Curso Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

CEU
Universidad
San Pablo

Licencia creative commons

Curso básico de estadística

Alfredo Sánchez Alberca (asalber@gmail.com).

Con esta licencia eres libre de:

- Copiar, distribuir y mostrar este trabajo.
- Realizar modificaciones de este trabajo.

Bajo las siguientes condiciones:

Reconocimiento. Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).

No comercial. No puede utilizar esta obra para fines comerciales.

Compartir bajo la misma licencia. Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Regresión no lineal

El ajuste de un modelo de regresión no lineal es similar al del modelo lineal y también puede realizarse mediante la técnica de mínimos cuadrados.

No obstante, en determinados casos un ajuste no lineal puede convertirse en un ajuste lineal mediante una sencilla transformación de alguna de las variables del modelo.

Transformación de modelos de regresión no lineales

- **Modelo logarítmico**: Un modelo logarítmico $y = a + b \log x$ se convierte en un modelo lineal haciendo el cambio $t = \log x$:

$$y = a + b \log x = a + bt.$$

- **Modelo exponencial**: Un modelo exponencial $y = ae^{bx}$ se convierte en un modelo lineal haciendo el cambio $z = \log y$:

$$z = \log y = \log(ae^{bx}) = \log a + \log e^{bx} = a' + bx.$$

- **Modelo potencial**: Un modelo potencial $y = ax^b$ se convierte en un modelo lineal haciendo los cambios $t = \log x$ y $z = \log y$:

$$z = \log y = \log(ax^b) = \log a + b \log x = a' + bt.$$

- **Modelo inverso**: Un modelo inverso $y = a + b/x$ se convierte en un modelo lineal haciendo el cambio $t = 1/x$:

$$y = a + b(1/x) = a + bt.$$

- **Modelo curva S**: Un modelo curva S $y = e^{a+b/x}$ se convierte en un modelo lineal haciendo los cambios $t = 1/x$ y $z = \log y$:

$$z = \log y = \log(e^{a+b/x}) = a + b(1/x) = a + bt.$$

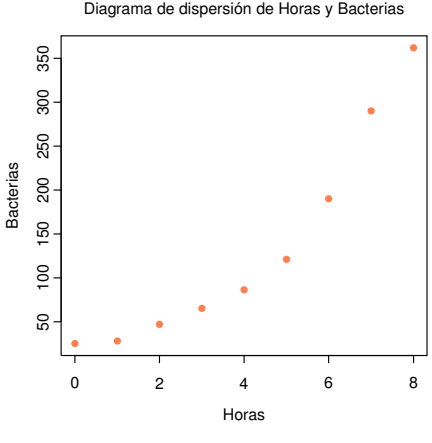
Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

El número de bacterias de un cultivo evoluciona con el tiempo según la siguiente tabla:

El diagrama de dispersión asociado es

Horas	Bacterias
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

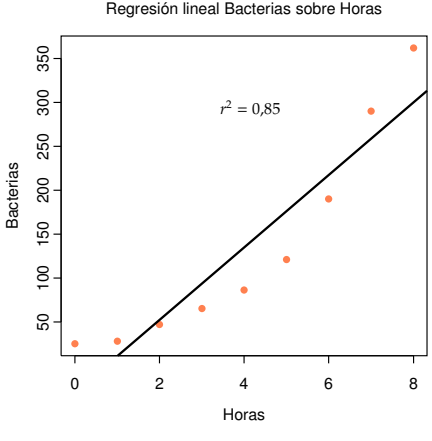


Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

Si realizamos un ajuste lineal, obtenemos la siguiente recta de regresión

Horas	Bacterias
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362



$$\text{Bacterias} = -30{,}18 + 41{,}27\,\text{Horas}$$

¿Es un buen modelo?

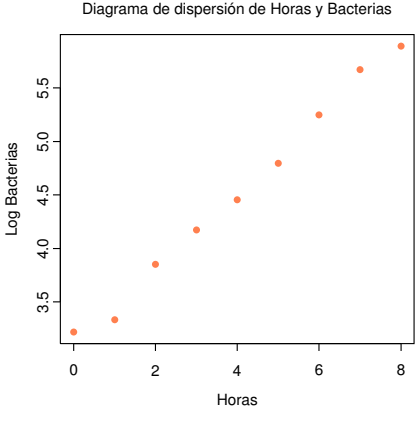
Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

Aunque el modelo lineal no es malo, de acuerdo al diagrama de dispersión es más lógico construir un modelo exponencial o cuadrático.

Para construir el modelo exponencial $y = ae^{bx}$ hay que realizar la transformación $z = \log y$, es decir, aplicar el logaritmo a la variable dependiente.

Horas	Bacterias	Log Bacterias
0	25	3,22
1	28	3,33
2	47	3,85
3	65	4,17
4	86	4,45
5	121	4,80
6	190	5,25
7	290	5,67
8	362	5,89



Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

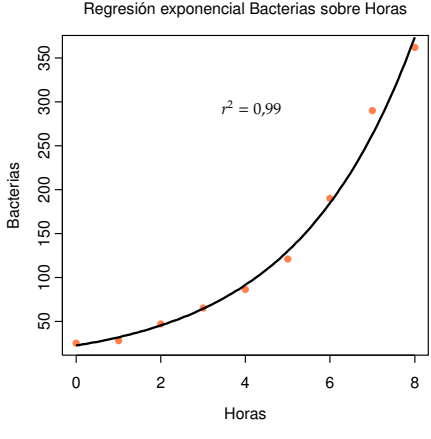
Ahora sólo queda calcular la recta de regresión del logaritmo de Bacterias sobre Horas

$$\text{Log Bacterias} = 3{,}107 + 0{,}352\,\text{Horas}.$$

Y deshaciendo el cambio de variable, se obtiene el modelo exponencial

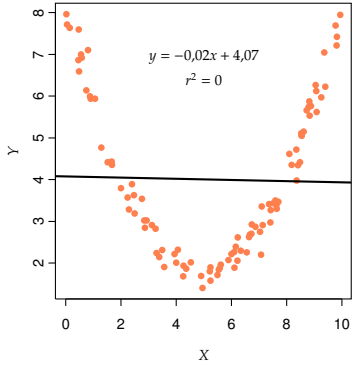
$$\text{Bacterias} = e^{3{,}107+0{,}352\,\text{Horas}},$$

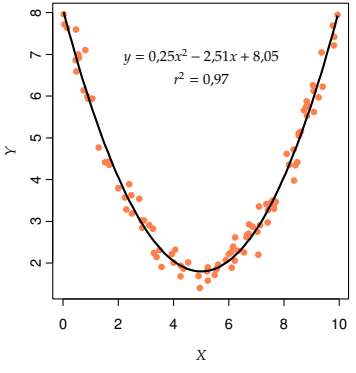
que, a la vista del coeficiente de determinación, es mucho mejor modelo que el lineal.



Interpretación de un coeficiente de determinación pequeño

Tanto el coeficiente de determinación como el de correlación hacen referencia a un modelo concreto, de manera que un coeficiente $r^2 = 0$ significa que no existe relación entre las variables del tipo planteado por el modelo, pero *eso no quiere decir que las variables sean independientes*, ya que puede existir relación de otro tipo.

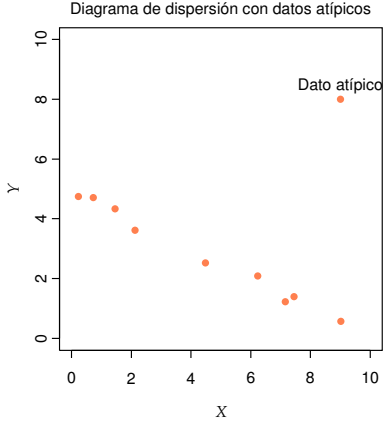




Datos atípicos en regresión

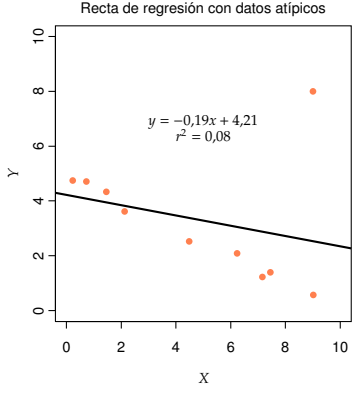
En un estudio de regresión es posible que aparezca algún individuo que se aleja notablemente de la tendencia del resto de individuos en la nube de puntos.

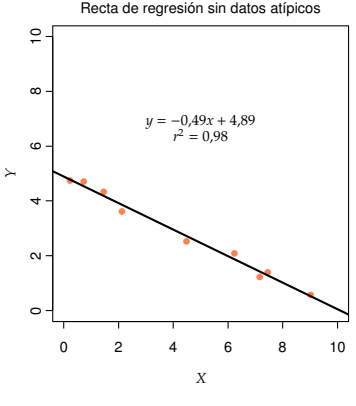
Aunque el individuo podría no ser un *dato atípico* al considerar las variables de manera separada, sí lo sería al considerarlas de manera conjunta.



Influencia de los datos atípicos en los modelos de regresión

Los datos atípicos en regresión suelen provocar cambios drásticos en el ajuste de los modelos de regresión, y por tanto, habrá que tener mucho cuidado con ellos.





Relaciones entre atributos

Los modelos de regresión vistos sólo pueden aplicarse cuando las variables estudiadas son cuantitativas.

Cuando se desea estudiar la relación entre atributos, tanto ordinales como nominales, es necesario recurrir a otro tipo de medidas de relación o de asociación. En este tema veremos tres de ellas:

- Coeficiente de correlación de Spearman.
- Coeficiente chi-cuadrado.
- Coeficiente de contingencia.

Coeficiente de correlación de Spearman

Cuando se tengan atributos ordinales es posible ordenar sus categorías y asignarles valores ordinales, de manera que se puede calcular el coeficiente de correlación lineal entre estos valores ordinales.

Esta medida de relación entre el orden que ocupan las categorías de dos atributos ordinales se conoce como coeficiente ce correlación de Spearman, y puede demostrarse fácilmente que puede calcularse a partir de la siguiente fórmula

Definición (Coeficiente de correlación de Spearman)

Dada una muestra de n individuos en los que se han medido dos atributos ordinales X e Y, el coeficiente de correlación de Spearman se define como:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre el valor ordinal de X y el valor ordinal de Y del individuo i.

Interpretación del coeficiente de correlación de Spearman

Como el coeficiente de correlación de Spearman es en el fondo el coeficiente de correlación lineal aplicado a los órdenes, se tiene:

$$-1 \leq r_s \leq 1,$$

de manera que:

- Si $r_s = 0$ entonces no existe relación entre los atributos ordinales.
- Si $r_s = 1$ entonces los órdenes de los atributos coinciden y existe una relación directa perfercta.
- Si $r_s = -1$ entonces los órdenes de los atributos están invertidos y existe una relación inversa perfecta.

En general, cuanto más cerca de 1 o -1 esté r_s, mayor será la relación entre los atributos, y cuanto más cerca de 0, menor será la relación.

Cálculo del coeficiente de correlación de Spearman

Ejemplo

Una muestra de 5 alumnos realizaron dos tareas diferentes X e Y, y se ordenaron de acuerdo a la destreza que manifestaron en cada tarea:

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	2	-1	1
Alumno 4	3	1	2	4
Alumno 5	4	5	-1	1
$\sum$			0	8

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8}{5(5^2-1)} = 0{,}6,$$

lo que indica que existe bastante relación directa entre las destrezas manifestadas en ambas tareas.

Cálculo del coeficiente de correlación de Spearman

Ejemplo con empates

Cuando hay empates en el orden de las categorías se atribuye a cada valor empatado la media aritmética de los valores ordinales que hubieran ocupado esos individuos en caso de no haber estado empatados.

Si en el ejemplo anterior los alumnos 4 y 5 se hubiesen comportado igual en la primera tarea y los alumnos 3 y 4 se hubiesen comportado igual en la segunda tarea, entonces se tendría

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	1,5	$-0{,}5$	0,25
Alumno 4	3,5	1,5	2	4
Alumno 5	3,5	5	$-1{,}5$	2,25
$\sum$			0	8,5

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8{,}5}{5(5^2-1)} = 0{,}58.$$

Relación entre atributos nominales

Cuando se quiere estudiar la relación entre atributos nominales no tiene sentido calcular el coeficiente de correlación de Spearman ya que las categorías no pueden ordenarse.

Para estudiar la relación entre atributos nominales se utilizan medidas basadas en las frecuencias de la tabla de frecuencias bidimensional, que para atributos se suele llamar *tabla de contingencia*.

Ejemplo En un estudio para ver si existe relación entre el sexo y el hábito de fumar se ha tomado una muestra de 100 personas. La tabla de contingencia resultante es

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

Si el hábito de fumar fuese independiente del sexo, la proporción de fumadores en mujeres y hombres sería la misma.

Frecuencias teóricas o esperadas

En general, dada una tabla de contingencia para dos atributos X e Y,

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_q	n_x
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	n_{x_1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	n_{x_i}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	n_{x_p}
n_y	n_{y_1}	$\cdots$	n_{y_j}	$\cdots$	n_{y_q}	n

si X e Y fuesen independientes, para cualquier valor y_j se tendría

$$\frac{n_{1j}}{n_{x_1}} = \frac{n_{2j}}{n_{x_2}} = \cdots = \frac{n_{pj}}{n_{x_p}} = \frac{n_{1j} + \cdots + n_{pj}}{n_{x_1} + \cdots + n_{x_p}} = \frac{n_{y_j}}{n},$$

de donde se deduce que

$$n_{ij} = \frac{n_{x_i} n_{y_j}}{n}.$$

A esta última expresión se le llama *frecuencia teórica* o *frecuencia esperada* del par (x_i, y_j).

Coeficiente chi-cuadrado χ^2

Es posible estudiar la relación entre dos atributos X e Y comparando las frecuencias reales con las esperadas:

Definición (Coeficiente chi-cuadrado χ^2)

Dada una muestra de tamaño n en la que se han medido dos atributos X e Y, se define el coeficiente χ^2 como

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{x_i} n_{y_j}}{n}\right)^2}{\frac{n_{x_i} n_{y_j}}{n}},$$

donde p es el número de categorías de X y q el número de categorías de Y.

Por ser suma de cuadrados, se cumple que

$$\chi^2 \geq 0,$$

de manera que $\chi^2 = 0$ cuando los atributos son independientes, y crece a medida que aumenta la dependencia entre las variables.

Cálculo del coeficiente chi-cuadrado χ^2

Ejemplo

Siguiendo con el ejemplo anterior, a partir de la tabla de contingencia

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

se obtienen las siguientes frecuencias esperadas:

Sexo	Si	No	n_i
Mujer	$\frac{40\cdot 38}{100} = 15{,}2$	$\frac{40\cdot 62}{100} = 24{,}8$	40
Hombre	$\frac{60\cdot 38}{100} = 22{,}8$	$\frac{60\cdot 62}{100} = 37{,}2$	60
n_j	38	62	100

y el coeficiente χ^2 vale

$$\chi^2 = \frac{(12-15{,}2)^2}{15{,}2} + \frac{(28-24{,}8)^2}{24{,}8} + \frac{(26-22{,}8)^2}{22{,}8} + \frac{(34-37{,}2)^2}{37{,}2} = 1{,}81,$$

lo que indica que no existe gran relación entre el sexo y el hábito de fumar.

Coeficiente de contingencia

El coeficiente χ^2 depende del tamaño muestral, ya que al multiplicar por una constante las frecuencias de todas las casillas, su valor queda multiplicado por dicha constante, lo que podría llevarnos al equívoco de pensar que ha aumentado la relación, incluso cuando las proporciones se mantienen. En consecuencia el valor de χ^2 no está acotado superiormente y resulta difícil de interpretar.

Para evitar estos problemas se suele utilizar el siguiente estadístico:

Definición (Coeficiente de contingencia)

Dada una muestra de tamaño n en la que se han medido dos atributos X e Y, se define el *coeficiente de contingencia* como

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Interpretación del coeficiente de contingencia

De la definición anterior se deduce que

$$0 \leq C \leq 1,$$

de manera que cuando $C = 0$ las variables son independientes, y crece a medida que aumenta la relación.

Aunque C nunca puede llegar a valer 1, se puede demostrar que para tablas de contingencia con k filas y k columnas, el valor máximo que puede alcanzar C es $\sqrt{(k-1)/k}$.

Ejemplo En el ejemplo anterior el coeficiente de contingencia vale

$$C = \sqrt{\frac{1{,}81}{1{,}81 + 100}} = 0{,}13.$$

Como se trata de una tabla de contingencia de 2×2, el valor máximo que podría tomar el coeficiente de contingencia es $\sqrt{(2-1)/2} = \sqrt{1/2} = 0{,}707$, y como $0{,}13$ está bastante lejos de este valor, se puede concluir que no existe demasiada relación entre el hábito de fumar y el sexo.