Máquinas de Aprendizaje
Tarea 1

Felipe Araya Barrera
Sebastián Vergara Miranda

Regresión Lineal Ordinaria (LSS)

Descripción conjunto de datos

El dataset se compone de 97 registros (pacientes), cada uno de los cuáles está descrito por 9 variables. Tales variables son:

- **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**: Indica si existe invasión de la vesícula seminal o no.
- **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.
- **lpsa**: Logaritmo del nivel de antígeno prostático específico (PSA).

Construcción del modelo de regresión

- Primero, es necesario normalizar los datos, ya que las variables tienen unidades de medida y escalas diferentes.

- Normalizar permite realizar comparaciones válidas entre variables.

- 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.

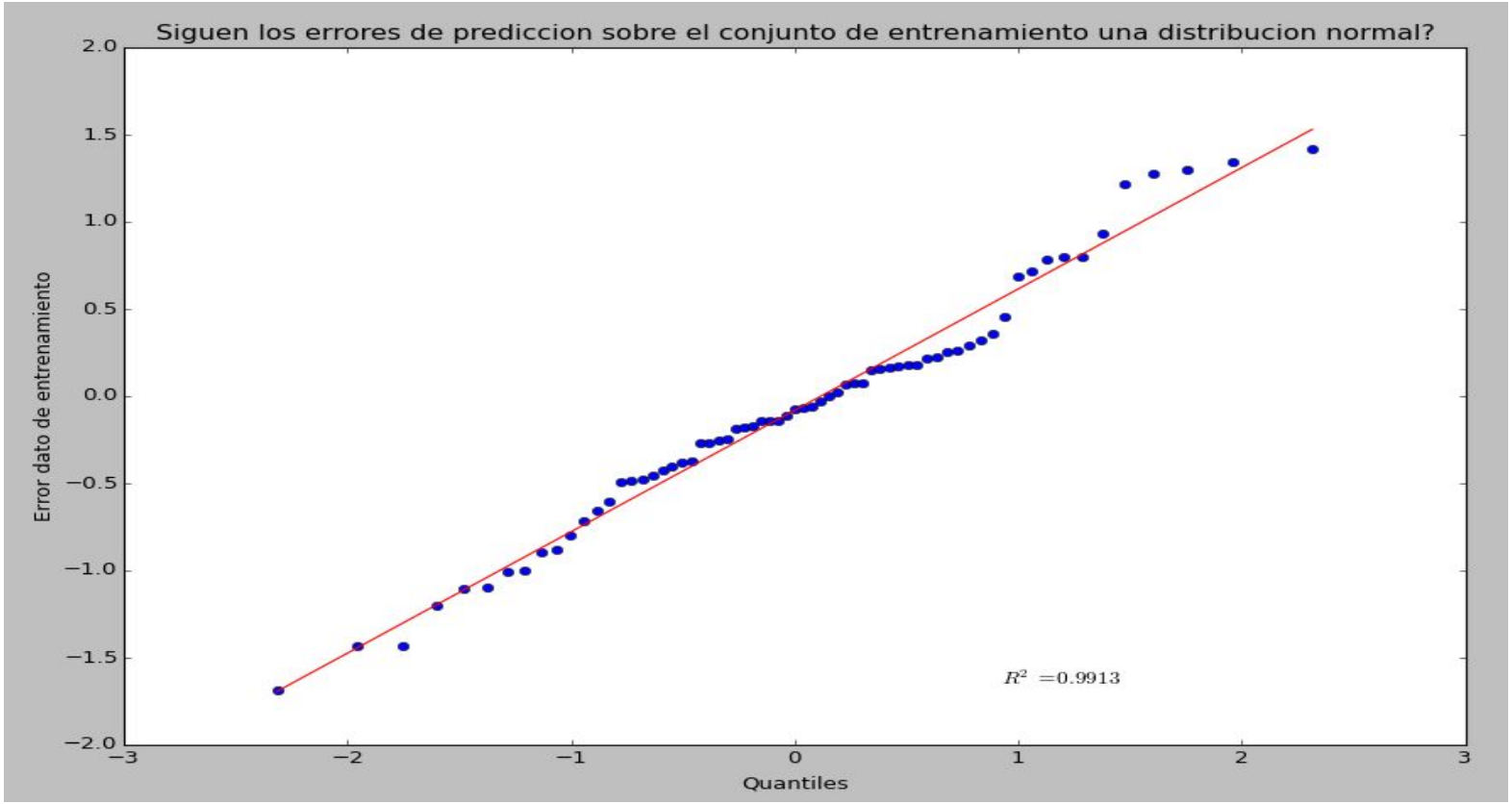
Pesos y Z-score

Variable	Peso	Error Estándar (SEM)	Z-score
lcavol	0.68	0.13	5.22
lweight	0.26	0.14	1.92
age	-0.14	0.12	-1.14
lbph	0.21	0.12	1.69
svi	0.30	0.12	2.44
lcp	-0.29	0.12	-2.33
gleason	-0.02	0.12	-0.18
pgg45	0.27	0.13	2.08

Estimación del error de predicción

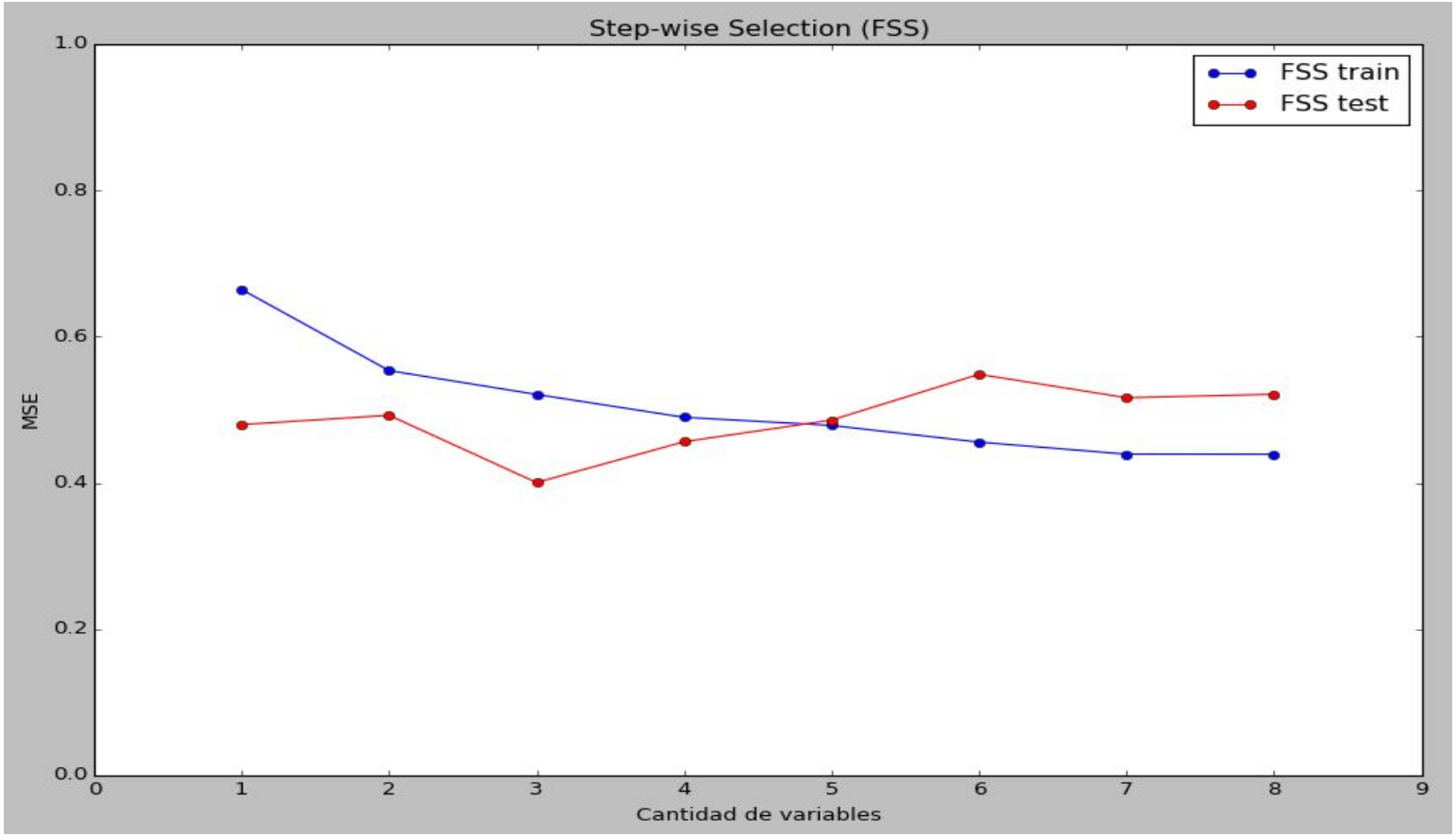
	LSS	K = 5	K = 10
MSE	0.52	0.96	0.76

Errores de predicción: ¿Distribución normal?

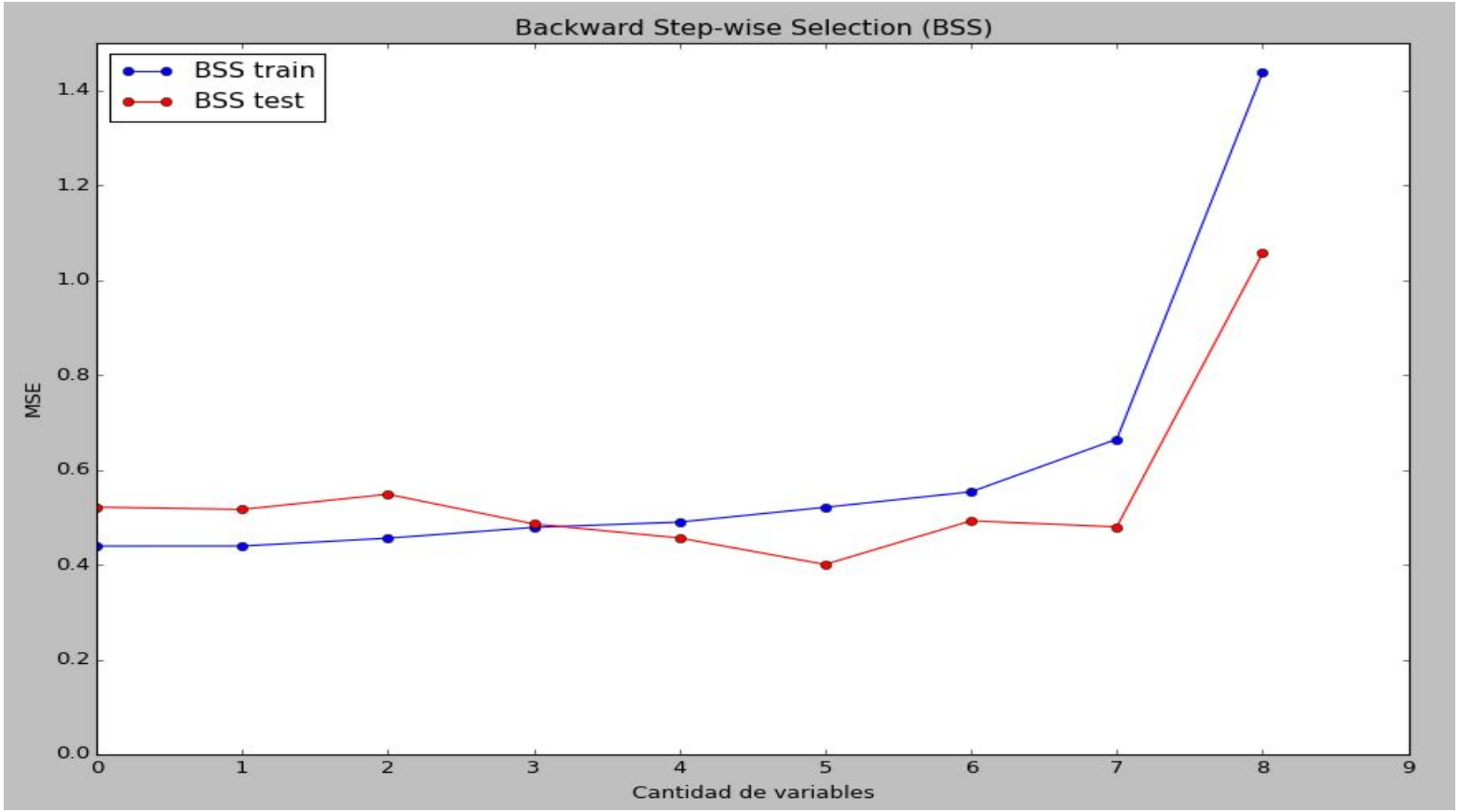


Selección de atributos

FSS

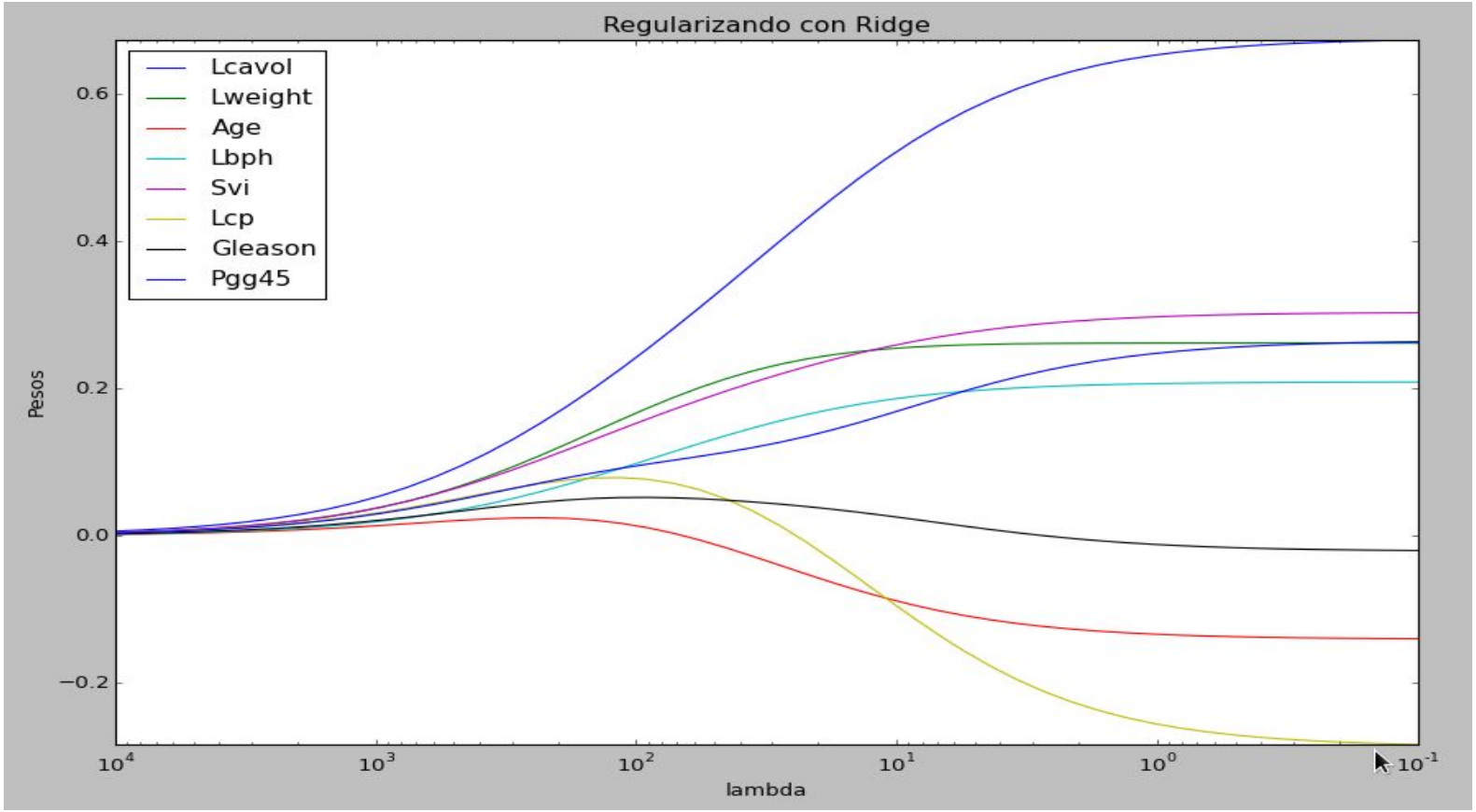

Step-wise Selection (FSS)
FSS train
FSS test
MSE
Cantidad de variables
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9

BSS

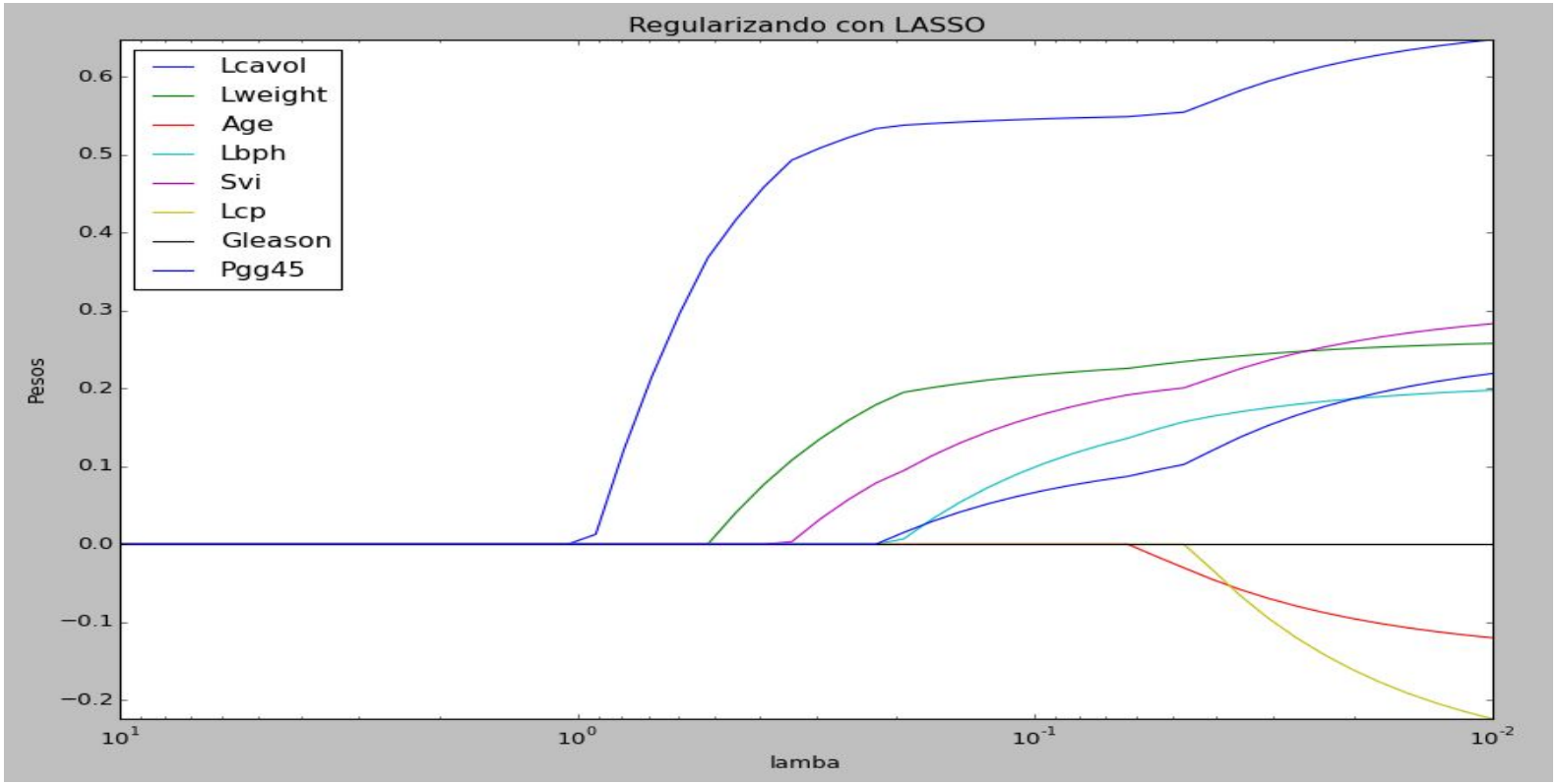

Backward Step-wise Selection (BSS)
BSS train
BSS test
MSE
Cantidad de variables
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5
6
7
8
9

Regularización

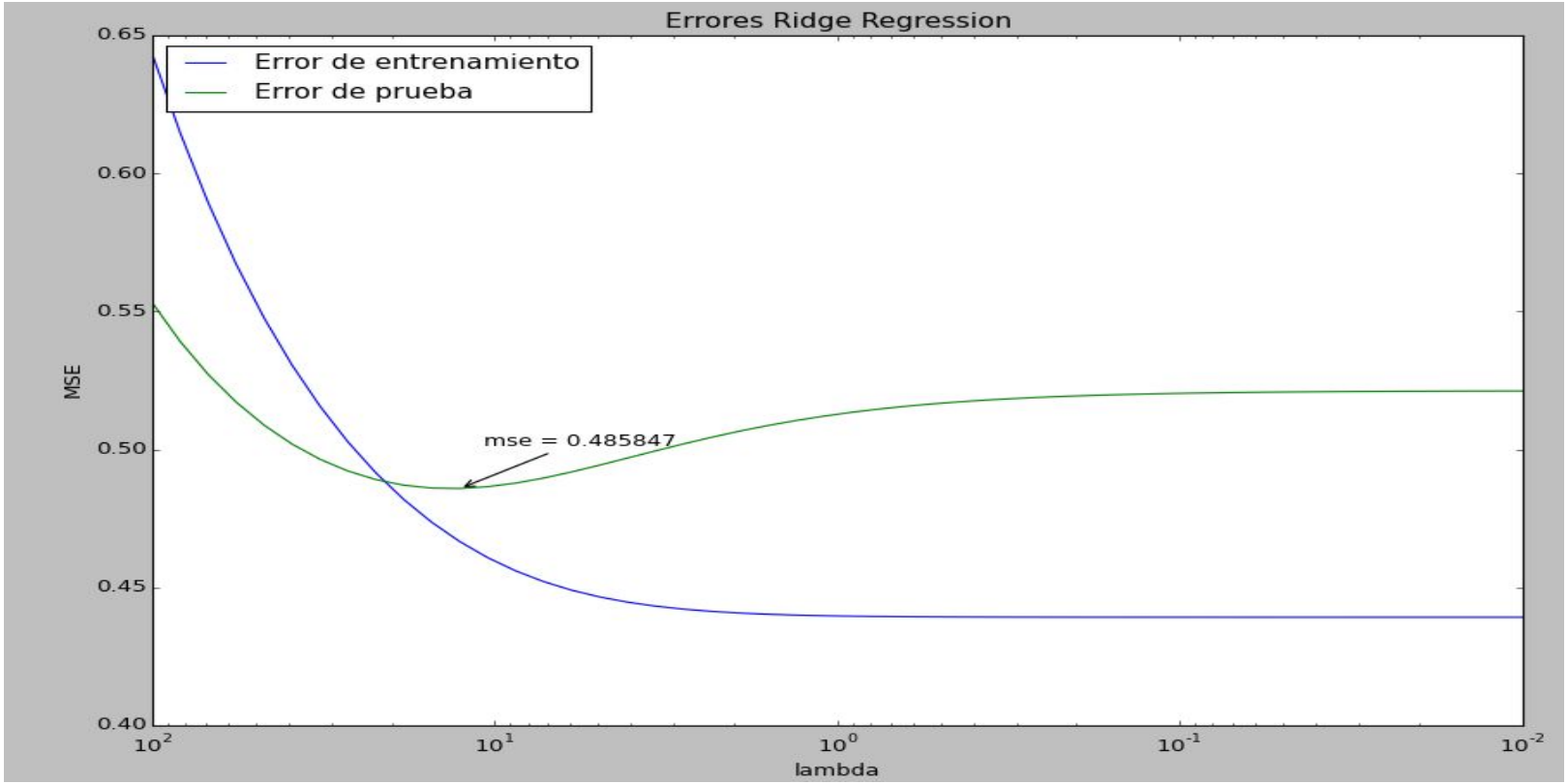
Ridge Regression: pesos v/s lambda



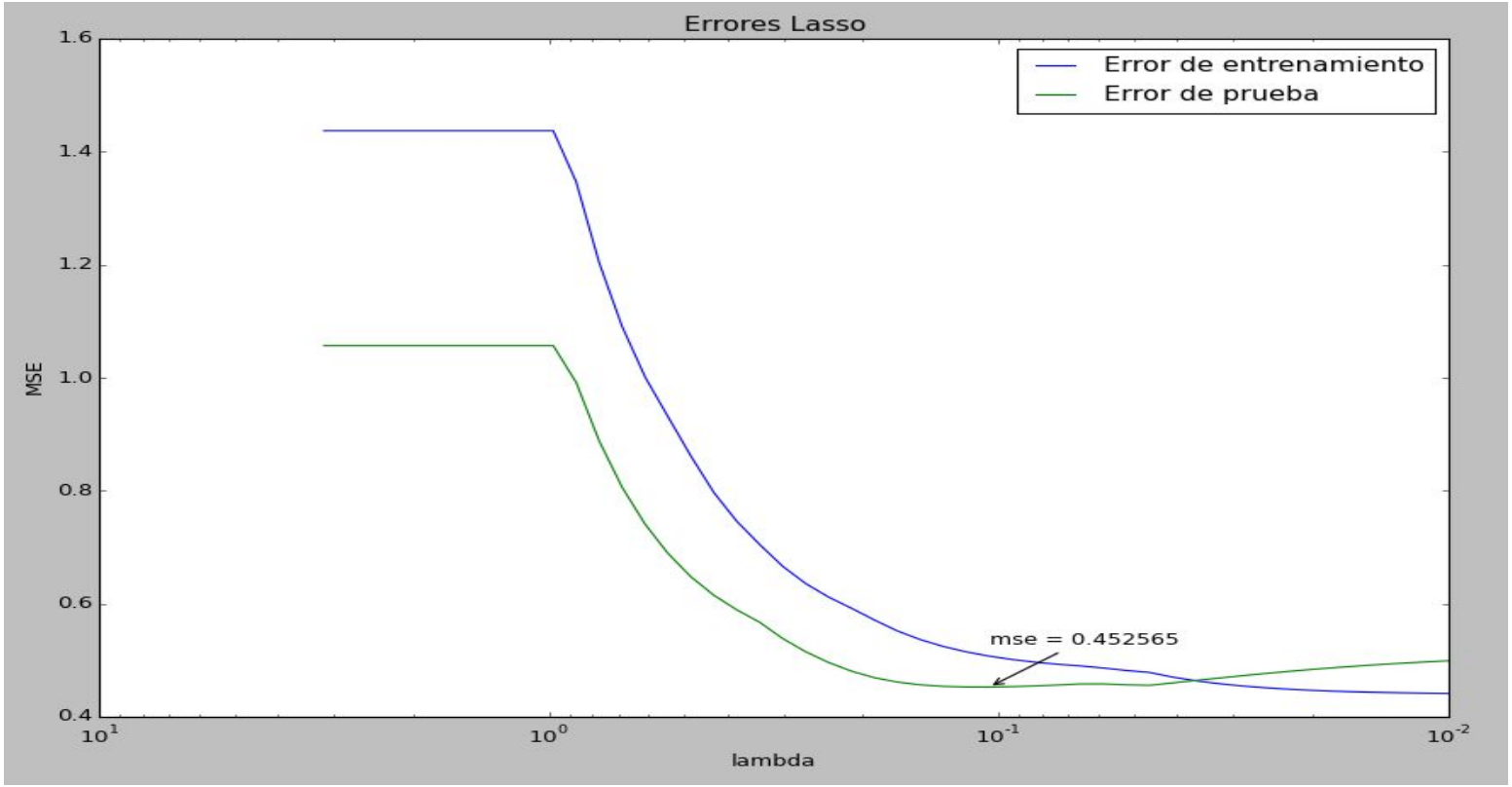
Lasso: pesos v/s lambda



Ridge Regression: Comparación de errores



Lasso: Comparación de errores



Estimación de parámetros de regresión

	Ridge	Lasso
lambda	2.3	0.01
MSE	0.752	0.759

Predicción de utilidades de películas

Construcción del modelo

- Matriz dispersa: Posibilita ahorro de memoria, pues sólo se almacenan valores no nulos de la matriz en ella.
- En base a resultados anteriores, se decidió usar Lasso para construir el modelo.
- Sin embargo, el máximo valor obtenido para coeficiente de determinación es 0.58, con $\lambda = 10^6$. Luego, todos los coeficientes del modelo son cero, por lo que modelo no es útil.
- Resultados análogos con Ridge Regression.
- Se prueba con LS: Se obtiene coeficiente 0.59.