MACHINE LEARNING
TAREA 1

ESTEBAN HERNÁNDEZ
CRISTIAN YAÑEZ

REGRESIÓN LINEAL ORDINARIA

Objetivo

- **Predecir el nivel de antígeno prostático específico (PSA)**

Datos

- lcavol: Logaritmo del volumen de cáncer presente
- lweight: Logaritmo del peso de la próstata
- age: Edad
- 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). **(Target)**

REGRESIÓN LINEAL ORDINARIA

- Se separan los datos en un conjunto de **entrenamiento** y un conjunto de **pruebas**
- Se **normalizan** los datos
  - Estandariza las unidades y escalas.
  - Permite eliminan los efectos de la media y la varianza de cada variable
  - Esto solo se aplica a los predictores.

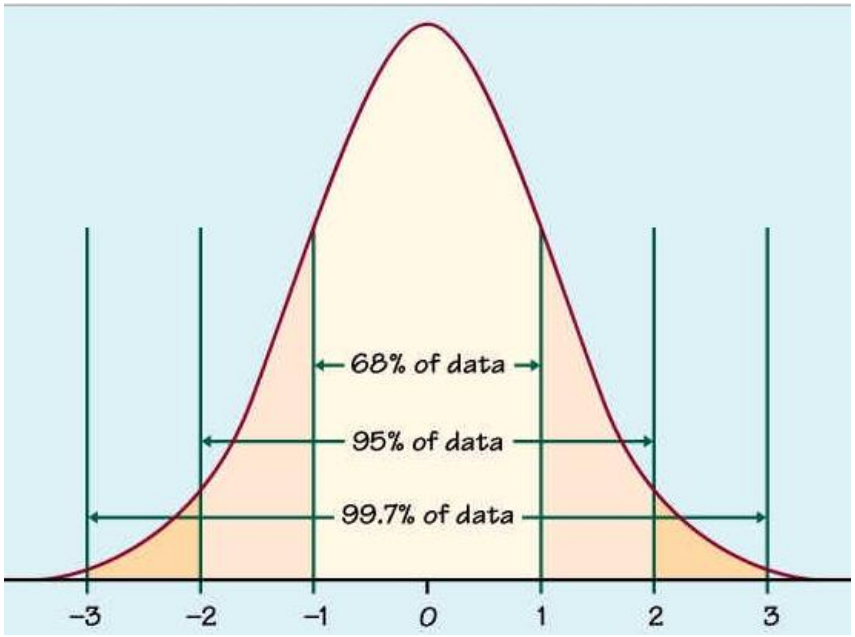
REGRESIÓN LINEAL ORDINARIA

Z-Score:

- los predictores con mayor correlación de predicción son **lcavol, lweight y svi.**

Predictor	Peso	Z-score
lcavol	0.5966	5.5912
lweight	0.2723	3.3793
age	-0.1456	-1.7054
lbph	0.1893	2.1973
svi	0.1794	1.7224
lcp	-0.1591	-1.2219
gleason	0.1008	0.8234
pgg45	0.1149	0.8875
intercept	2.4001	31.7233

- Con una significancia del 5% -> [-2,2]



REGRESIÓN LINEAL ORDINARIA

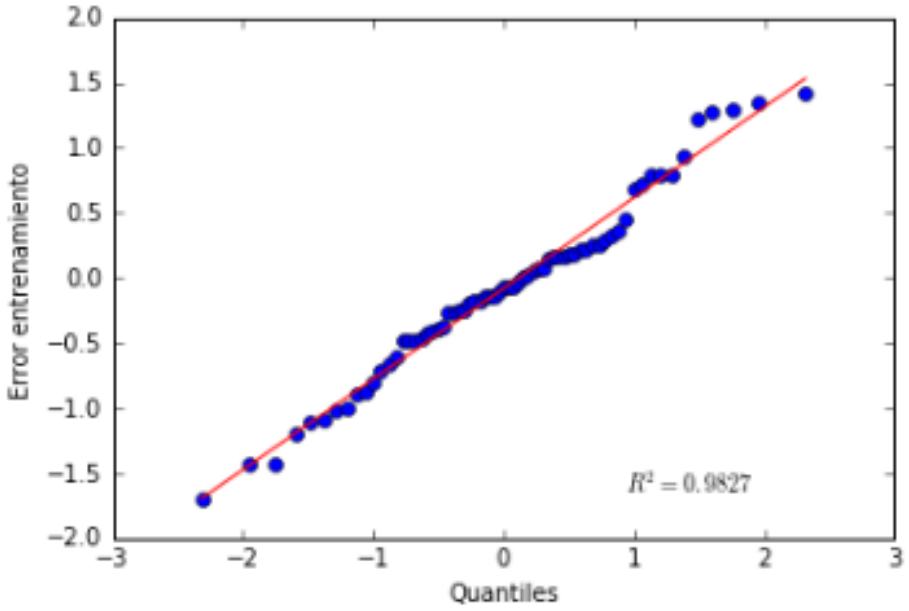
- **Cross validation:**

```
Sin Cross Validation = 0.5096
Con Cross validation
k = 5    MSE = 0.9565
k = 10   MSE = 0.7572
```

Es posible que el modelo este sobre-ajustado a los datos de entrenamiento.

REGRESIÓN LINEAL ORDINARIA

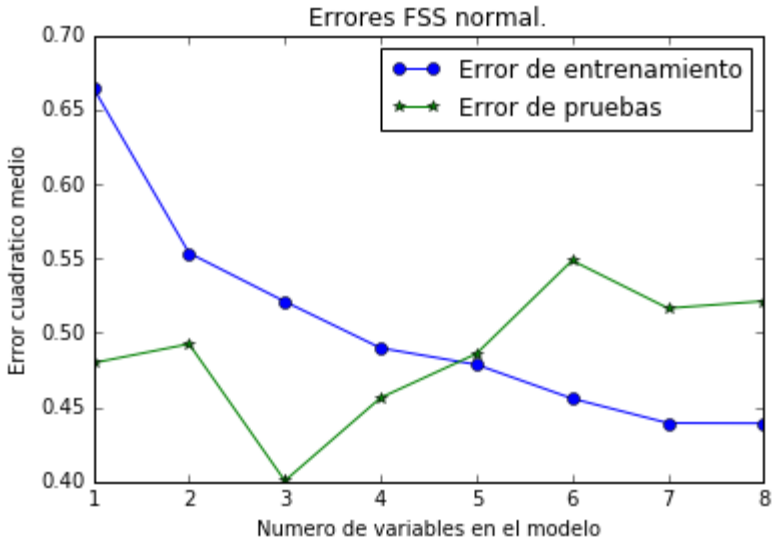
Q – Q Plot:

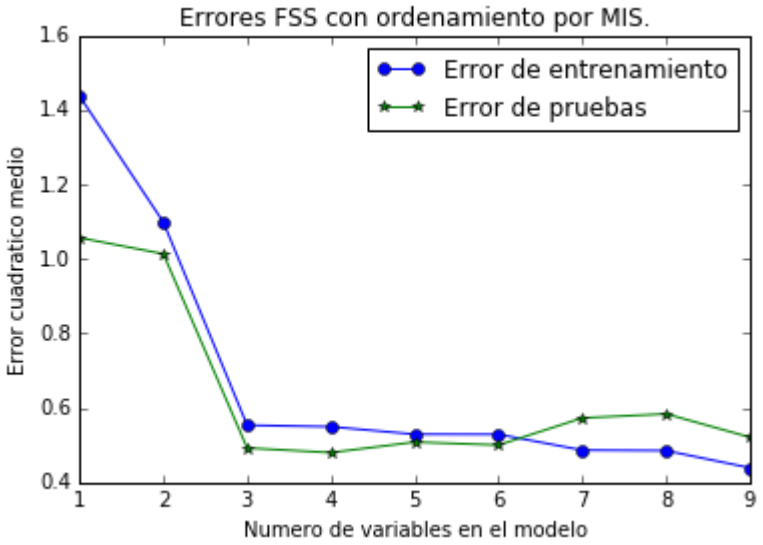


Es correcto señalar que los errores siguen una distribución normal.

SELECCIÓN DE ATRIBUTOS

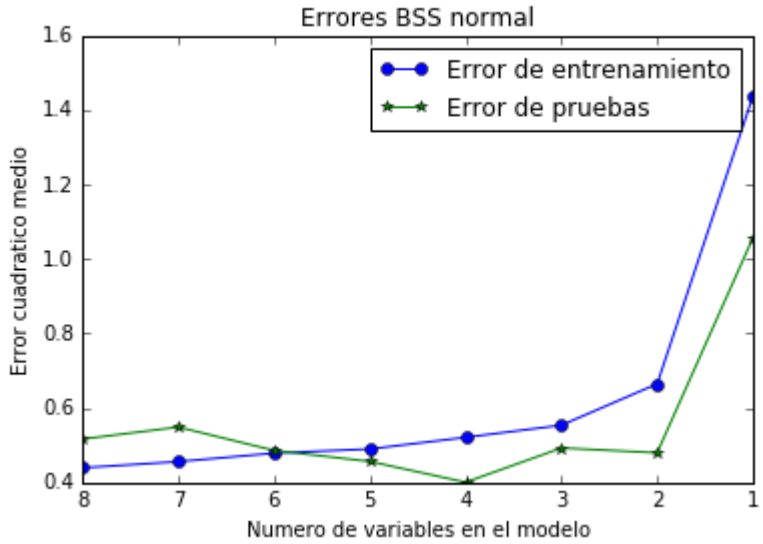
- **FSS (Forward Step-wise Selection) + MIS (Mutual information score)**

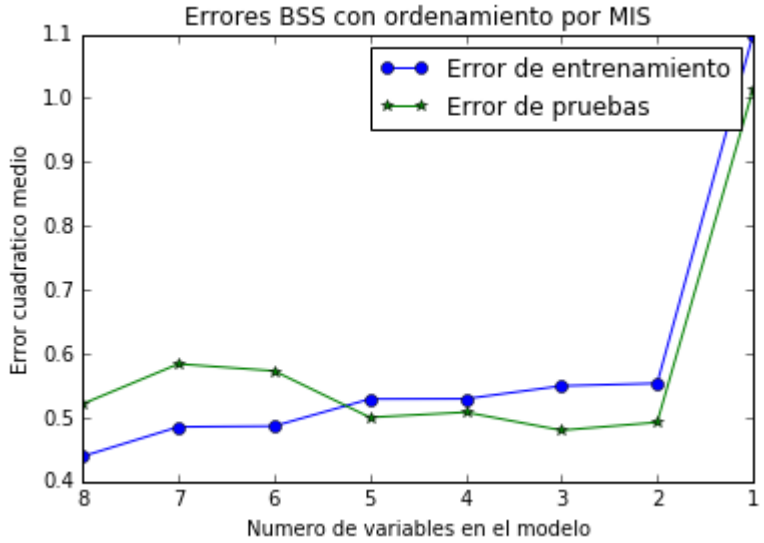




SELECCIÓN DE ATRIBUTOS

BSS (Backward Step-wise Selection) + MIS (Mutual information score)

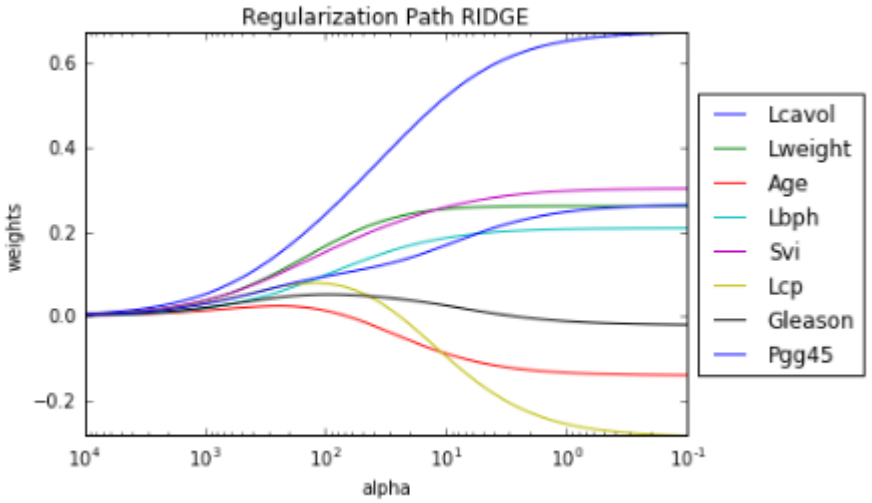




REGULARIZACIÓN

RIDGE Regression

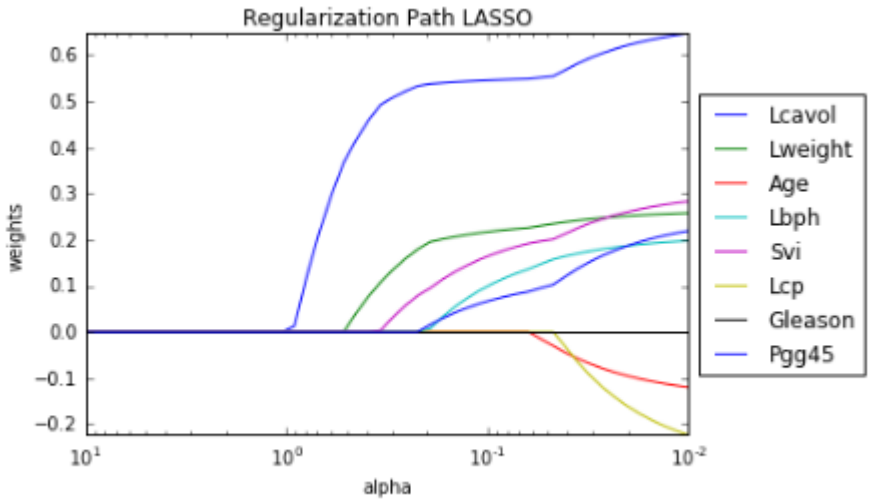
- Para λ muy grandes, es difícil diferenciar los pesos.
- las variables Lcavol, Svi, Lweight y Lbph son las que poseen mayores pesos y de menor varianza



REGULARIZACIÓN

LASSO regresion

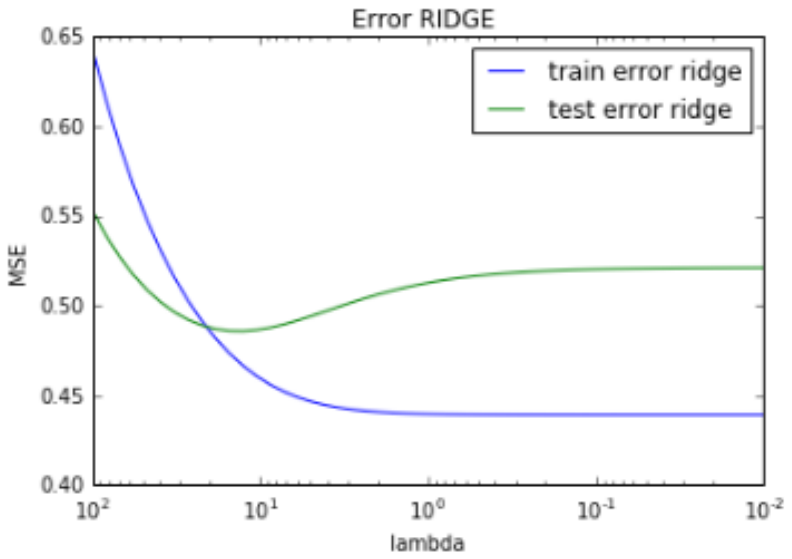
- Lasso funciona solo para valores de λ menores a 1
- La tendencia del modelo permite diferenciar que las variables con un mayor peso son Lcavol, Svi y Lweight
- Se puede percibir cierta ventaja al utilizar la regularización Lasso, ya que esta permite la diferenciación de pesos entre las variables de una forma mas clara



REGULARIZACIÓN

Error RIDGE

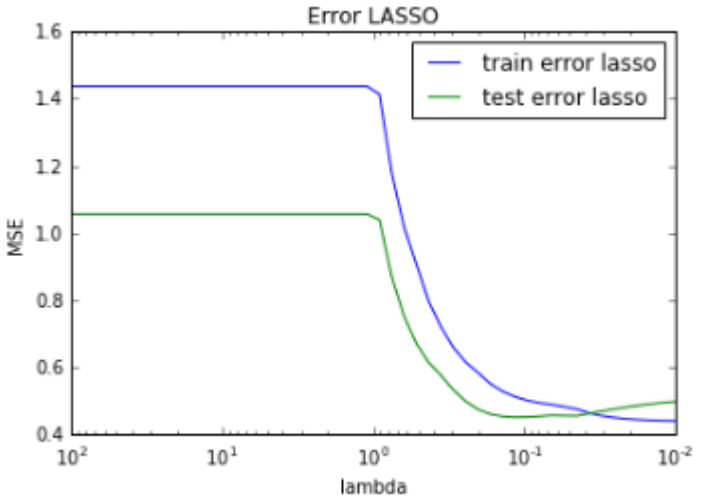
- el menor Error de test se encuentra aproximadamente en λ=20
  - en donde el MSE es aproximadamente 0.49
- Empieza a ocurrir un fenómeno de Overfitting



REGULARIZACIÓN

Error LASSO

- λ mayores a 1 no se puede obtener una buena conclusión
- Error de entrenamiento permanece mayor al de prueba hasta aproximadamente en λ=0.03
- Aproximadamente en λ=0.2 se encuentra el menor error cuadrático medio de prueba, el cual es aproximadamente de 0.45.



REGULARIZACIÓN

Estimación parámetro de regularización

- λ el cual minimiza el MSE:
  - λ Ridge : 2.33 MSE Ridge = 0.752
  - Λ Lasso: 0.01 MSE Lasso = 0.759

PREDICCIÓN DE UTILIDADES DE PELÍCULAS

Objetivo

- **Predecir el volumen de utilidades (en dólares) obtenidas por el estreno (al público, en USA) de una película**

- **Datos**

- **Meta**
  - Origen de la película, presupuesto, puntos de proyección, genero, calificación MPAA, actores con óscar y variable si se estrenó en vacaciones/feriado
- **Texto**
  - A partir de las crıticas publicadas para cada película se construyen características que corresponden a la frecuencia de palabras, parejas de palabras y tríos de palabras obtenidas de un vocabulario

PREDICCIÓN DE UTILIDADES DE PELÍCULAS

Modelo:

- ***ElasticNet*** **(***"Movie Reviews and Revenues: An Experiment in Text Regression"***)**

$$\theta = argmin \frac{1}{2n} \sum_{i=1}^{n} (y_i - (\beta_0 + x_i^T \beta))^2 + \lambda P(\beta)$$

$$P(\beta) = \sum_{j=1}^{p} (\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

Los mejores parametros encontrados fueron:

$\alpha = 0$ y $\lambda = 2.27584592607$

Correlación obtenida : 0.573593