KSCHOOL

Del Dato al Conocimiento

Introducción a la Modelización Estadística

Antonio Pita Lozano

Máster en Data Science

Sobre mi

KSCHOOL — *Del Dato al Conocimiento*

Director de Operaciones y Soluciones en *Synergic Partners*

19 años de experiencia laboral:

synergic partners — A Telefónica COMPANY — everis — an NTT DATA Company — BCC Banco de Crédito Cooperativo — Liberbank — cajamar — Universidad de Murcia

Mejor Científico de Datos de España 2016 (1ª edición)

DATA SCIENCE AWARDS SPAIN 2016

RECONOCIENDO EL TALENTO ANALÍTICO EN ESPAÑA

EFPA España — European Financial Planning Association

Asesor financiero europeo (EFA) Nº 9520

kaggle

Kaggler:
466/515.453

19 años de experiencia docente:

Universidad de Almería — School — CIFF Centro Internacional de Formación Financiera — Afi Escuela de Finanzas — UOC — Universitat Oberta de Catalunya — Grupo IOE — Academia Nota — EOI Escuela de organización industrial — ie School of Human Sciences & Technology — coursera — escuelaformaciónfinanciera

En continuo aprendizaje:

- Licenciado en Matemáticas (2001)
- DEA en Algebra (2003)
- Máster en Administración de Empresas y Marketing (2009)
- Máster en Derecho Comunitario (2010)
- Master en Asesoramiento Financiero (2012)
- Experto Universitario en Estadística Aplicada (2013)
- Máster en Visual Analytics and Big Data (2014)
- Machine Learning Specialization (2016)

https://antoniopitablog.wordpress.com

Modelización Estadística
Regresión Lineal

1. Componentes
2. Supuestos de la Regresión Lineal
3. Interpretación de Coeficientes
4. Evaluación de los modelos
5. Multicolinealidad y Análisis de Residuos
6. Interpretación en R
7. Análisis de Cambios Estructurales

Modelización Estadística

Las técnicas de modelización estadística buscan encontrar la distribución estadística que mejor representa a los datos para estudiar las relaciones entre las variables.

Principales técnicas:

- Regresión Lineal (LM) – Series Temporales
- Modelos Lineales Generalizados (GLM) – Logística, Poisson, GAM…
- Modelos Lineales Robustos
- Modelos Gráficos Probabilísticos (PGM)—Redes Bayesianas, Redes de Markov

Regresión Lineal

Objetivo

Estimar la relación entre una variable dependiente (variable explicada) y varias variables independientes (variables explicativas) mediante una expresión lineal en coeficientes con el objetivo de contrastar teorías y estimar efectos entre las variables.

Desarrollo

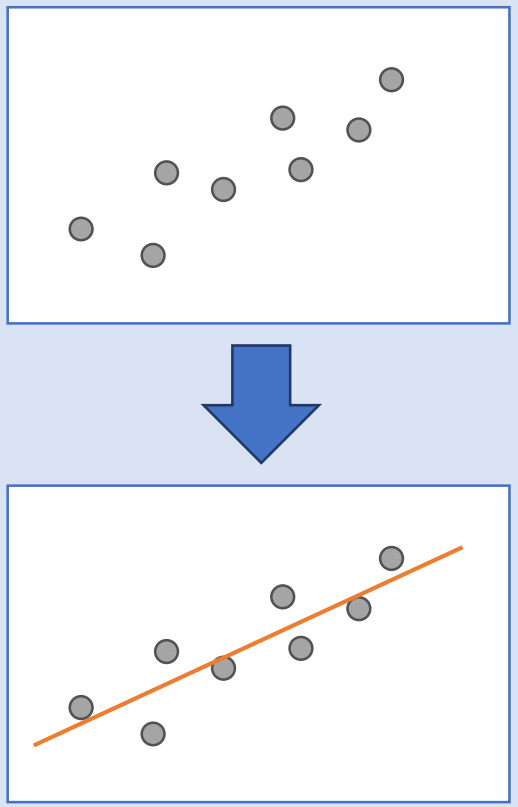
Para la estimación de una relación líneal es necesario establecer el modelo a estimar, que será una combinación lineal de los regresores. La diferencia con la variable dependiente (que debe ser numérica real) se denomina residuo:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

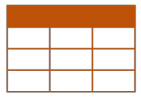
La estimación se realiza utilizando el estimador MCO y las estimaciones se obtienen resolviendo un sistema de ecuaciones.

La solución obtenidas es el mínimo global que minimiza la suma de los residuos al cuadrado

(*) en esta sesión sólo se considerará el modelo con término constante



Regresión Lineal: Componentes



Datos

Modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

Función de coste

$$g(\beta_0, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2$$

Algoritmo estimador

Mínimos Cuadrados Ordinarios (MCO)

Regresión Lineal: Teorema de Gauss-Markov

Supuestos ampliados de la Regresión Lineal Múltiple (RLM)

- RLM 1. Modelo lineal en parámetros
- RLM 2. Muestreo aleatorio
- RLM 3. Media condicionada nula

$$E(\varepsilon_i | x_{11}, x_{21}, \cdots, x_{(k-1)N}, x_{kN}) = E(\varepsilon_i) \quad \forall i = 1, \cdots, N$$

- RLM 4. No multicolinealidad perfecta
- RLM 5. Homocedasticidad

$$V(\varepsilon_i | x_{11}, x_{21}, \cdots, x_{(k-1)N}, x_{kN}) = V(\varepsilon_i) = \sigma^2$$

Los estimadores son los **estimadores lineales insesgados óptimos** (ELIO)

Insesgado: La esperanza del estimador coincide con el valor poblacional

Óptimo: El estimados es el de menor varianza entre los insesgados lineales

Regresión Lineal: Asunciones débiles

Supuestos de la Regresión Lineal Múltiple (RLM)

- RLM 1. Modelo lineal en parámetros
- RLM 2. Muestreo aleatorio
- RLM 3. Media condicionada nula

$$E(\varepsilon_i | x_{11}, x_{21}, \cdots, x_{(k-1)N}, x_{kN}) = E(\varepsilon_i) \quad \forall i = 1, \cdots, N$$

- RLM 4. No multicolinealidad perfecta

→ Los estimadores son **insesgados** (no tienen sesgo)

Supuestos débiles la Regresión Lineal Múltiple (RLM)

- RLM 1. Modelo lineal en parámetros
- RLM 3'. Exogeneidad débil

$$E(\varepsilon_i | x_{1i}, x_{2i}, \cdots, x_{ki}) = E(\varepsilon_i) \ \forall i \ \text{con} \ i = 1, \cdots, N$$

- RLM 4. No multicolinealidad perfecta

→ Los estimadores son **consistente** (en el límite no tienen sesgo)

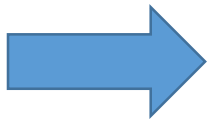
Big Data → **Sesgo despreciable y varianza casi nula**

Regresión Lineal: Interpretación de Coeficientes

Los coeficientes son efectos **ceteris paribus** si se cumple el supuesto de media condicionada nula:

$$E(\varepsilon/x_1, x_2, \ldots, x_k) = E(\varepsilon) = 0$$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$



β_i

Incremento de y al incrementar 1 unidad de x_i manteniendo el resto de variables dependientes constantes.

Regresión Lineal: Modelos en Niveles y Logaritmos

Se pueden estudiar efectos diferentes utilizando otras construcciones lineales

Modelo	Variable Dependiente	Variable Independiente	Interpretación
Regresión Level-Level $y = \beta_0 + \beta_1 x + \epsilon$	y	x	Un aumento de 1 unidad en x se corresponde con un aumento de beta unidades en y. **(efecto marginal)**
Regresión Log-Level $ln(y) = \beta_0 + \beta_1 x + \epsilon$	ln(y)	x	Un aumento de 1 unidad en x se corresponde con un aumento del 100*beta% en y. **(semielasticidad)**
Regresión Level-Log $y = \beta_0 + \beta_1 \cdot ln(x) + \epsilon$	y	ln(x)	Un aumento del 1% en x se corresponde con un aumento de beta/100 unidades en y.
Regresión Log-Log $ln(y) = \beta_0 + \beta_1 \cdot ln(x) + \epsilon$	ln(y)	ln(x)	Un aumento del 1% en x se corresponde con un aumento de beta% en y. **(elasticidad)**

Regresión Lineal: Descomposición de la Varianza

- STC= Suma total de cuadrados $$STC = \sum_{i=1}^{N} (y_i - \bar{y})^2$$
- SCE = Suma cuadrados de los errores $$SCE = \sum_{i=1}^{N} e_i^2$$
- SCR = Suma cuadrado de la regresión $$SCR = \sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

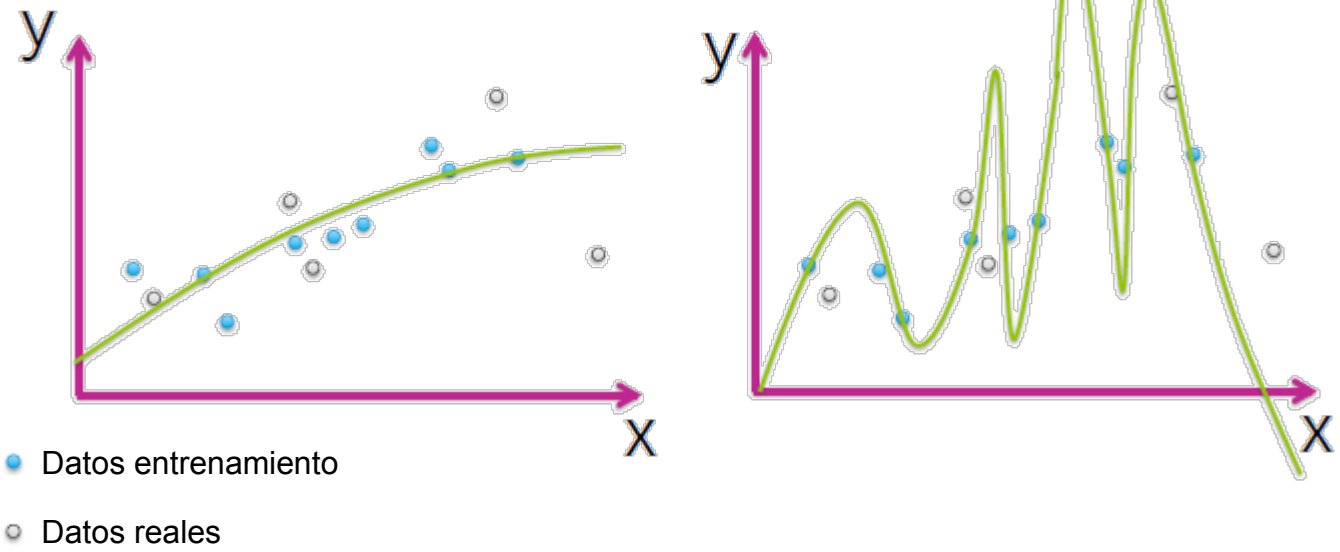
STC = SCR + SCE

Varianza Total | Varianza Explicada | Varianza Residual

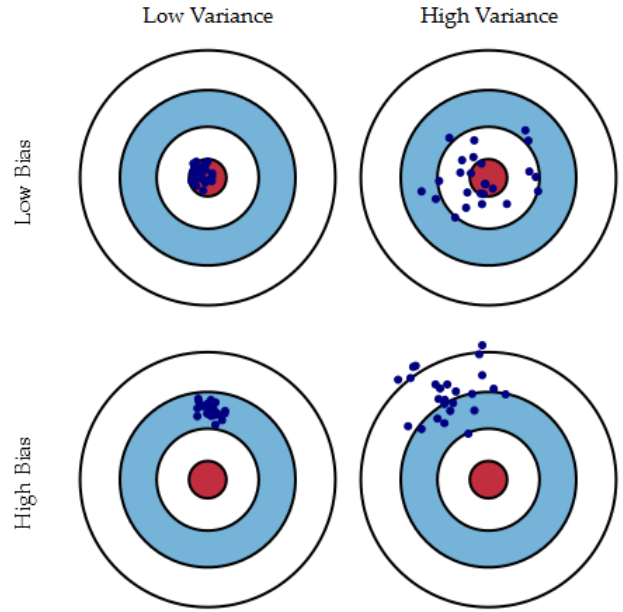
Coeficiente de Determinación $$R^2 = \frac{SCR}{STC}$$

Regresión Lineal: Overfitting

Con el objetivo de aumentar la varianza explicada por la regresión, podemos introducir nuevas variables, pero tenemos que tener cuidado que no produzcan sobre-ajustar (overfitting) y aumento de la varianza de los estimadores.



Bias - Variance Tradeoff



Regresión Lineal: Comparativa de Modelos

Comparativa de Modelos:

Existen diferentes técnicas para comparar modelos lineales en función a sus características.

Estas métricas tratan de penalizar la mejora del modelo al introducir una variable irrelevante.

RELATIVOS

R^2

$$R^2 = 1 - \frac{SCE}{STC} = \frac{SCR}{STC}$$

R^2 ajustado

$$R_a^2 = 1 - \frac{N-1}{N-K}\frac{SCE}{STC}$$

Se elige el modelo con valor más alto

ABSOLUTOS

BIC Bayesian Information Criterion

-2 Ln(likelihood) + 2K* Ln(N)

Se elige el modelo con valor más bajo

AIC Akaike Information Criterion

-2 Ln(likelihood) + 2K

Se elige el modelo con valor más bajo

COMPARATIVO

Contraste F

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-K)}$$

Se elige el modelo NR si se rechaza el test

Regresión Lineal: Multicolinealidad

Cuando dos de los variables explicativas están muy correlacionados entre sí puede provocar el aumento de la varianza de los coeficientes estimados del modelo y la no convergencia del estimador.

Como detectarlo:

- Coeficientes demasiado elevados sin interpretación.
- Falta de convergencia de los parámetros.
- Correlación elevada entre variables explicativas.

Es mas complicado identificar la colinealidad entre mas de dos variables

Técnicas para mitigar la multicolinealidad
- Regression rich lasso
- Reduccion de dimensionalidad

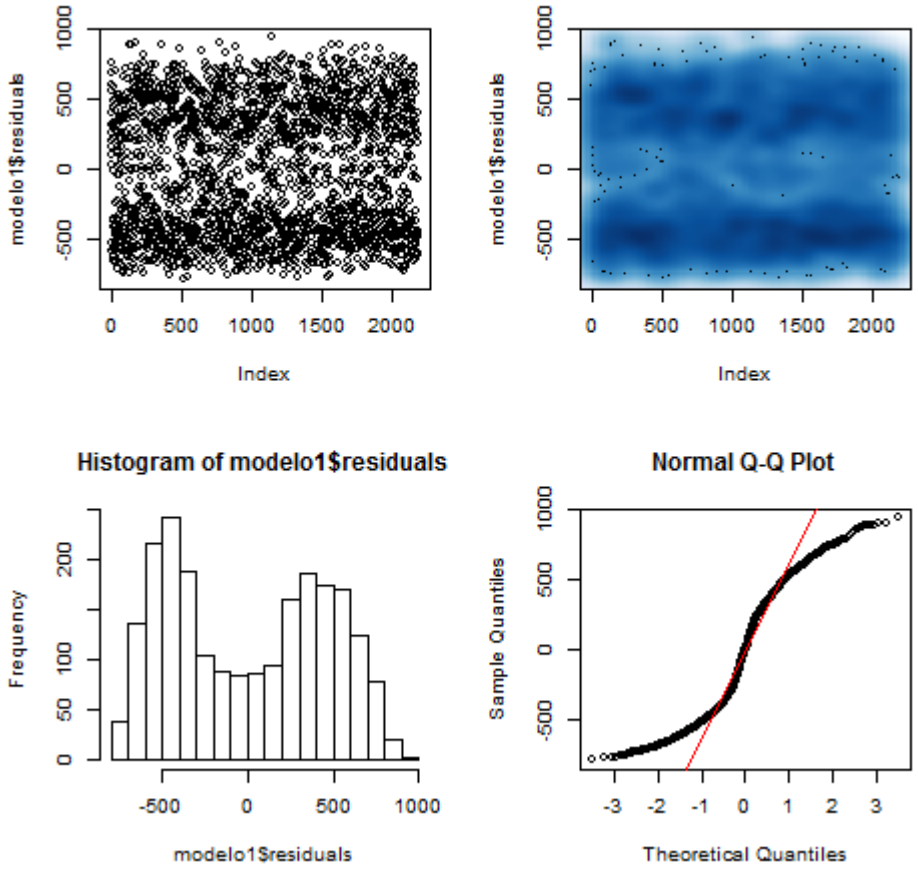
Regresión Lineal: Análisis de los residuos

Análisis de Residuos:

Si la distribución de los residuos no es una normal con media cero existen factores a introducir en el modelo que aportan información.

Esto disminuye la varianza y aparentemente las estimaciones son más precisas.

Sino tiene media cero,
Hay algún elemento que influye en Y y que no hemos capturado



Regresión Lineal: Interpretación en R


```
Call:
lm(formula = Cantidad ~ Precio, data = Ventas)

Residuals:
    Min      1Q  Median      3Q     Max
-560.1 -127.6  -14.7   123.9   629.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3304.878     15.778   209.5   <2e-16 ***
Precio      -251.793      2.494  -101.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 192.5 on 2188 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.8232
F-statistic: 1.019e+04 on 1 and 2188 DF,  p-value: < 2.2e-16
```

Coeficientes estimados

Contraste t:

$$t = \frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$$

Determinación:

$$R^2 = 1 - \frac{SCE}{STC} = \frac{SCR}{STC}$$

$$R_a^2 = 1 - \frac{N-1}{N-K}\frac{SCE}{STC}$$

Contraste F:

$$F = \frac{(SCE_R - SCE_{NR})/q}{SCE_{NR}/(N-K)}$$

Donde NR es el modelo no restringido y R el modelo restringido. N el tamaño total de la muestras, K el número de parámetros del modelo NR y q el número de restricciones. En ejemplo: q=K y SCE(R)=

Regresión Lineal: Análisis de Cambios Estructurales

Existe cambio estructural cuando los valores numéricos de los parámetros poblacionales no son iguales en submuestras diferentes.

$$y_i = \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + \ldots + \beta_k^{(1)} x_{ki} + \varepsilon_i$$

$$y_i = \beta_0^{(2)} + \beta_1^{(2)} x_{1i} + \ldots + \beta_k^{(2)} x_{ki} + \varepsilon_i$$

Test de Chow

$$H_0 : \beta_0^{(1)} = \beta_0^{(2)}, \ldots, \beta_k^{(1)} = \beta_k^{(2)}$$

$$H_1 : \text{No } H_0$$

- Hay el mismo efecto de una variable en una ciudad y en un pueblo?
- También están asociados a cambios de normativa
- Si las fórmulas para las dos poblaciones (estructuras) no son iguales, la estructura puede influir.
- Atención: Las igualdad de variables se mide estadísticamente

¡ ES UN CONTRASTE F!

KSCHOOL

Del Dato al Conocimiento

Introducción a la Modelización Estadística

Antonio Pita Lozano

Máster en Data Science

https://www.linkedin.com/in/antoniopitalozano/

@anto_pita