Tema 2. Descripción de dos variables.

Eva Romero Ramos
evarom03@ucm.es

Universidad Complutense de Madrid

Outline

1. Tabla de correlación y tablas de contigencia.
2. Covarianza
3. Regresión y correlación

Outline

1. **Tabla de correlación y tablas de contigencia.**
2. Covarianza
3. Regresión y correlación

Motivación

Supongamos que queremos estudiar dos variables a la vez:

Ejemplo 1:

X/Y	1	2	3	4	$n_{i\cdot}$
5	1	2	1	3	7
10	2	1	3	2	8
15	3	2	1	2	8
$n_{\cdot j}$	6	5	5	7	23

Ejemplo 2:

	Emplead@	Desemplead@	*Total*
Mujer	105	15	120

Tablas de correlación

X/Y	y_1	$\cdots$	y_j	$\cdots$	y_k	n_x
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_h	n_{h1}	$\cdots$	n_{hj}	$\cdots$	n_{hk}	$n_{h\cdot}$
n_y	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

Por ejemplo, n_{11} muestra el número de veces que x_1 aparece conjuntamente con y_1; n_{12} es la frecuencia conjunta de x_1 y y_2, etc.

Distribuciones Marginales

- A partir de la distribución bidimensional, podría interesarnos estudiar solo una variable. En este sentido, de la distribución bidimensional se obtienen 2 distribuciones unidimensionales (la de X y la de Y).
- Para el i-ésimo valor de X, la frecuencia marginal es:

$$\boxed{n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ij} + \cdots + n_{ik} = \sum_{j=1}^{k} n_{ij}}$$

- Del mismo modo, la frecuencia marginal del j-ésimo valor de Y es:

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{ij} + \cdots + n_{hj} = \sum^{h} n_{ij}$$

Distribuciones condicionales

- Nos pueden interesar obtener otras distribuciones unidimensionales para una variable asociadas a una condición en relación con la otra variable.
- Por ejemplo, la distribución de X condicional a que Y tome el valor y_2. Los valores y frecuencias de esta distribución son:

$X/Y = y_2$	n_{y_2}
x_1	n_{12}
$\vdots$	$\vdots$
x_i	n_{i2}
$\vdots$	$\vdots$

En general:

$X/Y=y_j$	n_{y_j}
x_1	n_{1j}
$\vdots$	$\vdots$
x_i	n_{ij}
$\vdots$	$\vdots$
x_h	n_{hj}
	$n_{\cdot j}$

$Y/X=x_i$	n_{x_i}
y_1	n_{i1}
$\vdots$	$\vdots$
y_j	n_{ij}
$\vdots$	$\vdots$
y_k	n_{ik}
	$n_{i\cdot}$

Las frecuencias relativas de la distribución condicional de X dado algún valor de Y, y la distribución condicional de Y dado algún valor de X son

Relaciones de dependencia

- Dependencia funcional:
- Y depende funcionalmente de X si existe una función que relaciona los elementos de X y los elementos de Y: $Y = f(X)$
- Dependencia estadística:
  Y depende estadísticamente de X si las variables están relacionadas, pero la relación no puede expresarse mediante una función matemática.
  La dependencia estadística puede medirse gradualmente, ya que puede haber relaciones más débiles o más fuertes. Llamamos a esta relación **correlación** entre variables cuantitativas y **contingencia** entre variables cualitativas.
- Independencia:

- Dos variables son <u>estadísticamente independientes</u> cuando su freciencia relativa conjunta es igual al producto de las frecuencias relativas marginales:

$$\boxed{\frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N} \cdot \frac{n_{\cdot j}}{N} \qquad \forall i, j}$$

- En este caso, las frecuencias relativas condicionales son iguales a las frecuencias relativas marginales:

$$f_{i|j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{(n_{i\cdot} n_{\cdot j}) / N}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} = f_{i\cdot}$$

Outline

1. Tabla de correlación y tablas de contigencia.
2. **Covarianza**
3. Regresión y correlación

Momentos bidimensionales

- Momentos con respecto a cero:

$$\alpha_{rs} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r y_j^s n_{ij}$$

- Ejemplos:

$$\boxed{\alpha_{10} = \overline{x}} \qquad \boxed{\alpha_{01} = \overline{y}}$$

- Momentos con respecto a la media:

$$m_{rs} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \, (y_j - \overline{y})^s \, n_{ij}$$

Covarianza

Para analizar el grado de relación que presentan dos variables X e Y utilizaremos la covarianza.

Definición.- Covarianza
La covarianza es una medida del grado de variación conjunta entre dos variables estadísticas, respecto a sus medias. Se obtiene mediante la siguiente fórmula:

$$\boxed{\mathrm{COV(X,Y)} = \mathrm{m}_{11} = S_{XY} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}}$$

También se puede obtener la covarianza en función de los momentos con respecto al origen:

Covarianza

- El valor de la covarianza en caso de independencia estadística es $S_{xy} = 0$.
- Lo contrario es no necesariamente cierto, es decir, una covarianza nula no implica necesariamente independencia.
- Si las variables presentan una relación positiva (cuando una crece la otra también crece) la covarianza será positiva. Si la relación entre las variables es negativa, la covarianza también lo será.

Transformaciones lineales

- Consideremos las siguientes características de X e Y: $\overline{x}, \overline{y}, S_x^2, S_y^2, S_{xy}$
- Si les aplicamos transformaciones lineales como las siguientes:

$$x' = a_1 + b_1 x \qquad y' = a_2 + b_2 y$$

- ¿Cómo se comportarán las medias aritméticas, varianzas, desviaciones típicas y la covarianza ante estos cambios?
- Medias aritméticas:

$$\overline{x'} = a_1 + b_1 \overline{x} \qquad \overline{y'} = a_2 + b_2 \overline{y}$$

- Varianzas y desviaciones típicas:

$$S_x'^2 = b_1^2 S_x^2 \qquad S_x' = b_1 S_x$$

Ejemplo 3: Covarianza

X/Y	1	2	4	$n_{i\cdot}$	$x_i n_i$	$x_i^2 n_i$	$\sum y_j n_{ij}$	$\sum x_i y_j n_{ij}$
5	1	0	2	3	15	75	9	45
10	2	1	0	3	30	300	4	40
15	0	1	3	4	60	900	14	210
$n_{\cdot j}$	3	2	5	10	105	1275		295
$y_j n_{\cdot j}$	3	4	20	27				
$y_j^2 n_{\cdot j}$	3	8	80	91				

X/Y	1	2	4	$n_{i\cdot}$	$x_i n_i$	$x_i^2 n_i$	$\sum y_j n_{ij}$	$\sum x_i y_j n_{ij}$
5	1	0	2	3	15	75	9	45
10	2	1	0	3	30	300	4	40
15	0	1	3	4	60	900	14	210
$n_{\cdot j}$	3	2	5	10	105	1275		295
$y_j n_{\cdot j}$	3	4	20	27				
$y_j^2 n_{\cdot j}$	3	8	80	91				

$$S_{xy} = \alpha_{11} - \alpha_{10}\alpha_{01} = 1.15$$

$$\alpha_{10} = \frac{105}{10} = 10.5$$

$$\alpha_{01} = \frac{27}{10} = 2.7$$

Outline

1. Tabla de correlación y tablas de contigencia.
2. Covarianza
3. **Regresión y correlación**

Definición.- Regresión

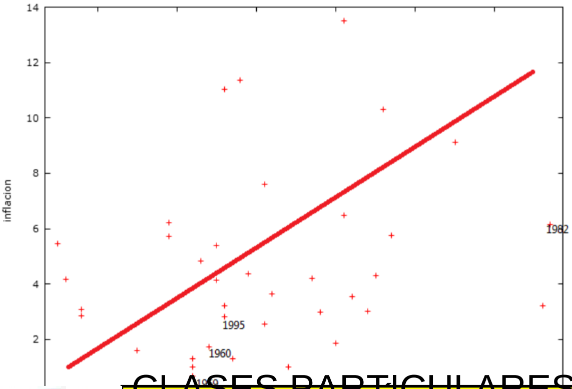
La regresión pretende encontrar la estructura de dependencia que mejor explique el comportamiento de una variable Y a la que denominaremos (variable dependiente, explicada o endógena) a partir de un conjunto de variables $X_1, X_2, \cdots, X_p$ (variables independientes, explicativas o exógenas) relacionadas con Y.

Definición.- Regresión lineal simple

La regresión lineal simple pretende encontrar la recta que mejor explica el comportamiento de la variable dependiente Y a partir del comportamiento de una única variable X.

Regresión

El gráfico de dispersión o nube de puntos representa cada par de valores de X e Y mediante un punto en el espacio eucliedo bidimensional.



Regresión

La ecuación del modelo de regresión lineal simple será:

$$\boxed{Y = a + bX + \epsilon}$$

A partir de la información de la muestra tendremos que encontrar los valores de a y b que consiguen minimizar las distancias entre la recta y los valores de las variables.

Utilizaremos para ello el método de mínimos cuadrados ordinarios, según el cual:

$$\boxed{b = \frac{S_{XY}}{S_X^2} = \frac{COV(X,Y)}{VAR(X)} = \frac{m_{11}}{m_{20}}} \qquad \boxed{a = \bar{y} - b\bar{x}}$$

Correlación

- Para medir el grado de dependencia entre dos variables usaremos el coeficiente de correlación lineal:

$$\boxed{r = \frac{S_{xy}}{S_x S_y}}$$

- El coeficiente de correlación lineal toma valores entre -1 y 1, $-1 \leq r \leq 1$

Correlación

- $r = 1$: Indica correlación positiva perfecta y todas las observaciones se sitúan sobre una recta. Es decir, existe una dependencia funcional reflejada en una recta creciente.
- $r = -1$: Indica correlación negativa perfecta, pero ahora la recta es decreciente.
- $r = 0$: Indica correlación nula, es decir, ausencia de relación líneal y aunque X varíe, Y no lo hace.
- Si $-1 < r < 0$: la correlación es negativa, es decir, las variables se relacionan aproximadamente en una línea recta decreciente, pero las observaciones no se encuentran necesariamente sobre la línea.
- Si $0 < r < 1$, la correlación es positiva, es decir, las variables se

Correlación

- Cuando las variables son estadísticamente independientes, su covarianza es cero. Por tanto, si las variables son independientes, también están incorreladas linealmente, es decir, $r = 0$.
- Sin embargo, dos variables pueden estar incorreladas linealmente y ser (incluso fuertemente) dependientes, ya que cuando $r = 0$ lo único que podemos decir es que la dependencia estadística lineal es nula, pero las variables pueden estar relacionadas mediante otro tipo de función (exponencial, hiperbólica, etc.)
- El valor absoluto del coeficiente de correlación permance invariante ante transformaciones lineales, pero r puede cambiar de signo, si la transformación cambia el sentido de la relación entre las variables.

Coeficiente de determinación

- El coeficiente de determinación se interpreta como el porcentaje de variación de la variable dependiente explicado por el modelo.
- En modelos de regresión lineal simple, se calcula simplemente como el cuadrado de coeficiente de correlación lineal:

$$\boxed{R^2 = \frac{(COV(X,Y))^2}{VAR(X) \cdot VAR(Y)}}$$

Ejemplo 4: Regresión y Correlación

x_i	y_j	n_{ij}	$x_i n_{ij}$	$y_j n_{ij}$	$x_i^2 n_{ij}$	$y_j^2 n_{ij}$	$x_i y_j n_{ij}$
2	1	6	12	6	24	6	12
2	4	7	14	28	28	112	56
3	2	4	12	8	36	16	24
3	5	2	6	10	18	50	30
5	4	1	5	4	25	16	20
		20	49	56	131	200	142

Ejemplo 4: Regresión y Correlación

x_i	y_j	n_{ij}	$x_i n_{ij}$	$y_j n_{ij}$	$x_i^2 n_{ij}$	$y_j^2 n_{ij}$	$x_i y_j n_{ij}$
2	1	6	12	6	24	6	12
2	4	7	14	28	28	112	56
3	2	4	12	8	36	16	24
3	5	2	6	10	18	50	30
5	4	1	5	4	25	16	20
		20	49	56	131	200	142

$$\alpha_{10} = \frac{49}{20} = 2.45 \qquad m_{20} = \frac{131}{20} - 2.45^2 = 0.5475$$

$$\alpha_{01} = \frac{56}{20} = 2.8 \qquad m_{02} = \frac{200}{20} - 2.8^2 = 2.16$$

142

$$S_{xy} = \alpha_{11} - \alpha_{10}\alpha_{01} = 0.24$$

Ejemplo 4: Regresión y Correlación

$$\alpha_{10} = \frac{49}{20} = 2.45 \qquad m_{20} = \frac{131}{20} - 2.45^2 = 0.5475$$

$$\alpha_{01} = \frac{56}{20} = 2.8 \qquad m_{02} = \frac{200}{20} - 2.8^2 = 2.16$$

$$\alpha_{11} = \frac{142}{20} = 7.1 \qquad S_{xy} = \alpha_{11} - \alpha_{10}\alpha_{01} = 0.24$$

$$b = \frac{COX(X,Y)}{var(X)} = \frac{0.24}{0.5475} = 0.4384$$

$$a = \bar{Y} - b\bar{X} = 2.8 - 0.4384 \cdot 2.45 = 1.726$$