Capítulo 2: Correlación y Regresión Simple

Fernando A. Crespo R.

30 de Septiembre de 2021

Índice

2.1 Covarianza y Correlación

- Hasta el momento hemos visto distribuciones conjuntas, medias y varianzas que entregan información útil de su distribuciones marginales. Pero ellas no nos entregan información de la relación entre dos variables, de como varían juntas.

Definición (2.1 Covarianza)

Sean X e Y v.a. que tienen distribución conjunta y cuyos primeros momentos y varianzas son $IE(X) = \mu_X$, $IE(Y) = \mu_Y$, $Var(X) = \sigma_X^2$ y $Var(Y) = \sigma_Y^2$. La covarianza de X e Y, que se denota por $Cov(X, Y)$, se define como:

$$Cov(X, Y) = IE\left[(X - \mu_X)(Y - \mu_Y)\right]. \tag{1}$$

-
- Se puede demostrar que si $\sigma_X^2 < \infty$ y $\sigma_Y^2 < \infty$, entonces existe la esperanza de (1) y $Cov(X, Y)$ será finita. $Cov(X, Y)$ puede tomar cualquier valor en IR.

2.1 Covarianza y Correlación

Definición (2.2 Correlación)

Si $0 < \sigma_X^2 < \infty$ *y* $0 < \sigma_Y^2 < \infty$, *entonces la correlación de* X *e* Y, *que se denota por* $\rho(X, Y)$, *se define como sigue:*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}. \tag{2}$$

▶

Teorema (2.1 Desigualdad de Schwartz)

Para cualquiera variables aleatorias U *y* V,

$$[IE(UV)]^2 \leq IE(U^2)IE(V^2). \tag{3}$$

▶

- De la Desigualdad de Schwartz se tiene $-1 \leq \rho(X, Y) \leq 1$.

Teorema (2.2)

Para cualquiera v.a. X *e* Y *tales que* $\sigma_X^2 < \infty$ *y* $\sigma_Y^2 < \infty$,

$$Cov(X, Y) = IE(XY) - IE(X)IE(Y). \tag{4}$$

2.1 Covarianza y Correlación

Teorema (2.3)

Si X e Y son v.a. independientes con $0 < \sigma_X^2 < \infty$ y $0 < \sigma_Y^2 < \infty$, entonces

$$Cov(X, Y) = \rho(X, Y) = 0. \tag{5}$$

- ▶
- ▶ Ejemplo 2.1.1: Variables aleatorias dependientes pero no correlacionadas. Suponga que la v.a. X puede tomar únicamente los tres valores -1, 0 y 1, que cada uno de estos tres valores tienen la misma probabilidad. Además, sea la v.a. Y definida por la relación $Y = X^2$.

Teorema (2.4)

Suponga que X es una v.a. tal que $0 < \sigma_X^2 < \infty$ y que $Y = aX + b$ para alguna constantes a y b, donde $a \neq 0$. Si $a > 0$, entonces $\rho(X, Y) = 1$. Si $a < 0$, entonces $\rho(X, Y) = -1$.

- ▶

2.1 Covarianza y Correlación

Teorema (2.5)

Si X e Y son v.a. con varianza finita, $\sigma_X^2 < \infty$ y $\sigma_Y^2 < \infty$, entonces

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y). \tag{6}$$

▶

2.1 Covarianza y Correlación

- Del Teorema 2.5 se obtiene:

$$Var(aX + bY + c) = a^2\,Var(X) + b^2\,Var(Y) + 2abCov(X, Y). \quad (7)$$

- También del Teorema 2.5 se obtiene:

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y). \quad (8)$$

Teorema (2.6)

Si $X_1, \ldots, X_n$ son v.a. tales que $Var(X_i) < \infty$ para $i = 1, \ldots, n$, entonces:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\,(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i, X_j). \quad (9)$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\,(X_i) + 2\sum_{i<j}\sum Cov(X_i, X_j). \quad (10)$$

-

2.1 Covarianza y Correlación

- Cálculos necesarios:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{11}$$

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{12}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{13}$$

$$\sigma_Y = \sqrt{Var(Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{14}$$

- Cálculo de covarianza:

$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \tag{15}$$

2.1 Covarianza y Correlación

- Cálculo correlación de Pearson:

$$r = r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{16}$$

2.3 Diagramas de Dispersión

- Es la gráfica para ver la relación de dos variables. Es la forma más intuitiva de ver la relaciones entre variables.
- Caso famoso: `https://es.wikipedia.org/wiki/Ley_de_elasticidad_de_Hooke`
- Caso famoso: `https://es.wikipedia.org/wiki/Ley_de_Hubble-Lemaitre`
- Caso famoso: `https://www.ligo.org/sp/science/Publication-GW170817Hubble/`

2.3 Diagramas de Dispersión

- Ejemplo 1: Veamos los datos en excel de rendimiento de automóviles respecto de su peso.

2.4 Prueba de Hipótesis de la Correlación

- $$\begin{aligned} H_0 : &\quad \rho = 0, \\ H_1 : &\quad \rho \neq 0 \end{aligned} \tag{17}$$
- Para calcular se tiene:
$$s_r = \sqrt{\frac{1-r}{n-2}}, \tag{18}$$
donde s_r es el error estándar del coeficiente de correlación, r es la correlación empírica, y n el número de observaciones pareadas.
- el estadístico:
$$t = \frac{r-\rho}{s_r}, \tag{19}$$
donde r es la correlación empírica, ρ es el valor hipotético, s_r el error estándar calculado con (18).
- t tiene una distribución $t-$student de $n-2$ grados de libertad.
- t se rechaza si $t > c$ o $t < -c$ con $F(c) = \frac{\alpha}{2}$.

2.4 Prueba de Hipótesis de la Correlación

- Una correlación entre dos variables no significa que una variable causa u ocasiona a la otra.
- Correlación 0 no significa que no exista relación entre variables, veamos un ejemplo gráfico.

2.5 Ecuaciones Lineales

- Cuando se menciona una correlación es para predecir una variable por la otra.
- Estudiar el valor que cambia, o denominada *variable dependiente*, habitualmente designada como y, por una variable independiente denominada x.
- La idea base es trazar una recta que permita ver la relación en el gráfico de dispersión.
- Para ello se plantea el modelo:

$$y = \beta_0 + \beta_1 x, \tag{20}$$

donde: β_0 es la ordenada al origen (o intercepción en y), β_1 pendiente de la recta.

- Este es un modelo que propone una relación entre las variables, si se permite algún error aleatorio a (20) se se denomina modelo estadístico:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{21}$$

con ϵ el error.

2.5 Ecuaciones Lineales

- La distribución de ϵ determina el grado de relación entre las variables independientes y dependiente.
- Para ello se hacen los siguientes supuestos:
  1. La distribución de probabilidad de ϵ es normal.
  2. La varianza de la distribución de ϵ es constante para todos los valores de x.
  3. La media de la distribución de probabilidad de ϵ es 0. Esta suposición implica que el valor medio de y para un valor de x es $IE(y) = \beta_0 + \beta_1 x$.
  4. Los valores de ϵ son independientes entre sí.

2.6 Método de Mínimos Cuadrados

- Para determinar la recta, se utiliza mínimos cuadrados:

$$\min \sum_{i=0}^{n} (y - \beta_0 - \beta_1 x)^2. \tag{22}$$

- al resolver (22), se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=0}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=0}^{n} X_i^2 - n\bar{X}^2} \tag{23}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{24}$$

2.6 Método de Mínimos Cuadrados

- Ejemplo 2.1, costo de mantenimiento anual de buses (USD) por tiempo de operación (años).
  1. Ver diagrama de dispersión.
  2. Ver si hay relación entre variables.
  3. Calcule el coeficiente de correlación.
  4. Pruebe el coeficiente de correlación para nivel 0.05.
  5. ¿Se puede usar la regresión lineal para analizar el costo?
  6. determine la ecuación del modelo lineal.
  7. Calcule el costo de mantenimiento anual para un bus con 5 años de operación.

2.7 Residuos

- Se denominan residuos a:

$$\hat{\epsilon} = y - \hat{y}. \tag{25}$$

  Donde y es el valor real e $\hat{y}$ valor estimado. Este es un valor empírico, es la estimación del error.

- Definimos como **error estándar de la estimación**, a la dispersión de los valores observados de y alrededor de la recta estimada:

$$s_{y,x} = \sqrt{\frac{\sum_{i=0}^{n}(y - \hat{y})^2}{n-2}}. \tag{26}$$

2.8 Predicción e Intervalo de Confianza

- La estimación puntual no proporciona información sobre la distancia que se encuentra respecto del parámetro poblacional. Para ello se requiere usar:

$$\hat{y} \pm t s_{\hat{y},x}. \tag{27}$$

con t la distribución de t student con $n-2$ grados de libertad, y $s_{\hat{y},x}$ el **error estándar de la estimación del pronóstico**.

- El error estándar de la predicción y:

$$s_{\hat{y},x} = s_{y,x} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=0}^{n} (X_i - \bar{X})^2}}. \tag{28}$$

X_p el valor dado de X, $\bar{X}$ la media de X, suma de cuadrados total para la variable X.

2.8 Predicción e Intervalo de Confianza

- Se puede pensar en estimar el valor medio de un número grande de experimentos para un valor de x:

$$\hat{y} \pm t s_{\hat{\mu},x}. \tag{29}$$

con t la distribución de t student con $n-2$ grados de libertad, y $s_{\hat{mu},x}$ **error estándar de distribución muestral del estimador de** y.

- La desviación estándar de la estimación:

$$s_{\hat{\mu},x} = s_{y,x} \sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=0}^{n}(X_i - \bar{X})^2}}. \tag{30}$$

X_p el valor dado de X, $\bar{X}$ la media de X, suma de cuadrados total para la variable X.

2.9 Coeficiente de determinación simple

- El coeficiente de determinación simple, R^2, mide el porcentaje de variabilidad en y que puede ser explicada por la variable predictora x:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=0}^{n}(Y_i - \bar{Y})^2}. \tag{31}$$

2.10 Prueba de Hipótesis de Análisis de la Regresión

- e quiere ver la hipótesis con dos colas para:

$$\begin{aligned} H_0: &\quad \beta_1 = 0, \\ H_1: &\quad \beta_1 \neq 0 \end{aligned} \tag{32}$$

- El error estándar del estimador se estima como:

$$s_b = \frac{s_{y,x}}{\sqrt{\sum_{i=0}^{n}(X_i - \bar{X})^2}}. \tag{33}$$

- El estadístico es:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_b}, \tag{34}$$

  donde β_1 es el valor hipotético.
- t tiene una distribución $t-$student de $n-2$ grados de libertad.
- t se rechaza si $t > c$ o $t < -c$ con $F(c) = \frac{\alpha}{2}$.