive actualmente.

- `PctSameHouse85`: Porcentaje de población que vive en la misma casa que hace 5 años.
- `PctSameCity85`: Porcentaje de población que vive en la misma ciudad que hace 5 años.
- `PctSameState85`: Porcentaje de población que vive en la mismo estado que 5 hace años.
- `LemasSwornFT`: Número de policías juramentados a tiempo completo.
- `LemasSwFTPerPop`: Policías juramentados a tiempo completo por cada 100.000 habitantes.
- `LemasSwFTFieldOps`: Policías juramentados a tiempo completo haciendo trabajo de campo (no administrativo).
- `LemasSwFTFieldPerPop`: Policías juramentados a tiempo completo haciendo trabajo de campo (no administrativo) por cada 100.000 habitantes.
- `LemasTotalReq`: Número total de solicitudes para ser policía.
- `LemasTotReqPerPop`: Número total de solicitudes para ser policía por cada 100.000 habitantes.
- `PolicReqPerOffic`: Número total de solicitudes para ser policía por cada oficial de policía.
- `PolicPerPop`: Oficiales de policía por cada 100.000 policías.
- `RacialMatchCommPol`: Una medida de la comparación entre la proporción racial en la policía y en la población. Cuanto mayor sea el valor, mayor es la coincidencia de estas proporciones.
- `PctPolicWhite`: Porcentaje de policías caucásicos.
- `PctPolicBlack`: Porcentaje de policías afroamericanos.
- `PctPolicHisp`: Porcentaje de policías hispánicos.
- `PctPolicAsian`: Porcentaje de policías asiáticos.
- `PctPolicMinor`: Porcentaje de policías que pertenecen a alguna minoría.
- `OfficAssgnDrugUnits`: Número de oficiales asignados a grupos especiales antidrogas.

- `NumKindsDrugsSeiz`: Número de diferentes tipos de drogas incautadas.
- `PolicAveOTWorked`: Media de las horas extra que hace la policía.
- `LandArea`: Área terrestre en millas cuadradas.
- `PopDens`: Densidad de población en personas por milla al cuadrado.
- `PctUsePubTrans`: Porcentaje de personas que utilizan el transporte público para sus desplazamientos.
- `PolicCars`: Número de coches de policía.
- `PolicOperBudg`: Presupuesto destinado a la policía.
- `LemasPctPolicOnPatr`: Porcentaje de policías de tiempo completo juramentados en patrulla.
- `LemasGangUnitDeploy`: Unidad antibandas desplegada (no es realmente una variable numérica pues toma el valor 0 si no está desplegada, 0.5 si está parte del tiempo y 1 si está desplegada permanentemente).
- `LemasPctOfficDrugUn`: Porcentaje de policías destinados a unidades antidroga.
- `PolicBudgPerPop`: Presupuesto destinado a la policía por población.
- `ViolentCrimesPerPop`: Número total de crímenes por cada 100.000 habitantes.

Estos datos tienen la particularidad de que están normalizados al intervalo $[0, 1]$, es decir, se ajustaron los datos para que todos estuvieran en dicho intervalo, siendo 1 el mayor de todos los datos y 0 el menor, y el resto se ajustaron de manera proporcional. Además, estos datos presentan un problema que no podemos ignorar y es que faltan los datos de 1675 ciudades de 22 variables. Para evitar este problema, podemos afrontarlo de dos maneras: eliminar las variables de las que nos faltan estos datos y tener así una muestra de 100 variables o eliminar las ciudades y todos sus datos del modelo con lo que pasaríamos a tener 319 datos de cada variable. En este capítulo vamos a abordar ambos casos para ver qué resultados obtenemos, pues ambos casos son interesantes.

5.1. Ejemplo 1

Para empezar, consideremos el caso en el que eliminamos las variables.

5.1.1. Mínimos cuadrados

Si aplicamos un modelo de regresión a todas las variables, obtendríamos sus coeficientes y muchos de ellos no serían significativos, por lo que directamente vamos a buscar un modelo de regresión utilizando un método forward con el Criterio de Información de Akaike para seleccionar variables y quedarnos sólo con aquellas que realmente influyan:

```
>
> Call:
> lm(formula = Y ~ PctKids2Par + RacePctWhite + HousVacant + PctUrban +
>     PctWorkMom + NumStreet + MalePctDivorce + PctIlleg + NumbUrban +
>     PctPersDenseHous + RacePctBlack + AgePct12t29 + MedOwnCostPctIncNoMtg +
>     PctPopUnderPov + PctWRetire + MedRentPctHousInc + RentLowQ +
>     MedRent + PctWWage + WhitePerCap + MalePctNevMarr + PctEmploy +
>     PctEmplManu + TotalPctDiv + PerCapInc + PctWInvInc + LemasPctOfficDrugUn +
>     MedOwnCostPctInc + PctVacMore6Mos + PctVacantBoarded + HispPerCap +
>     PctWFarmSelf + IndianPerCap + PctLess9thGrade + PctLargHouseFam +
>     AgePct12t21 + AsianPerCap, data = com1)
>
> Residuals:
>      Min       1Q   Median       3Q      Max
> -0.49563 -0.07187 -0.01348  0.04632  0.74777
>
> Coefficients:
>                       Estimate Std. Error t value Pr(>|t|)
> (Intercept)            0.754375   0.103560   7.284 4.66e-13 ***
> PctKids2Par           -0.344702   0.071189  -4.842 1.39e-06 ***
> RacePctWhite          -0.079034   0.045192  -1.749 0.080475 .
> HousVacant             0.268889   0.051731   5.198 2.23e-07 ***
> PctUrban               0.040023   0.009164   4.367 1.32e-05 ***
> PctWorkMom            -0.137158   0.025998  -5.276 1.47e-07 ***
> NumStreet              0.186394   0.040202   4.636 3.78e-06 ***
> MalePctDivorce         0.346558   0.086733   3.996 6.69e-05 ***
> PctIlleg               0.147034   0.040642   3.618 0.000305 ***
> NumbUrban             -0.213173   0.064223  -3.319 0.000919 ***
> PctPersDenseHous       0.250009   0.049892   5.011 5.90e-07 ***
> RacePctBlack           0.175664   0.039701   4.425 1.02e-05 ***
> AgePct12t29           -0.284100   0.078797  -3.605 0.000319 ***
> MedOwnCostPctIncNoMtg -0.080871   0.020752  -3.897 0.000101 ***
> PctPopUnderPov        -0.155248   0.044688  -3.474 0.000524 ***
> PctWRetire            -0.105415   0.031140  -3.385 0.000725 ***
> MedRentPctHousInc      0.054839   0.028724   1.909 0.056387 .
> RentLowQ              -0.242711   0.053528  -4.534 6.13e-06 ***
> MedRent                0.265462   0.058928   4.505 7.03e-06 ***
```

```
> PctWWage              -0.226305   0.050899  -4.446 9.23e-06 ***
> WhitePerCap           -0.321982   0.131061  -2.457 0.014107 *
> MalePctNevMarr         0.151484   0.046637   3.248 0.001181 **
> PctEmploy              0.199442   0.056800   3.511 0.000456 ***
> PctEmplManu           -0.031490   0.018775  -1.677 0.093656 .
> TotalPctDiv           -0.305082   0.097359  -3.134 0.001753 **
> PerCapInc              0.217361   0.145640   1.492 0.135743
> PctWInvInc            -0.167964   0.054768  -3.067 0.002193 **
> LemasPctOfficDrugUn    0.029512   0.014893   1.982 0.047657 *
> MedOwnCostPctInc      -0.047327   0.027356  -1.730 0.083775 .
> PctVacMore6Mos        -0.059783   0.022267  -2.685 0.007319 **
> PctVacantBoarded       0.044872   0.019949   2.249 0.024600 *
> HispPerCap             0.042397   0.022849   1.856 0.063674 .
> PctWFarmSelf           0.039122   0.018935   2.066 0.038949 *
> IndianPerCap          -0.035509   0.019082  -1.861 0.062914 .
> PctLess9thGrade       -0.057763   0.030931  -1.868 0.061981 .
> PctLargHouseFam       -0.088468   0.039740  -2.226 0.026116 *
> AgePct12t21            0.091530   0.050374   1.817 0.069370 .
> AsianPerCap            0.026418   0.018398   1.436 0.151194
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.1323 on 1956 degrees of freedom
> Multiple R-squared:  0.6833,Adjusted R-squared:  0.6773
> F-statistic: 114.1 on 37 and 1956 DF,  p-value: < 2.2e-16
```

Aquí vemos los distintos coeficientes de las variables que el criterio de Akaike considera que mejoran el modelo. Para ver un poco mejor estas relaciones podemos ver los gráficos que nos resultan al comparar la tasa de criminalidad con algunas de las variables más influyentes en la figura 5.1. Podemos ver que en todos ellos, la variable Y tiende a ser más pequeña para valores grandes o pequeños de según que variable explicativa.

5.1.2. Regresión Ridge

Si aplicamos la Regresión Ridge a estos datos, obtenemos los siguientes coeficientes para las variables:

```
> 100 x 1 sparse Matrix of class "dgCMatrix"
>                                s0
> (Intercept)          0.4560933553
> Population          -0.0297095171
> HouseholdSize        0.0301235478
```

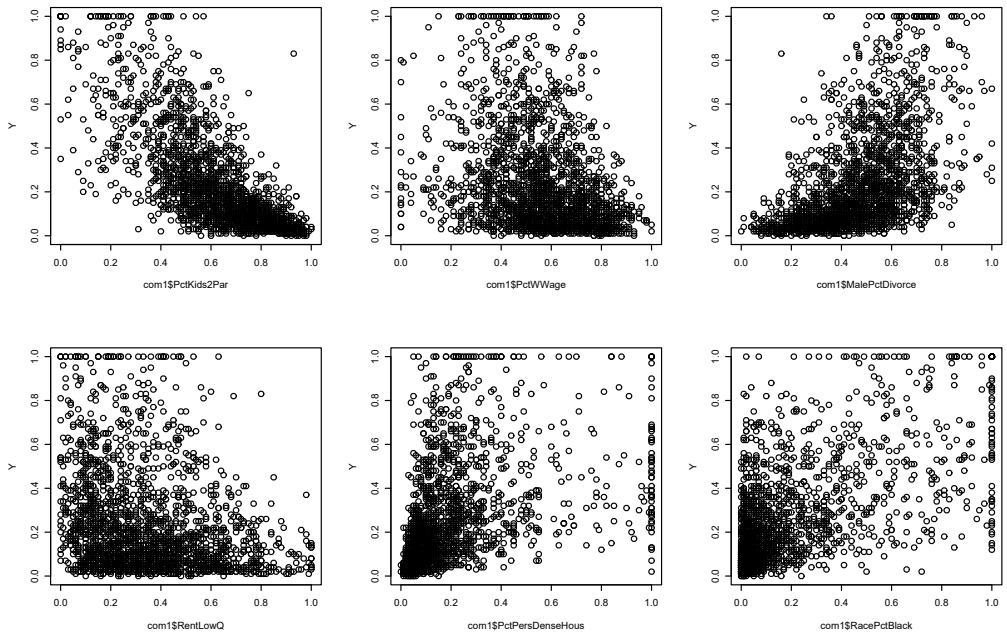

Figura 5.1: *Gráficas de algunas de las variables más significativas.*

```
> RacePctBlack           0.1394661301
> RacePctWhite          -0.0867297971
> RacePctAsian          -0.0218510370
> RacePctHisp            0.0190683341
> AgePct12t21            0.0031591461
> AgePct12t29           -0.1009316326
> AgePct16t24           -0.0152915864
> AgePct65up             0.0315360400
> NumbUrban             -0.0313472169
> PctUrban               0.0340435962
> medIncome              0.0164747207
> PctWWage              -0.0549940842
> PctWFarmSelf           0.0172727052
> PctWInvInc            -0.1113746633
> PctWSocSec             0.0379442434
> PctWPubAsst            0.0208559384
> PctWRetire            -0.0631919797
> medFamInc              0.0125101068
> PerCapInc             -0.0171698418
> WhitePerCap           -0.0337287719
> BlackPerCap           -0.0167040208
> IndianPerCap          -0.0260198756
```

```
> AsianPerCap            0.0241958130
> HispPerCap             0.0375842656
> NumUnderPov            0.0037737786
> PctPopUnderPov        -0.0619172926
> PctLess9thGrade       -0.0438614497
> PctNotHSGrad           0.0221032637
> PctBSorMore            0.0052265548
> PctUnemployed         -0.0281305453
> PctEmploy              0.0613736260
> PctEmplManu           -0.0322719396
> PctEmplProfServ       -0.0081482053
> PctOccupManu           0.0229215658
> PctOccupMgmtProf       0.0088727500
> MalePctDivorce         0.0894344450
> MalePctNevMarr         0.0686444635
> FemalePctDiv          -0.0258859780
> TotalPctDiv            0.0145045143
> PersPerFam             0.0304058052
> PctFam2Par            -0.0706730092
> PctKids2Par           -0.1095997072
> PctYoungKids2Par      -0.0552189284
> PctTeen2Par           -0.0303182063
> PctWorkMomYoungKids    0.0070611483
> PctWorkMom            -0.0849898684
> NumIlleg              -0.0134138536
> PctIlleg               0.1361176631
> NumImmig              -0.0886849967
> PctImmigRecent         0.0118910616
> PctImmigRec5          -0.0122629561
> PctImmigRec8          -0.0100405664
> PctImmigRec10          0.0037955495
> PctRecentImmig        -0.0021540449
> PctRecImmig5           0.0040933010
> PctRecImmig8           0.0222787947
> PctRecImmig10          0.0238714317
> PctSpeakEnglOnly      -0.0003991345
> PctNotSpeakEnglWell   -0.0372559224
> PctLargHouseFam       -0.0132397323
> PctLargHouseOccup     -0.0218004602
> PersPerOccupHous       0.0668886401
> PersPerOwnOccHous     -0.0354646028
> PersPerRentOccHous    -0.0025077079
> PctPersOwnOccup       -0.0349330674
> PctPersDenseHous       0.0755670054
```

```
> PctHousLess3BR          0.0500387656
> MedNumBR                0.0068305438
> HousVacant              0.1074051934
> PctHousOccup           -0.0638724084
> PctHousOwnOcc           0.0053917391
> PctVacantBoarded        0.0537397542
> PctVacMore6Mos         -0.0487201052
> MedYrHousBuilt         -0.0012566415
> PctHousNoPhone          0.0153193681
> PctWOFullPlumb         -0.0098691024
> OwnOccLowQuart         -0.0180747249
> OwnOccMedVal            0.0019599523
> OwnOccHiQuart           0.0068554825
> RentLowQ               -0.0817244717
> RentMedian              0.0153057645
> RentHighQ               0.0224687718
> MedRent                 0.0623737422
> MedRentPctHousInc       0.0542602826
> MedOwnCostPctInc       -0.0228732594
> MedOwnCostPctIncNoMtg -0.0721592885
> NumInShelters           0.0964445455
> NumStreet               0.1766345150
> PctForeignBorn          0.0325562281
> PctBornSameState       -0.0130301748
> PctSameHouse85          0.0064991058
> PctSameCity85           0.0325443440
> PctSameState85          0.0051978116
> LandArea                0.0325079827
> PopDens                 0.0020749578
> PctUsePubTrans         -0.0206410313
> LemasPctOfficDrugUn     0.0326679659
```

Como habíamos visto, ninguno es 0 ni cerca de 0 pero sí que podemos apreciar que en general, todos los coeficientes han sido restringidos y tienen valores más pequeños respecto a los que teníamos con el método forward.

Podemos mirar los grados de libertad para ver si hay variables dependientes:

```
> [1] 97.12862
```

Prácticamente tenemos 97, con lo que decimos que hay dos variables dependientes pero no sabemos cuáles son exactamente. Ésto también lo podemos ver con los autovalores de la matriz $X^\intercal X$ y el λ óptimo:

```
d

>  [1] 3.260364e+04 2.024893e+03 1.428944e+03 6.467918e+02 5.596041e+02
>  [6] 3.656448e+02 3.018701e+02 2.529087e+02 2.121765e+02 1.636579e+02
> [11] 1.488432e+02 1.121376e+02 1.095212e+02 1.034174e+02 8.558863e+01
> [16] 7.933181e+01 7.022297e+01 5.928931e+01 5.610885e+01 5.159809e+01
> [21] 4.929214e+01 4.831691e+01 4.761663e+01 4.336018e+01 3.898502e+01
> [26] 3.619007e+01 3.279486e+01 3.146984e+01 3.103643e+01 2.825156e+01
> [31] 2.739566e+01 2.706171e+01 2.557416e+01 2.267855e+01 2.113164e+01
> [36] 1.959286e+01 1.888499e+01 1.786156e+01 1.658575e+01 1.550341e+01
> [41] 1.474965e+01 1.426010e+01 1.295637e+01 1.178402e+01 1.107090e+01
> [46] 9.965745e+00 9.387954e+00 8.579411e+00 8.348259e+00 7.476259e+00
> [51] 7.334302e+00 6.967718e+00 6.181137e+00 6.066422e+00 5.811703e+00
> [56] 5.567866e+00 5.374594e+00 5.018406e+00 4.738926e+00 4.495809e+00
> [61] 4.288182e+00 4.130475e+00 3.954652e+00 3.676062e+00 3.448108e+00
> [66] 3.309311e+00 3.094778e+00 2.859792e+00 2.830746e+00 2.713138e+00
> [71] 2.498795e+00 2.373833e+00 2.201400e+00 2.084359e+00 1.936723e+00
> [76] 1.763900e+00 1.579989e+00 1.544875e+00 1.437436e+00 1.345496e+00
> [81] 1.262276e+00 1.190804e+00 9.931818e-01 8.867672e-01 7.601078e-01
> [86] 7.030199e-01 5.253055e-01 5.007320e-01 4.347332e-01 3.514039e-01
> [91] 3.080724e-01 2.983507e-01 2.763008e-01 1.504412e-01 1.385055e-01
> [96] 1.171689e-01 5.813454e-02 5.672343e-02 4.148618e-02

l

> [1] 0.01887681
```

El λ óptimo que tenemos es más pequeño que cualquiera de los autovalores, por lo que no está arreglando ningún autovalor aunque sí que hace que los tres últimos, y por tanto más pequeños, sean un poco más grandes. Este efecto es el que se ve reflejado en los grados de libertad pues al resto de autovalores apenas afecta.

5.1.3. Lasso

Para seleccionar el λ óptimo para el Lasso, comparamos el logaritmo del λ con el error cuadrático medio, con lo que tenemos la gráfica 5.2. Si nos fijamos en el número de variables con coeficiente distinto de 0, vemos que depende bastante del λ que escojamos.

Si obtenemos el λ óptimo y lo utilizamos para aplicar el Lasso, obtenemos los siguientes coeficientes:

```
> 100 x 1 sparse Matrix of class "dgCMatrix"
>                                  s0
```

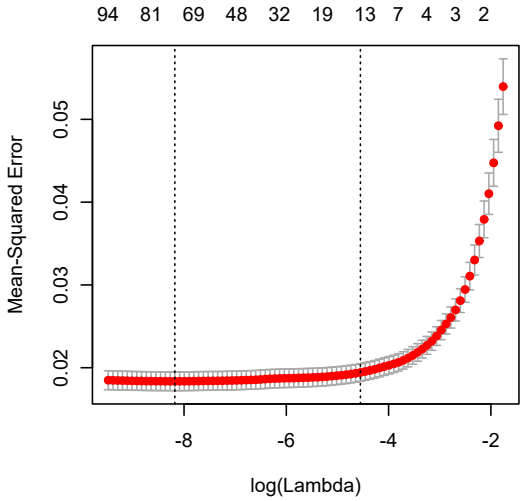

Figura 5.2: *Gráfico para encontrar el λ óptimo respecto al error cuadrático medio. Encima de la gráfica tenemos el número de variables con coeficiente distinto de 0 que tendríamos en el modelo en función de λ.*

```
> (Intercept)          0.5637582064
> Population           .
> HouseholdSize        .
> RacePctBlack         0.1965037448
> RacePctWhite        -0.0316380717
> RacePctAsian         .
> RacePctHisp          0.0535066052
> AgePct12t21          0.0404314609
> AgePct12t29         -0.2664091740
> AgePct16t24          .
> AgePct65up           0.0054847994
> NumbUrban           -0.0770334318
> PctUrban             0.0378515289
> medIncome            .
> PctWWage            -0.1660026357
> PctWFarmSelf         0.0362471524
> PctWInvInc          -0.1599632363
> PctWSocSec           0.0574472608
> PctWPubAsst          .
> PctWRetire          -0.0865105865
> medFamInc            0.0717434757
> PerCapInc            .
> WhitePerCap         -0.1589535329
```

```
> BlackPerCap           -0.0190149719
> IndianPerCap          -0.0296910950
> AsianPerCap            0.0233285150
> HispPerCap             0.0439200542
> NumUnderPov            .
> PctPopUnderPov        -0.1410286107
> PctLess9thGrade       -0.0584571859
> PctNotHSGrad           0.0072357541
> PctBSorMore            0.0389270753
> PctUnemployed         -0.0094685399
> PctEmploy              0.1619269583
> PctEmplManu           -0.0451120551
> PctEmplProfServ        .
> PctOccupManu           0.0483415395
> PctOccupMgmtProf       0.0315291101
> MalePctDivorce         0.1600908643
> MalePctNevMarr         0.1656200424
> FemalePctDiv          -0.1115486968
> TotalPctDiv            .
> PersPerFam             .
> PctFam2Par             .
> PctKids2Par           -0.2724223440
> PctYoungKids2Par      -0.0299212959
> PctTeen2Par            .
> PctWorkMomYoungKids    0.0319231239
> PctWorkMom            -0.1514885489
> NumIlleg              -0.0563170953
> PctIlleg               0.1410223356
> NumImmig              -0.1144375590
> PctImmigRecent         0.0056519245
> PctImmigRec5          -0.0000348122
> PctImmigRec8          -0.0003009410
> PctImmigRec10          .
> PctRecentImmig         .
> PctRecImmig5           .
> PctRecImmig8           0.0072666851
> PctRecImmig10          .
> PctSpeakEnglOnly       .
> PctNotSpeakEnglWell   -0.0991262071
> PctLargHouseFam       -0.0676689121
> PctLargHouseOccup     -0.0423302954
> PersPerOccupHous       0.3368638640
> PersPerOwnOccHous     -0.1306365586
> PersPerRentOccHous    -0.0837483469
```

```
> PctPersOwnOccup       -0.0855285394
> PctPersDenseHous       0.1731380631
> PctHousLess3BR         0.0763018922
> MedNumBR               0.0144319768
> HousVacant             0.1600433672
> PctHousOccup          -0.0534622762
> PctHousOwnOcc          .
> PctVacantBoarded       0.0509755264
> PctVacMore6Mos        -0.0631035391
> MedYrHousBuilt        -0.0074172923
> PctHousNoPhone         0.0128959266
> PctWOFullPlumb        -0.0101624386
> OwnOccLowQuart        -0.0462126447
> OwnOccMedVal           .
> OwnOccHiQuart          0.0016780454
> RentLowQ              -0.2019782030
> RentMedian             .
> RentHighQ              .
> MedRent                0.2383399594
> MedRentPctHousInc      0.0510279486
> MedOwnCostPctInc      -0.0409682306
> MedOwnCostPctIncNoMtg -0.0846631850
> NumInShelters          0.1125981067
> NumStreet              0.1874613873
> PctForeignBorn         0.0982852676
> PctBornSameState       .
> PctSameHouse85         .
> PctSameCity85          0.0245277883
> PctSameState85         0.0019756564
> LandArea               0.0241346852
> PopDens                .
> PctUsePubTrans        -0.0346328601
> LemasPctOfficDrugUn    0.0278660403
```

Vemos que este modelo considera influyentes 74 variables frente a las 38 que obteníamos con el modelo forward. A lo largo de este trabajo hemos defendido que el LASSO es mejor a la hora de seleccionar variables y sin embargo, estamos viendo que con el método forward obtenemos menos variables. Esto se debe a que si nos fijamos en el λ óptimo que hemos obtenido y su logaritmo:

```
>   [,1]                  [,2]
>   "lambda óptimo"       "log(lambda)"
> x "0.00028031462193663" "-8.17959793584325"
```

Vemos que el λ óptimo se encuentra prácticamente en el extremo derecho del intervalo. La diferencia si escogemos otro λ' en el extremo izquierdo del intervalo es bastante pequeña, por lo que podríamos considerar un modelo con alrededor de 15 variables y la diferencia en el error sería mínima. Sin embargo, consideraremos el modelo creado a partir del λ óptimo pues es donde tenemos el menor error .

Aquí nos podemos fijar en la variable `TotalPctDiv` que en el otro modelo sí influía, en este queda descartada. Podemos pensar que ésto se debe a que en el otro modelo esta variable realmente no era importante y está siendo sustituída por otras variables que sí consideramos en el modelo grande pero vemos que después de ella aún entraron unas cuantas más, y si nos fijamos en la significación, es una variable que podríamos considerar bastante influyente. Sin embargo, el criterio de Lasso descarta que esta variable tenga algún efecto sobre la variable respuesta en favor de otras variables.

5.1.4. Least Angle Regression

Si aplicamos el método de Least Angle Regression, las variables del modelo ordenadas por importancia quedan como:

```
>  [1] "PctKids2Par"           "PctIlleg"
>  [3] "RacePctWhite"          "HousVacant"
>  [5] "TotalPctDiv"           "NumStreet"
>  [7] "PctPersDenseHous"      "MalePctDivorce"
>  [9] "PctVacantBoarded"      "PctUrban"
> [11] "PctWorkMom"            "PctHousOccup"
> [13] "LemasPctOfficDrugUn"   "RacePctBlack"
> [15] "FemalePctDiv"          "AgePct12t29"
> [17] "MedRentPctHousInc"     "MedOwnCostPctIncNoMtg"
> [19] "PctForeignBorn"        "PctWInvInc"
> [21] "AsianPerCap"           "PctVacMore6Mos"
> [23] "PctEmplManu"           "HispPerCap"
> [25] "PctWRetire"            "IndianPerCap"
> [27] "PctSameCity85"         "NumInShelters"
> [29] "PctHousLess3BR"        "MedRent"
> [31] "PctWWage"              "NumImmig"
> [33] "PctYoungKids2Par"      "PctPopUnderPov"
> [35] "PctWFarmSelf"          "PctLess9thGrade"
> [37] "RentLowQ"              "PctRecImmig10"
> [39] "PctEmploy"             "BlackPerCap"
> [41] "PctPersOwnOccup"       "MalePctNevMarr"
> [43] "PctUsePubTrans"        "NumIlleg"
> [45] "MedOwnCostPctInc"      "WhitePerCap"
> [47] "PctRecImmig8"          "RacePctHisp"
```

```
> [49] "MedYrHousBuilt"        "PctWOFullPlumb"
> [51] "PctUnemployed"         "LandArea"
> [53] "PctWSocSec"            "PersPerOccupHous"
> [55] "AgePct12t21"           "NumbUrban"
> [57] "PctLargHouseFam"       "PctEmplProfServ"
> [59] "PctHousNoPhone"        "MedNumBR"
> [61] "PctNotSpeakEnglWell"   "PctBSorMore"
> [63] "PctLargHouseOccup"     "PctImmigRecent"
> [65] "OwnOccLowQuart"        "PersPerOwnOccHous"
> [67] "PctOccupManu"          "PctWorkMomYoungKids"
> [69] "PersPerRentOccHous"    "PctOccupMgmtProf"
> [71] "PctSameState85"        "medFamInc"
> [73] "AgePct65up"            "RacePctAsian"
> [75] "PctNotHSGrad"          "PctSpeakEnglOnly"
> [77] "OwnOccHiQuart"         "PctImmigRec5"
> [79] "PctSameHouse85"        "PopDens"
> [81] "RentHighQ"             "PctImmigRec10"
> [83] "PctTeen2Par"           "NumUnderPov"
> [85] "PctImmigRec8"          "PctRecentImmig"
> [87] "PersPerFam"            "PerCapInc"
> [89] "AgePct16t24"           "PctWPubAsst"
> [91] "medIncome"             "HouseholdSize"
> [93] "PctRecImmig5"          "PctHousOwnOcc"
> [95] "OwnOccMedVal"          "PctBornSameState"
> [97] "Population"            "RentMedian"
> [99] "PctFam2Par"
```

Curiosamente, la variable `TotalPctDiv` que según el método forward era bastante influyente aunque había otras más importantes y que según Lasso no influía, aquí aparece como la quinta variable más importante de todo el modelo. Con esto vemos que el que determinemos si una variable influye o no en el modelo depende del criterio que utilicemos y del método, pero no podemos asegurar que uno tenga más razón que otro. Lo que sí que podemos hacer es compararlos para ver qué método es mejor a la hora de predecir lo que pasará, que es realmente lo que buscamos. Ésto lo haremos en la última parte, aunque no consideraremos Least Angle Regression pues el modelo que obtenemos es el mismo que para mínimos cuadrados con todas las variables. Lo importante en este método es el orden de entrada de las variables que nos sirve como orden de importancia en el modelo.

5.1.5. Comparación

En esta última parte vamos a considerar el primer ejemplo pues tiene más observaciones. Para comparar mejor los distintos métodos que hemos visto, vamos a quitar 200 obser-

vaciones aleatorias del modelo, crearemos los modelos con mínimos cuadrados, Regresión Ridge y Lasso. Haremos sus intervalos para la predicción de una nueva observación para los 200 datos que hemos quitado y veremos si la tasa de criminalidad real pertenece o no al intervalo y así veremos cual es mejor a la hora de predecir. Repetimos este proceso 100 veces y hacemos la media del porcentaje de aciertos obteniendo así la tabla:

Método utilizado	Mínimos cuadrados	Regresión Ridge	LASSO
Nivel de confianza del 75 %	81.255 %	81.17 %	80.545 %
Nivel de confianza del 95 %	93.65 %	93.26 %	93.175 %

Hemos hecho este proceso para un intervalo de confianza del 75 % y hemos obtenido esos porcentajes de acierto, los cuales están por encima del nivel de confianza que tenemos, pero todos son parecidos. En el caso del nivel de confianza del 95 %, los porcentajes que tenemos son más pequeños que el nivel de confianza. Sin embargo, algo que apreciamos en ambos casos es que el porcentaje de acierto de mínimos cuadrados es mayor que Regresión Ridge y éste es mayor que el de Lasso. Ésto se debe a que, al aplicar las penalizaciones al método de mínimos cuadrados, estamos sacrificando algo de la precisión que tiene este método. Además, podemos ver que Lasso parece sacrificar aún más precisión, debido a que el cálculo de sus coeficientes no es tan sencillo como en los otros casos.

De lo anterior deducimos que el método de mínimos cuadrados tiene buenas propiedades y es el más preciso de todos a la hora de predecir una nueva observación. Sin embargo, también hemos visto que en el caso de variables dependientes, este método no es capaz de darnos unos coeficientes fiables. Aquí es donde entran los métodos presentados en este trabajo que nos dan unos coeficientes más precisos evitando este problema sacrificando un poco de precisión, o en el caso de Least Angle Regression, nos dan los coeficientes ordenados por orden de influencia en la variable respuesta, también esquivando el problema de variables dependientes. Con estos métodos, podemos obtener una mejor aproximación de los coeficientes y un mejor modelo de regresión que es lo que buscamos con este trabajo.

5.2. Ejemplo 2

Ahora vamos a considerar el caso en el que eliminamos las observaciones en las que nos faltan datos, con lo que pasaríamos a tener un modelo con 122 variables explicativas y 319 observaciones, bastantes menos que en el caso anterior.

5.2.1. Mínimos cuadrados

Si volvemos a buscar un modelo forward para ver qué variables entrarían al modelo, tenemos:

```
>
> Call:
> lm(formula = Y ~ PctKids2Par + NumStreet + PctEmplManu + PctWInvInc +
>     PolicReqPerOffic + LemasPctOfficDrugUn + OtherPerCap + PctBornSameState +
>     PersPerRentOccHous + PctSameState85 + PctUsePubTrans + RacialMatchCommPol +
>     PersPerFam + MedOwnCostPctInc + PctPolicAsian + TotalPctDiv +
>     PctNotHSGrad + AgePct12t21 + PctOccupManu + PctForeignBorn,
>     data = com)
>
> Residuals:
>      Min       1Q   Median       3Q      Max
> -0.45793 -0.09911 -0.01614  0.08236  0.41760
>
> Coefficients:
>                     Estimate Std. Error t value Pr(>|t|)
> (Intercept)          1.87484    0.22400   8.370 2.27e-15 ***
> PctKids2Par         -1.08535    0.10103 -10.743  < 2e-16 ***
> NumStreet            0.19538    0.04753   4.111 5.10e-05 ***
> PctEmplManu         -0.28586    0.08176  -3.497 0.000543 ***
> PctWInvInc          -0.64800    0.16143  -4.014 7.56e-05 ***
> PolicReqPerOffic     0.12367    0.04717   2.622 0.009198 **
> LemasPctOfficDrugUn -0.08319    0.03402  -2.445 0.015045 *
> OtherPerCap          0.12025    0.07682   1.565 0.118576
> PctBornSameState    -0.36152    0.10814  -3.343 0.000934 ***
> PersPerRentOccHous  -0.48626    0.13555  -3.587 0.000391 ***
> PctSameState85       0.22458    0.09081   2.473 0.013956 *
> PctUsePubTrans      -0.16022    0.04428  -3.619 0.000348 ***
> RacialMatchCommPol  -0.12340    0.04721  -2.614 0.009402 **
> PersPerFam           0.47646    0.16091   2.961 0.003313 **
> MedOwnCostPctInc    -0.21299    0.07657  -2.782 0.005751 **
> PctPolicAsian        0.06636    0.04634   1.432 0.153188
> TotalPctDiv         -0.34012    0.11543  -2.947 0.003466 **
> PctNotHSGrad        -0.50971    0.13518  -3.771 0.000196 ***
> AgePct12t21         -0.18953    0.10000  -1.895 0.059024 .
> PctOccupManu         0.30447    0.12579   2.420 0.016098 *
> PctForeignBorn       0.13309    0.06520   2.041 0.042120 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.1488 on 298 degrees of freedom
```

```
> Multiple R-squared:  0.7282,Adjusted R-squared:   0.71
> F-statistic: 39.92 on 20 and 298 DF,  p-value: < 2.2e-16
```

Vemos que en este caso el método forward selecciona menos variables para formar parte del modelo, lo cual se puede deber al menor número de observaciones o a que las variables que descartamos en el primer ejemplo explican mejor la variable respuesta, pues por ejemplo, la variable `PolicReqPerOffic` había sido eliminada para el otro ejemplo y aquí es de las primeras en entrar a formar parte del modelo.

5.2.2. Regresión Ridge

La Regresión Ridge se comporta de manera análoga al otro ejemplo, restringiendo los valores de los coeficientes hacia 0 pero sin hacer que ninguno llegue a anularse del todo.

Es más interesante ver los grados de libertad que tenemos en este caso al tener más variables:

```
> [1] 76.33583
```

Tenemos sobre 76 grados de libertad, es decir, tenemos alrededor de 46 variables que serían dependientes. Podemos pensar que ésto es extraño, pues en el anterior ejemplo teníamos 97 grados de libertad, pero ésto puede deberse a que al haber más variables, algunas variables que antes no eran dependientes, sí que lo sean de alguna nueva variable o a que al haber menos observaciones, es más fácil que dos variables se parezcan y tengan efectos parecidos en la variable respuesta. Si vemos los autovalores de $X^\intercal X$ y el λ óptimo:

```
d

>   [1] 6.751411e+03 3.310641e+02 2.963378e+02 1.457352e+02 1.076662e+02
>   [6] 7.899147e+01 5.961196e+01 5.164962e+01 4.334169e+01 3.430570e+01
>  [11] 2.985842e+01 2.444336e+01 2.173814e+01 1.907117e+01 1.694054e+01
>  [16] 1.666467e+01 1.396912e+01 1.362073e+01 1.271511e+01 1.245265e+01
>  [21] 1.107276e+01 1.035262e+01 9.453406e+00 8.700460e+00 8.427822e+00
>  [26] 7.309624e+00 7.098792e+00 6.369915e+00 5.752349e+00 5.634851e+00
>  [31] 5.213077e+00 4.698603e+00 4.620977e+00 4.371056e+00 4.280984e+00
>  [36] 3.763187e+00 3.579108e+00 3.324329e+00 3.172287e+00 2.899109e+00
>  [41] 2.825981e+00 2.669568e+00 2.430909e+00 2.354661e+00 2.305195e+00
>  [46] 2.232536e+00 2.126094e+00 2.015417e+00 1.862591e+00 1.843619e+00
>  [51] 1.665208e+00 1.516935e+00 1.447088e+00 1.357896e+00 1.270873e+00
>  [56] 1.154511e+00 1.132011e+00 1.083867e+00 1.045764e+00 9.337073e-01
>  [61] 9.241624e-01 8.506109e-01 8.078843e-01 7.966854e-01 7.469328e-01
>  [66] 7.050845e-01 6.670442e-01 6.214566e-01 6.096061e-01 5.861279e-01
>  [71] 5.489321e-01 5.329927e-01 5.122719e-01 4.363749e-01 4.119782e-01
```

```
>  [76] 3.607045e-01 3.498168e-01 3.330701e-01 3.202710e-01 2.952868e-01
>  [81] 2.855247e-01 2.738219e-01 2.636199e-01 2.403370e-01 2.177143e-01
>  [86] 2.004889e-01 1.927065e-01 1.845068e-01 1.745235e-01 1.522989e-01
>  [91] 1.345950e-01 1.303834e-01 1.227295e-01 1.204170e-01 1.141251e-01
>  [96] 1.103628e-01 9.146213e-02 8.622036e-02 8.162227e-02 7.849304e-02
> [101] 7.120651e-02 6.639947e-02 6.354592e-02 5.637272e-02 4.672519e-02
> [106] 4.438049e-02 4.057538e-02 3.541944e-02 3.018446e-02 2.440120e-02
> [111] 2.170522e-02 1.767574e-02 1.562684e-02 1.409917e-02 1.314679e-02
> [116] 1.129816e-02 8.259237e-03 7.731541e-03 5.617162e-03 2.528735e-03
> [121] 1.774247e-03 3.309519e-05

l

> [1] 0.3109443
```

Vemos que en este caso, el λ que hemos escogido sí que corrige bastantes autovalores bajos de la matriz y ésto causa esa descenso en los grados de libertad.

5.2.3. Lasso

De manera análoga al otro ejemplo, obtenemos la gráfica 5.3. Si nos fijamos en los números que están encima, vemos que el intervalo con los mejores λ tiene asociados unos números muy pequeños de variables con coeficientes distintos de 0. Ésto nos indica que tendremos pocas variables en el modelo.

Si obtenemos el λ óptimo y lo utilizamos para aplicar el Lasso, obtenemos los siguientes coeficientes:

```
> 123 x 1 sparse Matrix of class "dgCMatrix"
>                                  s0
> (Intercept)            8.231608e-01
> Population             .
> HouseholdSize          .
> RacePctBlack           .
> RacePctWhite          -2.379550e-01
> RacePctAsian           .
> RacePctHisp            .
> AgePct12t21            .
> AgePct12t29            .
> AgePct16t24            .
> AgePct65up             .
> NumbUrban              .
> PctUrban               .
```

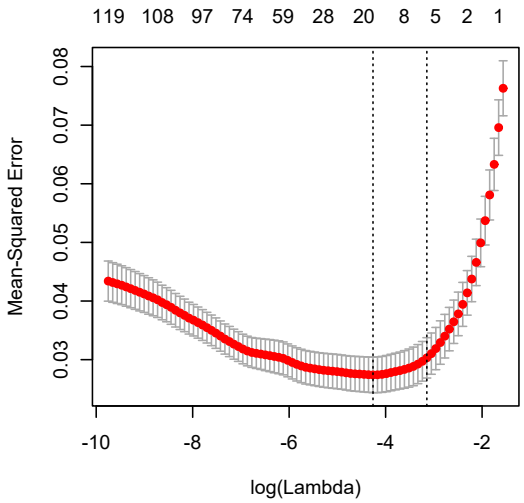

Figura 5.3: *Gráfico para encontrar el λ óptimo respecto al error cuadrático medio. Encima tenemos el número de variables con coeficiente distinto de 0 que tendríamos en el modelo para cada λ.*

```
> medIncome              .
> PctWWage               .
> PctWFarmSelf           .
> PctWInvInc            -1.579734e-01
> PctWSocSec             .
> PctWPubAsst            .
> PctWRetire             .
> medFamInc              .
> PerCapInc              .
> WhitePerCap            .
> BlackPerCap            .
> IndianPerCap           .
> AsianPerCap            .
> OtherPerCap            .
> HispPerCap             .
> NumUnderPov            .
> PctPopUnderPov         .
> PctLess9thGrade        .
> PctNotHSGrad           .
> PctBSorMore            .
> PctUnemployed          3.997975e-03
> PctEmploy              .
> PctEmplManu           -5.753988e-02
```

```
> PctEmplProfServ         .
> PctOccupManu            .
> PctOccupMgmtProf        .
> MalePctDivorce          .
> MalePctNevMarr          .
> FemalePctDiv            .
> TotalPctDiv             6.363810e-02
> PersPerFam              .
> PctFam2Par             -2.080818e-01
> PctKids2Par            -3.392593e-01
> PctYoungKids2Par        .
> PctTeen2Par             .
> PctWorkMomYoungKids     .
> PctWorkMom              .
> NumIlleg                .
> PctIlleg                5.177847e-05
> NumImmig                .
> PctImmigRecent          .
> PctImmigRec5            .
> PctImmigRec8            .
> PctImmigRec10           .
> PctRecentImmig          .
> PctRecImmig5            .
> PctRecImmig8            .
> PctRecImmig10           .
> PctSpeakEnglOnly        .
> PctNotSpeakEnglWell     .
> PctLargHouseFam         .
> PctLargHouseOccup       .
> PersPerOccupHous        .
> PersPerOwnOccHous       .
> PersPerRentOccHous      .
> PctPersOwnOccup         .
> PctPersDenseHous        .
> PctHousLess3BR          4.778221e-02
> MedNumBR                .
> HousVacant              .
> PctHousOccup           -1.719584e-02
> PctHousOwnOcc           .
> PctVacantBoarded        3.924570e-03
> PctVacMore6Mos          .
> MedYrHousBuilt          .
> PctHousNoPhone          .
> PctWOFullPlumb          .
```

```
> OwnOccLowQuart         .
> OwnOccMedVal           .
> OwnOccHiQuart          .
> RentLowQ               .
> RentMedian             .
> RentHighQ              .
> MedRent                .
> MedRentPctHousInc      .
> MedOwnCostPctInc       .
> MedOwnCostPctIncNoMtg  .
> NumInShelters          .
> NumStreet              1.591734e-01
> PctForeignBorn         .
> PctBornSameState       .
> PctSameHouse85         .
> PctSameCity85          .
> PctSameState85         .
> LemasSwornFT           .
> LemasSwFTPerPop        .
> LemasSwFTFieldOps      .
> LemasSwFTFieldPerPop   .
> LemasTotalReq          .
> LemasTotReqPerPop      .
> PolicReqPerOffic       3.901968e-02
> PolicPerPop            .
> RacialMatchCommPol     .
> PctPolicWhite          .
> PctPolicBlack          .
> PctPolicHisp           .
> PctPolicAsian          .
> PctPolicMinor          .
> OfficAssgnDrugUnits    .
> NumKindsDrugsSeiz      .
> PolicAveOTWorked       .
> LandArea               .
> PopDens                .
> PctUsePubTrans         .
> PolicCars              4.546114e-02
> PolicOperBudg          .
> LemasPctPolicOnPatr    .
> LemasGangUnitDeploy    1.212495e-02
> LemasPctOfficDrugUn    .
> PolicBudgPerPop        .
```

Es bastante notorio el bajo número de variables que Lasso incluye en el modelo, te-

niendo, al contrario que en el ejemplo anterior, menos variables que el método forward, y muchas menos que en el ejemplo anterior. En este caso vemos como Lasso realmente actúa seleccionando variables y descartando 107 variables de las 122 que tenemos.

5.2.4. Least Angle Regression

Si aplicamos el método de Least Angle Regression, las variables del modelo ordenadas por importancia quedan como:

```
>   [1] "PctKids2Par"           "RacePctWhite"
>   [3] "NumStreet"             "FemalePctDiv"
>   [5] "PctWInvInc"            "TotalPctDiv"
>   [7] "PolicCars"             "HousVacant"
>   [9] "PctFam2Par"            "PctHousLess3BR"
>  [11] "PctEmplManu"           "PolicReqPerOffic"
>  [13] "PctHousOccup"          "LemasGangUnitDeploy"
>  [15] "PctVacantBoarded"      "PctIlleg"
>  [17] "PctUnemployed"         "PctPolicWhite"
>  [19] "RacialMatchCommPol"    "PctPolicAsian"
>  [21] "LemasPctOfficDrugUn"   "RacePctBlack"
>  [23] "OtherPerCap"           "PctBornSameState"
>  [25] "MedOwnCostPctIncNoMtg" "PctWWage"
>  [27] "AgePct12t29"           "PctWFarmSelf"
>  [29] "PctUsePubTrans"        "MedOwnCostPctInc"
>  [31] "HispPerCap"            "PctLess9thGrade"
>  [33] "IndianPerCap"          "PctSameState85"
>  [35] "RentLowQ"              "RacePctHisp"
>  [37] "PctImmigRec10"         "PctWRetire"
>  [39] "PctForeignBorn"        "PolicAveOTWorked"
>  [41] "LandArea"              "PersPerRentOccHous"
>  [43] "NumUnderPov"           "PctWOFullPlumb"
>  [45] "NumImmig"              "PctImmigRecent"
>  [47] "BlackPerCap"           "PctImmigRec5"
>  [49] "AsianPerCap"           "WhitePerCap"
>  [51] "PopDens"               "LemasPctPolicOnPatr"
>  [53] "NumInShelters"         "LemasTotalReq"
>  [55] "PctOccupManu"          "NumKindsDrugsSeiz"
>  [57] "PersPerFam"            "PctNotHSGrad"
>  [59] "PctUrban"              "MedRent"
>  [61] "PctEmplProfServ"       "OwnOccLowQuart"
>  [63] "PctWPubAsst"           "MedRentPctHousInc"
>  [65] "MalePctDivorce"        "HouseholdSize"
>  [67] "LemasTotReqPerPop"     "PctOccupMgmtProf"
>  [69] "MedNumBR"              "PctPolicHisp"
```

```
>  [71] "OwnOccMedVal"          "PctNotSpeakEnglWell"
>  [73] "PolicOperBudg"         "PctWorkMom"
>  [75] "RacePctAsian"          "NumIlleg"
>  [77] "PctSameCity85"         "PctEmploy"
>  [79] "PctVacMore6Mos"        "PctSpeakEnglOnly"
>  [81] "PctPolicMinor"         "PolicBudgPerPop"
>  [83] "OfficAssgnDrugUnits"   "AgePct12t21"
>  [85] "PctWSocSec"            "PctPopUnderPov"
>  [87] "PctWorkMomYoungKids"   "AgePct65up"
>  [89] "PctPersDenseHous"      "PerCapInc"
>  [91] "PctLargHouseOccup"     "medFamInc"
>  [93] "PctHousOwnOcc"         "LemasSwFTPerPop"
>  [95] "MalePctNevMarr"        "PctRecentImmig"
>  [97] "LemasSwFTFieldPerPop"  "PersPerOwnOccHous"
>  [99] "PctHousNoPhone"        "PctPolicBlack"
> [101] "NumbUrban"             "MedYrHousBuilt"
> [103] "RentHighQ"             "PctRecImmig10"
> [105] "PctYoungKids2Par"      "PersPerOccupHous"
> [107] "PctBSorMore"           "PctTeen2Par"
> [109] "PctSameHouse85"        "LemasSwornFT"
> [111] "OwnOccHiQuart"         "PctRecImmig8"
> [113] "LemasSwFTFieldOps"     "RentMedian"
> [115] "AgePct16t24"           "PctImmigRec8"
> [117] "PctPersOwnOccup"       "medIncome"
> [119] "PctLargHouseFam"       "Population"
> [121] "PolicPerPop"           "PctRecImmig5"
```

En este caso también podemos ver que entre las primeras variables se encuentra `PolicCars`, una variable que no teníamos en el otro ejemplo y que resulta ser bastante importante.

En general, vemos que los modelos parecen ser mucho más sencillos en este ejemplo debido a que algunas de las variables que eliminamos estaban relacionadas con la Policía y es lógico que estos datos afecten a la tasa de criminalidad, pero también se debe a la gran disminución de observaciones que tenemos, pues esto nos hace perder información.

Bibliografía

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Faraway, J. (2016). R package. Basado en el libro "Linear Models with R"publicado en Agosto de 204.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The elements of statistical learning*. Springer.

Hastie, T. (2019). R package.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An introduction to statistical learning*. Springer.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. (2017). R package.