EYP1113 - Probabilidad y Estadística

Laboratorio 6

Profesoras: Natalia Venegas y Pilar Tello

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2019

Contenido I

Múltiples Variables Aleatorias

Para el par de variables aleatorias X e Y se define la función de distribución de probabilidad acumulada como:

$$F_{X,Y}(x, y) = P(X \leq x,\ Y \leq y)$$

La cual satisface la axiomática fundamental de probabilidades:

- $F_{X,Y}(-\infty, -\infty) = 0$
- $F_{X,Y}(-\infty, y) = 0$
- $F_{X,Y}(x, -\infty) = 0$
- $F_{X,Y}(x, +\infty) = F_X(x)$
- $F_{X,Y}(+\infty, y) = F_Y(y)$
- $F_{X,Y}(+\infty, +\infty) = 1$

Si las variables aleatorias X e Y son discretas, la función de distribución de probabilidad conjunta es

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

siendo su función de distribución de probabilidad acumulada igual a

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P(X = x_i, Y = y_j),$$

con $(x_i, y_j) \in \Theta_{X,Y}$.

Ahora, si las variables aleatorias X e Y son continuas, la función de densidad de probabilidad conjunta se define como:

$$f_{X,Y}(x,y)\,dx\,dy = P(x < X \leq x + dx, y < Y \leq y + dy)$$

Entonces,

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du.$$

Si las derivadas parciales existen, entonces

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x,y)$$

Para variables aleatorias discretas X e Y, la probabilidad de $(X = x)$ puede depender de los valores que puede tomar Y (viceversa).

Con base a lo visto en probabilidades, se define la función de distribución de probabilidad condicional como:

$$p_{X|Y=y}(x) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad p_Y(y) > 0$$

De manera similar, se tiene que

$$p_{Y|X=x}(y) = P(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \quad p_X(X) > 0$$

La distribución marginal de una variable aleatoria se puede obtener aplicando el teorema de probabilidades totales.

Para determinar la distribución marginal de X, $p_X(x)$, tenemos que

$$\begin{aligned} p_X(x) &= \sum_{y \in \Theta_Y} p_{X|Y=y}(x) \cdot p_Y(y) \\ &= \sum_{y \in \Theta_Y} p_{X,Y}(x, y) \end{aligned}$$

De la misma forma se tiene que

$$p_Y(y) = \sum_{x \in \Theta_X} p_{X,Y}(x, y)$$

Múltiples Variables Aleatorias

Considere un grupo de 20 niños, según su edad y género de acuerdo a la tabla siguiente:

	Edad					
Sexo	9	10	11	12	13	14
0	1	0	4	2	1	1
1	1	3	1	1	3	2

Donde Sexo = 1 representa a una mujer y Sexo = 0 representa a un hombre.

1. Suponga que el experimento consiste en seleccionar un niño al azar. Encontrar la función de probabilidad conjunta de (X, Y) donde X es una variable aleatoria que registra el sexo del niño e Y es una variable aleatoria que registra la edad del niño.
2. Sea B el evento definido como: "la edad es un número par y es mujer". Calcule $P(B)$.
3. Encontrar la función de probabilidad de la variable aleatoria X.
4. Encontrar la función de probabilidad de la variable aleatoria Y.
5. Encontrar la función de probabilidad condicional de $Y \mid X$.
6. ¿Son independientes las variables aleatorias X e Y?
7. Supongamos que el precio del uniforme de una niño depende de su género. Si es hombre, el precio del uniforme es $2500 + 120\text{Edad}$ y si es mujer, el precio es $3000 + 150\text{Edad}$. Calcular el costo esperado de un uniforme.

La cantidad de huevos que pone un insecto tiene distribución Poisson de parámetro λ. La probabilidad que tiene cada huevo de sobrevivir es p. Asuma que la supervivencia de los distintos huevos son independientes.
Sea:
Y : Cantidad de huevos que pone un insecto.
X : Cantidad de huevos que sobrevive del insecto.
Luego:

$$Y \sim Poisson(\lambda)$$

$$X \mid Y = y \sim Binomial(y, p)$$

Obtenga la función de probabilidad conjunta utilizando $p = 0.6$ y $\lambda = 15$. Grafique su función.

Grafique la función de densidad conjunta de la variable aleatoria (X, Y) dada por:

$$f_{X,Y}(x, y) = \alpha\beta e^{-\alpha x - \beta y}, \qquad x, y, \alpha, \beta > 0$$

Introducción al Análisis de Regresión

Introducción al Análisis de Regresión

Cuando hay dos o más variables puede existir algún tipo de relación entre ellas.

En presencia de aleatoriedad la relación puede no ser única, por tanto se requiere de una descripción probabilística.

Cuando la relación probabilística es descrita en términos de la media y varianza de una de ellas en función de la otra, estamos frente a un Análisis de Regresión.

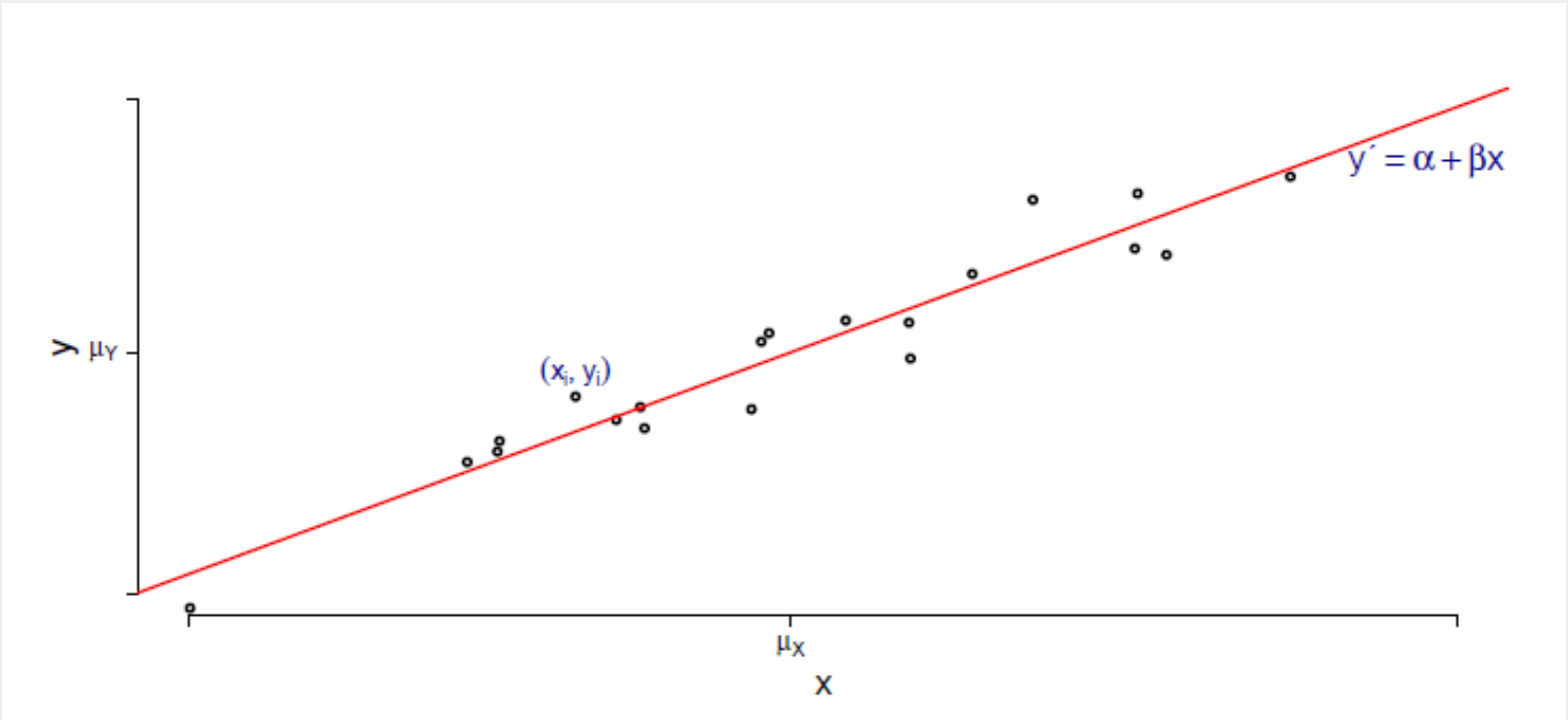
La relación lineal o no lineal obtenida por medio de un análisis de regresión no implica causalidad.

El grado de relación lineal entre dos variables puede ser medido por el coeficiente de correlación estadística.

Fundamentos del Análisis de Regresión Lineal

Fundamentos del Análisis de Regresión Lineal

Cuando un conjunto de pares de datos de dos variables, digamos *X* e *Y*, son graficados en dos dimensiones, tal gráfico se denomina “diagrama de punto” (o scatter).



Del gráfico se desprende, que ha medida que x crece, y tiende a incrementarse, o viceversa.

Sin embargo,conocido un valor de X, digamos dado $X = x$, no tengo información exacta sobre el valor de Y.

En términos promedio, se puede establecer una relación lineal entre X e Y, es decir:

$$\mathsf{E}(Y \mid X = x) = \alpha + \beta\, x$$

Estamos frente a un modelo de regresión lineal, donde α y β son los coeficientes de regresión (intercepto y pendiente, respectivamente).

Esta relación se conoce como la ecuación de regresión, y representa la regresión de *Y* sobre *X*.

Los coeficientes de regresión α y β deben ser estimados a partir de los datos.

En el diagrama de dispersión, es de esperarse que la varianza de *Y* dependa de los valores de *X*.

En general, la varianza condicional puede variar con x. Sin embargo, en esta primera etapa consideramos la varianza constante, es decir,

$$\text{Var}(Y \mid X = x) = \sigma^2$$

Se puede apreciar que “la mejor recta” es aquella que minimiza las distancias entre los puntos y ésta. Es decir, aquella que minimiza $|y_i - y_i'|$ para todo los pares de puntos.

Estimación de los parámetros α y β

Basados en una muestra de tamaño n, es decir, $(x_1, y_1),\ldots,(x_n, y_n)$. El error total absoluto puede representarse por

$$\Delta^2 = \sum_{i=1}^{n}(y_i - y_i')^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

Entonces, los estimadores de mínimos cuadrados de α y β están dados por

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

La ecuación de regresión estimada por mínimos cuadrados queda como

$$\mathsf{E}(Y \mid X = x) = \hat{\alpha} + \hat{\beta}\, x$$

Si X e Y son variables aleatorias, también puede realizarse la regresión de X sobre Y.

Esta nueva ecuación tiene un punto en común con la regresión de Y sobre X, que corresponde al "centro de gravedad" $(\overline{x}, \overline{y})$, donde se interceptan ambas rectas de regresión.

La ecuación de regresión predice el valor medio de Y como una función de X, siendo relevante la varianza de Y condicional al valor de X.

En este caso, hemos supuesto varianza constante con para todo x.

Un estimador insesgado para la varianza esta dado por:

$$
\begin{aligned}
s^2_{Y\,|\,x} &= \frac{\Delta^2}{n-2} \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - y_i')^2 \\
&= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right]
\end{aligned}
$$

Coeficiente de determinación ajustado

La razón de la varianza condicional relativa a la varianza original es una medida de la reducción de la varianza de Y al descontar la cantidad de variación de la varianza con X.

Esta reducción se representa por

$$r^2 = 1 - \frac{s^2_{Y\,|\,x}}{s^2_Y}$$

valor que se relaciona con el coeficiente de correlación ρ.

Análisis de Correlación

Análisis de Correlación

Intuitivamente, si $s_{Y|X}$ es cercano a cero, diremos que la ecuación provee buen predictor de Y para valores dados de X.

Sin embargo, una mejor medida de la relación lineal entre dos variables aleatorias X e Y es el coeficiente de correlación definido como

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

En otras palabras, el coeficiente de correlación es una medida de la calidad del ajuste de la recta de regresión.

Una forma de poder visualizar la correlación entre variables numéricas es utilizando el comando `pairs`.

Análisis de Correlación

Ejecutemos el siguiente código para visualizar las correlaciones de la base de datos `iris` que viene por defecto en R:

```
panel.cor=function(x, y, ...){
par(usr = c(0, 1, 0, 1))
txt=as.character(format(cor(x, y), digits=2))
text(0.5, 0.5, txt, cex = 6* abs(cor(x, y)))
}
pairs(iris[1:4], upper.panel=panel.cor)
```

Análisis de Correlación

Para (x_1, y_1),...,(x_n, y_n) observados, se define el estimador del coeficiente de correlación como

$$\hat{\rho}_{X,Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X \cdot s_Y}$$

Se puede mostrar que

$$\hat{\rho} = \hat{\beta} \frac{s_X}{s_Y} \quad \text{y} \quad \hat{\rho}^2 = 1 - \frac{(n-2)}{(n-1)} \frac{s^2_{Y|x}}{s^2_Y}$$

El coeficiente ρ varia entre -1 y +1.

Un valor próximo a ± 1 implica una fuerte asociación lineal.

En cambio, si $\rho \approx$ o diremos que no existe asociación lineal.

Consideremos dos variables aleatorias X e Ycon distribución conjunta Normal-Bivariada

$$f_{X,Y}(x,y) = \frac{1}{2\,\pi\,\sigma_X\,\sigma_Y\,\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\,\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}$$

Recuerde que

$$X \sim \text{Normal}(\mu_X,\, \sigma_X), \quad Y \sim \text{Normal}(\mu_Y,\, \sigma_Y)$$

$$Y \mid X = x \sim \text{Normal}\left(\mu_Y + \frac{\rho\,\sigma_Y}{\sigma_X}(x-\mu_X),\, \sigma_Y\,\sqrt{(1-\rho^2)}\right)$$

```
library(rgl)
f.xy = function(x, y, mu.x = 0, mu.y = 0, s.x = 1, s.y = 1,
rho  = 0){
n.r = length(x)
n.c = length(y)
M = matrix(NA, ncol = n.c, nrow = n.r)
for(i in 1:n.r){
M[i,] = dnorm(x[i], mean = mu.x, sd = s.x) * dnorm(y, mean = mu.y
+ rho*s.y*(x[i]-mu.x)/s.x, sd = s.y*sqrt(1-rho^2))
}
M
}
x = seq(-5,5,0.1)
y = seq(-5,5,0.1)
z = f.xy(x, y, rho = 0)
rgl.surface(x = x, y = z*10, z = y, color = "red", back="lines")
```

Notemos que en este caso

$$\begin{aligned} \mathrm{E}(Y \mid X = x) &= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \\ &= \left(\mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \cdot \mu_X\right) + \rho \frac{\sigma_Y}{\sigma_X} \cdot x \\ &= \alpha + \beta x \end{aligned}$$

y

$$\mathrm{Var}(Y \mid X = x) = \sigma_Y^2 (1 - \rho^2)$$

Aplicación

Descargue la `Data06.txt` de webcursos. Esta base contiene el peso, edad e índice de grasas para 25 individuos.

1. Construya un diagrama de dispersión de todas las posibles combinaciones de las variables.
2. Obtenga por definición los valores estimados $\hat{\alpha}$ y $\hat{\beta}$, para el modelo que relaciona la edad y las grasas.
3. Trace la recta de regresión. (Utilice la función **abline(a =** $\hat{\alpha}$**, b =** $\hat{\beta}$**)**)
4. Obtenga por definición el valor de $s_{Y|X}$ y el coeficiente de determinación r^2. Comente.
5. Utilizando **summary(lm(X∼Z))** compare los resultado obtenidos en (b) y (d).