EL MODELO LINEAL

Clase 1

Introducción

Sergio Camiz

LIMA - Marzo-Mayo 2025

¿Quién soy?

- Licenciado en Matemáticas en 1969 (Università di Roma)
- Profesor desde 1975
- Doctor en Análisis de datos desde 2002 (Université Paris Dauphine)
- 170 artículos en revistas, libros o actas de congresos
- Fotógrafo profesional en los años 1970
- Página web: www.camiz.it

- Música, Montaña, Vela...

Programa

1. Introducción: Análisis de datos, estadística y modelización
2. Regresión lineal simple, su análisis, estimación y propiedades
3. Regresión lineal múltiple, estimación y análisis de la varianza
4. Propiedades estadísticas, análisis de la varianza, inferencia
5. Partición de la regresión, estadísticas, selección del modelo
6. Falta de ajuste, diagnósticos, residuos, puntos aberrantes
7. Análisis de la varianza, análisis de la varianza de dos vías
8. Análisis de la covarianza, pruebas post-hoc

Estructura del examen

Para el examen, se evalúan los resultados de los ejercicios realizados durante el curso y se requiere, al final, un trabajo en forma de artículo científico. Este trabajo consistirá en un estudio sobre datos encontrados en internet, en el trabajo o en una investigación personal, que requiera para su desarrollo una regresión múltiple.

El alumno es libre de elegir el problema y los datos que prefiera, pero el archivo no puede utilizar series de tiempo, ya que en este caso los datos no son independientes, y deben tener al menos 30 observaciones y 5 variables.

Bibliografía

- Camiz, S., 2010. *El modelo lineal*. Notas de curso. Lima.
- Faraway, J., 2005. *Linear models with R*. Chapman & Hall.
- Fahrmeir, L., Kneib, T. y S. Lang, 2013. *Regression: Models, Methods, and Applications*. Berlin, Springer.
- Fox, J. y S. Weisberg, 2011. *An R Companion to Applied Regression*, Thousand Oaks (California), Sage.
- Guttman, I., 1982. *Linear Models: An Introduction*. New York, Wiley.
- Langrand, C. y L.M. Pinzón, 2009. *Análisis De Datos. Métodos y ejemplos*, Bogotá, Escuela Colombiana de Ingeniería Julio Garavito.
- Montgomery, D.C. y E.A. Peck, 1982. *Introduction to Linear Regression Analysis*. New York, J. Wiley.

Asuntos de la Clase 1

- Herramientas
- Ciencia y estadística
- Ejemplo

Linux

Se recomienda el empleo de software libre, que es gratuito, bien actualizado y que se encuentra en todos los sistemas.

- *Linux* el sistema operativo más estable y poderoso. Hay miles de distribuciones: Debian, Ubuntu, Linux Mint, ...
- *Linux Mint Cinnamon* el más parecido a Windows, fácil.
- *Libre Office* el equivalente de Microsoft Office.
- *LaTeX* para la escritura de textos científicos.
- *Texstudio* entorno de escritura para LaTeX.
- *R* el mejor sistema de cálculo y programación estadísticos.
- *Rstudio* entorno de desarrollo para R.
- *Rmarkdown* sistema en R para desarrollar procedimientos comentados, informes, etc.

El método científico

- El conocimiento científico no es estructuralmente diferente del conocimiento común, sino por el mayor control sobre lo que se hace y el rigor lógico del pensamiento consecuente.
- Esto se basa en observaciones repetidas de un fenómeno, con el fin de entender tanto los aspectos comunes como las diferencias entre las observaciones.
- Las observaciones producen datos, cuyo estudio comparado permite generar hipótesis sobre el funcionamiento del fenómeno mismo.
- *"La Estadística tiene como objeto la recolección de datos para su análisis e interpretación"* (M. Carbon). Por lo tanto, se trata de la herramienta básica para la investigación científica.

Colección y análisis de datos, estadística y modelización

- Cómo hacer una *observación* depende fuertemente del enfoque científico y de la cultura del investigador.
- Las observaciones tienen que ser repetidas de la misma manera, ya que un cambio podría afectar a los resultados.
- Los datos observados se tratan de diferente manera, según la etapa de la investigación:
  - *exploratoria*: se busca una síntesis informativa sobre estructuras y relaciones.
  - *confirmatoria*: se intenta determinar la fiabilidad de los resultados encontrados y se espera una *inferencia*.
  - *modelización*: descripción, simulación y predicción.

La etapa exploratoria

- Se recolectan datos.
- Se estudian con *estadísticas descriptivas* para:
  - síntesis de la información de cada variable.
  - control de calidad de los datos.
- Se realizan *análisis exploratorios* (*geométricos*) para:
  - "hacer hablar a los datos" (Benzécri, 1982).
  - buscar estructuras y relaciones.
  - calcular estadísticas, parámetros que los describen.
  - buscar *factores* para ordenar los datos.
  - buscar *particiones* para estructurar los datos en clases.
- Se formulan hipótesis.

La etapa confirmatoria

- Se deben probar hipótesis o tomar decisiones.
- Se recolectan datos con *muestras*, tomadas de la *población de referencia* de acuerdo con un *plan experimental* compatible con las tareas que se deciden.
- Se estudian los datos con técnicas *estadísticas*, para:
  - *probar* hipótesis.
  - *confirmar* las hipótesis establecidas.
  - *estimar* parámetros.
  - definir *intervalos de confianza*.
- Se realiza *inferencia estadística* a la población de referencia.

La modelización

- Un modelo matemático es una simplificación de la realidad, que solo explica una parte del fenómeno que se está estudiando.
- Esto tiene que ser compatible con los resultados conseguidos en las etapas anteriores.
- Un modelo *teórico* explica el fenómeno desde el punto de vista de las relaciones causales.
- Un modelo *estadístico* solo muestra la intensidad de las relaciones que se estudian, sin considerar la causalidad.
- Empleando el modelo se puede *simular* el fenómeno, hacer previsiones y simular datos.

En conclusión, se trata de una adquisición cultural importante.

- La secuencia de las tres etapas puede conducir a la construcción de un modelo sin los riesgos de utilizar un modelo definido *a priori*, sin una crítica eficaz (Benzécri *et al.*, 1982).
- En la práctica, tal vez no se espere un *modelo teórico*, sino un *modelo estadístico* que no pretenda explicar el fenómeno, sino que solo permita estimar valores que de otro modo no serían conocidos.

Ejemplo: si se encuentra una relación fuerte entre el consumo de electricidad y la radiación solar, se puede evaluar la radiación a través del consumo de electricidad, aun cuando el consumo de electricidad no tiene efecto sobre la actividad del sol.

La estructura correspondiente a una tabla de datos es una hoja electrónica, donde:

- por filas están las unidades.
- por columnas están los caracteres.
- en cada celda hay un dato.
- los márgenes tienen las etiquetas de las unidades y de los caracteres.

En *R*, la estructura correspondiente es un `data.frame`, una especie de matriz con columnas que pueden ser de diferente tipo (cuantitativo, cualitativo, lógico, etc.).

Lo que se observa se organiza formalmente en una *tabla de datos*.

- Llamaremos a cada objeto examinado una *unidad estadística* o un *individuo*.
- Llamaremos *caracteres* o *variables* a los diversos elementos que se decide observar de manera repetitiva sobre las unidades estadísticas.
- Llamaremos *dato* a la modalidad con la cual un carácter se observó sobre una unidad.
- Llamaremos *observación* al conjunto de datos de los diversos caracteres observados sobre la misma unidad.
- Llamaremos *muestra* al conjunto de observaciones homogéneas realizadas sobre un conjunto de unidades.

Ejemplo

Para 13 países de América Latina, se supone una relación lineal entre la *tasa de urbanización* (x) y un *indicador de desarrollo humano* (IDH), creado por el economista paquistaní Mahbub ul Haq en 1990, dependiendo del producto bruto, la esperanza de vida, la alfabetización y la escolarización (y). Los datos deben estar en el data frame `SudAmerica`.

País	*x = Tasa de urbanización*	*y = IDH*
Argentina	86	0.833
Bolivia	51	0.394
Brasil	75	0.739
Chile	86	0.863
Colombia	70	0.758
Ecuador	56	0.641
Guyana	35	0.539
Panamá	54	0.731
Paraguay	48	0.637
Perú	70	0.600
Suriname	48	0.749
Uruguay	86	0.880
Venezuela	84	0.824

Lo tenemos como archivo `SudAmerica.csv`:

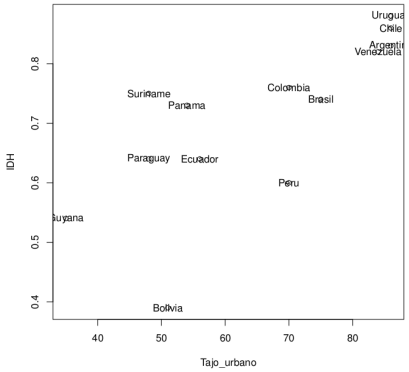
```
"","Tajo_urbano","IDH"
"Argentina",86,0.833
"Bolivia",51,0.394
"Brasil",75,0.739
"Chile",86,0.863
"Colombia",70,0.758
"Ecuador",56,0.641
"Guyana",35,0.539
"Panama",54,0.731
"Paraguay",48,0.637
"Peru",70,0.600
"Suriname",48,0.749
"Uruguay",86,0.880
"Venezuela",84,0.824
```


Lo podemos cargar en R, una vez que lo tenemos en una carpeta <dir> (atención: en Windows, el carácter \ debe escribirse como //):

```
> setwd("<dir>")
> SudAmerica <- read.csv("SudAmerica.csv",
                         row.names=1)
> attach(SudAmerica)
> lab<-rownames(SudAmerica)
> plot(Tajo_urbano,IDH)
> text(Tajo_urbano,IDH,lab)
> dev.copy(pdf,"SudAmerica1.pdf")
> dev.off()
```

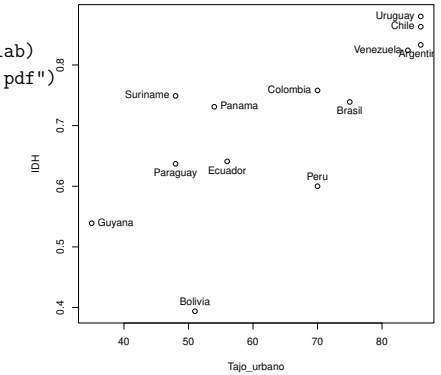
Las dos instrucciones finales permiten de guardar el gráfico en .pdf.

Ya conseguimos esta imagen:



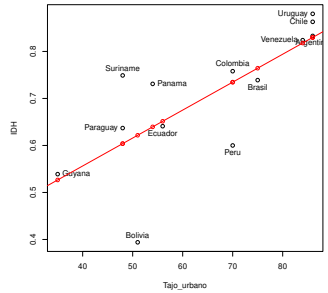
Que se puede mejorar con los parámetros `pos = c(1,2,3,4)`, `offset = x.x`, o empleando el comando `identify` en lugar de `text`, pues permite localizar las etiquetas:

```
> plot(Tajo_urbano,IDH)
> identify(Tajo_urbano,IDH,lab)
> dev.copy(pdf,"SudAmerica2.pdf")
> dev.off()
```



El modelo lineal

Nuestro objetivo es trazar una recta a través de los puntos. Pero, ¿cuál recta?



La recta de regresión y los valores estimados se dibujan así:

```
attach(SudAmerica)
lmSA=lm(IDH~Tajo_urbano,data=SudAmerica)
abline(lmSA,col="red")
points(Tajo_urbano,lmSA$fitted.values,col="red")
```

Con las cuatro instrucciones:

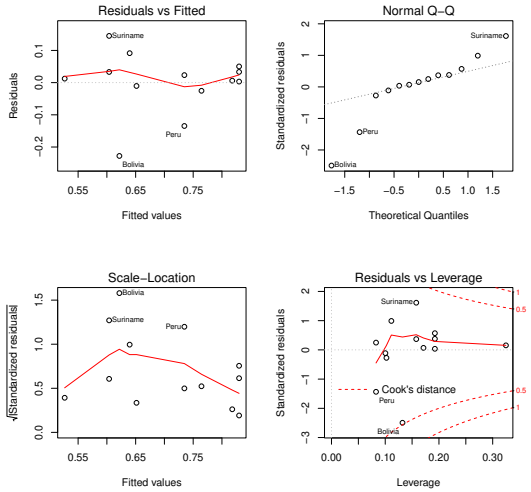
```
lmSA=lm(IDH~Tajo_urbano,data=SudAmerica); lmSA
summary(lmSA)
par(mfrow=c(2,2))
plot(lmSA)
```

ya se consiguen los siguientes resultados, que serán estudiados a continuación:


```
Call:
lm(formula = IDH ~ Tajo_urbano, data = SudAmerica)

Residuals:
     Min       1Q  Median      3Q     Max
-0.23188 -0.01133 0.01236 0.03579 0.14579

Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 0.321479  0.109569   2.934  0.01359 *
Tajo_urbano 0.005890  0.001624   3.626  0.00398 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09892 on 11 degrees of freedom
Multiple R-squared: 0.5445, Adjusted R-squared: 0.5031
F-statistic: 13.15 on 1 and 11 DF, p-value: 0.003982
```


Aquí los resultados más relevantes son

- los parámetros estimados,
- su error estándar y
- la estadística t de Student y
- su probabilidad de ser relevante,
- el error estándar de los residuos,
- el R^2 múltiple que mide cuanto el modelo explica,
- la estadística F y
- su probabilidad.



Los gráficos de residuos proporcionados automáticamente sirven para la evaluación de la calidad del modelo:

- *arriba, izquierda* Residuos contra valores predichos: su distancia delrepresenta cuando el modelo no explica, la curva cuanto son lineales.
- *arriba, derecha* Desvío a la normalidad de los residuos, como distancia de la recta.
- *bajo, izquierda* Raíz de residuos estandarizados, indicando una distribución parecida o no.
- *bajo, derecha* Búsqueda de valores anómalos fuera de las lineas rojas punteadas.

Todos estos aspectos estarán estudiados en detalle en el curso.