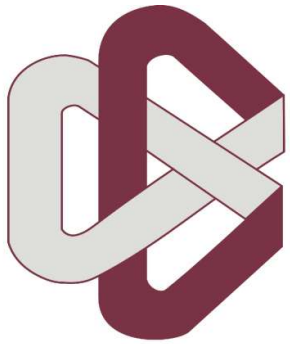
Modelos estadísticos: una reflexión sobre su concepción, ventajas y usos

Eloísa Díaz Francés Murguía
(diazfran@cimat.mx)
Area de Probabilidad y Estadística
Abril, 2021

en colaboración con
Víctor Alvarado Estrella (DEMAT, UG)
y tras comunicación personal con
Rafael González de Gouveia (exalumno CIMAT)



CIMAT

(NY: 1928 – Berkeley: 2005)

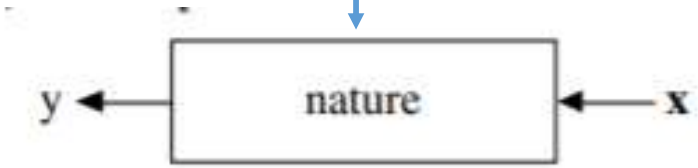
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

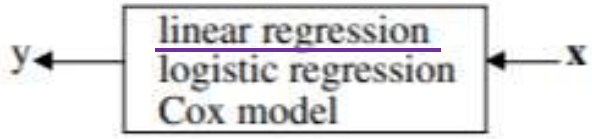
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

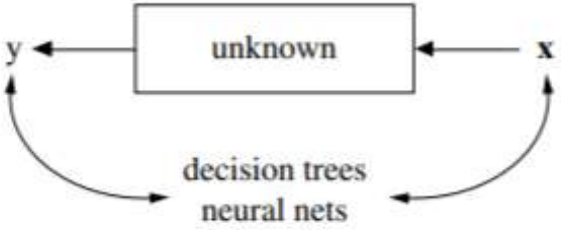
CAJA NEGRA:



The Data Modeling Culture



The Algorithmic Modeling Culture



?

"The goals in statistics are to use data to predict and to get information about the underlying data mechanism.

….. The emphasis needs to be on the problem and on the data."

David Sprott
(1930-2013)

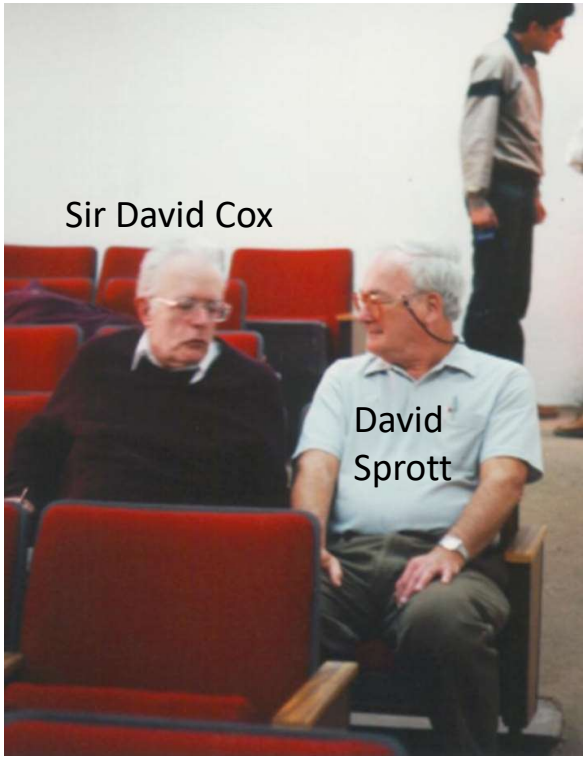
Modelación estadística de un fenómeno aleatorio de interés (George Box 1980, David Sprott 2000 y Sir David Cox, 2001)

1. Recabar datos del fenómeno,
   explorarlos y entenderlos
2. Plantear modelo estadístico F(x;θ)
3. Estimar (puntual, región e intervalo)
4. Combinar experimentos?
5. Validar modelo
6. Selección y comparación del mejor modelo

ITERAR

George Box
(1919-2013)

Jorge Argáez



Auditorio del CIMAT

Ejemplo: Modelo de Regresión Lineal Simple

- Variable respuesta: Y
- Variables explicativas: $X_1, \ldots, X_p$
- Un modelo condicional: $Y / X \sim N(f(x_1, \ldots, x_p; \beta), \sigma^2)$

Usualmente: $f(x_1, \ldots, x_p; \beta) = \mathrm{X}'\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Carl Friedrich Gauss, 1777-1855

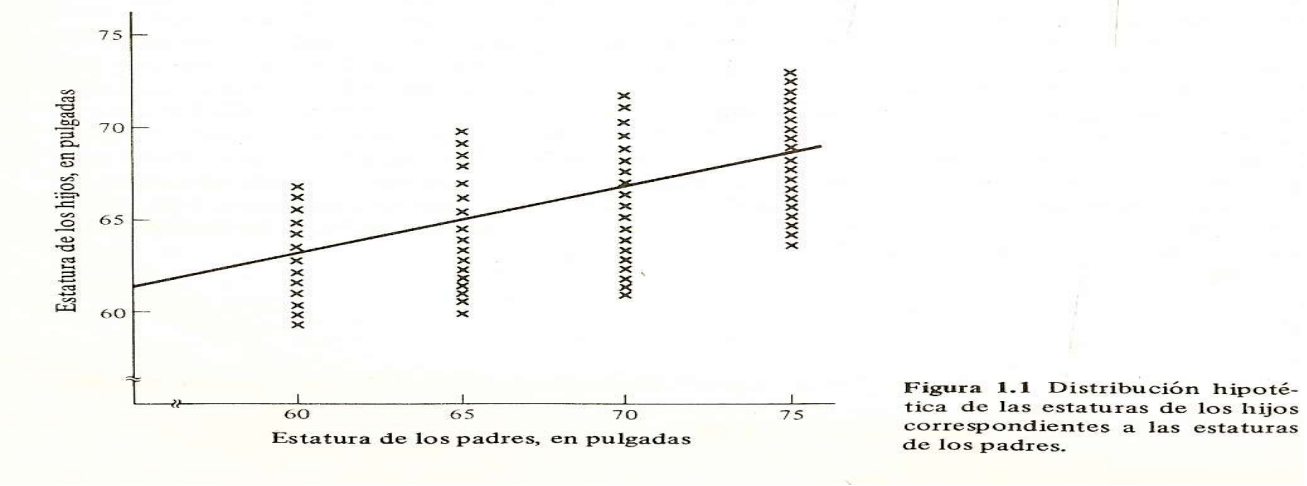


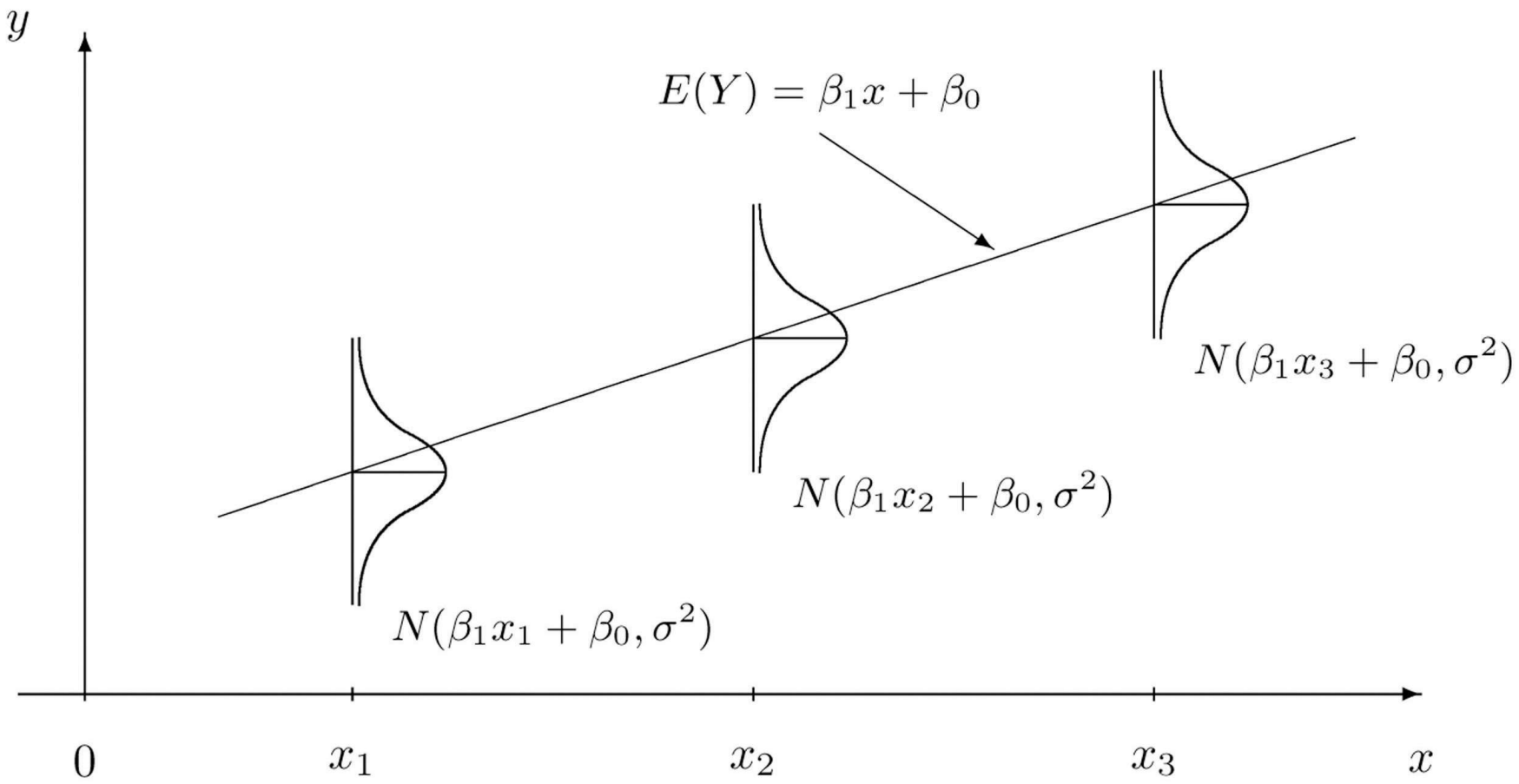
Figura 1.1 Distribución hipotética de las estaturas de los hijos correspondientes a las estaturas de los padres.

Un modelo condicional:

$$Y \mid X \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

Resume e incluye a muchos modelos aislados:

$$Y \sim N\left(\mu_x, \sigma_x^2\right)$$



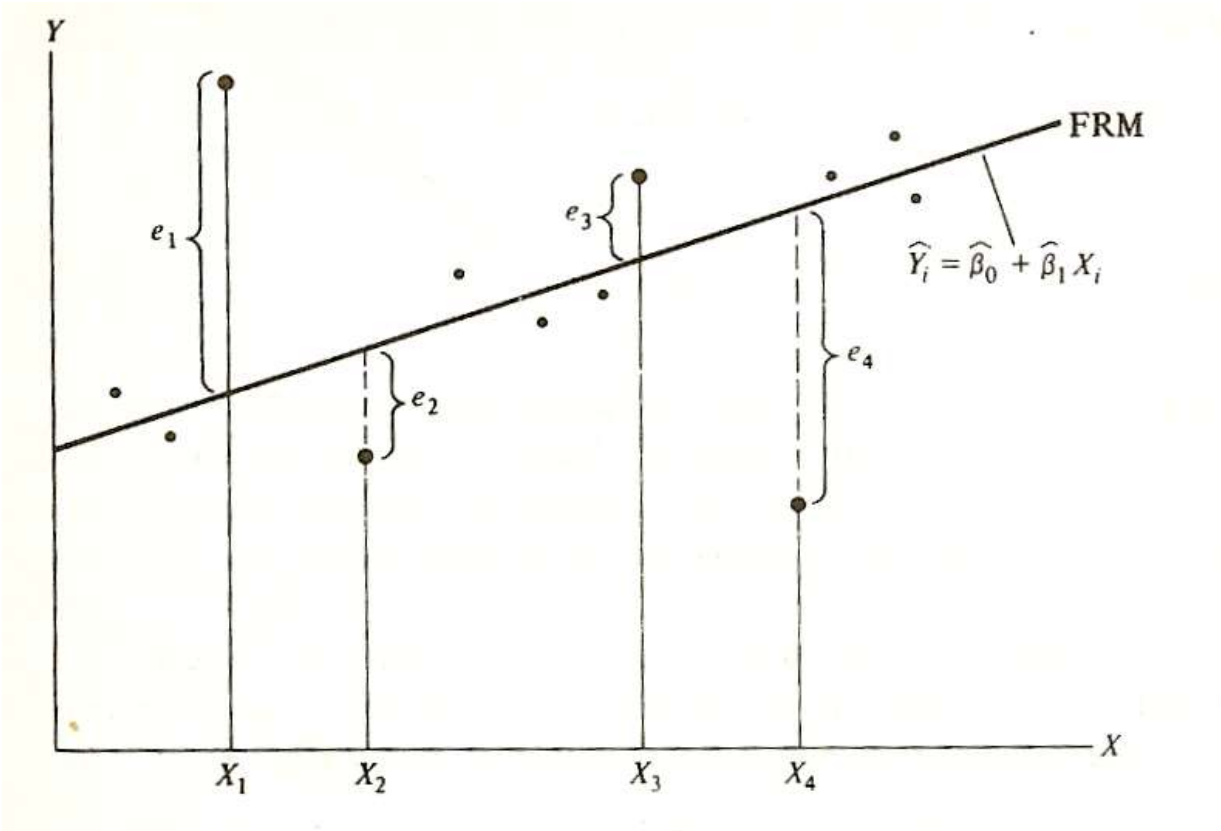
La ‘misma’ solución con dos enfoques:

Legendre: Algorítmico. Sólo caso particular normal: Mínimos cuadrados

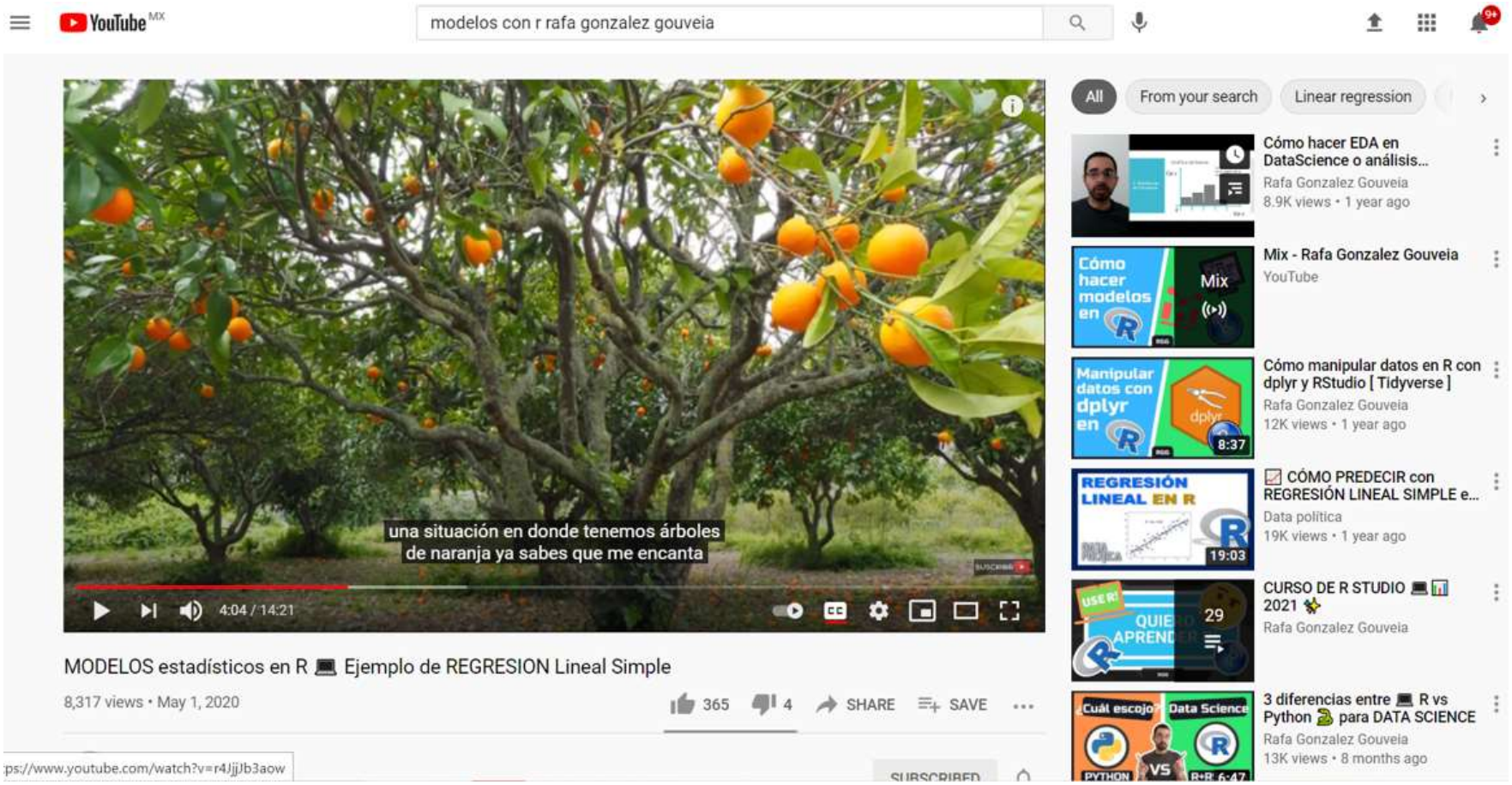
Densidad conjunta para los n modelos normales para y1,….,yn:

$$f(\vec{y}) = \frac{1}{\sigma_x^n} \exp\left[-\frac{1}{2\sigma_x^2} \sum_{i=1}^{n} (y_i - \mu_x)^2 \right]$$

Gauss: Gran visionario propuso modelo general usado hoy en día



Ejemplo de Circunferencias de Naranjos por edad



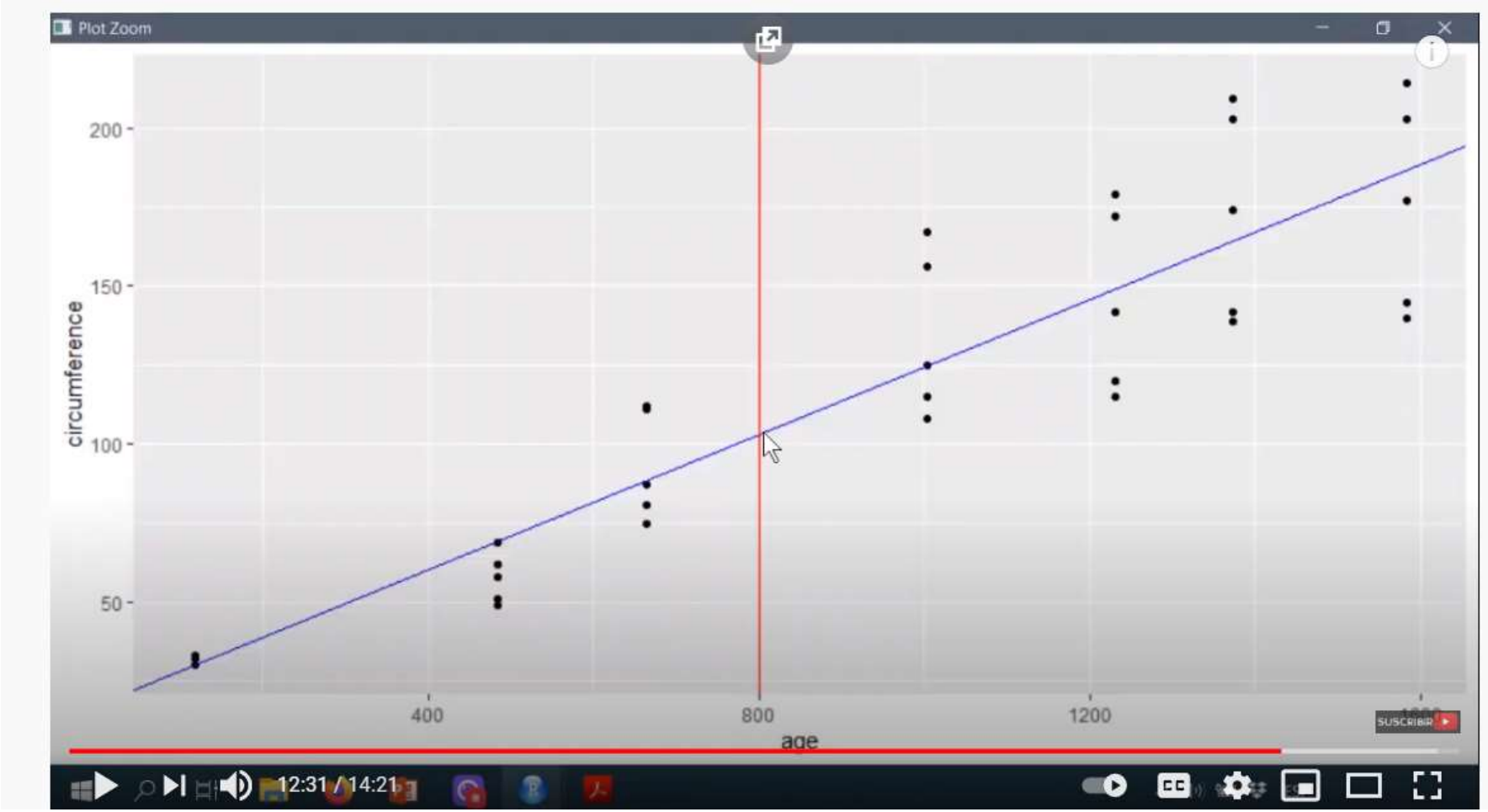
M.C. Rafael González de Gouveia (exalumno CIMAT)

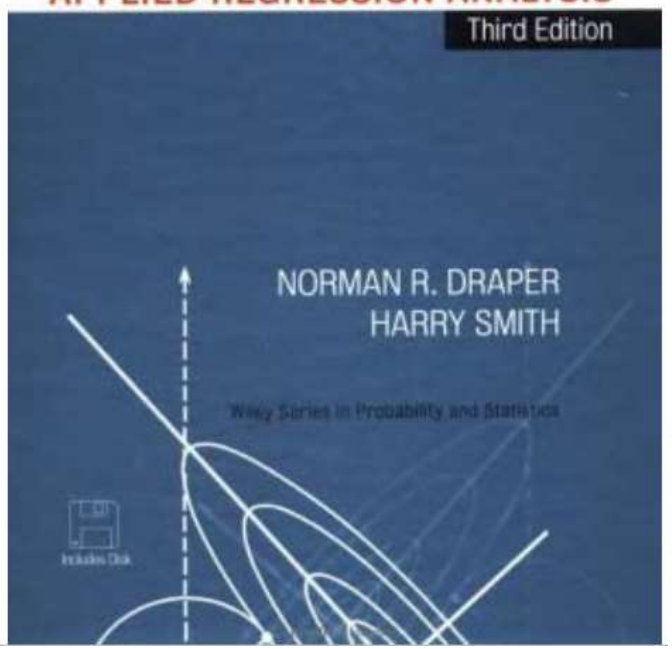
NARANJO:

Plantado desde semilla son 7 años a dar frutos.

Sembrando rama puede dar fruta de 3 a 5 años.

Naranjos adultos miden de 9 a 15 metros de altura.


Plot Zoom
200
150
100
50
circumference
400
800
1200
age
SUSCRIBIR
12:31 / 14:21



24 An Introduction to Nonlinear Estimation

N. The data below arose from five orange trees grown at Riverside, California, during the period 1969–1973. The response w in the body of the table is the trunk circumference in millimeters, and the predictor variable t is the time in days, with an arbitrary origin taken on December 31, 1968. Fit the models given in Eqs. (24.7.2), (24.7.5), (24.7.8), and (24.7.10) to these data. Based on a visual inspection of the fitted models, which seems to be most useful?

	Response w for Tree No.				
t	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

Observa
ramas
y altura de
bifurcación

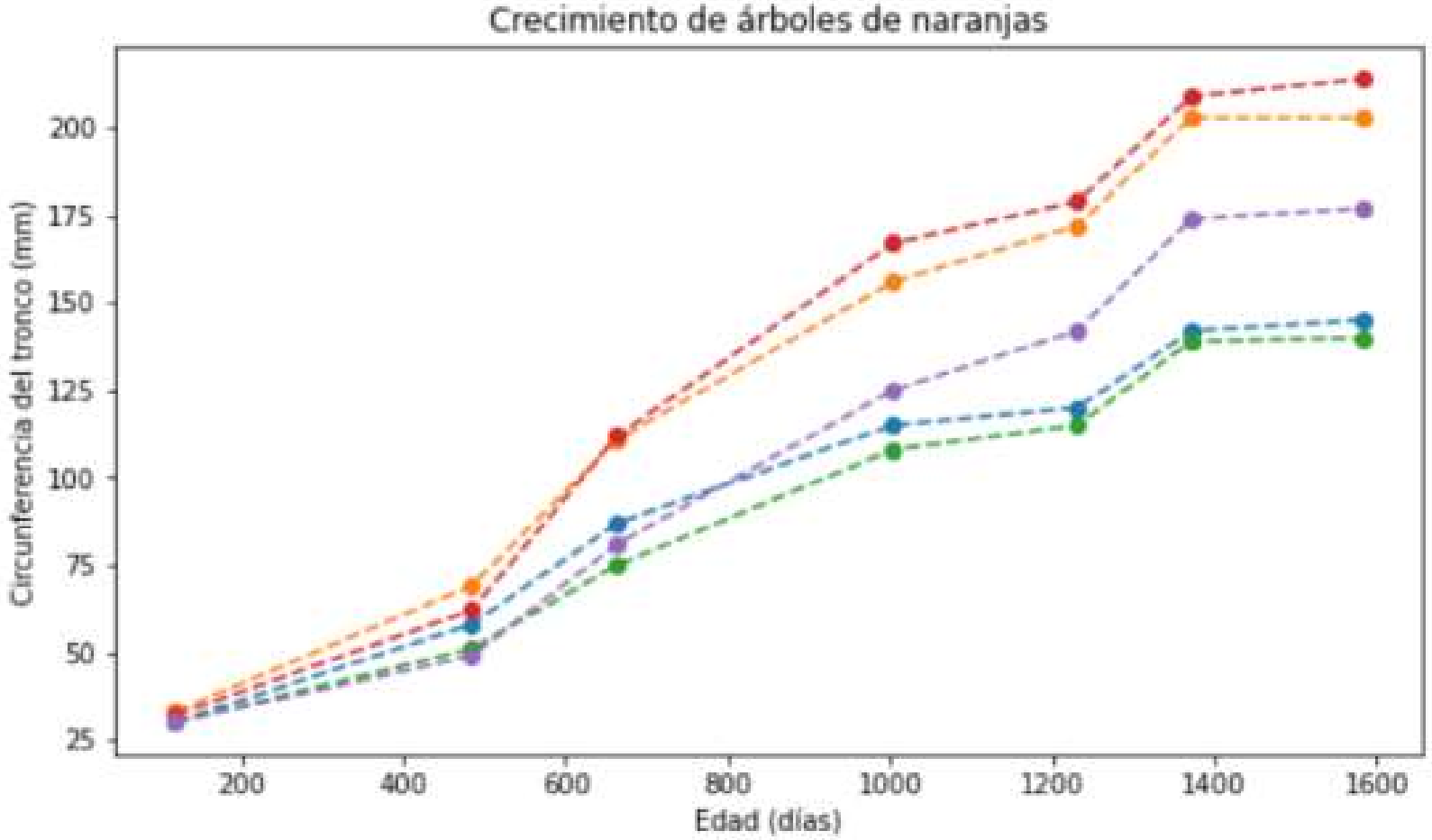


Figura 1.10: Trayectorias de crecimiento de los árboles de naranjas.

Modelos a considerar:

1. Y|X Normal con media lineal $\beta_0 + \beta_1 x$ y varianza constante (**?**)
2. Y|X Normal con media lineal y varianza creciente (lineal o curveada)

Modelando parámetros escala normales
(las desviaciones estándar):

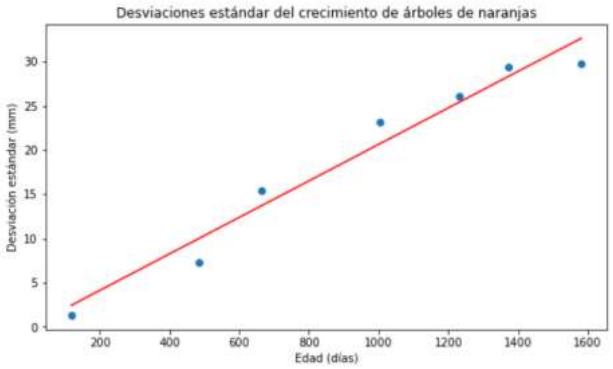


Figura 1.13: Modelo lineal para las desviaciones estándar empíricas.

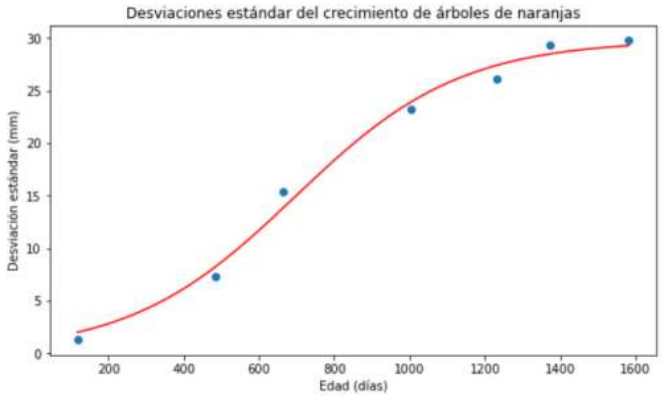


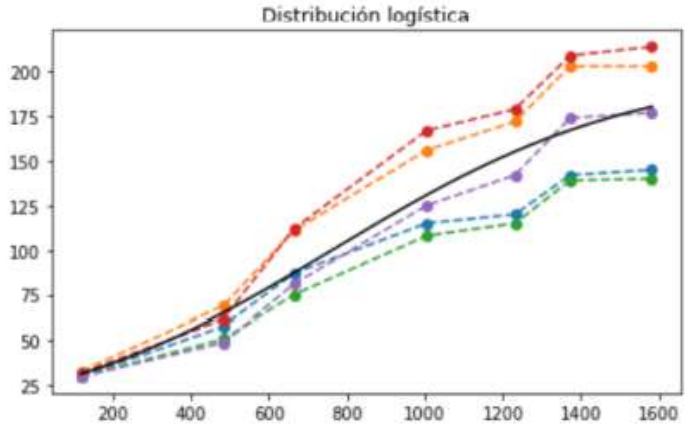
Figura 1.15: Modelo logístico para las desviaciones estándar empíricas.

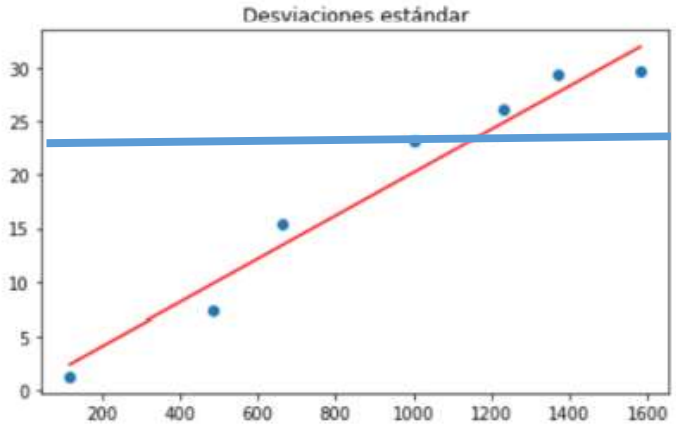
Más modelos:

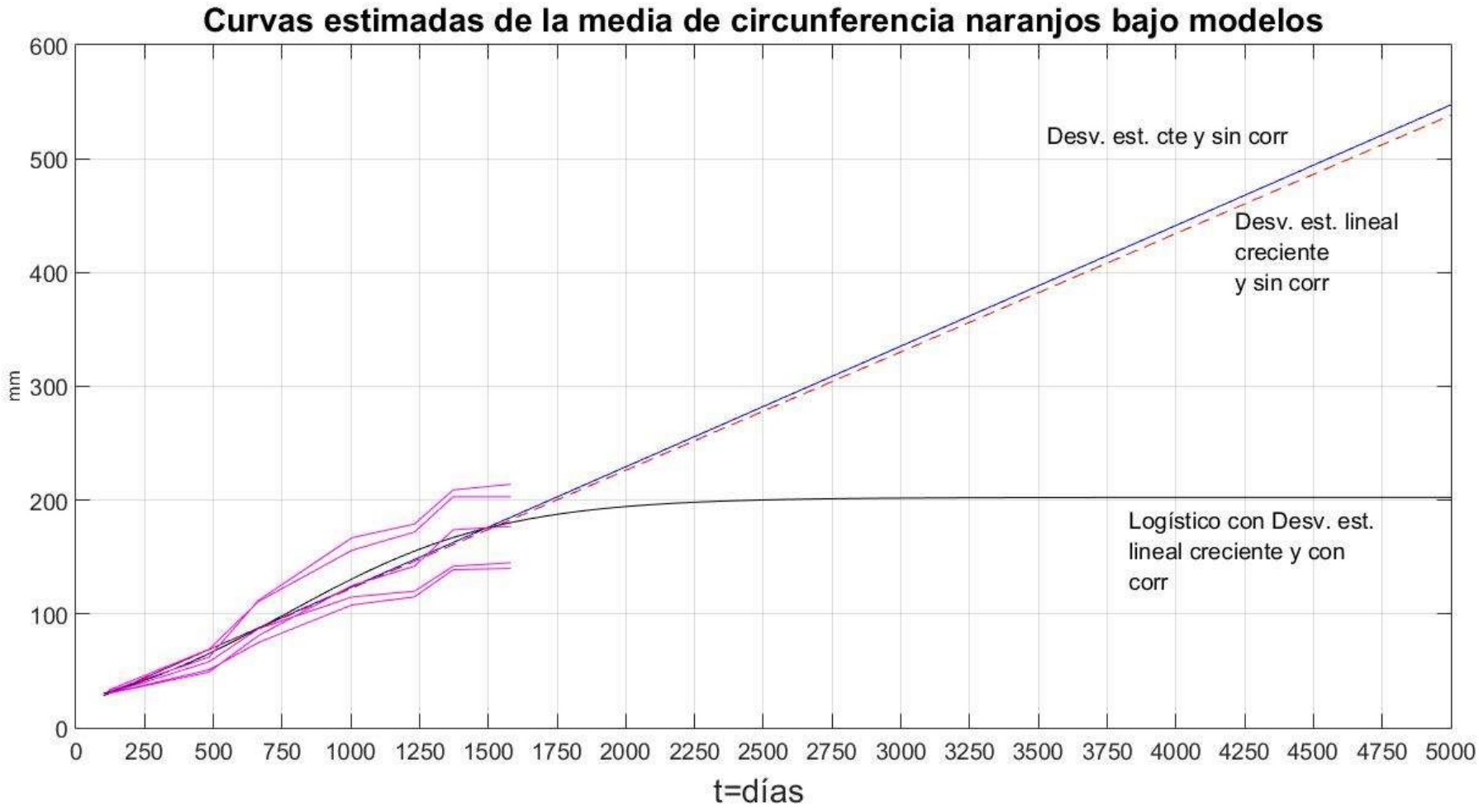
3. Curva sigmoidal Logística para la media:

$$f\left(t;a,b,c\right)=\frac{a}{1+\exp\left[-\left(\frac{t-b}{c}\right)\right]}$$

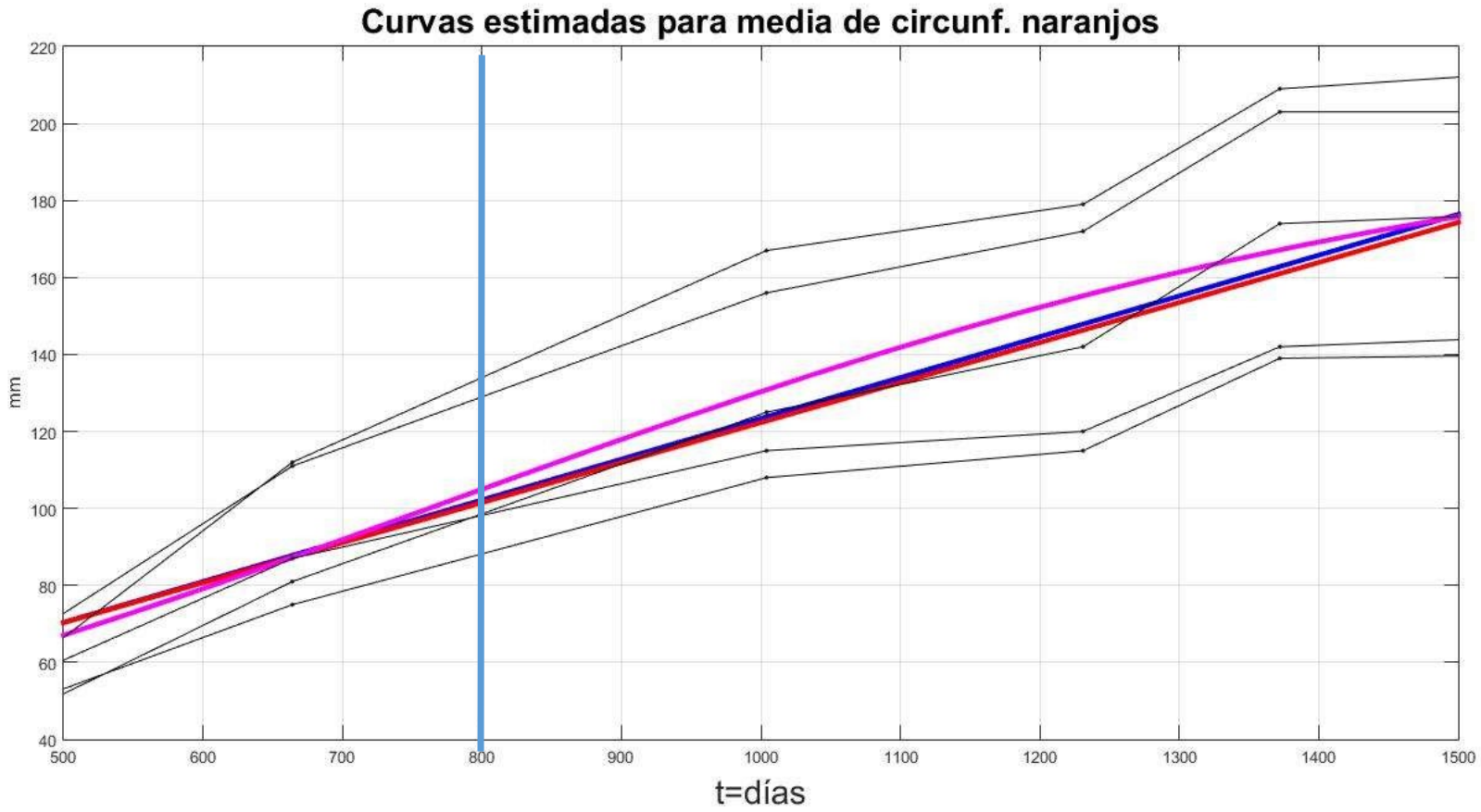
Con escala normal creciente (lineal) y con correlación alta: $\hat{\rho}=0.998.$





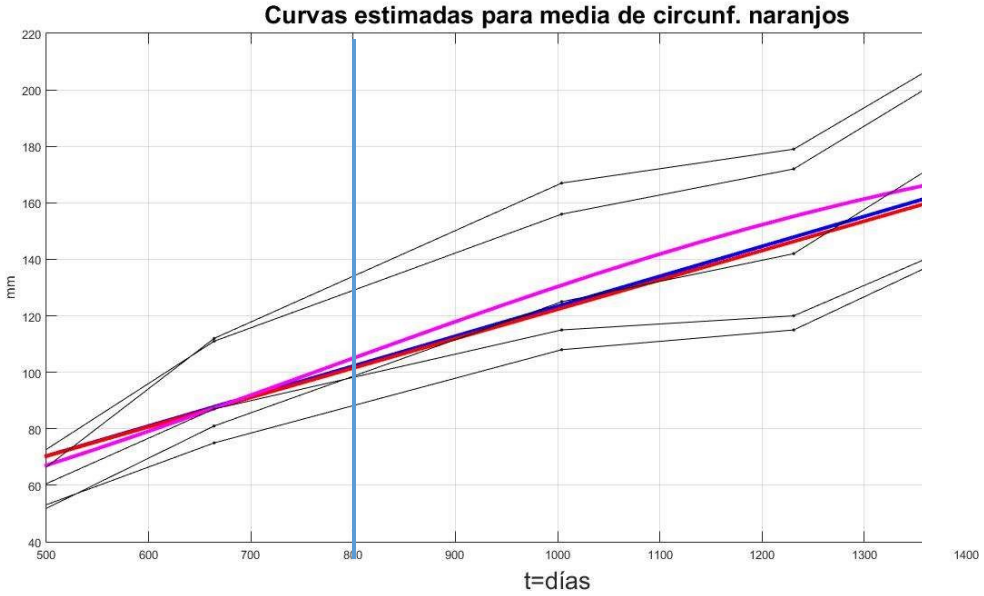

Curvas estimadas de la media de circunferencia naranjos bajo modelos
600
500
400
300
200
100
0
mm
0
250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
t=días
Desv. est. cte y sin corr
Desv. est. lineal
creciente
y sin corr
Logístico con Desv. est.
lineal creciente y con
corr

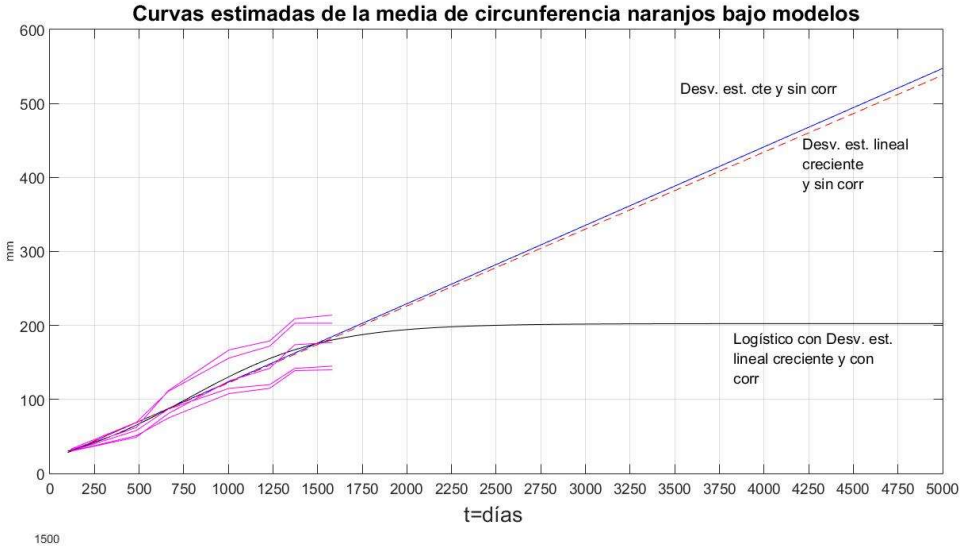
Para predecir media u observación en t=800:



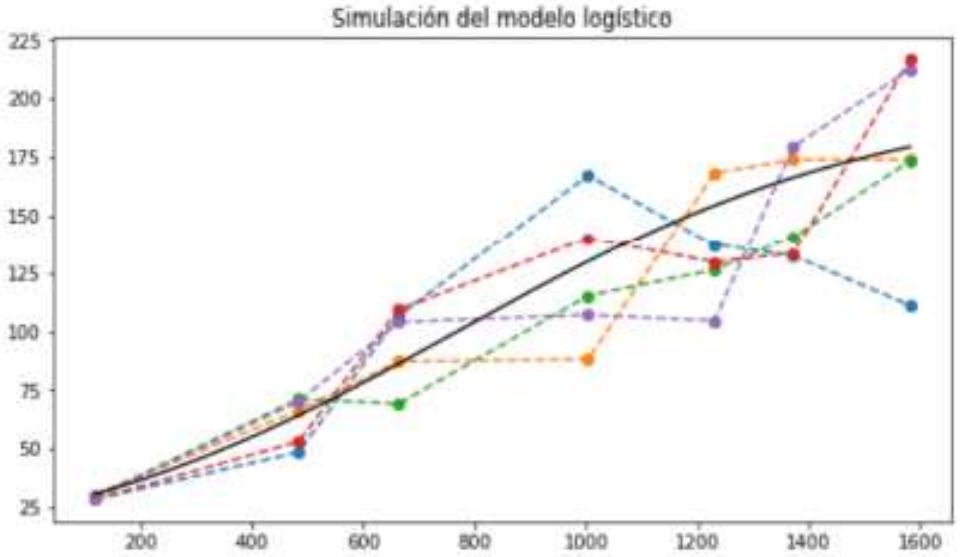
Intervalos de predicción de circunferencia de naranjos

AIC	Modelo	Estimación en t=800	Predicción de Media en t=800	Predicción de naranjo en t=800	Estimación en t=3650	Predicción de Media en t=3650	Predicción de naranjo en t=3650
324	Media Lineal y escala contante	102.81	[94.4, 111.2]	[53.8, 151.8]	407.11	[360.5, 453.8]	[339.9,474.3]
299	Media Lineal y escala creciente	102.12	[95.8, 108.4]	[66.3,138.2]	400.69	[367.9, 434.3]	[236.9, 565.4]
276	Media Logística, escala creciente, con correlación alta	104.98	[91.9, 118.2]	[70.7, 139.3]	202.37	[165.9, 256.4]	[55.5, 356]

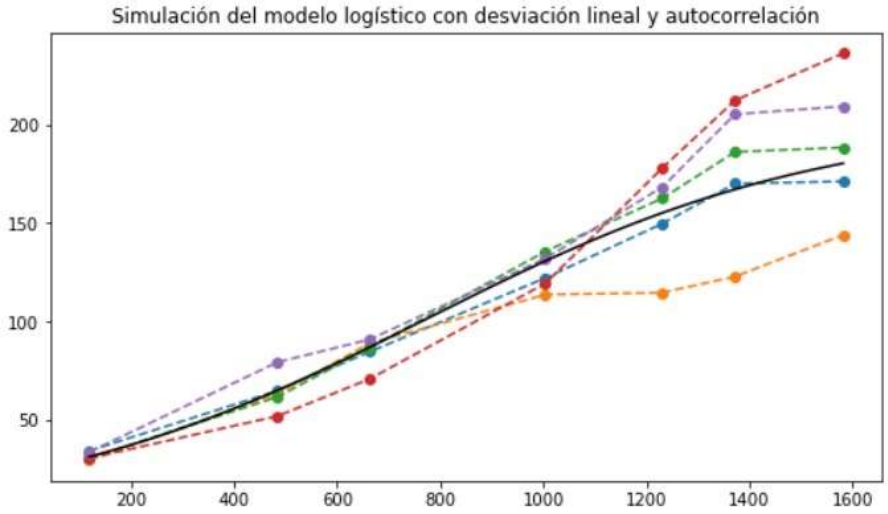




Simulaciones de naranjos con modelo Logístico:



Sin correlación



Con correlación

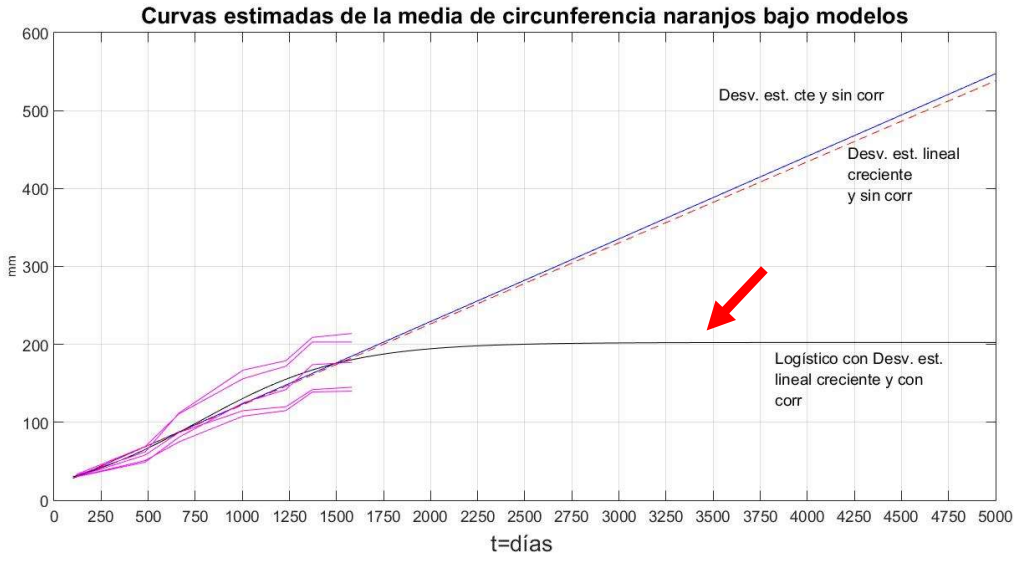
Selección de buen modelo para naranjos:

Para la media de circunferencia de rama del naranjo en el tiempo:

una curva logística.

Con escala normal creciente (lineal) como función del tiempo:

CON CORRELACION MUY ALTA



Conclusión:

Un buen analista de datos debe conocer ambos extremos del espectro abarcado entre:

Modelación estadística con planteo y manejo fino de modelos de probabilidad a la medida

Algoritmo computacional sofisticado

Con ello modelará eficientemente y aprovechará lo mejor de los dos mundos en la proporción conveniente para cada problema de interés.