Tarea-Examen 5-6: Selección de modelos y regularización

Curso Avanzado de Estadística. Profa. Guillermina Eslava Gómez.

Aldo Sayeg Pasos Trejo. César Cossio Guerrero.

Posgrado en Ciencias Matemáticas. Universidad Nacional Autónoma de México.

13 de mayo de 2020

1. Problema 1

Para el conjunto de datos "Boston", que consta de 13 variables numéricas y un variable categórica, buscamos predecir la variable `crim` como función de las otras 13 variables y sus interacciones a pares, lo que nos da $13 + \binom{13}{2} + 1 = 91$ variables predictoras.

Notemos que las escalas de los datos son muy distintas entre cada variable, por lo se estandarizaron todas las variables numéricas para que tengan promedio 0 y error estándar 1. La variable categórica, al ser una indicadora, se dejó igual.

Posteriormente, se buscó cual era el valor de l tal que al aplicar una transformación Box-Cox a `crim` se maximizaba la probabilidad de tener un comportamiento normal. Se encontro que $l = -0.11$ y se realizó la transformación a la variable respuesta.
Probaremos siete modelos: tres sin regularizar, tres regularizados y uno utilizando un random forest regressor.

1.1. Modelo normal

Antes de escoger los modelos, realizamos un ajuste con un modelo normal de mínimos cuadrados que usa todas las variables. La tabla 1 muestra as estadísticas de dicho modelo.

R^2	aic	bic	Aparent error	Non apparent error	Non apparent error std
0.9373	1164.0844	1548.6992	0.4078	0.7629	0.0596

Tabla 1: Estadísticas del modelo base. Los errores no aparentes se calcularon mediante validación cruzada con $k = 5$ y 100 repeticiones

1.2. Modelos no regularizdos

Para los modelos regularizados, el primero se obtubo utilizando selección stepwise empezando con las 91 variables originales y removiendo variables para minimizar el AIC.

El segundo se obtubo de igual manera, pero minizando el BIC en lugar del AIC. Para el tercer modelo no regularizado, se busco el mejor subconjunto en un procedimiento hacia adelante, empezando por el mejor modelo de una variable y añadiendo variables hasta llegar al mejor modelo de 20 variables. Cabe señalar que la búsqueda exaustiva era imposible en términos de la potencia computacional, ya que hay alredor de 10^{11} modelos con entre 1 y 8 variables.

Las tablas `drop1` de los modelos, que muestran que son optimos en el sentido de que quitar una variable no mejora su AIC, corresponden a las tablas 8, 10 y 12 contenidas en el anexo a este reporte. Cabe mencionar que todos estos modelos incluían también una constante.

1.3. Modelos regularizados

Se seleccionaron tres modelos utilizando el paquete `glmnet`. Uno pare regresión Ridge, otro para Lasso y otro para Elastic Net con $\alpha = 0.5$. Se analizaron distintos valores de regularización lambda. La gráfica 1 muestra el error cuadrático medio como función del parámetro de regularización λ para estos modelos, en un análisis con validación cruzada para $k = 200$ y $n = 1$

La constante λ se selecciono como `lambda_1se`, la mayor λ tal que su error cuadrático medio esta a una desviación estándar del mínimo. Este valor se escogió pensando en que el modelo tuviera la menor cantidad posible de variables para aumentar su capacidad descriptiva y para tener un modelo que sea computacionalmente más sencillo de manejar. Además, el error estándar era suficientemente pequeño por lo que la λ escogida aproximaba bien el error mínimo con menos variables.

1.4. Random forest

Se costruyó un modelo de random forest con $Nboot = 100$ árboles distintos, tal que cada uno trabajaba con una muestra obtenida por bootstrap de los datos originales. El criterio para ramificar un árbol era el error cuadrático medio. Por sugerencia de los autores del paquete, se tomaban en cuenta las $m = 91$ variables a la hora de buscar la mejor ramificación en cada árbol. Esto contrasta con el valor sugerido de $m = 91/3 \approx 30$ para regresión. Sin embargo, ese valor sugerido se encontró mediante análisis empírico, por lo que tomar las 91 variables es igual de arbitrario que la selección mencionada.

1.5. Conclusiones

La tabla 2 sintetiza las características propias de cada modelo.

	model	Dfs	details
1	step (AIC)	52	-
2	step (BIC)	32	-
3	Best subset	20	Selección hacia adelante (forward), iniciando en el mejor modelo de 1 variable y finalizando en el mejor con 20
4	Lasso	25	Se usó lambda_1se = 0.0286
5	Ridge	92	Se usó lambda_1se = 0.5232
6	Elastic Net	31	Se usó lambda_1se = 0.0521
7	Random Forest	91	NBoot = 100, m = 91

Tabla 2: Detalles de cada modelo

Para revisar explícitamente las variables que usan los modelos, referimos al lector a la figura 1. . La tabla 3 sintetiza los errores y las estadísticas relevantes de cada modelo

	model	Dfs	R^2	Aparent error	Non aparent error	Non aparent error std
1	step (AIC)	52	0.9199	0.4245	0.5671	0.0255
2	step (BIC)	32	0.9107	0.4733	0.5564	0.0143
3	Best subset	20	0.8860	0.6043	0.6641	0.0129
4	Lasso	25	0.8871	0.5983	0.6749	0.0126
5	Ridge	92	0.8824	0.6234	0.7304	0.0125
6	Elastic Net	31	0.8881	0.5934	0.6772	0.0129
7	Random Forest	91	0.9929	0.0378	0.3039	0.0123

Tabla 3: Errores para cada modelo

Los errores no aparentes se calcularon mediante validación cruzada. Para el modelo de random forest, se usó $k = 5$ y 20 repeticiones. Para todos los demás modelos se tomó de igual manera $k = 5$ y se hicieron 100 repeticiones.

Para a los modelos 1 y 2, obtenidos al hacer selección de variables mediante stepwise utilizando el AIC y el BIC, respectivamente, como criterios a minimizar, también se calculó el error no aparente al incluir. La tabla 4 muestra dichos valores y los compara con las tasas no aparentes obtenidas sin incluir la selección de variables (que a su vez se muestran en la tabla 3)

	model	Selection outside CV	Selection outside CV std	Selection inside CV	Selection inside CV std
1	step (AIC)	0.5671	0.0255	1.5123	0.1992
2	step (BIC)	0.5564	0.0143	1.3623	0.3950

Tabla 4: Error cuadrático medio al incluir el proceso de selección en la validación. Debido a las limitaciones computacionales, para el error incluyendo el proceso de selección se realizaron 20 iteraciones solamente

Pensando puramente en el poder predictivo, el modelo que utiliza Random Forest es el que presenta el R^2 más alto y menores errores tanto aparentes como no aparentes. Sin embargo, debido a su uso de todas las variables, no es el más óptimo para hacer descripción por el alto número de variables. En ese sentido, el modelo obtenido por step usando el criterio BIC tiene un R^2 más alto que los regularizados y errores aparentes muy bajos, y cuenta con tan solo 32 variables, mucho más manejables para descripción. Consideramos que es un mejor modelo predictivo.

2. Problema 2

La base de datos con la que se trabajó es Riboflavin. Esta consta de $n = 71$ observaciones en $p = 4089$ dimesiones que corresponden a la expresión de los genes de distintas sepas de *Bacillus subtilis* en relación con su producción de vitamina riboflavin, (B-2).

2.1. Selección de modelos

Como primer paso se seleccionaron 3 modelos por Lasso, Ridge y Elasticnet ($\alpha = 0.5$) con base en los valores de error calculados por 5-fold Cross Validation para una y para 500 repeticiones. Esto dado que el parámetro de *tunning* λ depende de una semilla inicial, por lo cual se desarrolló un programa capaz de calcular otro valor óptimo de λ además de λ_{min} y λ_{1se}.

Se realizó la selección de modelos mediante el uso de la construcción de una tabla de errores generados por cross validation tanto para Lasso, Elasticnet, y Ridge. Además de los modelos seleccionados para los valores de λ_{min} y λ_{1se} se realizó el cálculo por Cross Validation con $k = 5$ y $B = 500$ repeticiones para seleccionar un valor óptimo de λ_{1se} que denominaremos como λ_{RCV}. El criterio de decisión fue buscar aquel modelo que presentará el menor error, como muestra la tabla 5.

Modelo	mse_{min}	mse_{1se}	mse_{RCV}	λ_{min}	λ_{1se}	λ_{RCV}	df_{min}	df_{1se}	df_{RCV}
Ridge	0.26	0.29	0.24	5.93	21.82	18.98	4089	4089	4089
Elasticnet	0.24	0.30	0.19	0.10	0.24	0.13	49	34	45
Lasso	0.18	0.22	0.18	0.039	0.08	0.05	40	29	34

Tabla 5: Se presentan los valores de los errores por Cross Validation. Las 3 primeras columnas corresponden a los errores de MSE, las siguientes 3 a los valores de λ, y las últimas 3 al número de variables de cada modelo.

De la tabla 5, podemos notar que el valor de λ_{RCV} tiene una cierta ventaja respecto al error de

λ_{1se} y de λ_{min} ya que lo disminuye o lo mantiene. Y para los casos de Elasticnet y Lasso además conserva las mismas característcas de baja dimenionalidad que el modelo correspondiente a λ_{1se}. Como conlusión se puede pensar en esta metodología como la acción de escoger el modelo cuya dimensionalidad siga siendo pequeña y el error disminuya lo más que se pueda.

Por otra parte, se agregan las figuras que resultaron de realizar los cálculos por Cross Validation, tanto para una como para 500 repeticiones para la selección de los modelos, ver la figura 3. En ellas se puede apreciar que el error por Cross Validation podría tener valores más bajos para ciertas λ_{1se} sin perder la propiedad de tener pocas variables. Sin embargo, a falta de una solución sencilla para conocer la bondad de ajuste o alguna medida con el criterio BIC o la función de pérdida del ajuste no se nos ocurrieron más criterios para la selección de modelos.

También es bueno aclarar que las variables de cada modelo seleccionado no son necesariamente las variables predictoras verdaderas ya que hay un efecto considerable de multicolinealidad que detectamos. Sin embargo, y a pesar de que dicha tarea sale de los objetivos de este trabajo, implementamos *PCA* y *Hierarchical clustering* para agrupar y comprobar si podía encontrarse alguna relación entre los modelos seleccionados y los clusters, pero dicha tarea no dió algún resultado digno de presentarse en este trabajo. Otro aspecto importante es que por nuestra falta de conocimiento acerca del tema y la elevada cantidad de variables de cada modelo seleccionado nos orilló a omitir la presentación explícita de las variables obtenidas[1].

2.2. Cálculo de errores

e procedió a calcular tanto los errores aparentes como los no aparentes utilizando Cross Validation con $k = 5$, y 500 repeticiones. Dichos resultados se comparan con el modelo nulo, que en este caso se escogió como el modelo que solo cuenta con una constante (la media[2]), ver tabla 6. Podemos notar de la tabla 6 que todos los modelos obtenidos mediante Lasso, Ridge o Elastic Net

Modelo	Error aparente	Error no aparente	Error no aparente std
Ridge	0.079	0.29	0.027
Elasticnet	0.056	0.25	0.034
Lasso	0.056	0.24	0.028
Modelo nulo	0.83	0.86	0.018

Tabla 6: Se presentan los errores MSE calulados por 5-fold Cross Validation con 500 repeticiones para los 4 modelos seleccionados: Ridge, Elasticnet, Lasso, y el modelo nulo. El primer renglón cuenta con los errores aparentes o de entrenamiento, mientras que el segundo muestra los errores no aparentes o de validación.

tienen errores de predicción menores que los presentados por el modelo nulo. También podemos observar que Lasso y Elasticnet tienen la capacidad de hacer una reducción de variables predictivas significativa en el modelo, mientras que Ridge no posee esta habilidad pues no está diseñado para ello.

[1] A pesar de ello se pueden calcular de manera sencilla en el código de *R* anexado a esta tarea. O bien dado que cada valor de λ define un modelo se pueden obtener a partir de dicho valor.

[2] Utilizar una regresión múltiple fue inviable computacionalmente.

2.3. Conclusiones

Podemos enfatizar que la reducción de varaibles es muy notoria, pues se pasa de 4088 variables a solo contar con 34 o 45. También resulta de dicha reducción de dimensionalidad no conlleva un costo significativo en el error de predicción. Cabría también utilizar diferentes métodos de reducción de dimensionalidad a la par con estas metodología para hacer más evaluaciones y tener más modelos de donde poder seleccionar.

De la figura 3 podemos notar que el error que presenta el valor mínimo de λ con una repetición a veces resulta ser mayor que el valor de 1se para alguna otra repetición. O bien que para un mismo valor de lambda el error cambia, y este efecto, pensamos, puede deberse a la semilla con la que se realiza el cálculo del error.

Anexo 1: Tablas relevantes

Problema 1

Dropped	aic	Change	rss	Change	F_test_P_value
indus	1170.5042	-64.1465	244.7722	-29.9945	0.0000
chas	1109.5669	-3.2092	217.0002	-2.2225	0.0000
nox	1112.9807	-6.6230	218.4692	-3.6915	0.0000
rm	1116.2603	-9.9026	219.8898	-5.1121	0.0000
age	1118.1255	-11.7678	220.7018	-5.9241	0.0000
rad	1334.1591	-227.8015	338.2403	-123.4626	0.0000
tax	1152.4560	-46.0983	236.1954	-21.4177	0.0000
ptratio	1116.5640	-10.2063	220.0218	-5.2441	0.0000
black	1127.2471	-20.8894	224.7165	-9.9388	0.0000
zn:rad	1116.5953	-10.2376	220.0354	-5.2577	0.0000
zn:tax	1107.8273	-1.4696	216.2555	-1.4778	0.0000
zn:ptratio	1147.2111	-40.8534	233.7598	-18.9821	0.0000
zn:black	1110.9513	-4.5936	217.5948	-2.8171	0.0000
zn:lstat	1110.4378	-4.0801	217.3741	-2.5963	0.0000
zn:medv	1128.6895	-22.3318	225.3580	-10.5803	0.0000
indus:chas	1130.7156	-24.3579	226.2622	-11.4844	0.0000
indus:rm	1113.3641	-7.0064	218.6348	-3.8571	0.0000
indus:age	1112.1707	-5.8130	218.1198	-3.3421	0.0000
indus:dis	1121.0147	-14.6570	221.9657	-7.1879	0.0000
indus:rad	1148.5736	-42.2159	234.3901	-19.6124	0.0000
indus:lstat	1109.4576	-3.0999	216.9534	-2.1757	0.0000
chas:nox	1107.0549	-0.6972	215.9256	-1.1479	0.0000
chas:rad	1122.5000	-16.1423	222.6182	-7.8404	0.0000
chas:tax	1116.5466	-10.1889	220.0143	-5.2365	0.0000
chas:ptratio	1110.8793	-4.5216	217.5638	-2.7861	0.0000
chas:medv	1107.4374	-1.0797	216.0889	-1.3112	0.0000
nox:rm	1109.9112	-3.5535	217.1480	-2.3702	0.0000
nox:age	1111.4292	-5.0715	217.8004	-3.0227	0.0000
nox:dis	1107.2842	-0.9265	216.0235	-1.2458	0.0000
nox:rad	1111.5419	-5.1842	217.8489	-3.0712	0.0000
nox:black	1108.9016	-2.5439	216.7151	-1.9374	0.0000
nox:lstat	1108.0570	-1.6993	216.3537	-1.5760	0.0000
rm:age	1127.0312	-20.6736	224.6207	-9.8429	0.0000
rm:dis	1108.4313	-2.0736	216.5138	-1.7361	0.0000
rm:lstat	1108.9197	-2.5620	216.7229	-1.9451	0.0000
age:dis	1111.3901	-5.0324	217.7836	-3.0058	0.0000
age:black	1108.8151	-2.4574	216.6781	-1.9004	0.0000
age:lstat	1110.7735	-4.4158	217.5183	-2.7406	0.0000
age:medv	1131.2931	-24.9354	226.5206	-11.7428	0.0000
dis:rad	1137.3609	-31.0032	229.2533	-14.4755	0.0000
dis:tax	1126.3030	-19.9453	224.2976	-9.5199	0.0000

Continued on next page

Dropped	aic	Change	rss	Change	F_test_P_value
dis:ptratio	1147.1842	-40.8266	233.7474	-18.9697	0.0000
dis:black	1108.8309	-2.4732	216.6849	-1.9071	0.0000
dis:lstat	1107.4313	-1.0736	216.0863	-1.3086	0.0000
rad:tax	1162.1464	-55.7888	240.7624	-25.9847	0.0000
rad:ptratio	1124.8337	-18.4760	223.6473	-8.8695	0.0000
rad:lstat	1113.9155	-7.5578	218.8732	-4.0955	0.0000
rad:medv	1114.8278	-8.4701	219.2682	-4.4904	0.0000
tax:black	1111.4821	-5.1244	217.8232	-3.0454	0.0000
tax:medv	1117.8369	-11.4792	220.5760	-5.7983	0.0000
ptratio:lstat	1110.8969	-4.5392	217.5714	-2.7936	0.0000
lstat:medv	1109.1756	-2.8179	216.8325	-2.0548	0.0000

Tabla 8: `drop1` para modelo seleccionado mediante step con AIC

Dropped	aic	Change	rss	Change	F_test_P_value
indus	1178.8433	-57.4121	269.3091	-29.8352	0.0000
chas	1131.3009	-9.8698	245.1579	-5.6840	0.0000
nox	1127.2810	-5.8498	243.2180	-3.7440	0.0000
rm	1138.7234	-17.2922	248.7806	-9.3067	0.0000
age	1151.0509	-29.6197	254.9160	-15.4421	0.0000
rad	1345.7913	-224.3602	374.5767	-135.1028	0.0000
tax	1154.1394	-32.7082	256.4767	-17.0028	0.0000
ptratio	1135.5138	-14.0826	247.2076	-7.7336	0.0000
black	1147.3374	-25.9062	253.0520	-13.5781	0.0000
zn:rad	1141.6912	-20.2601	250.2441	-10.7702	0.0000
zn:ptratio	1170.8985	-49.4673	265.1137	-25.6397	0.0000
zn:medv	1137.9303	-16.4991	248.3910	-8.9171	0.0000
indus:chas	1141.2698	-19.8386	250.0357	-10.5618	0.0000
indus:dis	1141.5834	-20.1522	250.1908	-10.7168	0.0000
indus:rad	1154.5266	-33.0954	256.6730	-17.1991	0.0000
chas:rad	1135.7532	-14.3221	247.3246	-7.8507	0.0000
chas:tax	1137.0461	-15.6149	247.9573	-8.4834	0.0000
nox:rad	1129.6508	-8.2196	244.3597	-4.8858	0.0000
nox:black	1128.1628	-6.7316	243.6422	-4.1683	0.0000
rm:age	1133.9579	-12.5267	246.4486	-6.9747	0.0000
age:dis	1131.7354	-10.3042	245.3685	-5.8946	0.0000
age:medv	1141.6045	-20.1734	250.2012	-10.7273	0.0000
dis:rad	1158.0516	-36.6204	258.4674	-18.9934	0.0000
dis:tax	1135.9143	-14.4831	247.4033	-7.9294	0.0000
dis:ptratio	1158.4380	-37.0069	258.6648	-19.1909	0.0000

Continued on next page

Dropped	aic	Change	rss	Change	F_test_P_value
rad:tax	1164.7357	-43.3045	261.9043	-22.4304	0.0000
rad:ptratio	1141.9392	-20.5081	250.3667	-10.8928	0.0000
rad:lstat	1132.1784	-10.7473	245.5834	-6.1095	0.0000
rad:medv	1134.9231	-13.4919	246.9192	-7.4452	0.0000
tax:black	1131.6063	-10.1752	245.3059	-5.8320	0.0000
tax:medv	1137.2285	-15.7974	248.0467	-8.5728	0.0000
lstat:medv	1162.1166	-40.6855	260.5522	-21.0783	0.0000

Tabla 10: `drop1` para modelo seleccionado mediante step con BIC

Dropped	aic	Change	rss	Change	F_test_P_value
rad	1513.1417	-292.0245	546.7324	-240.9470	0.0000
indus:rad	1223.6706	-2.5534	308.5494	-2.7641	0.0000
nox:ptratio	1227.2537	-6.1365	310.7421	-4.9568	0.0000
indus:dis	1244.1984	-23.0812	321.3243	-15.5390	0.0000
age	1226.0415	-4.9243	309.9986	-4.2133	0.0000
rad:tax	1234.6310	-13.5138	315.3058	-9.5205	0.0000
lstat	1246.0270	-24.9098	322.4876	-16.7023	0.0000
nox	1273.6161	-52.4989	340.5591	-34.7738	0.0000
indus:nox	1219.1798	1.9374	305.8232	-0.0378	0.0000
indus:ptratio	1225.0491	-3.9319	309.3912	-3.6058	0.0000
ptratio:medv	1228.6463	-7.5291	311.5985	-5.8132	0.0000
nox:tax	1219.2931	1.8241	305.8916	-0.1063	0.0000
dis:rad	1248.8821	-27.7649	324.3124	-18.5271	0.0000
zn:nox	1227.6882	-6.5710	311.0090	-5.2237	0.0000
tax:lstat	1241.8156	-20.6984	319.8147	-14.0294	0.0000
ptratio:lstat	1227.4549	-6.3377	310.8657	-5.0803	0.0000
tax:ptratio	1231.3335	-10.2163	313.2577	-7.4724	0.0000
nox:rad	1235.7129	-14.5957	315.9807	-10.1953	0.0000
dis:ptratio	1228.7638	-7.6466	311.6709	-5.8856	0.0000
nox:age	1228.7421	-7.6249	311.6575	-5.8722	0.0000

Tabla 12: `drop1` para modelo seleccionado mediante mejor subconjunto

Anexo 2: Figuras relevantes

Problema 1

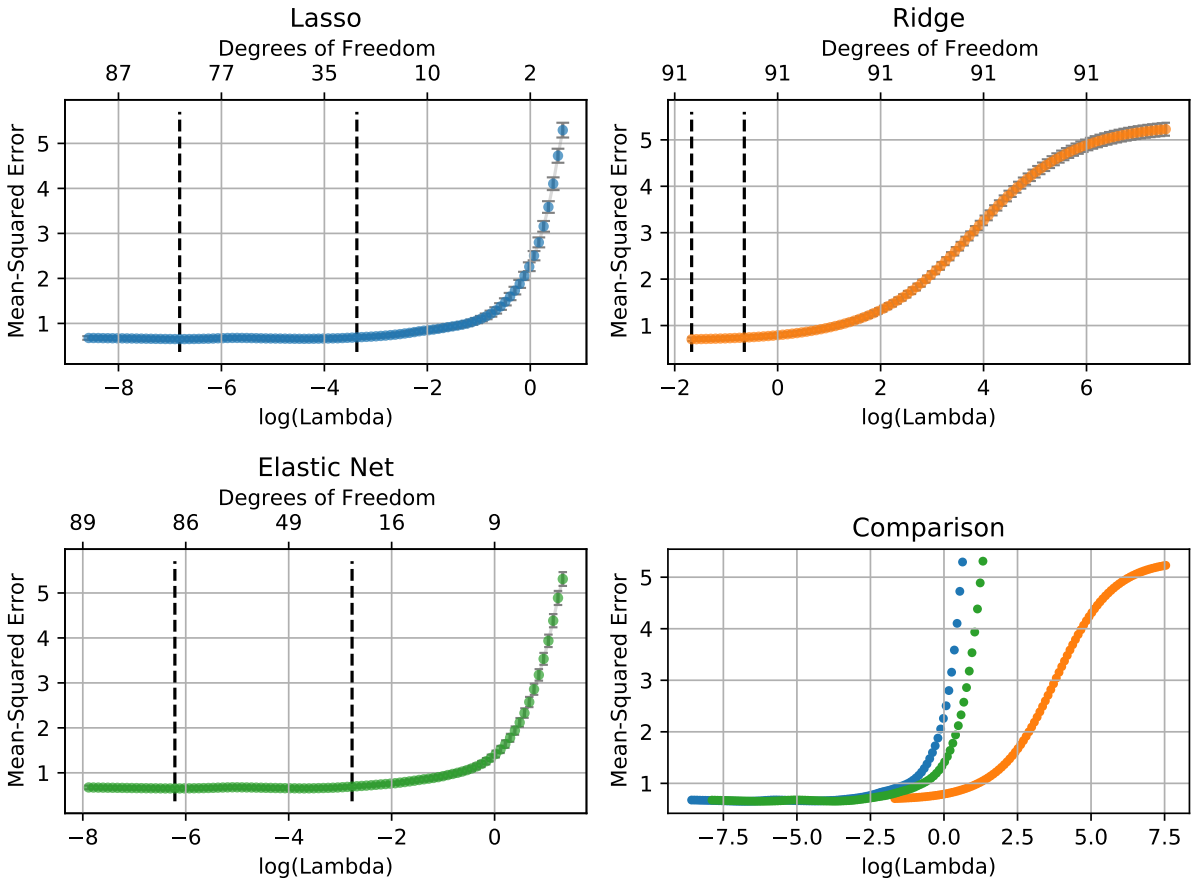


Figura 1: Error cuadrático medio para los modelos regularizados, encontrado por validación cruzada para $k = 20$. Las primer recta vertical es el valor de `lambda_min`, mientras la segunda coresponde a `lambda_1se`

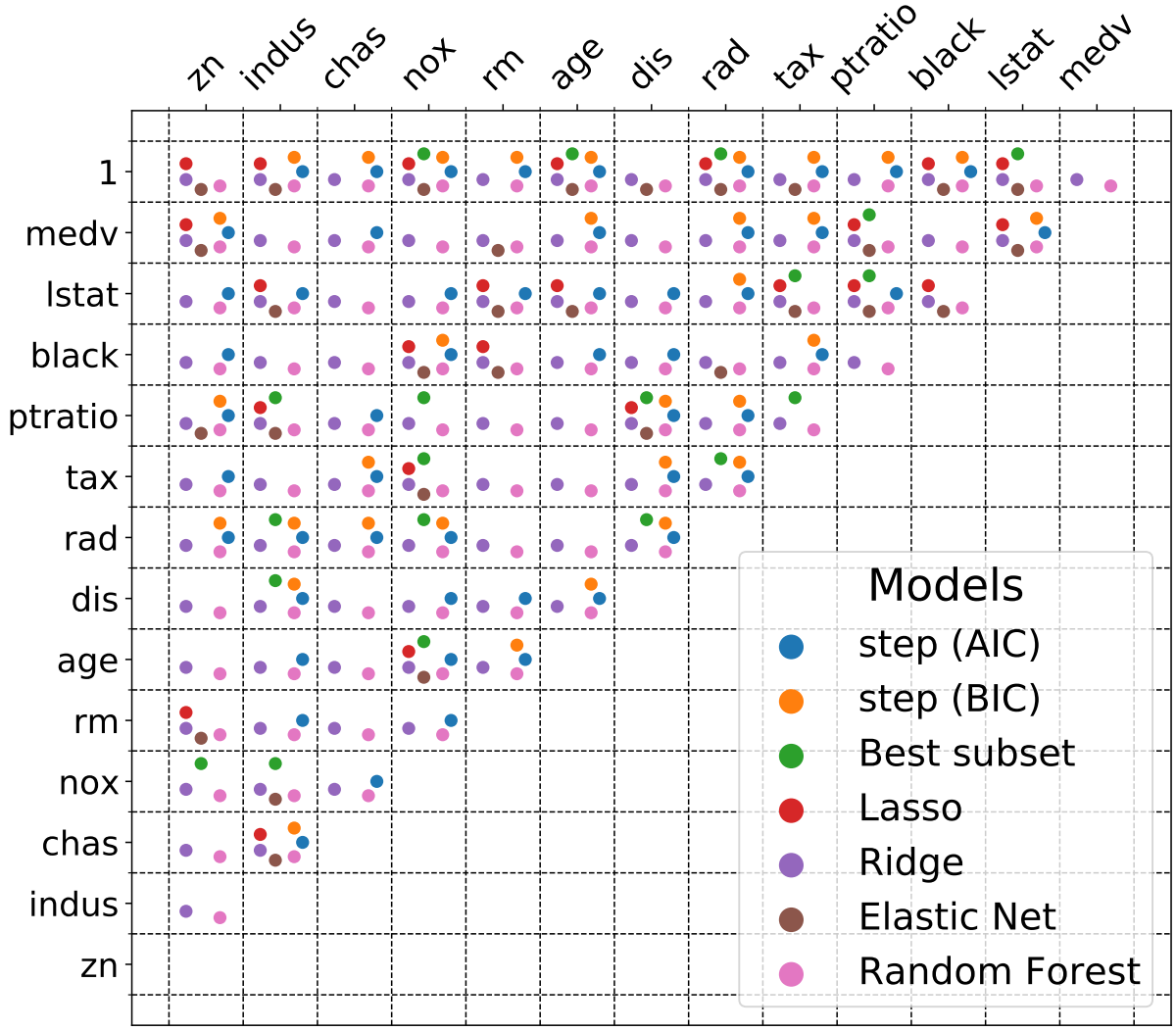
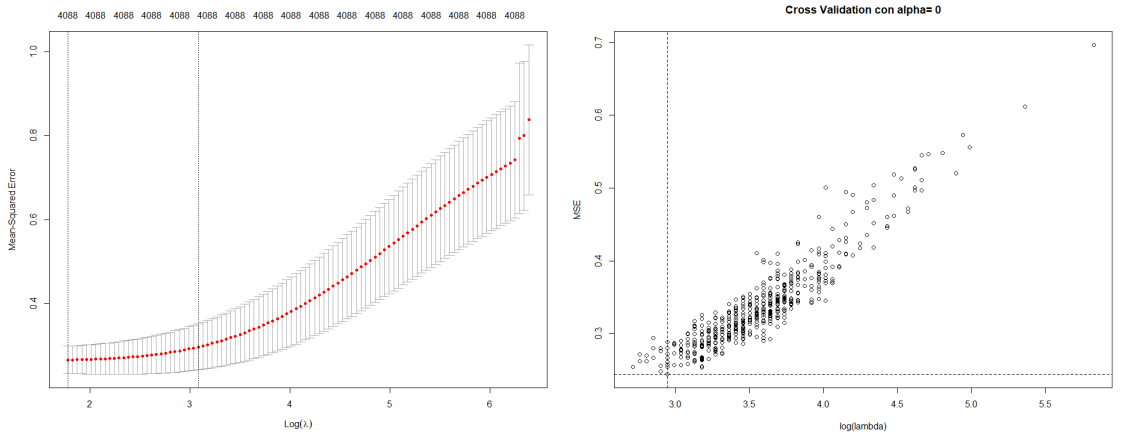


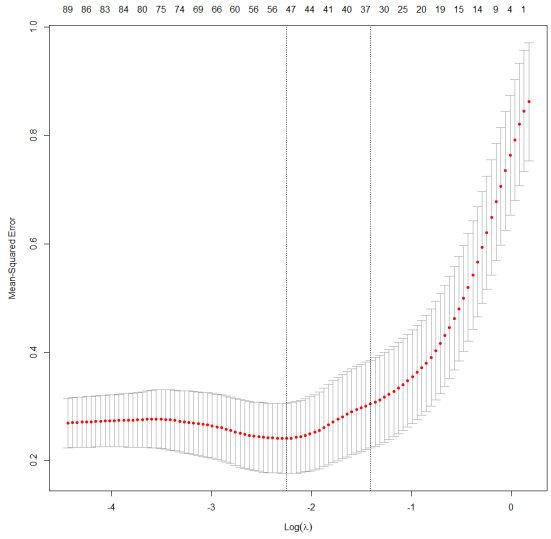
Figura 2: Variables para cada modelo. Un punto en una región indica que se incluyo la interacción entre esas variables. Un punto en el renglón con 1 indica que se incluyó ese efecto principal

Problema 2

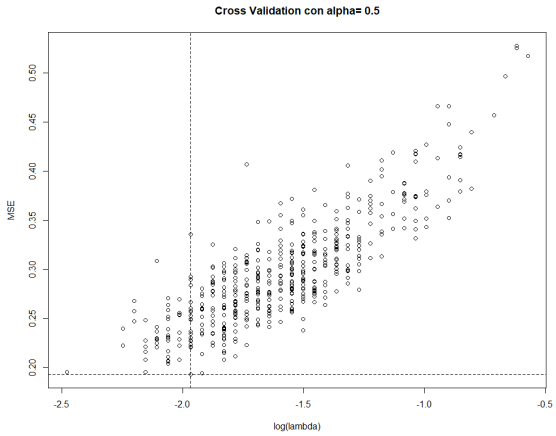


(a)

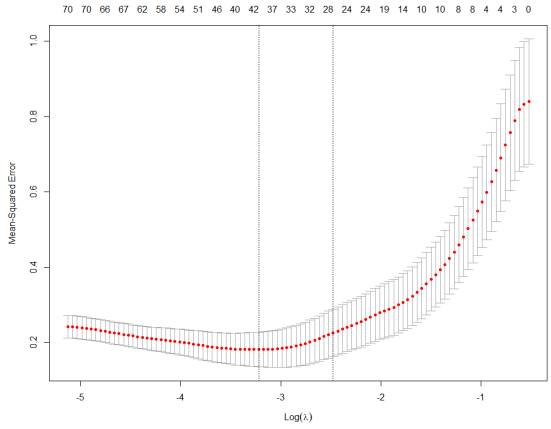
(b)



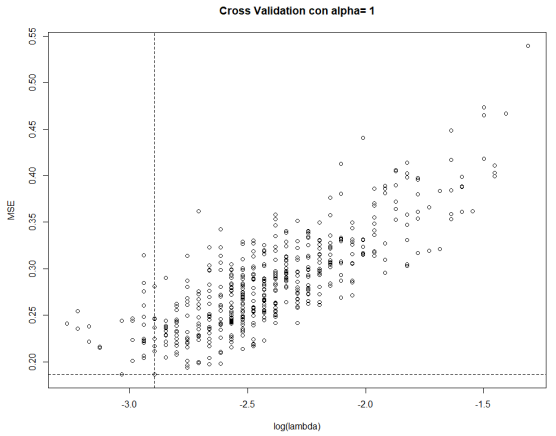
(c)



(d)



(e)



(f)

Figura 3: Perfiles de errores por Cross validataion para: (a) Ridge con 1 repetición; (b) Ridge con 500 repeticiones; (c) para Elasticnet con 1 repetición, (d) para Ridge con 500 repeticiones, (e) para Lasso con 1 repetición; (f) para Lasso con 500 repeticiones.