

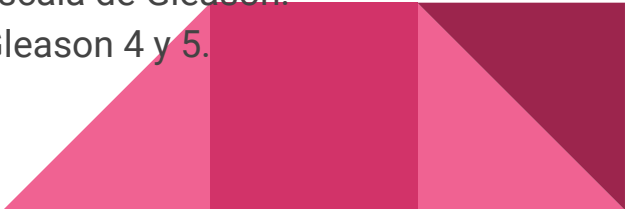
Machine Learning - Tarea 1

Felipe Berrios - Felipe Vásquez

Regresión Lineal Ordinaria (LSS)

Conjunto de Datos

El dataset se compone de 97 registros (pacientes), cada uno de los cuáles está descrito por 9 variables, las que se detallan a continuación:

- ❑ **lpsa**: Logaritmo del nivel de antígeno prostático específico (PSA). Variable a predecir.
 - ❑ **lcavol**: Logaritmo del volumen de cáncer presente.
 - ❑ **lweight**: Logaritmo del peso de la próstata.
 - ❑ **age**: Edad del paciente.
 - ❑ **lbph**: Logaritmo de la cantidad de hiperplasia benigna de próstata.
 - ❑ **svi**: Invasión de la vesícula seminal.
 - ❑ **lcp**: Logaritmo de la penetración capsular.
 - ❑ **gleason**: Medida del grado de agresividad del cáncer, en base a la escala de Gleason.
 - ❑ **pgg45**: Porcentaje que representa la presencia de los patrones de Gleason 4 y 5.
- 

Regresión lineal

- ❑ Antes de generar el modelo predictivo los datos deben ser normalizados, debido a que las variables tienen unidades de medida y escalas diferentes.
- ❑ Normalizar permite eliminar los efectos de la media y la varianza presentes en los datos.
- ❑ Como se quiere determinar si existe relación entre lpsa y el resto de variables para la detección del cáncer prostático se tiene que:
 - ❑ Predictores: lcavol, lweight, age, lbph, svi, lcp, gleason y pgg45.
 - ❑ Variable dependiente: lpsa.
- ❑ Entonces, el modelo de regresión lineal recibe como parámetro el conjunto de predictores (más el intercepto) y la variable que se desea predecir.



Pesos y Z-score

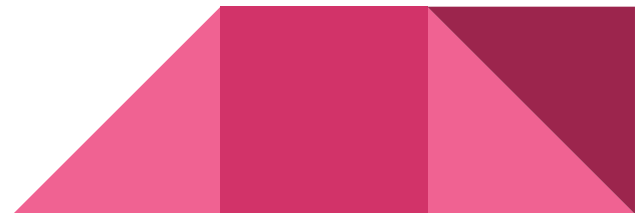
- ❑ Las variables más correlacionadas con lpsa son: **lcavol**, **lweight**, **svi**, de acuerdo al valor del Z-score.
- ❑ Con un nivel de significancia del 5%, las variables **age**, **lcp**, **gleason**, **pgg45**, no tienen relación con lpsa.

Variable	Peso	Z-score
lcavol	0.68	5.32
lweight	0.26	2.73
age	-0.14	-1.38
lbph	0.21	2.04
svi	0.30	2.44
lcp	-0.29	-1.85
gleason	-0.02	-0.14
pgg45	0.27	1.72
intercept	2.46	27.36

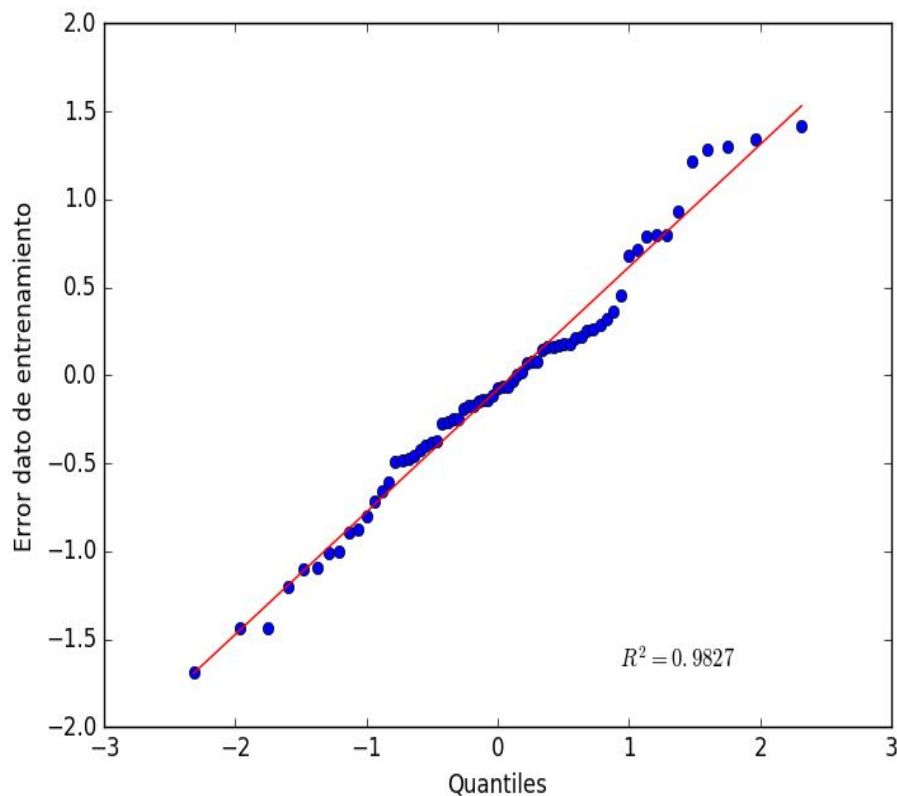
Estimación del error de predicción

Error de predicción usando validación cruzada:

	LSS	K = 5	K = 10
MSE	0.52	0.96	0.76



Error de predicción



- Se observa que los errores de entrenamiento vs los percentiles de una distribución normal se pueden representar como una recta.
- Por lo tanto, los datos siguen una distribución normal.

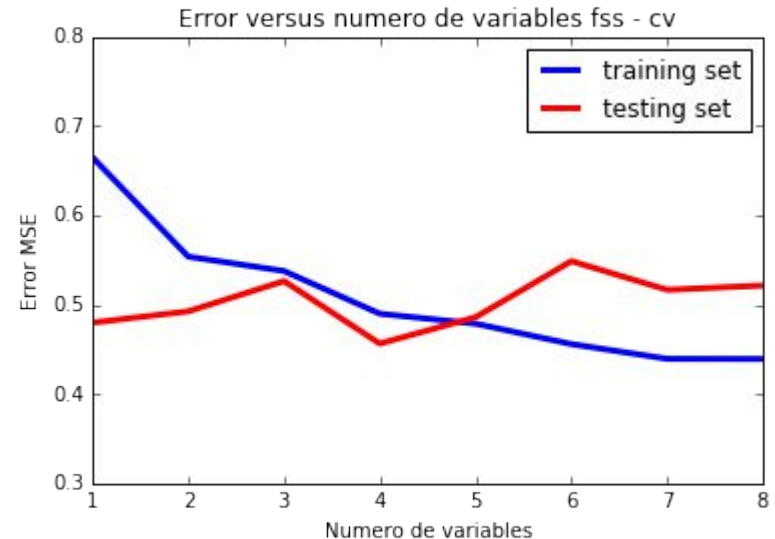
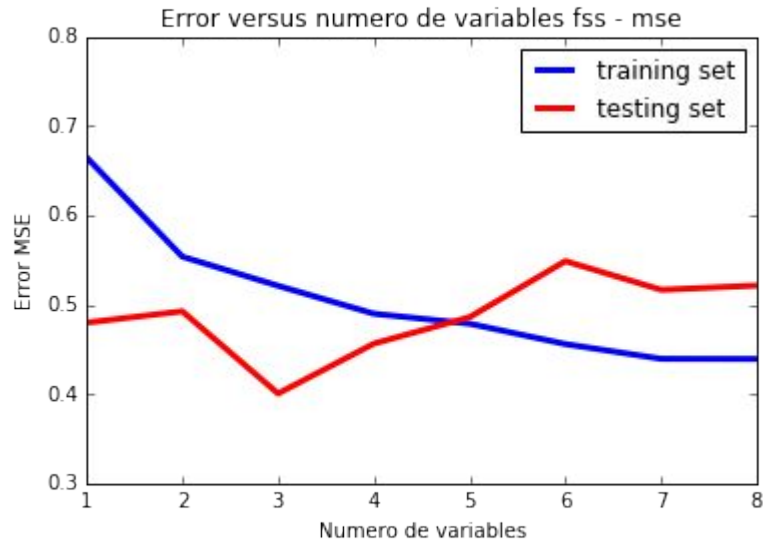
Selección de atributos

Forward Step-wise Selection (FSS)

Criterio de selección de atributos: K-fold cross validation con **K=10**

Orden seleccion mse: Lcavol, Lweight, Svi, Lbph, Pgg45, Lcp, Age, Gleason

Orden seleccion cv: Lcavol, Lweight, Lbph, Svi, Pgg45, Lcp, Age, Gleason



Interpretacion FSS

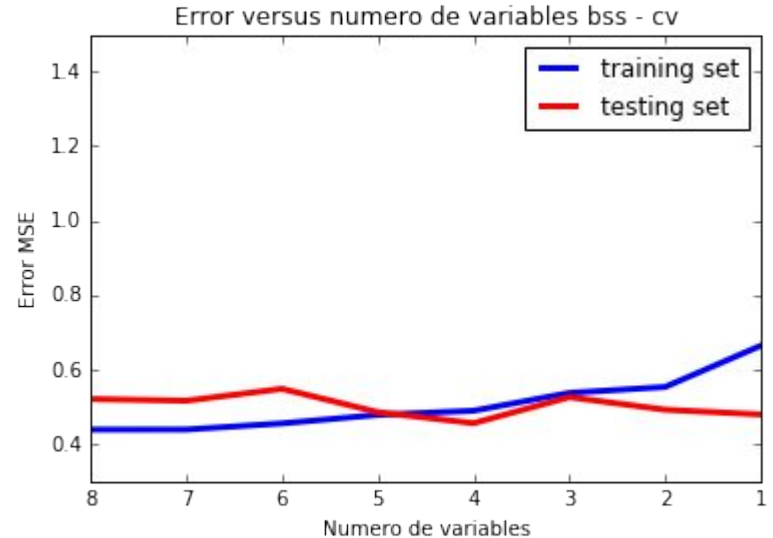
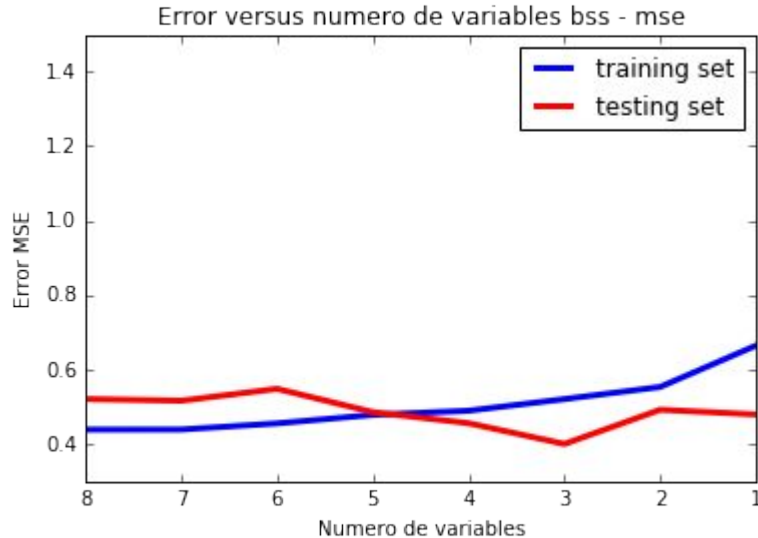
- A priori el grafico MSE parece indicar que con 3 variables: Lcavol, Lweight y Svi el error es mínimo.
- Cross Validation produce que se intercambie el orden de selección de la tercera y cuarta variable.
- Curva de testing set se aproxima mas a curva de traning set en Cross Validation.
- Mínimo error ocurre en 4 variables, luego comienza a aumentar produciendo overfitting debido a la complejidad del modelo.

Backward Step-wise Selection (BSS).

Criterio de selección de atributos: K-fold cross validation con **K=10**

Orden seleccion mse: Gleason, Age, Lcp, Pgg45, Lbph, Svi, Lweight, Lcavol

Orden seleccion cv: Gleason, Age, Lcp, Pgg45, Svi, Lbph, Lweight, Lcavol

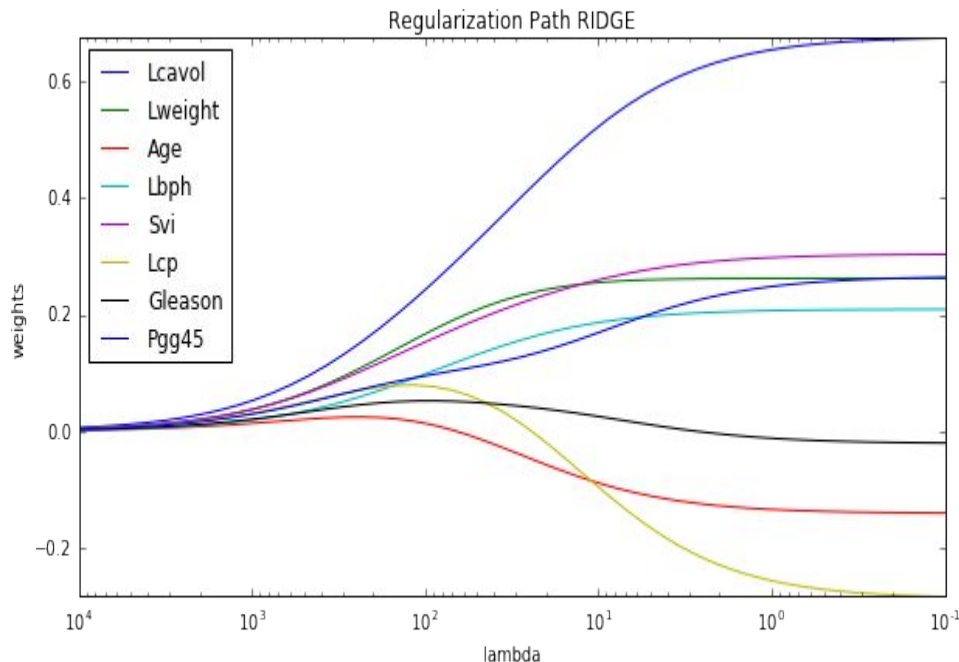


Interpretacion BSS

- Se obtiene el mismo resultado que para FSS respecto a las variables en el modelo, tanto para MSE como para Cross Validation respectivamente.
- Errores son los mismos.
- BSS trabaja con todas las variables desde un comienzo, por lo que puede ser computacionalmente más caro.
- Se utiliza el mismo criterio de selección de variables, el resultado es invariante. La selección es: L_{cavol} , L_{weight} , S_{vi} , L_{bph} .

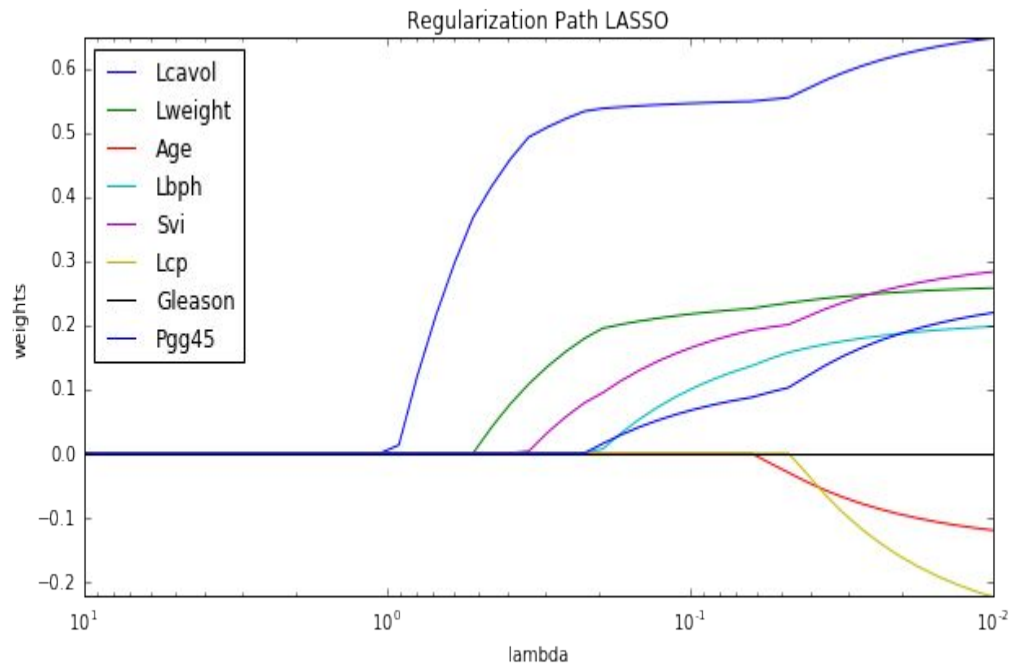
Regularización

Ridge Regression



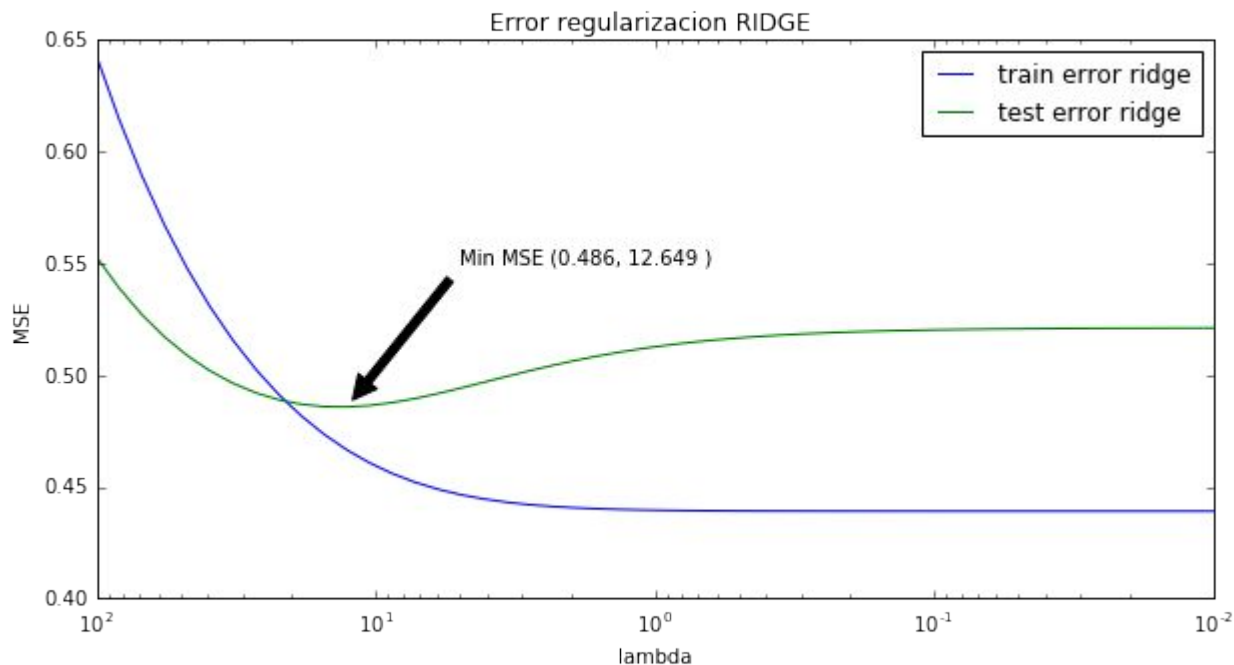
- ❑ No es bueno utilizar λ muy grande, no se diferencian pesos.
- ❑ Existe una clara separación, mayor peso en las 5 variables superiores.
- ❑ Variables Lcavol, Svi y Lweight tienen mayor preponderancia por sobre las demás.
- ❑ λ adecuado menor a 100 para lograr diferenciación.
- ❑ Ridge no elimina variables del modelo.

Lasso Regression



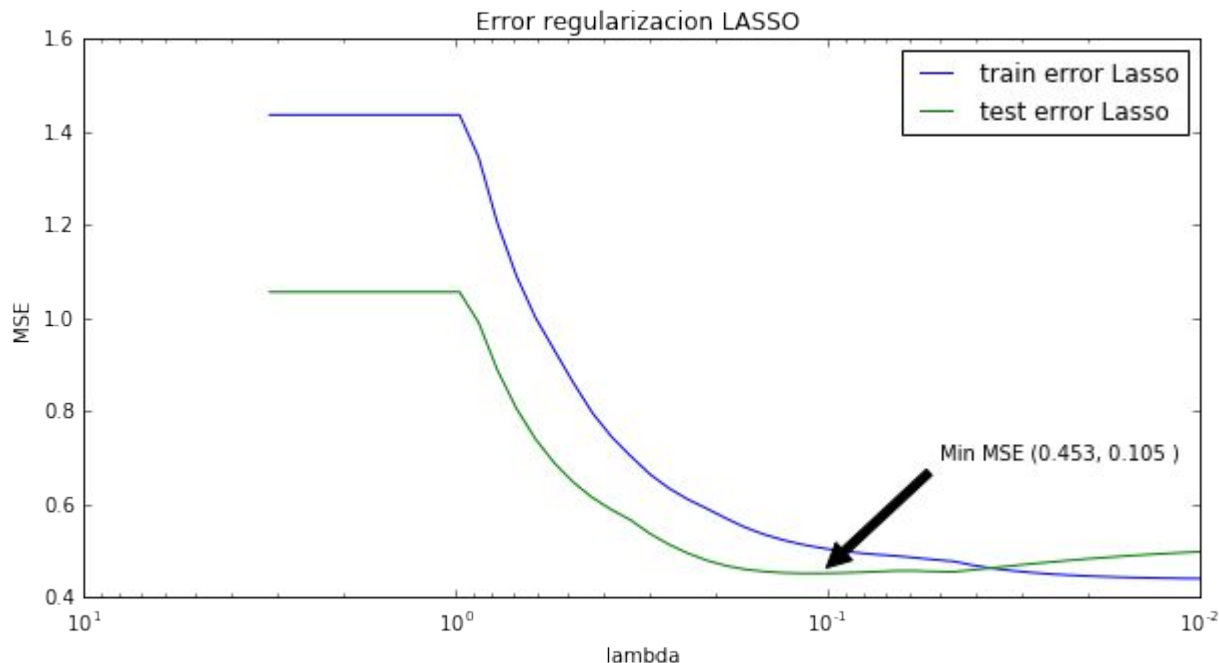
- ❑ Parámetro λ debe ser menor a 1, para que Lasso tenga efecto.
- ❑ Lasso si es capaz de eliminar variables del modelo al establecer sus coeficientes como cero.
- ❑ Coeficiente de variable Gleason siempre es cero.
- ❑ Coeficientes negativos indican que la variable tiene menor relevancia.
- ❑ Lasso y Ridge muestran la misma información en este caso. Lasso es mas facil de interpretar.

Error de entrenamiento y de prueba en Ridge



- ❑ Luego de intersección de curvas se produce sobreajuste
- ❑ λ Que produce mínimo MSE está en la zona de sobreajuste.
- ❑ Seleccionar un valor del parámetro de regularización mayor al punto de corte de las curvas.

Error de entrenamiento y de prueba en Lasso



- ❑ Lasso generaliza bastante bien para ciertos λ .
- ❑ En este caso, se puede utilizar el parámetro λ para el mínimo MSE, ya que ahí no ocurre overfitting.
- ❑ No deberían utilizarse parametros de regularización luego del punto de corte de las curvas

Estimación del parámetro de regularización

	Ridge	Lasso
lambda	2.3	0.01
MSE	0.752	0.759

- ❑ Resultados dependen de la partición utilizada en datos de entrenamiento y prueba.
- ❑ Sobreajuste debido a tamaño de datos de entrenamiento pequeño y numero de parametros grande.

Predicción de Utilidades de Películas

Construcción de modelo

- ❑ Matriz dispersa: Posibilita ahorro de memoria al almacenar solamente valores no nulos de la matriz en memoria.
- ❑ Se utiliza ElasticNet para construir el modelo.
- ❑ El máximo coeficiente de determinación encontrado es: 0.59 con un λ de 3.04 y un α de 0.25
- ❑ De acuerdo al alpha, se puede decir que la penalización de Ridge influye más en los resultados.

