

MODELOS LINEALES

PRESENTACIÓN DEL CURSO

Fernando Massa; Bruno Bellagamba

Martes 5 de marzo 2024



FACULTAD DE
CIENCIAS ECONÓMICAS
Y DE ADMINISTRACIÓN

UESTA INSTITUTO
DE ESTADÍSTICA



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



1 ASPECTOS GENERALES

2 REPASO DE INFERENCIA I

3 MODELOS LINEALES

4 CIGÜEÑAS

5 PRÓXIMA CLASE



Fernando Massa.

- Licenciado en Estadística 2009 (Udelar).
- Magíster en Ingeniería Matemática 2015 (Udelar).
- Cursando Doctorado en Estadística (UNR, Argentina).




Bruno Bellagamba.

- Bachiller.

- Clases teórico/prácticas martes y jueves de 10 a 12 en el salón 004.
- Clases grabadas durante la edición 2022.
- Talleres cada 2 semanas (aproximadamente).
- A lo largo del curso se desarrollarán 3 tópicos:
 - ① Regresión Lineal.
 - ② Análisis de Varianza (ANOVA).
 - ③ Modelos Lineales Generalizados (GLM).
- Se requiere haber aprobado Inferencia I y se sugiere refrescar conceptos de Álgebra Lineal.

- Se llevarán a cabo 3 instancias evaluatorias:
 - ① Una revisión durante el primer receso, (16 de mayo) (40 puntos).
 - ② Entregas de ejercicios a lo largo del curso (20 puntos).
 - ③ Elaboración de un proyecto final en grupos de 2 o 3 estudiantes (40 puntos).
- El curso se exonera totalmente:
 - ① Obteniendo al menos el 60% del puntaje total.
 - ② Obteniendo al menos el 40% de los puntos en cada instancia evaluatoria.
- El curso se exonera parcialmente (hay que rendir un oral):
 - ① Obteniendo entre 40% y 60% del puntaje total.
 - ② Obteniendo al menos el 40% de los puntos en cada instancia evaluatoria.
- Si se obtiene menos del 40% del total, se debe rendir examen completo.

- Los materiales vistos en clase (diapositivas, prácticos, scripts, etc) se harán disponibles en el EVA del curso. Adicionalmente, se contará con un calendario donde se presentarán los temas a tratar en cada clase.
- Estas diapositivas se usarán a modo de “guía” durante las clases pero se espera que los estudiantes profundicen los conocimientos en la bibliografía del curso.
- Durante los talleres proporcionarán *scripts* de  pero se alienta a que los estudiantes indaguen sobre maneras más eficientes, novedosas o elegantes para obtener los mismos (o mejores) resultados.



-  Carmona, Francesc (2003). *Modelos Lineales (notas de curso)*. Departament d'Estadística.
-  Faraway, Julian (2014). *Linear Models with R, second edition*. Chapman Hall/CRC.
-  Hosmer, David y Stanley Lemeshow (2010). *Applied Logistic Regression*. John Wiley Sons, Inc.
-  Peña, Daniel (2010). *Regresión y Diseño de Experimentos*. Alianza Editorial.
-  Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition*. John Wiley Sons, Inc.
-  Wooldridge, Jeffrey (2010). *Introducción a la Econometría. Un Enfoque Moderno, cuarta edición*. Cengage Learning.

SOBRE LAS DIAPOSITIVAS

Para aprovechar mejor estas diapositivas, o para usarlas como guía a la hora de prepararse para las evaluaciones, tenga en cuenta los siguientes aspectos.

TÍTULO

Cuando se presente una definición importante, la misma será resaltada en un recuadro como este.

A lo largo de las clases aparecerán los siguientes íconos:



Temario



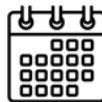
Objetivos



Ejemplo



Ejercicios



Próxima



Bibliografía

PROCESOS INFERENCIALES

- Estimación.
- Intervalos de confianza.
- Pruebas de hipótesis.

Todos se basan en realizar algún tipo de afirmación sobre una cierta población a partir de una muestra.

Los tres procedimientos inferenciales vistos en ese curso requerían del conocimiento de la distribución de probabilidad de los datos para luego adquirir cierto conocimiento sobre alguno de los parámetros de dicha distribución.

Sean X_1, X_2, \dots, X_n una MAS variables aleatorias i.i.d. con distribución $N(\mu, \sigma^2)$.

Maximizando la función de verosimilitud es posible obtener los estimadores \bar{x} y s^2 .

Luego, apelando a resultados del curso de Probabilidad I, se construyeron intervalos de confianza para μ y σ^2 empleando las distribuciones t y χ^2 .

Empleando distintos métodos se realizaron pruebas de hipótesis sobre estos parámetros. Apareció el concepto de “p-valor”.



Veamos un ejemplo:

Sean X_1, X_2, \dots, X_n una MAS variables aleatorias i.i.d. con distribución $N(\mu_x, \sigma^2)$ y Y_1, Y_2, \dots, Y_m una MAS de variables aleatorias i.i.d con distribución $N(\mu_y, \sigma^2)$.

Una prueba de hipótesis que se planteó en el contexto de la *comparación de grupos* fue la siguiente:

$$H_0) \quad \mu_x = \mu_y$$

$$H_1) \quad \mu_x \neq \mu_y$$

A partir del estadístico t para muestras (o grupos) independientes (con varianzas iguales o distintas) era posible contrastar esta hipótesis.



Continuando con el ejemplo:

Un experimento buscó determinar si dos métodos de estudio (intensivo o moderado) producían los mismos resultados en el resultado de un examen.

	Intensivo	Moderado
n	12	10
\bar{x}	46,31	42,79
s^2	6,44	7,52

El valor del estadístico de prueba fue de 3,128. ¿Entonces?

UN SALTO HACIA MODELOS LINEALES

Un aspecto interesante sobre este ejemplo es que tiene oculto un modelo lineal.
Supongamos que los datos pueden ser organizados en dos variables.

- Resultado del examen (valor numérico).
- Método de estudio (variable con dos categorías).

Reescribamos el modelo probabilístico, según los supuestos iniciales.

$$\mathbb{E}(Y) = \begin{cases} \mu_I & , \text{ si } X = \text{intensivo} \\ \mu_M & , \text{ si } X = \text{moderado} \end{cases} \quad \text{Var}(Y) = \sigma^2$$

Si decidimos codificar a la variable X con el valor 1 en los casos correspondientes a alumnos de estudio intensivo y 0 en moderado ...

UN SALTO HACIA MODELOS LINEALES

... entonces surge nuestro primer modelo lineal.

$$Y = \mu_M + (\mu_I - \mu_M)X + \varepsilon$$

Donde:

- Y es el resultado del examen.
- X es una variable *dummy* que indexa la modalidad de estudio.
- ε es el componente aleatorio del modelo. Y asumimos $\mathbb{E}(\varepsilon) = 0$ y $\text{Var}(\varepsilon) = \sigma^2$

Si adicionalmente, renombramos $\beta_0 = \mu_M$ y $\beta_1 = \mu_I - \mu_M$, entonces:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y la hipótesis que interesa contrastar es:

$$H_0) \quad \beta_1 = 0$$

$$H_1) \quad \beta_1 \neq 0$$

Pero, ¿Es realmente la misma situación?

$$\mathbb{E}(Y) = \mathbb{E}[\mu_M + (\mu_I - \mu_M)X + \varepsilon]$$

$$\mathbb{E}(Y) = \mu_M + (\mu_I - \mu_M)X + \mathbb{E}(\varepsilon)$$

$$\mathbb{E}(Y) = \mu_M + (\mu_I - \mu_M)X$$

$$\mathbb{E}(Y) = \begin{cases} \mu_I & , \text{ si } X = 1 \text{ (intensivo)} \\ \mu_M & , \text{ si } X = 0 \text{ (moderado)} \end{cases}$$

Además:

$$\mathbb{V}ar(Y) = \mathbb{V}ar[\mu_M + (\mu_I - \mu_M)X + \varepsilon]$$

$$\mathbb{V}ar(Y) = \mathbb{V}ar(\varepsilon)$$

$$\mathbb{V}ar(Y) = \sigma^2$$

Entonces sí, es exactamente la misma situación. Pero, como veremos durante el curso, permite manejar muchísimas situaciones más.

¿QUÉ SITUACIONES?

Continuando con el mismo ejemplo, ¿qué tal si nos interesara comparar un tercer método de estudio? (intensivo, moderado y no estudia).

Esta situación NO puede ser manejada con el estadístico t , salvo que se hagan varias pruebas, pero esto tiene un inconveniente, se “*infla*” el error de tipo I.

Empleando modelos lineales veremos que es posible comparar 2, 3 o el número de métodos de estudio que se desee sin inflar el error de tipo I.

¿QUÉ OTRAS SITUACIONES?

Otro tipo de situaciones que nos puede interesar es responder preguntas del estilo:

¿El resultado del examen puede ser predicho a partir del resultado de la primera revisión?

En este caso, ni siquiera estamos en presencia de una situación donde se comparan 2 grupos. Las variables que intervienen son ambas de naturaleza cuantitativa.

¿Es una caso donde el interés recae sobre el coeficiente de correlación?



Diagrama de dispersión entre las notas de la primera revisión y las notas del examen.

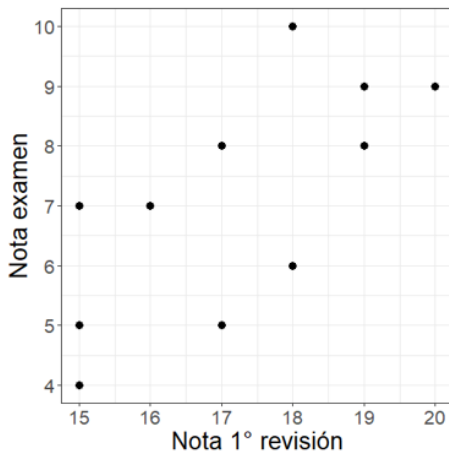


Diagrama de dispersion de las notas



El coeficiente de correlación de Pearson se define como:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Y su equivalente muestral es:

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$$

Para los datos del ejemplo, el valor de r es 0,723.

- ¿Este valor es alto o bajo?
- ¿Qué nos dice respecto de la relación entre las notas?
- ¿Es significativo?

¿QUÉ OTRAS SITUACIONES?

Resulta que la situación anterior también puede ser planteada como un modelo lineal de la forma.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

En este caso:

- Y es la nota del examen.
- X es la nota de la primera revisión.
- β_0 es la nota **promedio** cuando la nota de la primera revisión es 0.
- β_1 expresa el aumento **promedio** en la nota final por cada punto que aumenta la nota en la primera revisión.

¿QUÉ OTRAS SITUACIONES?

Analizando la situación mediante este modelo obtenemos las siguientes ventajas:

- Es más fácil analizar la relación entre X e Y .
- No se pierde la interpretación del coeficiente de correlación.
- Es posible contemplar situaciones más complejas.

No obstante, a diferencia de lo que sucede con el coeficiente de correlación debemos pensar que una de las variables *explica* o *predice* a la otra.



Este dicho tiene su origen en la observación de que aldeas con mayor número de cigüeñas, solían tener mayor número de nacimientos.

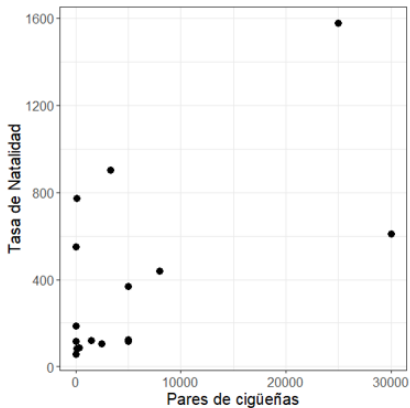
En el artículo "*Storks deliver babies ($p=0.0008$)*" del profesor Robert Mathews, se presentan los siguientes datos.

Country	Area (km ²)	Storks (pairs)	Humans (10 ⁶)	Birth rate (10 ³ /yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

¿LAS CIGÜEÑAS TRAEN A LOS BEBÉS?



Un diagrama de dispersión entre el número de cigüeñas y la tasa de nacimientos por cada 10.000 habitantes muestra la siguiente relación.



Tasa de Natalidad según cantidad de cigüeñas



Al ajustar el siguiente modelo de regresión simple.

$$\text{Tasa de natalidad} = \beta_0 + \beta_1 \text{Cigüeñas} + \varepsilon$$

Se obtienen las siguientes estimaciones:

	Estimación	Error Std.	estadístico t	p-valor
Constante	225,029	93,560	2,405	0,029
Cigüeñas	0,029	0,009	3,063	0,008

De donde se concluye que hay una relación significativa entre el número de cigüeñas y la tasa de natalidad. ¿Entonces?



La explicación radica en que ambas variables están asociadas a una tercera variable, en este caso *Superficie*. Entonces, volvamos a indagar sobre la relación entre la tasa de natalidad y el número de cigüeñas pero **ajustando** el análisis por la superficie del país.

$$\text{Tasa de natalidad} = \beta_0 + \beta_1 \text{Cigüeñas} + \beta_2 \text{Superficie} + \varepsilon$$

Se obtienen las siguientes estimaciones:

	Estimación	Error Std.	estadístico t	p-valor
Constante	-7,4117	56,7022	-0,1310	0,8980
Cigüeñas	0,0060	0,0057	1,0610	0,3070
Superficie	0,0016	0,0002	6,9640	<0.001

¿ALGO MÁS?

Sí, mucho más.

Resulta que el modelo anterior es un caso particular de uno mucho más general.

REGRESIÓN LINEAL MÚLTIPLE

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

A partir de este modelo es posible:

- Aumentar la precisión a la hora de predecir Y a partir de un conjunto de variables explicativas.
- Determinar el “*impacto*” de una variable X_j teniendo en cuenta otro conjunto de variables.
- Considerar otro montón de situaciones.

¿Y SI LA “Y” NO ES CUANTITATIVA?

En todos los ejemplos anteriores, la variable Y era de carácter cuantitativo, pero esto no tiene porqué restringir el análisis.

Sobre el final del curso generalizaremos el modelo anterior para considerar casos donde interese modelar variables que no sean cuantitativas.

- Variables cualitativas.
- Variables discretas (conteos).
- Otras.



Comenzaremos con el modelo de regresión simple.

- ¿Qué es un modelo lineal?
- Componentes del modelo.
- Situaciones en las que puede utilizarse.
- Estimación por mínimos cuadrados.
- Relación entre el coeficiente β_1 y el coeficiente de correlación.

¿Preguntas?

Muchas Gracias