Análisis Multivariado

Felipe Morales Apablaza

Universidad Adolfo Ibañez

October 9, 2019

Introducción

- ¿Qué harían si supieran que el dólar el día de mañana aumentará en 20 pesos su precio?

Introducción

- ¿Y si supieran que el día de mañana va a llover?

Introducción

- Si supiésemos el futuro podríamos tomar mejores decisiones el día de hoy.
- Es por esto que académicos y cientistas de datos, a través de teorías y/o algoritmos, han dedicado esfuerzos por predecir eventos futuros de interés.

Pronosticando en la práctica: Ventas

1968

Sales Forecasting Using Exponential Smoothing

Bruce Nicholas Anez
University of Rhode Island

Pronosticando en la práctica: Cambio Climático

Ahora o nunca: nueva predicción sobre el cambio climático

El nuevo informe de la ONU modificó el pronóstico del cambio climático. Si antes la meta era no aumentar 2 grados centígrados, ahora es no pasar de 1,5. Las consecuencias están a la vuelta de la esquina. Los expertos explican por qué medio grado en la temperatura sí importa.

Pero... ¿Cómo pronosticamos?

En extremo, podemos distinguir entre dos tipos de enfoques:

- **Modelos teóricos:** conjunto organizado de ideas que explican un fenómeno, deducidas a partir de la observación, la experencia o el razonamiento lógico.
- **Inteligencia Artificial:** uso de algoritmos que, sin necesariamente explicar los mecanismos, buscan pronosticar.

Modelos teóricos: Ecuación de Mincer

- En Microeconomía I les enseñaron que el salario de los trabajadores es igual a $PMgL$ (recuerde que $PMgL = w$).
- Si las personas son más productivas, entonces su salario más alto.
- Entonces, si la educación nos vuelve más productivos, podríamos pronosticar salarios con la información educacional de las personas.

Inteligencia Artificial: Pronosticando ventas de hamburguesas

En lugar de preocuparnos de la teoría que explica las ventas de hamburguesas podríamos preocuparnos de otras cosas para pronosticar:

- ¿Las ventas del mes pasado?
- ¿El promedio de ventas del año pasado?
- ¿Las ventas de la semana pasada?
- ¿Existe algún patrón que siguen las ventas? Quizás podríamos utilizar algoritmos para pronosticar.

Y... ¿Entonces?

- En la práctica, obviamente, se pueden realizar pronósticos utilizando una mezcla de ambos enfoques.
- En este tópico del curso se comenzará la discusión sobre métodos de pronóstico.

¿Cómo pronosticar?

- Supongamos que se nos ha asignado la tarea de entrevistar a los nuevos postulantes de un magíster de la UAI que nos esperan en la sala del lado.
- Si tuviesemos que pronosticar la edad del primer entrevistado cómo lo harían?
- Una alternativa razonable podría ser el promedio de los actuales estudiantes del magíster...

¿Cómo pronosticar?

- Supongamos que se nos ha asignado la tarea de entrevistar a los nuevos postulantes de un magíster de la UAI que nos esperan en la sala del lado.
- Si tuviesemos que pronosticar la edad del primer entrevistado cómo lo harían?
- Una alternativa razonable podría ser el promedio de los actuales estudiantes del magíster...

¿Cómo pronosticar?

- ¿Y si además les digo que escucha Marco Antonio Solis por las tardes?

¿Cómo pronosticar?

- ¿Y si también les digo que es calvo?

¿Cómo pronosticar?

- Si tenemos información de estas tres variables, quizás, puede que nuestro mejor pronosticador es el promedio de edad de los estudiantes del magíster de este año condicionado a que estamos hablando de estudiantes calvos fanáticos de Marco Antonio Solis.
- ¿Y si incluímos más variables? Deberíamos condicionar por estas variables extras a la hora de calcular el promedio

¿Cómo pronosticar?

- ¿Y si a la hora de condicionar por tantas variables a la hora de calcular un promedio no existen datos?
- Hay un método que incluso nos permite hacer este cálculo con individuos de características inexistentes
- A este método lo llamaremos *Mínimos Cuadrados Ordinarios (MCO).*
- ¡Este método nos permitirá utilizar toda la información disponible para calcular promedios incluso para individuos que no existen!

Enfoque de los Mínimos Cuadrados Ordinarios

- Denotaremos como Y nuestra variable de interés a pronosticar
- X corresponderá al vector de predictores que tenemos disponibles donde $X = (X_1 X_2 ... X_k)$
- u corresponderá a nuestro término del error y corresponderá a todos lo factores que afectan a Y que no son $X_1, X_2, ..., X_k$

Enfoque de los Mínimos Cuadrados Ordinarios

Vamos a suponer que la relación entre X y Y puede ser representada de la siguiente forma:

$$Y = f(X) + u$$

- Donde f es una función desconocida dependiente de $X_1, X_2, ..., X_k$
- En esta representación f representa la información sistemática que entrega X sobre Y.

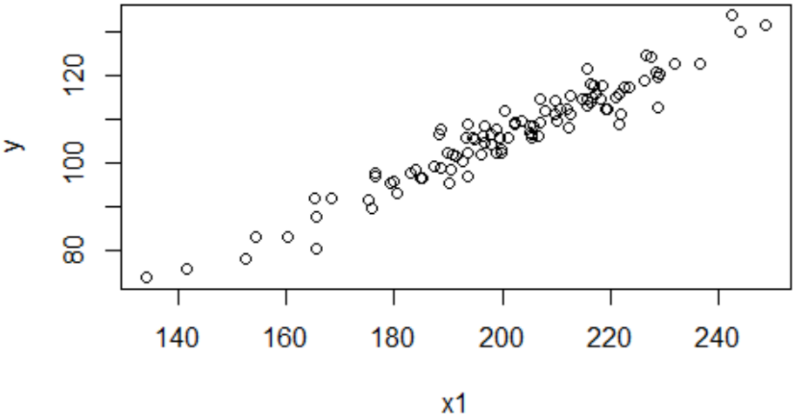
Ejemplo práctico: simulemos datos

```r
# Creando variables de forma aleatoria:

x1<- rnorm(100, mean=200, sd=20)
x2<- rnorm(100, mean=30, sd=15)
u<- rnorm(100, mean=0, sd=1)

# ¿Qué va a ser "Y"?
y<- 0.4 + 0.5*x1 + 0.2*x2 + u
```

Ejemplo práctico: simulemos datos

```r
# Pegando los datos:
datos<- data.frame(cbind(y,x1,x2,u))
colnames(datos)<- c("y","x1","x2","u")
plot(x=x1 , y=y , data=datos)
```

Ejemplo práctico: simulemos datos

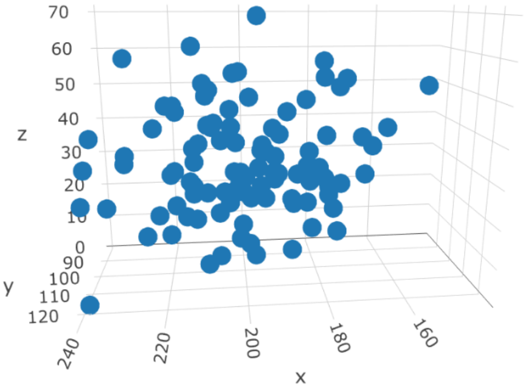


¿Y si ploteamos esto en términos tridimensionales?

Con el paquete *plotly* podemos graficar tridimensionalmente de forma sencilla.

```
# ¿Podemos incluir más?
library(plotly)
plot_ly(data=datos, x = x1, z = x2, y = y)
```

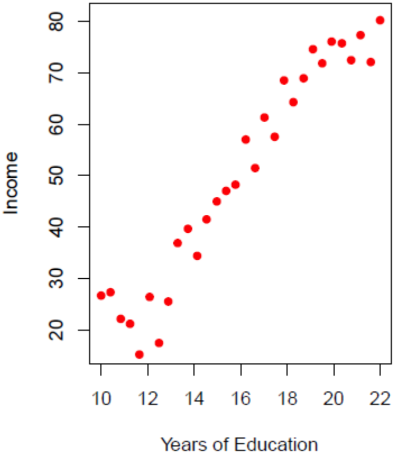
¿Y si ploteamos esto en términos tridimensionales?



Enfoque de los Mínimos Cuadrados Ordinarios

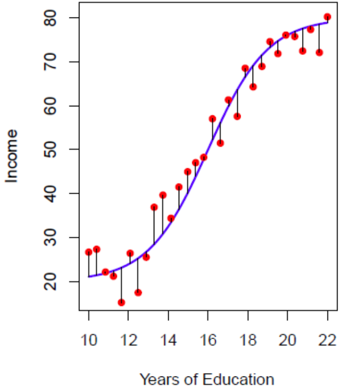
- Dado que la función f que relaciona a X con Y generalmente es una función desconocida, debemos estimarla con información observable.
- A la estimación de f la denotaremos como $\hat{f}$
- Por ejemplo, podríamos realizar una encuesta a 30 individuos y preguntarles su ingreso (income) y los años de educación.

Enfoque de los Mínimos Cuadrados Ordinarios



Enfoque de los Mínimos Cuadrados Ordinarios

- Nuestro objetivo será encontrar la curva f que nos identifica la relación entre ambas variables:



Enfoque de los Mínimos Cuadrados Ordinarios

- Para simplificar la discusión sobre f, vamos a suponer que f está dado por la siguiente forma funcional:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{1}$$

- donde β_0, β_1, ... , β_k corresponden a parámetros desconocidos de interés.
- Nuestro interés será estimar los β para encontrar una estimación sobre f
- Como verán más adelante el suponer linealidad en f no es tan restrictivo a la hora de realizar pronósticos.

Enfoque de los Mínimos Cuadrados Ordinarios

- Definiremos a nuestra función de predicciones estimada, $\widehat{f}(X)$, como $\widehat{f}(X) = \hat{y}$.
- Las predicciones de nuestro modelo se definirán como:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + ... + \widehat{\beta}_k X_{ik}$$

Nuestro error de predicción, $\hat{u}_i$, se definirá como:

$$\hat{u}_i = y_i - \hat{y}_i$$

- $\hat{u}_i$ medirá la diferencia entre los valores originales de la variable y_i con los pronósticos que realizamos, $\hat{y}_i$

Enfoque de los Mínimos Cuadrados Ordinarios

- Utilizaremos una muestra de N datos para estimar f
- Definiremos una *función de pérdida* para medir cuánto nos equivocamos realizando las predicciones para los N individuos.
- A esta función la denominaremos Suma de Cuadrado de Residuos (SCR) y estará definida como:

$$SCR = \sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

$$SCR = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - ... - \hat{\beta}_2 X_{ik})^2 \tag{3}$$

- El objetivo de MCO será elegir los $\hat{\beta}$ que minimizan la SCR.

Mínimos Cuadrados Ordinarios

- Matemáticamente es posible demostrar que:

$$\hat{\beta} = (X^T X)^{-1}(X^T y) \tag{4}$$

donde $\hat{\beta}$ es un vector de $(K + 1) * 1$.

- Si bien la estimación puede ser matemáticamente desafiante, el cálculo se puede realizar en R de forma muy sencilla.

Mínimos Cuadrados Ordinarios

lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments

`formula` an object of class "`formula`" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

Mínimos Cuadrados Ordinarios

Cargaremos dos paquetes de trabajo:

```
library(AER)
library(scales)
```

Y cargaremos la base de datos *CASchools*:

```
# load the `CASchools` dataset
data(CASchools)
```

Mínimos Cuadrados Ordinarios

- Esta base de datos contiene información de colegios en múltiples aspectos: número de profesores, estudiantes, precio del colegio, entre otras variables.

district	school	county	grades	students	teachers	calworks	lunch
75119	Sunol Glen Unified	Alameda	KK-08	195	10.900	0.5102	2.0408
61499	Manzanita Elementary	Butte	KK-08	240	11.150	15.4167	47.9167
61549	Thermalito Union Elementary	Butte	KK-08	1550	82.900	55.0323	76.3226
61457	Golden Feather Union Elementary	Butte	KK-08	243	14.000	36.4754	77.0492
61523	Palermo Union Elementary	Butte	KK-08	1335	71.500	33.1086	78.4270
62042	Burrel Union Elementary	Fresno	KK-08	137	6.400	12.3188	86.9565
68536	Holt Union Elementary	San Joaquin	KK-08	195	10.000	12.9032	94.6237
63834	Vineland Elementary	Kern	KK-08	888	42.500	18.8063	100.0000
62331	Orange Center Elementary	Fresno	KK-08	379	19.000	32.1900	93.1398
67306	Del Paso Heights Elementary	Sacramento	KK-06	2247	108.000	78.9942	87.3164
65722	Le Grand Union Elementary	Merced	KK-08	446	21.000	18.6099	85.8744
62174	West Fresno Elementary	Fresno	KK-08	987	47.000	71.7131	98.6056

Mínimos Cuadrados Ordinarios

Generaremos dos variables:

- STR : ratio de estudiantes por cada profesor.
- *score*: puntaje promedio del curso entre matemática y lectura.

```
# add student-teacher ratio
CASchools$STR <- CASchools$students/CASchools$teachers

# add average test-score
CASchools$score <- (CASchools$read + CASchools$math)/2
```

Mínimos Cuadrados Ordinarios

- Podemos predecir la relación entre el puntaje (*score*) y el ratio de estudiantes/profesores (STR):

$$score_i = \beta_0 + \beta_1 STR_i + u_i \tag{5}$$

- En R lo podemos hacer con la siguiente codificación:

```
linear_model <- lm(score ~ STR, data = CASchools)
```

Resultados de regresión

```
Call:
lm(formula = score ~ STR, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
STR          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,	Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Interpretación

Notemos los siguientes resultados:

- $\widehat{\beta}_0 = 698.9$: el $score_i$ predicho para un colegio con $STR = 0$ es de 698.
- $\widehat{\beta}_1 = -2.2798$: por cada aumento en una unidad de STR disminuirá en 2.2798 el $score$.

Predicciones

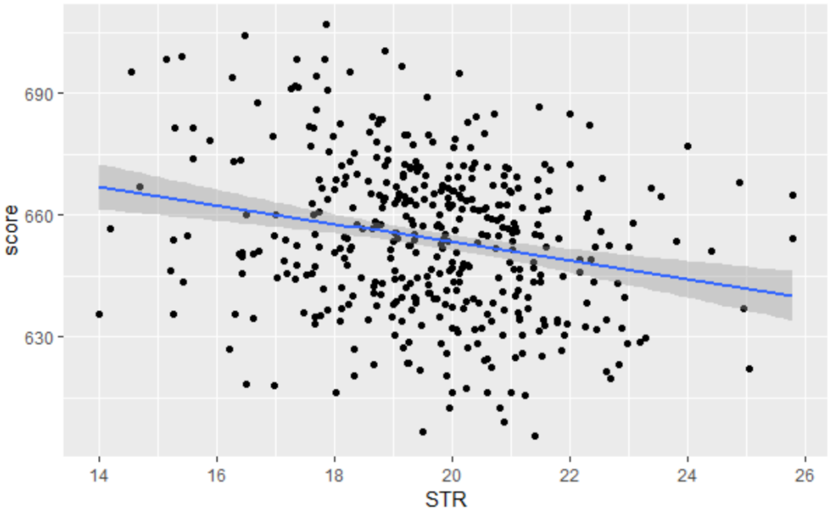
- Con la función *predict()* podemos utilizar nuestras estimaciones para realizar predicciones con el modelo.
- Generaremos una variable *predicciones* que será la predicción para cada observación con la que se realiza la muestra.

```
CASchools$predicciones<- predict(linear_model)
```

¿Cómo se ve esto gráficamente?

```r
# Estimaciones:
ggplot(CASchools, aes(x= STR, y=score)) +   geom_point() +   stat_smooth(method = lm)
```

¿Cómo se ve esto gráficamente?



¿Cuánto nos equivocamos?

También podemos calcular nuestros errores de predicción o residuos con la función *resid()*:

```
# Residuos:
CASchools$residual<-    resid(linear_model)
```

¿Y si incluimos más variables?

Podemos pronosticar con más variables que STR, por ejemplo, incluir variables como *income* y *expenditure*. Es decir, vamos a proponer un modelo como:

$$score_i = \beta_0 + \beta_1 STR_i + \beta_2 income_i + \beta_3 expenditure_i + u_i \quad (6)$$

- En R esta estimación la podemos hacer de forma sencilla:

```
linear_model2<- lm(score ~ STR + income + expenditure , data= CASchools)
```

¿Y si incluimos más variables?

```
Call:
lm(formula = score ~ STR + income + expenditure, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.926  -8.805   0.106   9.083  32.567 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 669.745072  13.973921  47.928  < 2e-16
STR          -1.325765   0.436846  -3.035  0.00256
income        1.894375   0.094534  20.039  < 2e-16
expenditure  -0.003495   0.001336  -2.616  0.00922

(Intercept) ***
STR         ** 
income      ***
expenditure ** 
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.26 on 416 degrees of freedom
Multiple R-squared:  0.5194,	Adjusted R-squared:  0.5159 
F-statistic: 149.9 on 3 and 416 DF,  p-value: < 2.2e-16
```

¿Y si incluimos más variables?

- ¿Cómo podemos pronosticar? Con la función *predict()* como lo hicimos anteriormente
- ¿La interpretación? Veamos...

Análisis Multivariado

Felipe Morales Apablaza

Universidad Adolfo Ibañez

October 9, 2019