Metodología de investigación cuantitativa

Relación entre variables

Pablo Geraldo Bastías
pdgerald@uc.cl

Propósito de la sesión

- Comprender el significado de distintas medidas de asociación entre variables
- Comprender el modelo de regresión lineal como herramienta descriptiva de la relación entre variables
- Comprender el uso y las limitaciones del modelo de regresión lineal para identificar relaciones causales

Estructura de la sección 1

Valor Esperado

El valor esperando (*aka* esperanza, media, promedio) es una medida de tendencia central de la distribución de una variable aleatoria.
La esperanza de una variable ($\mathbb{E}[X]$) **no es** una variable, sino una constante. Así, $\mathbb{E}[\cdot]$ no es una función, sino un operador.
En el caso discreto, tenemos que:

$$\mathbb{E}[X] = \sum_{x} xf(x)$$

donde $f(x)$ es la PMF de la variable aleatoria X.
La función de una variable aleatoria es también una variable aleatoria, por lo tanto, puede operarse directamente con ella.

Valor Esperado (R)

En primer lugar, generamos algunos números como ejemplo:

```r
# Adaptado de Carsey y Harden, 2014, p.10
y <- c(3,.5,5,2.5,5.5,6)
x <- c(1,2,3,4,5,6)
z <- c(1,1,1,2,2,2)
dat <- as.data.frame(cbind(y,x,z))
```

Valor Esperado (R)

El valor esperado de una variable corresponde a la suma de sus valores ponderadas por su probabilidad de ocurrencia:

```
# A mano
py <- 1/6 #Prob de Y: 1/6 cada valor (uniforme)
prom_y <- 3*py +.5*py + 5*py + 2.5*py + 5.5*py + 6*py
prom_y

## [1] 3.75

# Usando la función "mean"
mean(y)

## [1] 3.75
```

Propiedades

- El valor esperando de una constante es la misma constante:
  $\forall c \in \mathbb{R}, \mathbb{E}[c] = c$
- El valor esperado de una constante por una variable aleatoria, es la constante por la esperanza de la variable aleatoria:
  $\forall a \in \mathbb{R}, \mathbb{E}[aX] = a\mathbb{E}[X]$
- El valor esperado es una función lineal:
  $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$

Varianza

La varianza (σ_X^2 o $\mathbb{V}(X)$) caracteriza la dispersión de la distribución de una variable aleatoria. Al igual que la esperanza, la varianza es una constante.
La varianza es el promedio de las desviaciones de las observaciones en torno a su valor esperado, elevadas al cuadrado:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Alternativamente, la varianza puede expresarse como

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Desviación estándar

La desviación estándar (σ_X) es la raíz cuadrada de la varianza. Es igualmente una medida de dispersión, pero tiene la ventaja de que permanece en la escala de la variable original:

$$\sqrt{\mathbb{V}(X)} = \sigma(X)$$

Varianza (R)

```r
# A mano (¿Por qué se divide por 5 y no por 6?)
var_y = sum((y-mean(y))^2)/5
var_y
```

```
## [1] 4.475
```

```r
# Usando la función "var"
var(y)
```

```
## [1] 4.475
```

Desviación estándar (R)

```r
# A mano
sd_y = sqrt(var_y)
sd_y

## [1] 2.11542

# Usando la función "sd"
sd(y)

## [1] 2.11542
```

Propiedades

La varianza de una constante es 0 (¿intuitivo?), lo que implica:

- $\forall$ c *in* $\mathbb{R}$, $\mathbb{V}(X + c) = \mathbb{V}(X)$
- $\forall$ c $\in \mathbb{R}$, $\sigma(X + c) = \sigma(X)$
- $\forall$ a $\in \mathbb{R}$, $\mathbb{V}(aX) = a^2\mathbb{V}(X)$
- $\forall$ a $\in \mathbb{R}$, $\sigma(aX) = |a|\sigma(X)$

Estructura de la sección 2

Covarianza

La covarianza ($\sigma_{X,Y}$) es una generalización de la varianza para el caso bivariado, y expresa el grado en que dos variables se "mueven" conjuntamente.
Puede tomar valores positivos, expresando que ambas variables crecen o decrecen juntas, o negativos, expresando que lo hacen en sentido inverso.

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Alternativamente, la covarianza puede expresarse como

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covarianza (R)

```r
# A mano
cov_xy = sum((y-mean(y))*(x-mean(x)))/5
cov_xy

## [1] 2.75

# Usando la función "cov"
cov(x,y)

## [1] 2.75
```

Propiedades

- $\forall$ c, d $\in \mathbb{R}$, Cov(c,X) = Cov(X,c) = Cov(c,d) = 0
- Cov(X,Y) = Cov(Y,X)
- Cov(X,X) = V(X)
- Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)

Correlación

Al igual que la varianza, la covarianza no tiene un rango acotado de valores. De allí que resulta útil re-escalarla, de manera análoga a como la desviación estándar re-escala la varianza.
La correlación ($\rho(X, Y)$) se encuentra acotada en el rango [-1,1], y corresponde a:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$$

Correlación (R)

```r
# A mano
corr_xy = cov_xy/(sd(x)*sd(y))
corr_xy

## [1] 0.6948677

# Usando la función "cor"
cor(x,y)

## [1] 0.6948677
```

Independencia (2.0)

A partir de lo anterior, es posible derivar algunas propiedades adicionales de la independencia entre dos variables. Si X e Y son independientes, se cumple que:

- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Su covarianza es cero: $Cov(X, Y) = 0$
- Su correlación es cero: $\rho(X, Y) = 0$
- Sus varianzas son aditivas: $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$

Esperanza condicional

De manera similar al cálculo de la probabilidad condicional de un evento o la PMF (o PDF) de una variable aleatoria, se puede obtener la esperanza condicional de una variable en relación a otra.
Para dos variables aleatorias discretas X e Y, con PMF conjunta f, la esperanza condicional de Y dado $X = x$ corresponde a:

$$\mathbb{E}[Y|X = x] = \sum_{y} y f_{Y|X}(y|x)$$

Esperanza condicional (R)

```r
# E(Y|X<=4)
py_x = 1/4
prom_y_x4 = 3*py_x + .5*py_x + 5*py_x + 2.5*py_x
prom_y_x4
```

```
## [1] 2.75
```

Mejor predictor lineal (BLP)

La función de esperanza condicional (CEF) es la mejor aproximación al valor de Y dado una muestra de X, en el sentido de que miniminza el promedio de la desviación cuadrática de los valores de Y (es MMSE).
¿Qué ocurre si nos restringimos a las funciones lineales, de la forma $g(X) = a + bX$? Entonces obtenemos el *mejor predictor lineal* (BLP).

Mejor predictor lineal (BLP)

Dadas las variables aleatorias X e Y, el BLP de Y dado X es la función $g(X) = \alpha + \beta X$, donde:

$$\alpha = E[Y] - \frac{Cov(X, Y)}{V(X)} E[X]$$

$$\beta = \frac{Cov(X, Y)}{V(X)}$$

Mejor predictor lineal (BLP)

```r
# Usamos la forma Y = alfa + beta*X
beta = cov(x,y)/var(x)
beta
```

```
## [1] 0.7857143
```

```r
alfa = mean(y) - beta*mean(x)
alfa
```

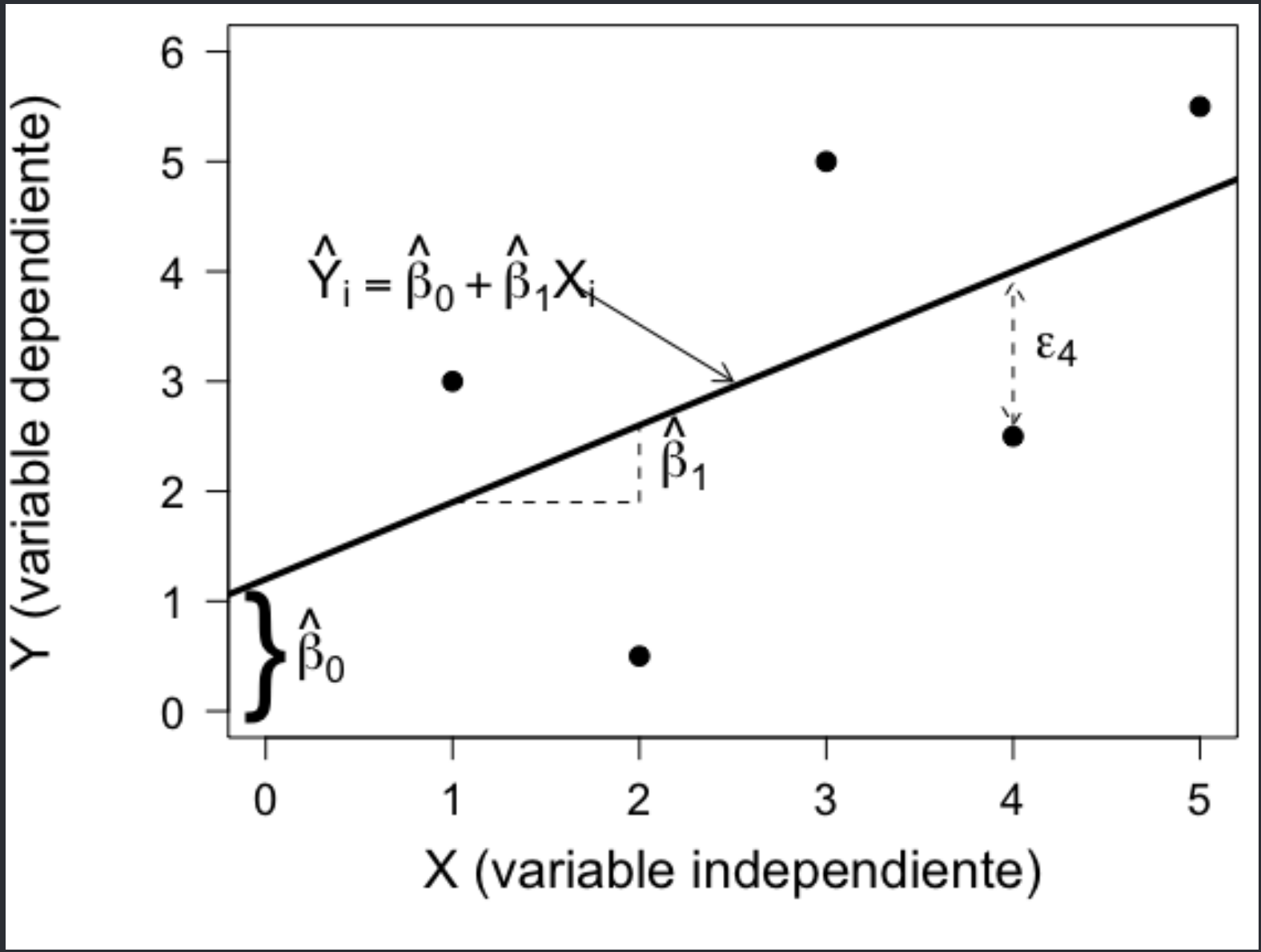
```
## [1] 1
```

Mejor predictor lineal (BLP)

```r
# Usamos la función "lm"
# lm es linear model
# coef pide  los coeficientes ("betas")
# round es para aproximarlos (3 digitos).
round(coef(lm(y~x)),3)

## (Intercept)           x
##       1.000       0.786
```

Mejor predictor lineal (BLP)



Estructura de la sección 3

Referencias

- Aronow, P., & Miller, B. (2015) Theory of Agnostic Statistics
- Capinski, M., & Zastawniak, T. (2000) Probability Through Problems.
- Carsey, T. y Harden, J. (2014) Monte Carlo Simulations and Resampling Methods for Social Science
- Medina, F. (2015) Razonamiento Estadístico, Magíster en Bioestadística, U. de Chile.
- Olea, R. (2013) Nivelación Probabilidad, Magíster en Sociología, PUC.
- Rincón, L. (2014) Introducción a la Probabilidad, UNAM.