



DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Regresión Polinomial

Trabajo Práctico 1

26 de abril de 2019

Reconocimiento De Patrones

Integrante	LU	Correo electrónico
Ingaramo, Pablo	544/15	pablo2martin@hotmail.com
Cuba, Kennedy	503/15	kennedy.wrc@hotmail.com



Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (++54 +11) 4576-3300

<http://www.exactas.uba.ar>

Índice general

1. Introducción teórica	1
1.1. Regresión polinomial	1
1.1.1. Definiciones	1
1.1.2. Objetivo	2
1.2. Boston Housing Dataset	2
2. Desarrollo	4
2.1. Regresión Polinomial	4
2.2. Boston Housing Dataset	4
3. Experimentación	5
3.1. variación de los grados de Polinomios	5
4. Conclusiones	17
4.1. Referencias	17

Resumen

El objetivo de este trabajo práctico es ver como se puede utilizar la regresión polinomial para estimar un conjunto de datos mediante polinomios.

Para tener una forma de medir y comparar los resultados, se presentan funciones de error y se analiza como afecta al vector que los minimiza.

Luego se presentará una conocida set datos ("Boston Housing Dataset") y la forma de calcular la matriz de correlación, junto con un breve analisis de la feature MEDV del set datos.

Capítulo 1

Introducción teórica

1.1. Regresión polinomial

Queremos predecir valores de la primera en función de valores observados de la segunda. El análisis de regresión consiste en ajustar un modelo a los datos, estimando coeficientes a partir de las observaciones, con el fin de predecir valores de la variable de respuesta a partir de una (regresión simple) o más variables (regresión múltiple) predictivas o explicativas.

El análisis de regresión se puede usar para:

- identificar a las variables predictivas relacionadas con una variable de respuesta
- describir la forma de la relación entre estas variables y para derivar una función matemática óptima que modele esta relación
- predecir la variable de respuesta a partir de la(s) explicativas o predictoras

1.1.1. Definiciones

- M = grado del polinomio interpolante.
- N = cantidad de datos.
- $X_i = (x_i^0, \dots, x_i^M)$.
- $W = (w_0, \dots, w_M)$.
- $Y(X, W) = XW^T$.
- $E_D(W) = (1/2) \sum_{i=1}^N (Y(X_i, W) - t_i)^2$

- $E(W) = (1/2) \sum_{i=1}^N (Y(X_i, W) - t_i)^2 + (\lambda/2) \|W\|_2$
- $E_{RMS}(W) = \sqrt{2E(W)/N}$.
- $E_{DRMS}(W) = \sqrt{2E_D(W)/N}$.

Dado que la función $E(W)$ es convexa en w , está posee un minimo y esté es único.

1.1.2. Objetivo

Buscar cual es el W que minimice los errores de $Y(X_i, W) - t_i$.

1.2. Boston Housing Dataset

The dataset contains a total of 506 cases.

Hay 14 atributos en cada caso del conjunto de datos.

- CRIM - tasa de criminalidad per cápita por ciudad
- ZN - proporción de tierra residencial dividida en zonas para lotes de más de 25,000 pies cuadrados.
- INDUS - proporción de acres de negocios no comerciales por ciudad.
- CHAS - Variable ficticia del río Charles (1 si el tramo limita con el río; 0 si no)
- NOX - concentración de óxidos nítricos (partes por 10 millones)
- RM - número medio de habitaciones por vivienda
- AGE - proporción de unidades ocupadas por sus propietarios construidas antes de 1940
- DIS - distancias ponderadas a cinco centros de empleo de Boston
- RAD - índice de accesibilidad a carreteras radiales
- TAX - tasa del impuesto sobre el valor total de la propiedad por cada 10,000
- PTRATIO - número de alumnos por docente y por ciudad
- B - $1000(B_k - 0.63)^2$ donde B_k es la proporción de negros por ciudad

- LSTAT - menor condición de la población
- MEDV - Valor medio de las viviendas ocupadas por sus propietarios en miles de dólares.

Capítulo 2

Desarrollo

El desarrollado de este trabajo pretende hacer un análisis de como afecta la concideración de diferentes funciones de error al vector solución y la variación de la media y varianza muestrales de los errores.

2.1. Regresión Polinomial

Se implementan funciones de error.

Se realiza un analisis de el valor medio y la desviación standard muestrales para cada grado de polinomio considerado.

Graficamos la influencia del grado del polinomio sobre el E_{RMS}

2.2. Boston Housing Dataset

Lo que se busca es mostrar las relaciones entre las features de este set de datos.

Para ellos graficaremos los scatter-plot de todos los features y la matriz de correlación.

Luego para aquellas features que presenten un coeficiente de correlación con la feature MEDV mayor a 0.5 en valor absoluto se ajustan con polinomios de grado 1, 2 y 3, se grafican los scatter-plot correspondientes y se calculan los E_{RMS} .

Capítulo 3

Experimentación

3.1. variación de los grados de Polinomios

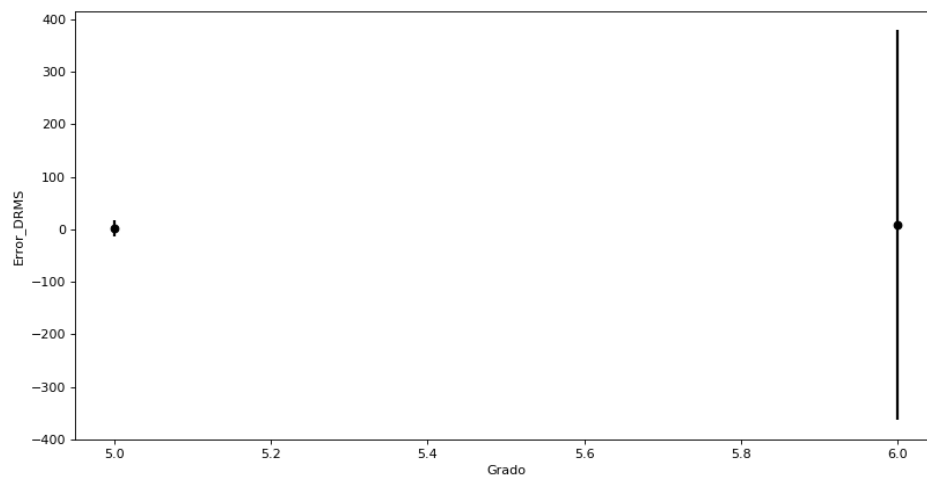


Figura 3.1:

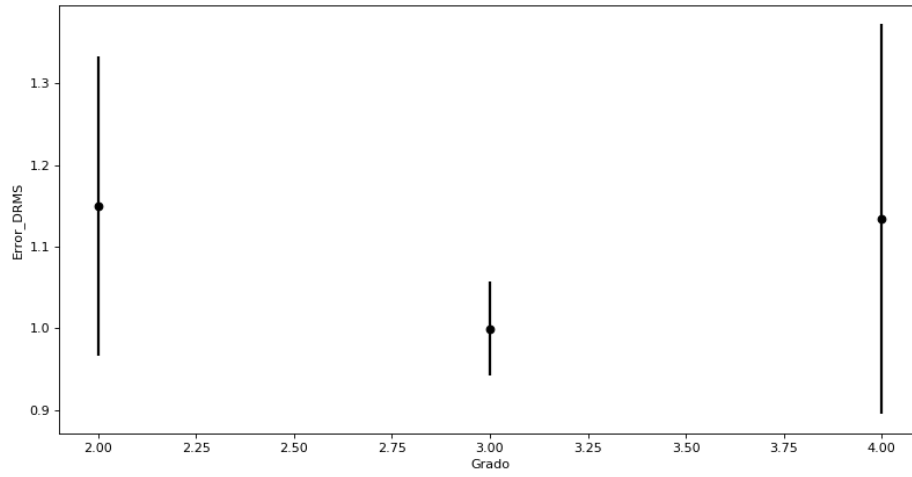


Figura 3.2:

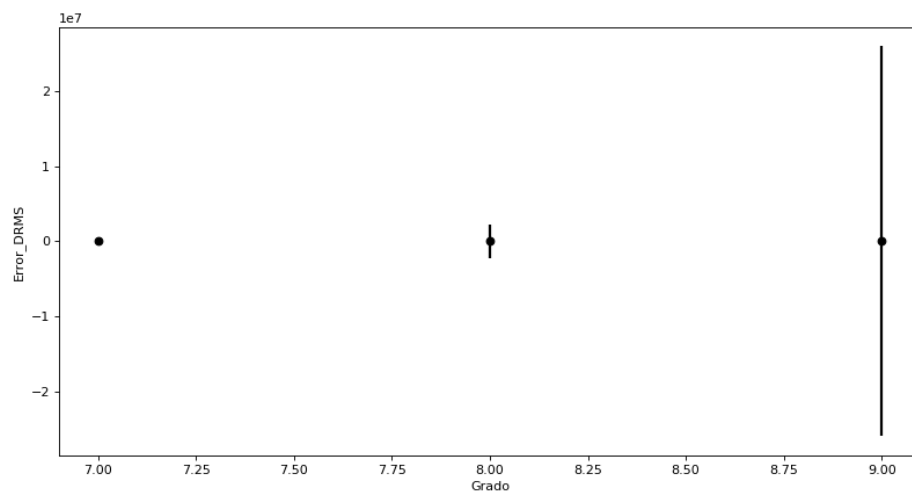


Figura 3.3:

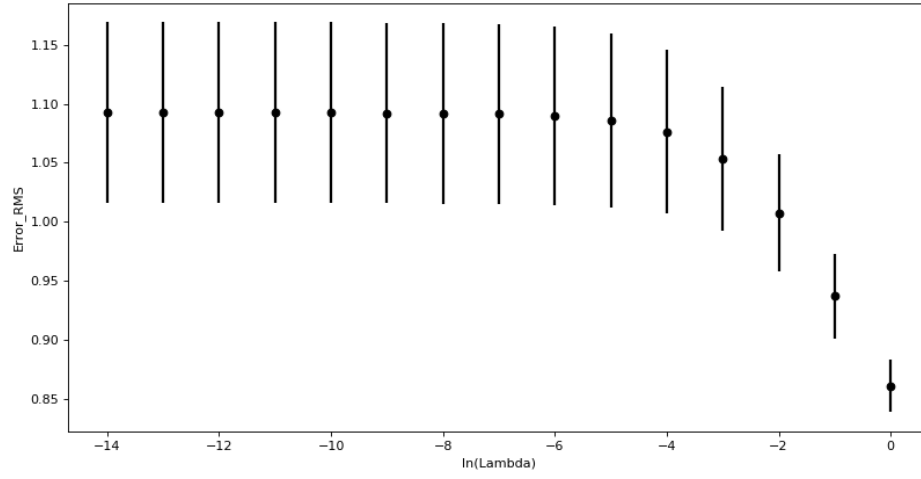


Figura 3.4:

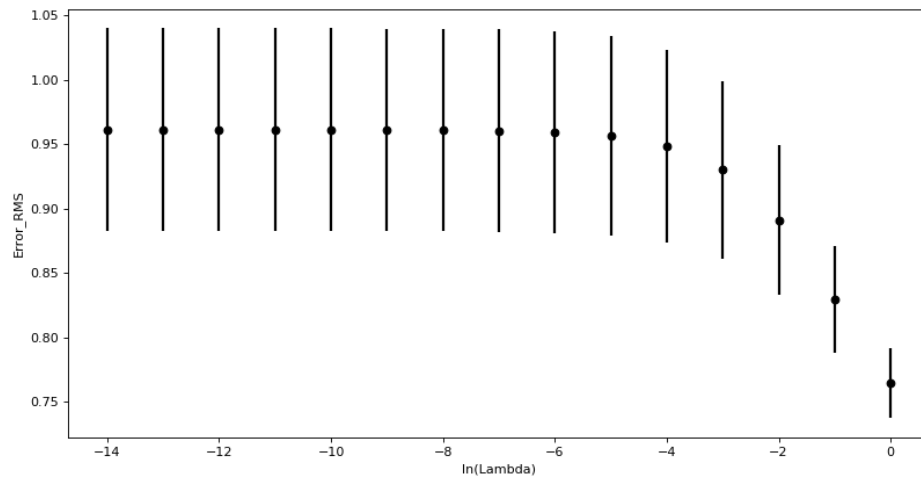


Figura 3.5:

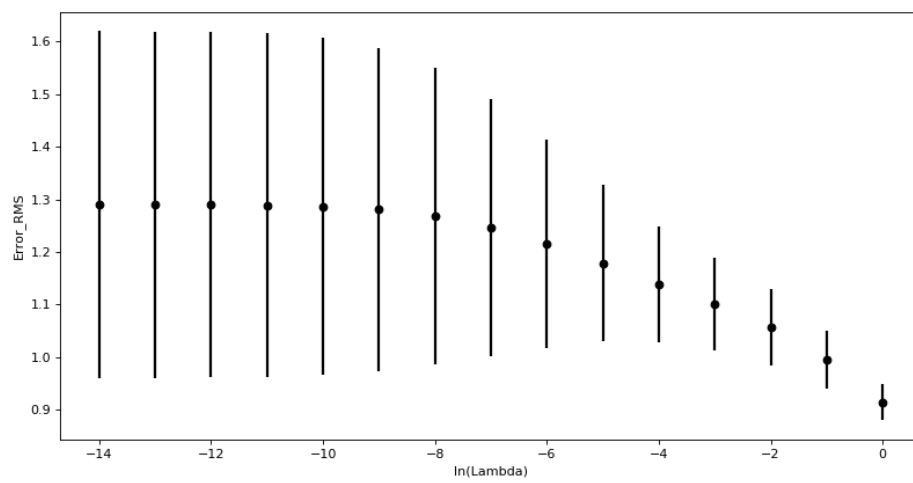


Figura 3.6:

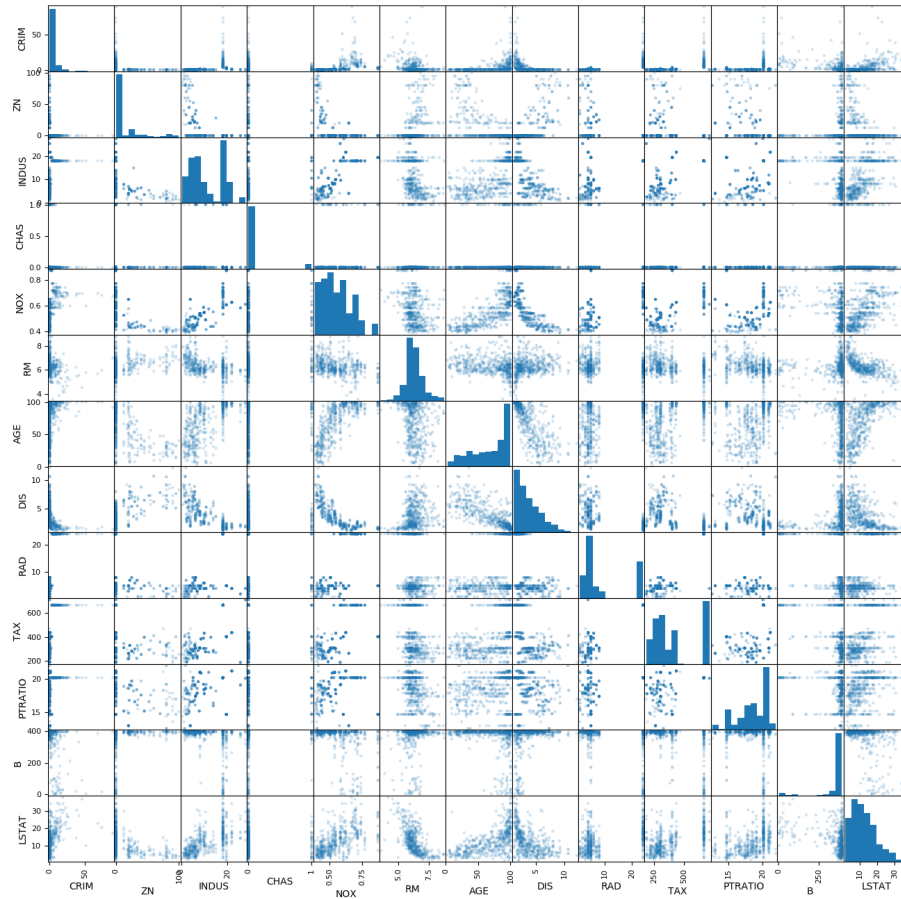


Figura 3.7: Scatter-plots de Boston Housing Dataset

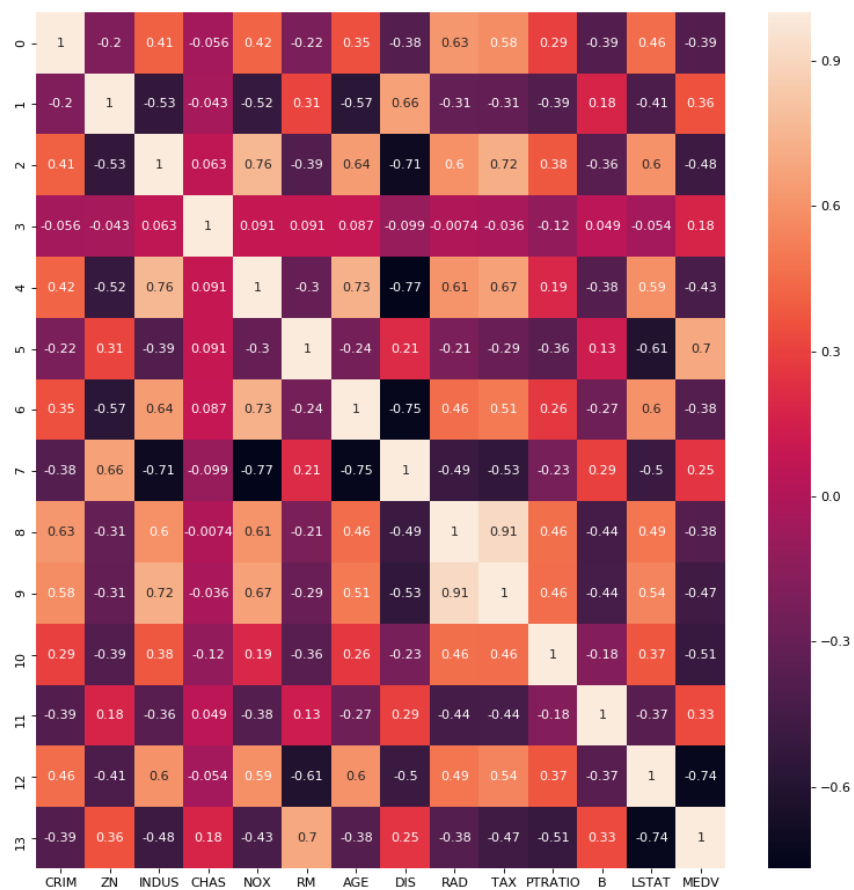


Figura 3.8: Matriz de correlación de Boston Housing Dataset

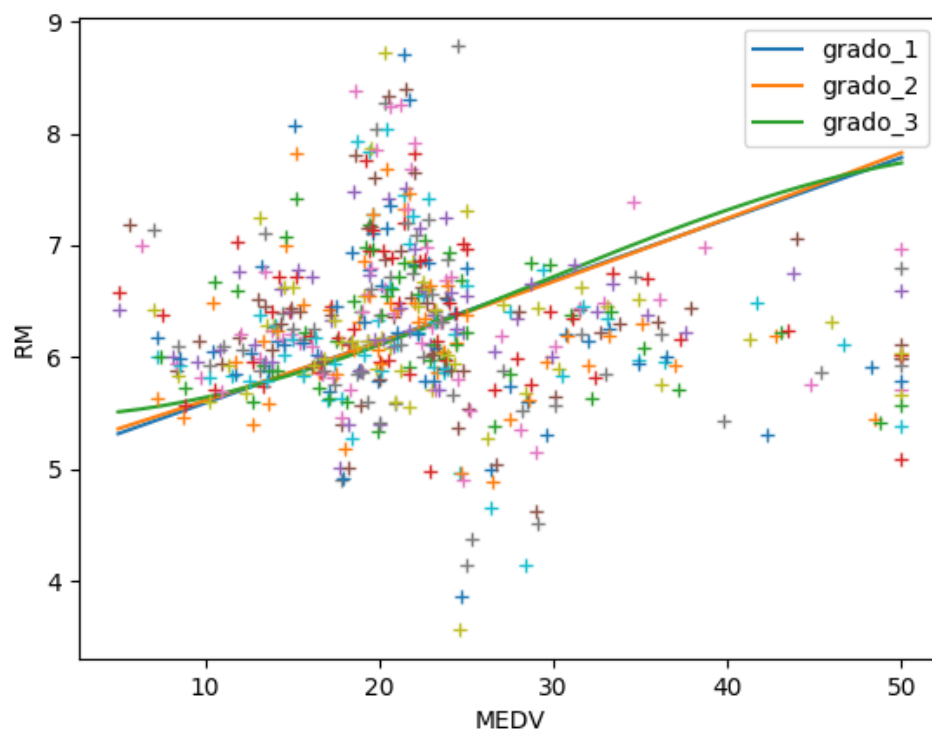


Figura 3.9: Estimación de la distribución de puntos entre MEDV y RM

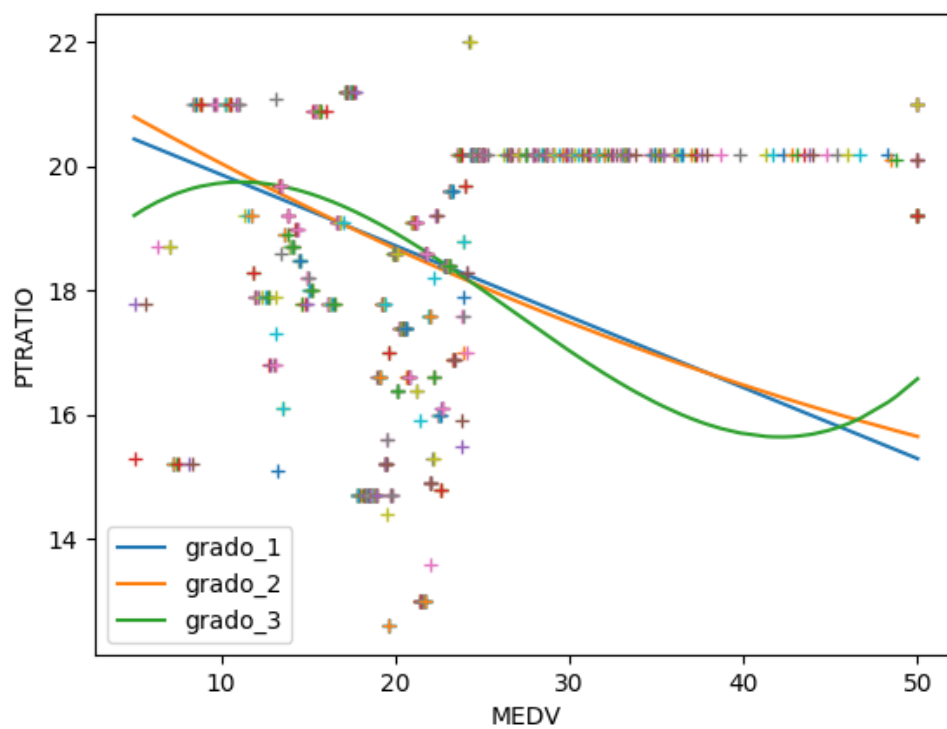


Figura 3.10: Estimación de la distribución de puntos entre MEDV y PTRATIO

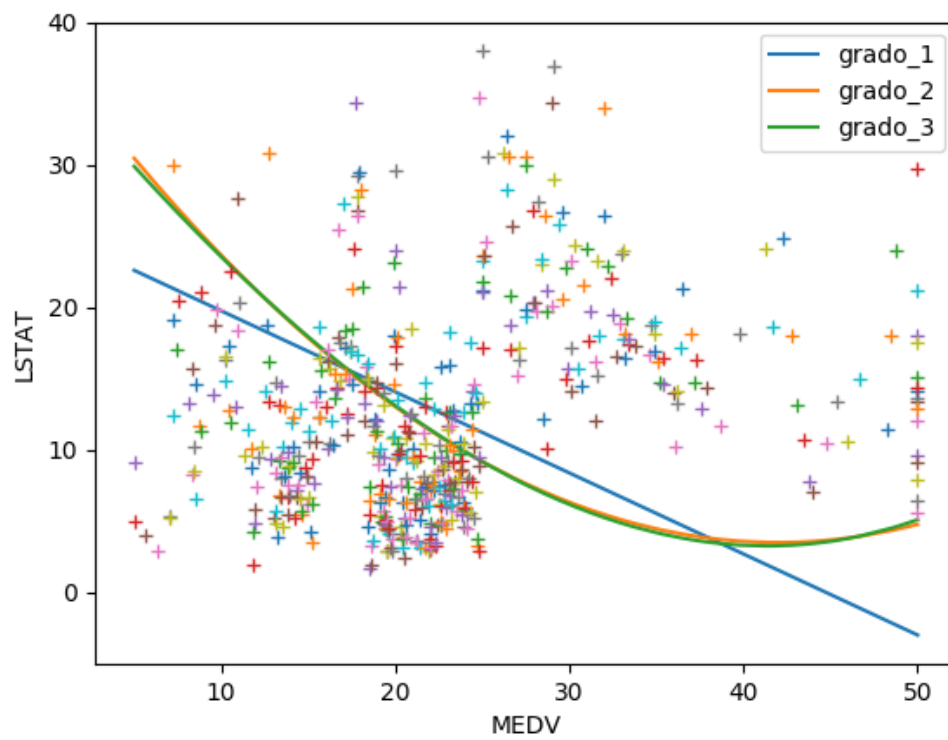


Figura 3.11: Estimación de la distribución de puntos entre MEDV y LSTAT

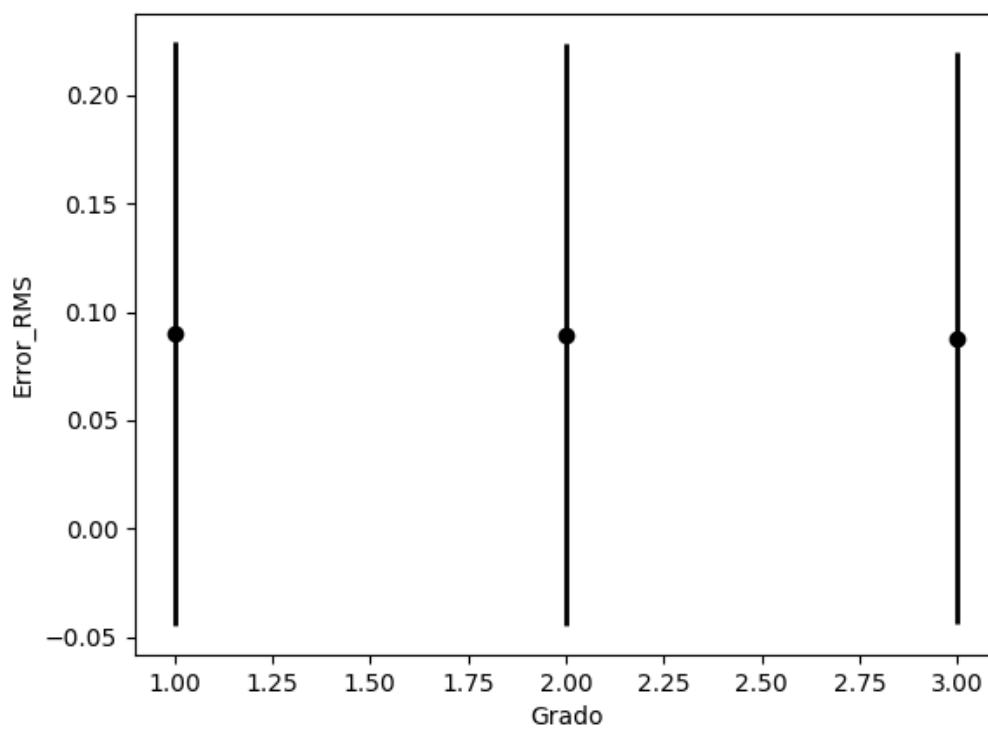


Figura 3.12: Media y Varianza al estimar la relación entre MEDV y RM

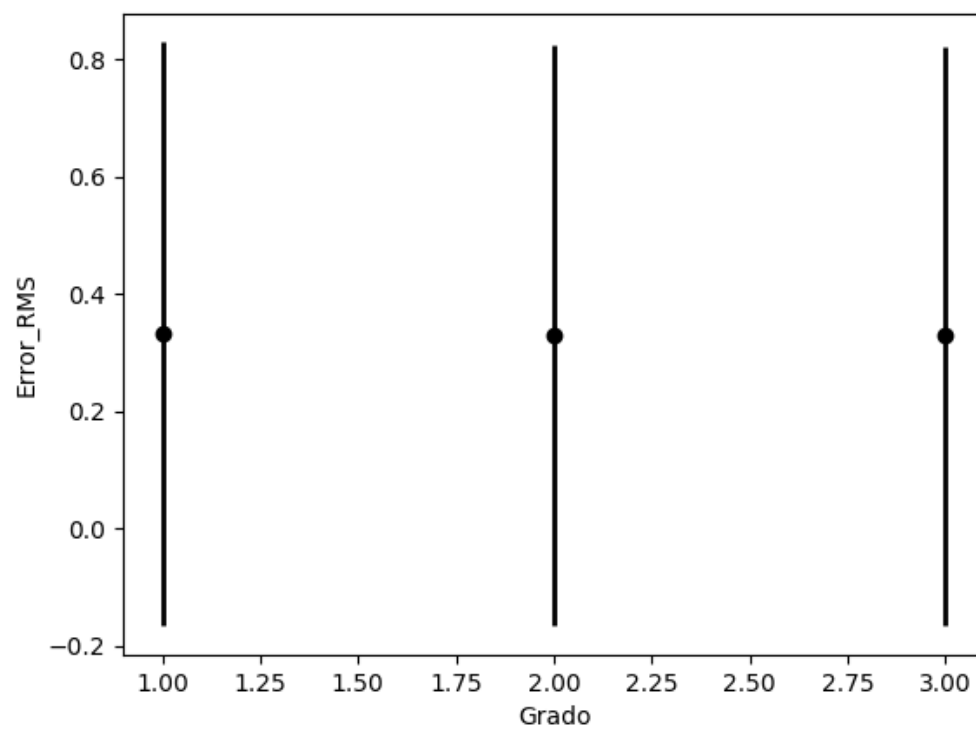


Figura 3.13: Media y Varianza al estimar la relación entre MEDV y PTRATIO

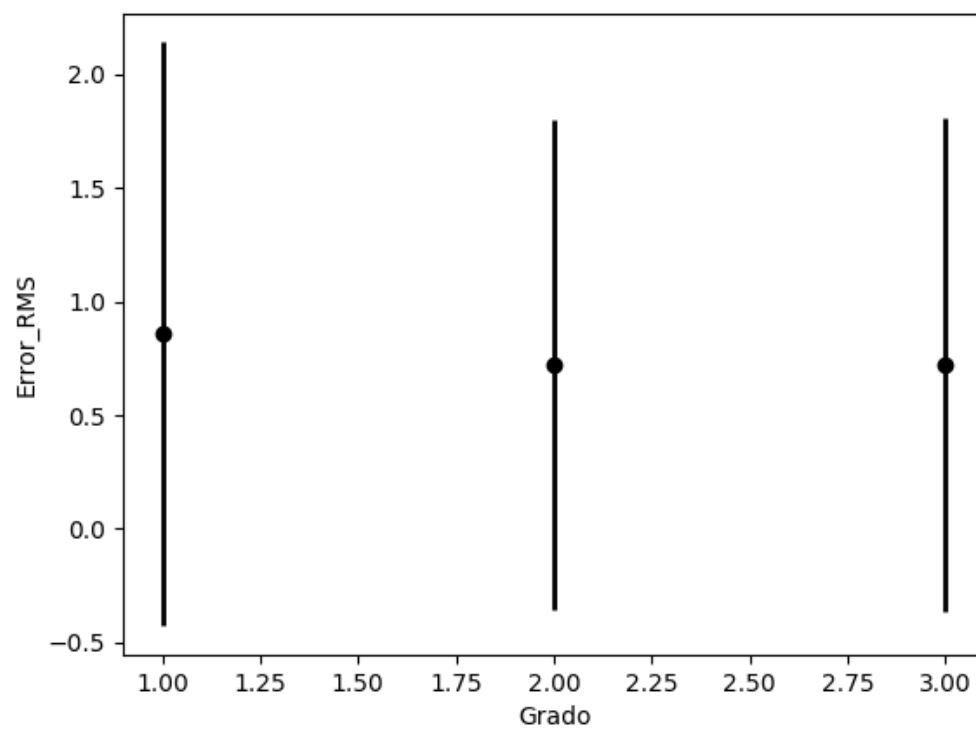


Figura 3.14: Media y Varianza al estimar la relación entre MEDV y LSTAT

Capítulo 4

Conclusiones

Luego de realizar todos los experimentos y haber entendido bien el funcionamiento, podemos realizar algunas conclusiones.

4.1. Referencias

1. Pattern Recognition and Machine Learning by C. Bishop, Springer 2006.