

Laboratorio de Datos

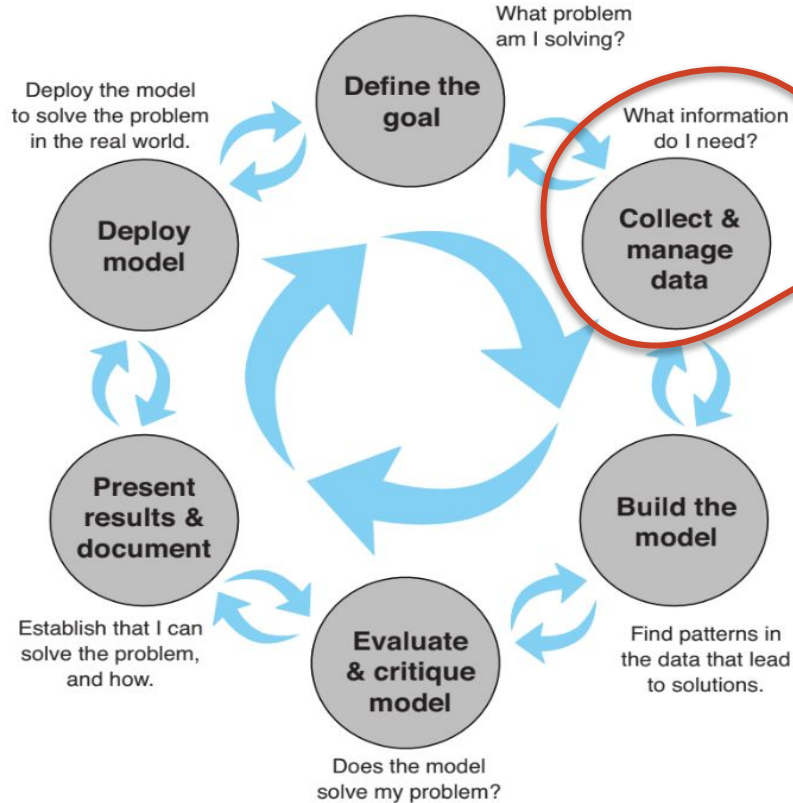
Regresión



Facultad de Ciencias Exactas y Naturales - UBA



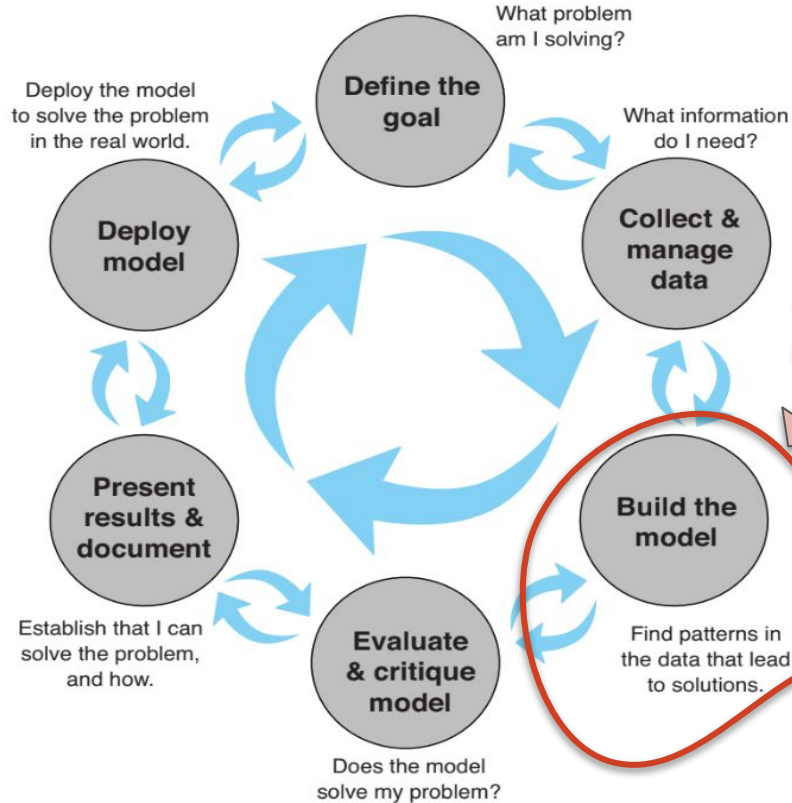
Recorrido de la materia (hasta ahora)



1º parte de la materia

- ✓ Lenguaje de programación (Python)
- ✓ Modelado conceptual de los datos (DER)
- ✓ Representación de los datos (modelo relacional)
- ✓ Formas de consultar los datos (AR/SQL)
- ✓ Recomendaciones para el diseño (Normalización)
- ✓ Calidad de datos
- ✓ Leyes acerca de la Protección de Datos

Recorrido de la materia (hasta ahora)



- ✓ *Visualización y Exploración de los datos*
- ✓ *Intro a Modelado: Clasificación y Regresión*
- ✓ *Clasificación: Árboles de decisión*
- ✓ *Regresión: Regresión Lineal Simple*

Laboratorio de Datos

Regresión Lineal y KNN

... por Manuela Cerdeiro (y modificaciones de P. Turjanski)

Algunos modelos lineales

- + Regresión lineal simple (vimos) $\beta_0 + \beta_1 \cdot x$
- + Regresión lineal múltiple $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$
- + Regresión polinomial $\beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$

¿Lineal en qué?

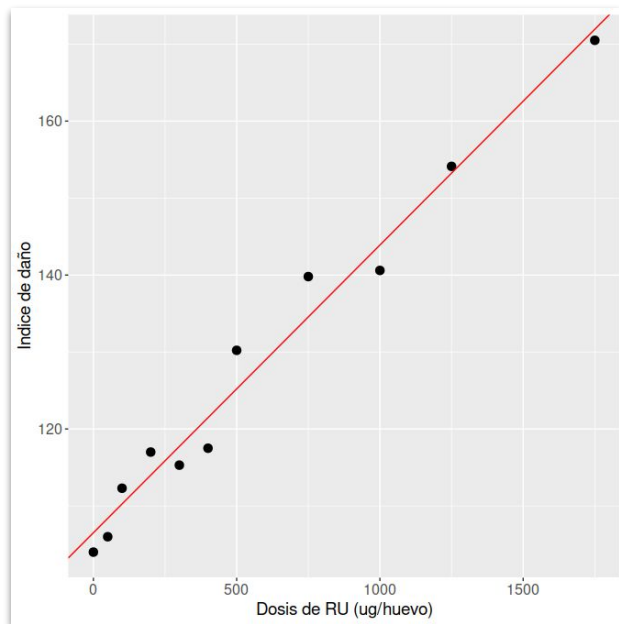
Algunos modelos lineales

- + Regresión lineal simple (vimos) $\beta_0 + \beta_1 \cdot x$
- + Regresión lineal múltiple $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$
- + Regresión polinomial $\beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$

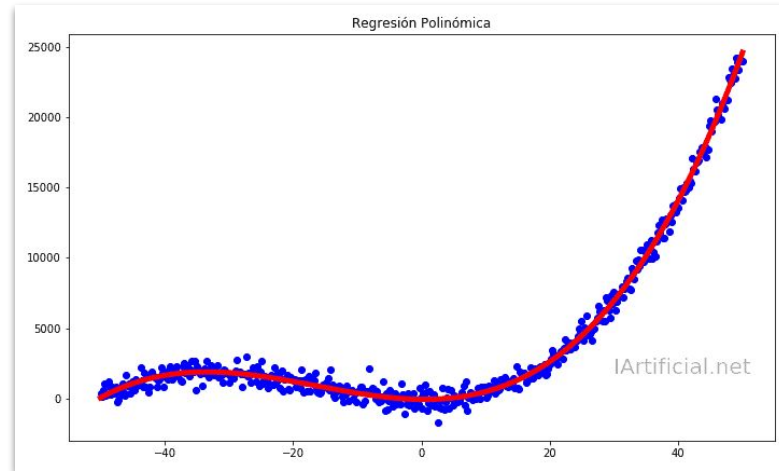
¿Lineal en qué?

En todos los casos, la función es **lineal** en los **parámetros** del modelo.

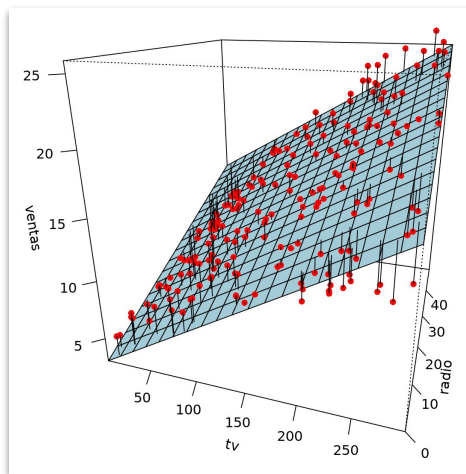
Algunos modelos lineales



$$\beta_0 + \beta_1 \cdot X$$



$$\beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$$



$$\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Dataset Properati

Dataset Properati



- Dataset real de publicaciones digitales de inmuebles
- Provisto en forma libre por Properati (<https://www.properati.com.ar/>)
- Objetivo: Comprender y predecir los precios de las casas de CABA en función de algunas variables relevantes del dataset de properati

	id	j_date	created_on	lat	lon	l1	l2	l3	l4	l5	l6	rooms	bedrooms	bathrooms	surface_total	surface_covered	price	currency	ice_peric	title	property_type	eration_ty
	zzzxIWgAZRd1DqhvURjBQ==	-06-08	2019-05-21	-34.3964	-58.6467	Argentina	Bs.As. G.B.A. Zona Norte	Tigre	Nordelta	nan	nan	2	1	2	62	48	19000	ARS	nan	DEPARTAMENTO.	Departamento	Alquiler
	zzzvna7PZmzmTT70jcwUvA==	-12-31	2019-03-28	-34.5795	-58.3999	Argentina	Capital Federal	Palermo	Palermo Chico	nan	nan	2	nan	2	74	65	390000	USD	nan	Av Scalabrini	Departamento	Venta
	zzzVQ7PJlCgKqKjZDRK6qw==	-12-31	2019-05-16	-34.674	-58.4338	Argentina	Bs.As. G.B.A. Zona Sur	Lanús	Lanús	nan	nan	1	nan	2	nan	242	25000	ARS	Mensual	Local - Lanús Oeste	Local comercial	Alquiler
	zzzNjKrEcFxfRe9lWCp/Gg==	-12-31	2019-05-03	-34.5653	-58.5311	Argentina	Bs.As. G.B.A. Zona Norte	General San Martín	Villa Maipu	nan	nan	2	nan	1	197	159	153000	USD	nan	Calle 57 Ber...	Casa	Venta
	zzz7MWEDLFkpvqS5yAqzQ==	-07-10	2019-04-03	-37.1657	-56.9846	Argentina	Buenos Aires Costa Atlántica	Cariló	nan	nan	nan	nan	nan	2	34	34	1700	USD	Mensual	LOCAL EN ALQUILER	Local comercial	Alquiler
	zzz2P4ROHH1q1tAFLKmjg==	-12-31	2019-03-20	nan	nan	Argentina	Bs.As. G.B.A. Zona Norte	nan	nan	nan	nan	nan	nan	nan	nan	nan	30000	USD	Mensual	Terreno - Parada Robles	Lote	Venta
	zzynfxkhG8n7UJBVCFf9g==	-05-01	2019-04-13	nan	nan	Argentina	Bs.As. G.B.A. Zona Sur	Lomas de Zamora	Banfield	nan	nan	3	nan	2	120	120	20000	ARS	Mensual	Duplex 3 amb...	Casa	Alquiler
	zzxFMCDgXvcX1Jf6fR/Nuw==	-05-12	2019-04-02	-34.4424	-58.5865	Argentina	Bs.As. G.B.A. Zona Norte	Tigre	Tigre	nan	nan	3	nan	1	57	57	11000	ARS	Mensual	ALQUILER DE...	Departamento	Alquiler
	zzwQnhNQeYncszjywnh3g==	-07-19	2019-06-03	-38.693	-62.2191	Argentina	Buenos Aires Interior	Bahía Blanca	nan	nan	nan	nan	nan	nan	788	nan	95000	USD	nan	Patagonia - ...	Lote	Venta
	zzvJfCB2qdmDp5cQDWQNSA==	-07-12	2019-02-02	-31.3961	-64.2444	Argentina	Córdoba	Córdoba	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	COCHERA EN A...	Cochera	Alquiler
	zzuo6RASaHwlnMnyo2B74W==	-07-10	2019-06-12	-34.6614	-58.6714	Argentina	Bs.As. G.B.A. Zona Oeste	Ituzaingó	nan	nan	nan	nan	nan	nan	90	90	10000	ARS	Mensual	LOCAL EN ALQUILER	Local comercial	Alquiler
	zzt6pvh0SuoBjYnJa4Iwncg==	-01-16	2019-01-02	-34.5683	-58.6986	Argentina	Bs.As. G.B.A. Zona Norte	San Miguel	Bella Vista	nan	nan	5	nan	3	nan	148	35000	ARS	Mensual	Casa - San Miguel	Casa	Alquiler
	zzsrfKxROXKJJu1svdjOJYA==	-06-07	2019-01-24	nan	nan	Argentina	Bs.As. G.B.A. Zona Norte	Tigre	nan	nan	nan	nan	3	3	211	211	325000	USD	nan	Casa en vent...	Casa	Venta
	zzskohbhiFBYUaPOMQhbzw==	-07-02	2019-04-29	-27.3676	-55.9836	Argentina	Misiones	Posadas	nan	nan	nan	8	3	3	440	200	220000	USD	nan	Vende Chalet...	Casa	Venta
	zzrlop3HrxAgIT0VJ9wRg==	-06-18	2019-01-17	-34.6401	-58.5013	Argentina	Capital Federal	Villa Luro	nan	nan	nan	2	1	nan	54	47	118000	USD	Mensual	DEPARTAMENTO EN VENTA	Departamento	Venta
	zzrWrrQYdOQwCAyoHZ93Vw==	-06-12	2019-05-17	-31.414	-64.1631	Argentina	Córdoba	Córdoba	nan	nan	nan	4	2	nan	60	60	70000	USD	nan	B° Gral Paz ...	Departamento	Venta
	zzqDyT23dtx6pe3mRqbG+A==	-02-19	2019-02-18	-34.4026	-58.6685	Argentina	Bs.As. G.B.A. Zona Norte	Tigre	Nordelta	Barrio...	nan	nan	2	1	82	70	163000	USD	Mensual	Duplex de 3a...	Casa	Venta
	zzpIYLLLAvzkog6wi8tLA==	-12-31	2019-05-07	-35.1312	-58.3857	Argentina	Bs.As. G.B.A. Zona Sur	San Vicente	nan	nan	nan	nan	nan	nan	nan	nan	38000	USD	Mensual	Terreno - San Vicente	Lote	Venta
	zzoxa7co2rLS2lwcec5XtQ==	-05-20	2019-03-19	-32.9467	-60.6427	Argentina	Santa Fe	Rosario	nan	nan	nan	3	2	1	60	60	110000	USD	nan	RIOJA Y CORRIENTES	Departamento	Venta
	zzoO9vTn5xQ1GV2yq2f4ag==	-07-19	2019-06-03	nan	nan	Argentina	Santa Fe	Rosario	nan	nan	nan	nan	nan	nan	300	nan	790000	ARS	nan	VENTA DE TER...	Lote	Venta
	zzmqHNXY1La5HuLJ2kzHw==	-12-31	2019-06-18	-34.6424	-58.4539	Argentina	Capital Federal	Flores	nan	nan	nan	4	3	2	220	200	189000	USD	nan	VENTA CASA 4...	Casa	Venta
	zzmadp0Wnz8VDHej+BxnsG==	-12-31	2019-05-12	-32.9233	-60.6769	Argentina	Santa Fe	Rosario	nan	nan	nan	1	nan	1	33	33	nan	nan	nan	Monoambiente en Venta	Departamento	Venta
	zzmSr8a2J+njZ/vlcAwDwg==	-12-31	2019-02-24	-31.5394	-60.6388	Argentina	Santa Fe	Monte Vera	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	Mensual	Campo Horticola	Lote	Venta
	zzlZruiCAU7xiPJkkUxdA==	-04-26	2019-02-07	-34.5806	-58.563	Argentina	Bs.As. G.B.A. Zona Norte	General San Martín	Villa Libertad	nan	nan	2	nan	1	60	nan	9200	ARS	Mensual	Diego Pombo ...	Departamento	Alquiler
	zzlNUymwQGXhRx2Aon1KtQ==	-08-17	2019-04-26	nan	nan	Argentina	Bs.As. G.B.A. Zona Norte	nan	nan	nan	nan	nan	nan	1	35	nan	18000	ARS	nan	Independenci...	Local comercial	Alquiler
	zzl1xBpbXk09/RVovbq8Bw==	-07-13	2019-04-09	nan	nan	Argentina	Santa Fe	Rosario	nan	nan	nan	nan	nan	nan	300	nan	12000	USD	nan	TERRENO EN B...	Lote	Venta
	zzkrGFUAzMN80uQwlQxghw==	-02-27	2019-02-13	-37.9545	-57.5644	Argentina	Buenos Aires Costa Atlántica	Mar del Plata	nan	nan	nan	3	2	3	866	215	260000	USD	Mensual	Chalet en Ve...	Casa	Venta
	zzkM1o5h8aeJojaZjgPXuA==	-12-31	2019-03-17	-34.6019	-58.3853	Argentina	Capital Federal	Tribunales	nan	nan	nan	nan	nan	2	243	243	360000	USD	nan	Lavalle y Mo...	Oficina	Venta
	zzidQWeC1GchzV6hhHm0lw==	-03-21	2019-03-20	-38.0018	-57.5474	Argentina	Buenos Aires Costa Atlántica	Mar del Plata	Centro	nan	nan	2	1	1	nan	nan	76000	USD	Mensual	DEPARTAMENTO EN VENTA	Departamento	Venta
	zzhGQVJXsmVBdGsGMSU4zQ==	-12-31	2019-05-02	-34.5992	-58.4015	Argentina	Capital Federal	Recoleta	nan	nan	nan	nan	nan	4	nan	nan	5500	ARS	Mensual	Alquiler hab...	Casa	Alquiler temporal

1. Recolección y Exploración de datos



- Descargamos el dataset (original) denominado

`ar_properties.csv`

- Removemos columnas que no vamos utilizar
- Limpiamos y filtramos un poco los datos
- Generamos el dataset

`data_selec.csv`

-> Alquiler en \$ (eliminando valores extremos)

-> Sólo algunos barrios de CABA

2. División de datos



- Cargamos el dataset con datos limpios denominado

`data_alq_caba.csv`

- Dividimos el dataset en train(80%) y test(20%)

- Guardamos ambos dataframes

`data_alq_caba_train.csv`

`data_alq_caba_test.csv`

3. Modelos con Train



- Cargamos el dataset con datos limpios denominado

`data_alq_caba_train.csv`

- Proponemos distintos modelos para *predecir el precio de alquiler*

i. Modelo Lineal Simple

ii. Modelo Lineal Múltiple

iii. Modelo Polinomial

¿Cuál es el mejor modelo?

Error Cuadrático Medio (MSE por Mean Squared Error)

El error cuadrático medio (de cualquier modelo) mide el promedio de los errores al cuadrado, es decir:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2$$

Y_i son los valores observados (reales)

\hat{Y}_i son los valores estimados por el modelo

Evaluación

¿Cómo sé si mi modelo realmente mejora al agregar parámetros?

```
##  
# -----  
#  
#      Modelo Polinomial 1  
# -----  
#  
# Quinta propuesta: Modelo Polinomial tomando a:  
# X1   = surface_covered (variable predictora)  
# Y    = price           (variable a predecir)  
# grado = 1
```

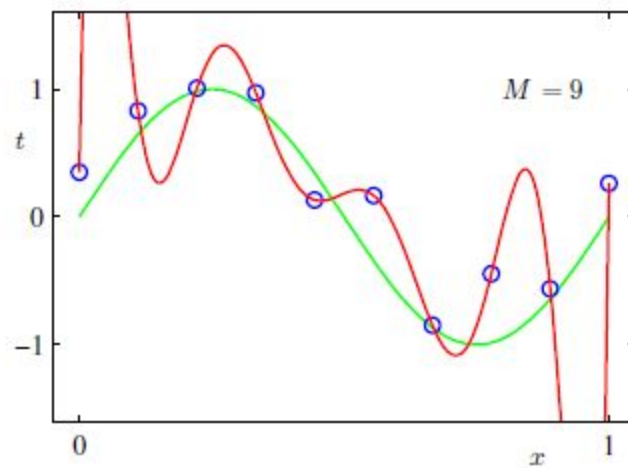
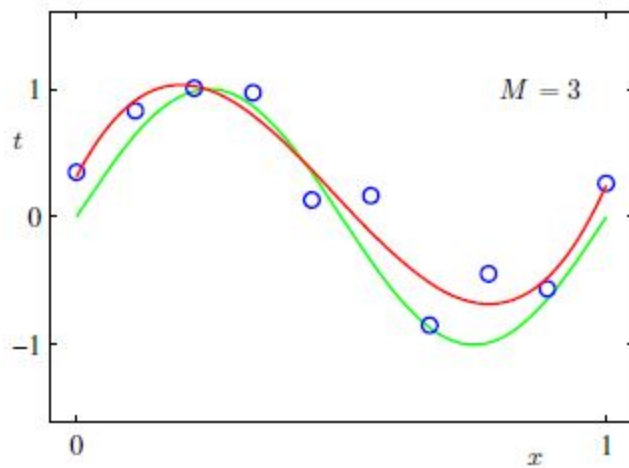
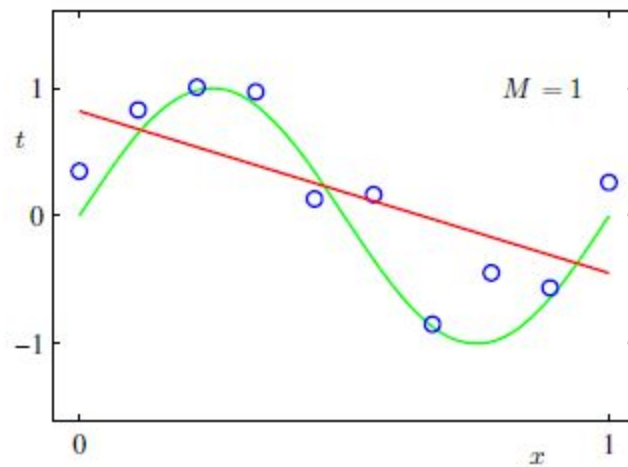
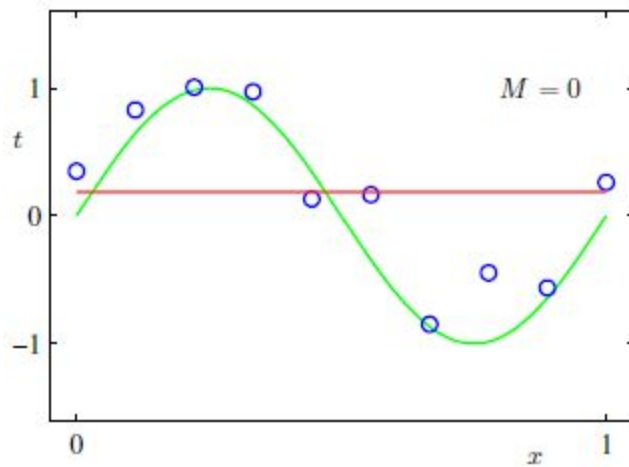
Coefficients

```
-----  
intercept : 6092.354503831017  
pendientes: [ 0.          255.20987921]  
R^2 (train): 0.54  
MSE (train): 49942831.35
```

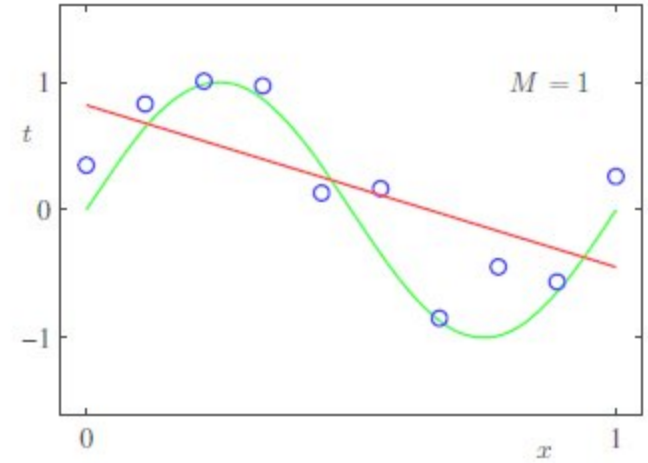
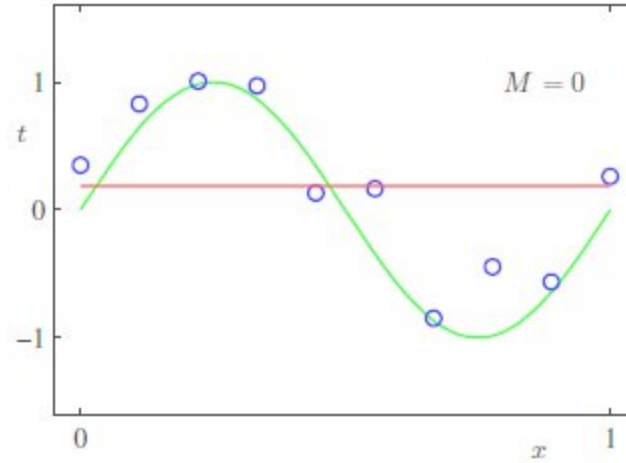
```
##  
# -----  
#  
#      Modelo Polinomial 1  
# -----  
#  
# Quinta propuesta: Modelo Polinomial tomando a:  
# X1   = surface_covered (variable predictora)  
# Y    = price           (variable a predecir)  
# grado = 3
```

Coefficients

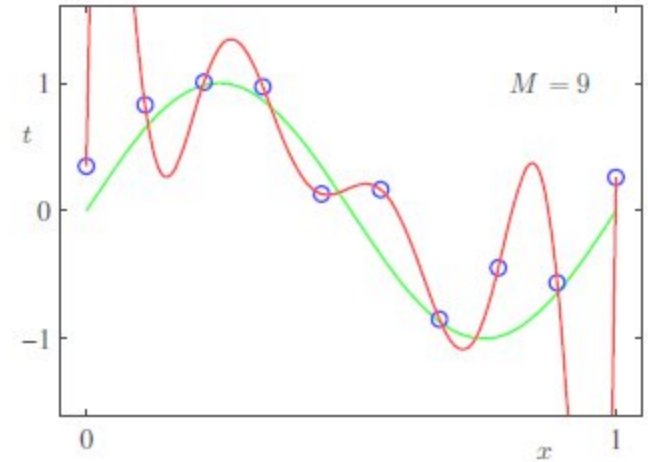
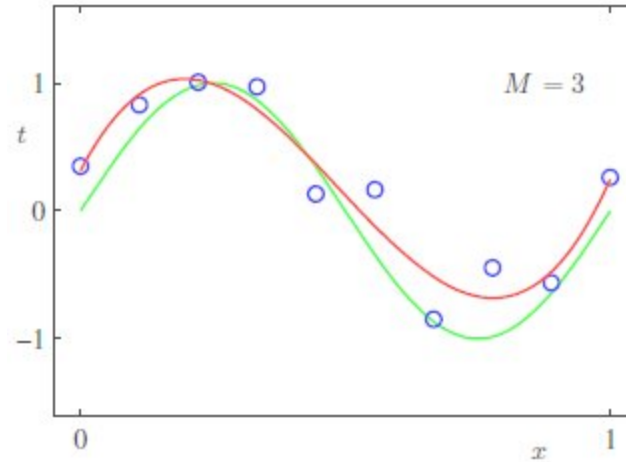
```
-----  
intercept : -312.1975409025872  
pendientes: [ 0.00000000e+00  4.58020308e+02 -1.36455538e+00  1.07500462e-03]  
R^2 (train): 0.56  
MSE (train): 48077867.99
```



Si los parámetros del modelo son insuficientes, el modelo no llega a explicar lo suficiente, hablamos de sub-ajuste.



Si los parámetros del modelo son demasiados, el modelo se adapta a los datos de manera excesiva, perdiendo así capacidad explicativa y predictiva, hablamos de sobre-ajuste.



Evaluación



Volvemos a la pregunta ...

¿Cómo sé si mi modelo realmente mejora al agregar parámetros?

Evaluación



Volvemos a la pregunta ...

¿Cómo sé si mi modelo realmente mejora al agregar parámetros?

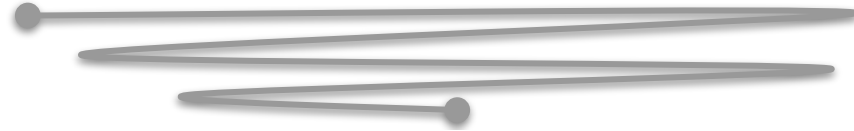
Evaluamos sobre datos nuevos

Dataset

Training

Testing

4. Evaluación de los modelos utilizando datos de Test



- Cargamos el dataset con datos limpios denominado

`data_alq_caba_test.csv`

- Evaluamos los modelos propuestos ...

i. Modelo Lineal Simple

ii. Modelo Lineal Múltiple

iii. Modelo Polinomial

} ¿Cuál es el mejor modelo?

Conclusiones



- Vimos varios modelos en los que se pretende explicar o predecir una variable continua a partir de otras variables
- Comparamos modelos según una medida de su bondad de ajuste, en este caso el error cuadrático medio o R^2 . Hay muchas más
- Vimos que no siempre más parámetros dan un mejor modelo
- Vimos que para evaluar cuán bueno es el modelo, hay que ver cómo se desempeña con datos nuevos, distintos a los que usamos para entrenarlo

Ooootro tema

¿Cuánto medirá de adulto?



Basado en una clase de Mariela Sued

Información



Es varón



La mamá es bajita, mide 156 cm

¿Cuánto medirá de adulto?

- + Sin información → ¿Qué podemos decir?

ESTIMAMOS:

¿Cuánto medirá de adulto?

- + Sin información → ¿Qué podemos decir?

Necesitamos **datos**.
Completemos columna
“Altura”



¿Cuánto medirá de adulto?

¿Promediamos?

ESTIMAMOS: 174

¿Cuánto medirá de adulto?

+ Sin información 

+ Es varón →

Completemos
columna "sexo"





¿Cuánto medirá de adulto?



¿Promediamos entre varones?

ESTIMAMOS: 176

¿Cuánto medirá de adulto?

- + Sin información → 
- + Es varón → 
- + Es varón y la mamá bajita →

¿Cuánto medirá de adulto?

- + Sin información → 
- + Es varón → 
- + Es varón y la mamá bajita (G, M, B) → [Completemos](#)
[columna](#)
["contextura mamá"](#)



¿Cuánto medirá de adulto?

¿Promediamos entre varones de mamás bajitas?

ESTIMAMOS: 172.8

¿Cuánto medirá de adulto?

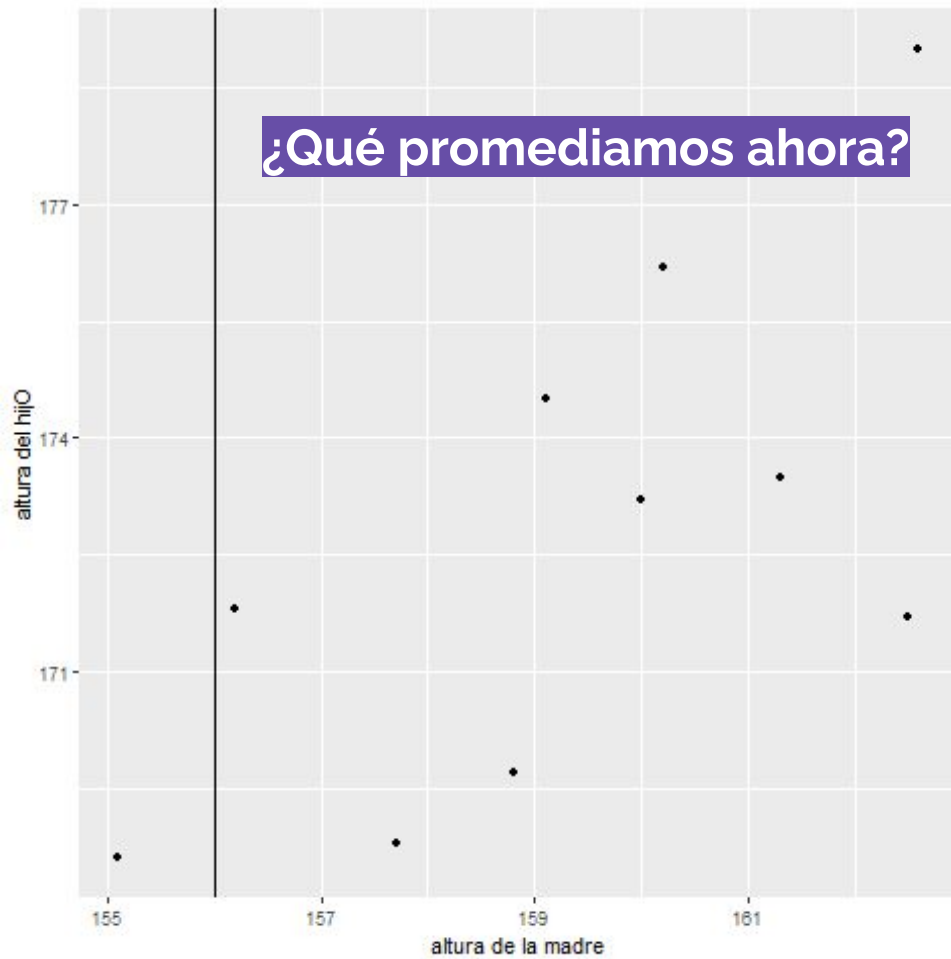
- + Sin información → ✓
- + Es varón → ✓
- + Es varón y la mamá bajita → ✓
- + Es varón y la mamá mide 156 →

Completemos columna
“altura mamá”



¿Cuánto medirá de adulto?

¿Qué promediamos ahora?



Una posibilidad: KNN

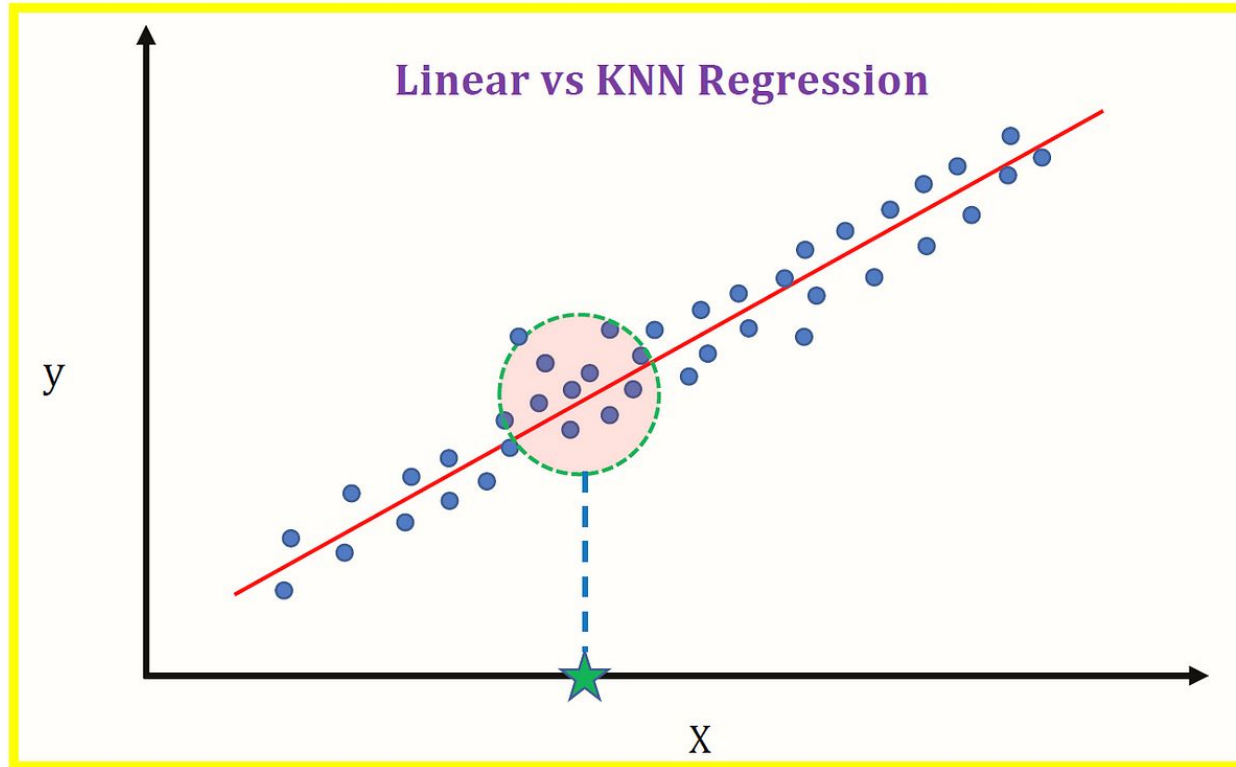
Idea: Promediamos los valores de casos parecidos

kNN: k nearest neighbors - k vecinos más cercanos

Ej. Consideramos los 5 valores más **cercanos*** al valor nuevo (altura madre).
Promediamos las alturas de esos 5 varones

*Cercanos: en la o las variables explicativas,
y con la distancia que consideremos.

Modelo de kNN



KNN con sklearn



Conclusiones



- Vimos otro modelo que permite predecir una variable continua a partir de otras variables
- Comparamos modelos según una medida de su bondad de ajuste, en este caso el R^2
- Vimos que no siempre más parámetros dan un mejor modelo