Análisis de regresión lineal múltiple

Sesión N° 4
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López

Universidad Diego Portales

Objetivos de la sesión

- Introducir el análisis de regresión lineal simple y múltiple
- Revisar la ejecución de modelos de regresión en R

Contenidos Sesión N°4

Introducción a regresión lineal

Regresión lineal múltiple

Supuestos del modelo de regresión

Introducción a regresión lineal

Varianza

Varianza: es el promedio de las distancias de los casos al promedio, tomando en cuenta los signos (eleva al cuadrado todas las distancias al promediarlas)

$$S^2 = \frac{\sum_{i=1}^{n} (X_j - \overline{X})^2}{n-1}$$

Desviación estándar: Es la raíz cuadrada de la varianza. Es la que mejor da cuenta de la dispersión (es decir de las distancias de los casos al promedio

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

Covarianza

La covarianza da cuenta del a variación conjunta de dos variables respecto de sus medias.
Puede tomar valores positivos, dando cuenta de una relación directa (por ejemplo a mayor educación, mayores ingresos) o valores negativos, dando cuenta de una relación inversa (por ejemplo a menores horas de trabajo, mayor satisfacción con la vida).

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

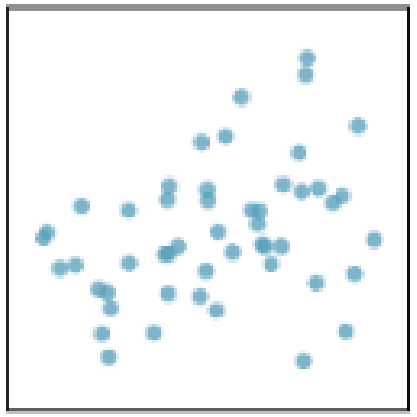
Correlación

La correlación (correlación de Pearson) corresponde a un valor estandarizado de la covarianza que puede tomar valores ente -1 y 1.

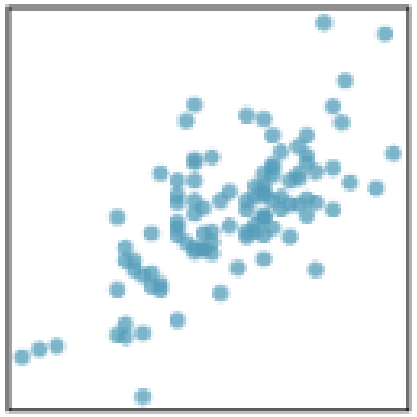
$$rXY = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi - \bar{X})^2}\ \sqrt{\sum_{i=1}^{n}(Yi - \bar{Y})^2}}$$

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

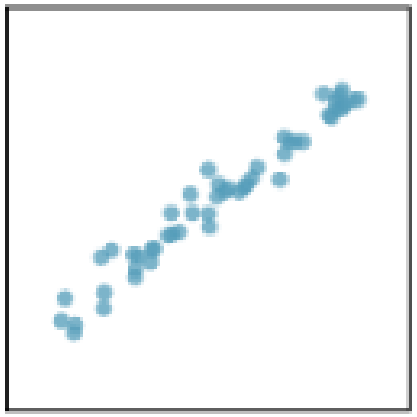
Correlación



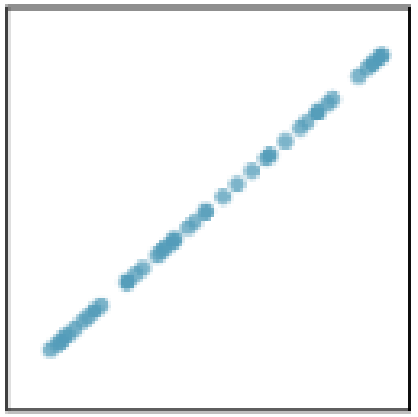
R = 0.33



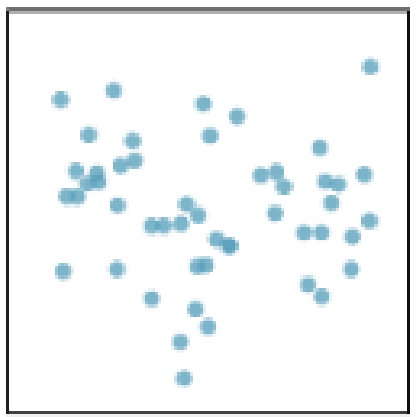
R = 0.69



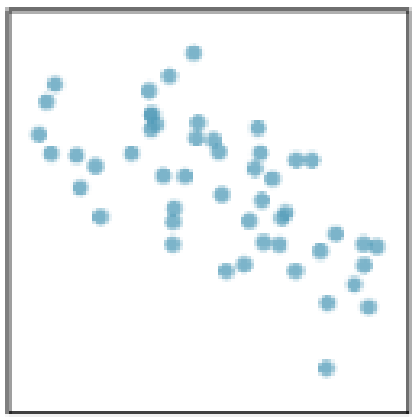
R = 0.98



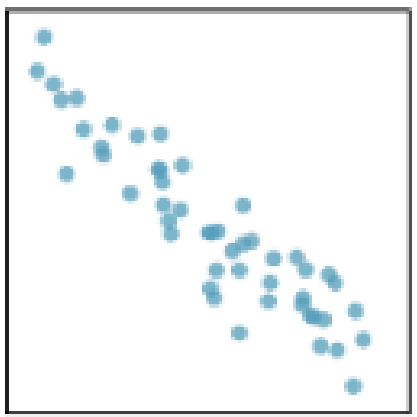
R = 1.00



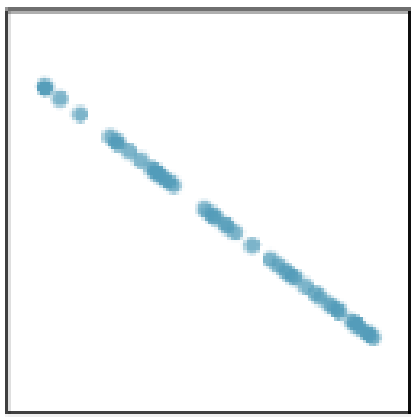
R = –0.08



R = –0.64



R = –0.92



R = –1.00

Regresión lineal

La regresión lineal es una técnica de análisis estadístico que nos permite estimar los efectos de ciertas variables (independientes o predictores) en una variable cuantitativa (dependiente o predicha).

Nos permite conocer la variación de una variable dependiente respecto a la variación de una o más variables independientes. Esto nos puede permitir predecir el valor que asume la variable dependiente á partir del valor de las independientes.

Al igual que las pruebas estadística bivariadas revisadas en la clases pasadas, nos permite hacer inferencia estadística, es decir, determinar si las relaciones observadas en el modelo de regresión son estadísticamente significativas.

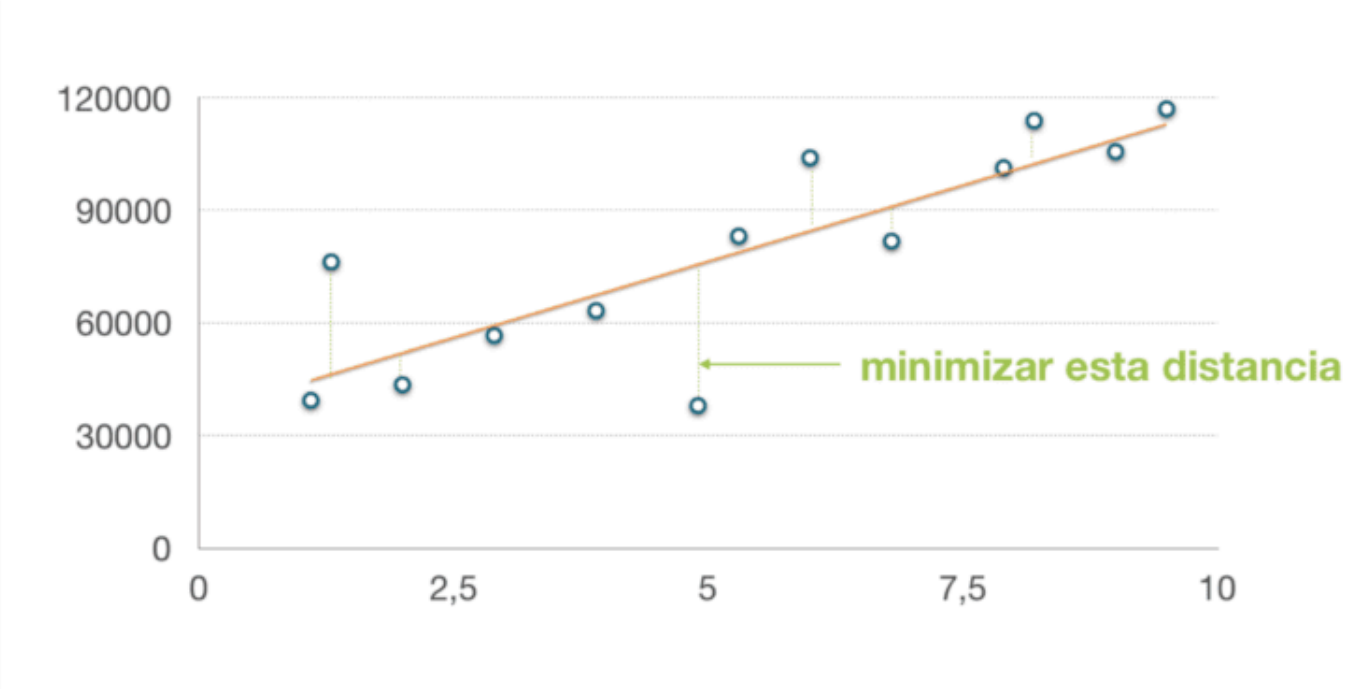
Recta de regresión

La covariación de Y respecto de X puede expresarse a partir de una recta.

$$Y = \alpha + \beta X$$

Y = Valor estimado de la variable dependiente
α = El intercepto (valor que asume Y cuando X es 0)
β = el coeficiente de regresión o pendiente es decir, el efecto en Y de un aumento de 1 en X
X = la variable independiente.

Estimación de mínimos cuadrados ordinarios



Para estimar dicha recta se utiliza la estimación de mínimos cuadrados ordinarios (Ordinary Least Squares OLS). Esta busca minimizar al suma de los residuos al cuadrado, siendo estos últimos la diferencia entre los valores predichos por el modelo (la recta) y cada valor observado.

Coeficientes de la ecuación

Constante o intercepto Valor esperado de Y cuando la variable independiente tiene el valor 0

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Coeficiente de regresión (β) Cambio en Y por cada aumento en una unidad de X, indica la relación entre X e Y

$$b_1 = \frac{Cov(XY)}{VarX}$$

Coeficiente de determinación R2

Estadístico de ajuste que describe la proporción de la varianza de Y que se relaciona con las variables independientes del modelo.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

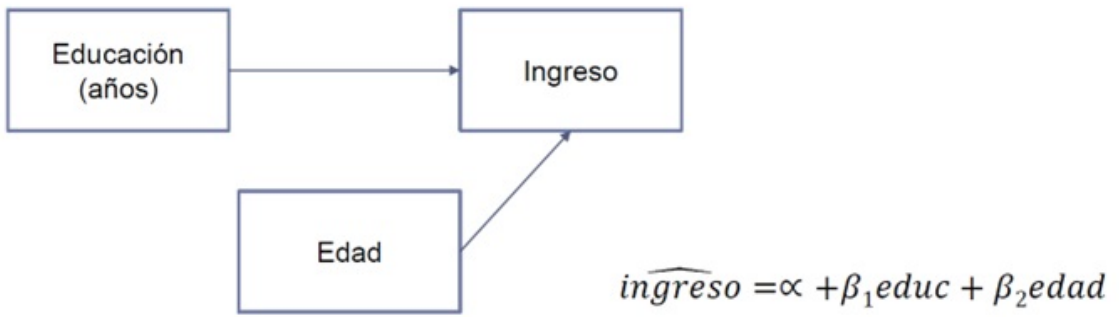
$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals} / (n - K)}{SS_{total} / (n - 1)}$$

Regresión lineal múltiple

Regresión lineal múltiple

Corresponde a una extensión del modelo de regresión lineal donde agregamos más variables independientes.

$$\hat{y}_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$



Regresión lineal múltiple

Constante o intercepto
Valor esperado de Y cuando las variables independientes tienen valor 0

Coeficiente de regresión (β)
Cambio en Y por cada aumento en una unidad de X, controlando por las demás variables en el modelo. Indica la relación parcial entre X e Y

Coeficiente de determinación R2
Igual interpretación que en el modelo de regresión simple Describe la proporción en la varianza de Y que es explicada por el modelo de regresión (en este caso, por todas las variables incluidas)

Control estadístico

En una regresión lineal múltiple buscamos dar cuenta de la relación entre una variable x1 sobre y considerando además el efecto de x2 sobre y.
Así, podemos estimar el efecto, por ejemplo, de la edad en los ingresos independiente de la educación, es decir, manteniendo esta última constante (controlando por).
Cuando no existe relación entre los predictores, la inclusión de otras variables no afecta el valor de los coeficientes de regresión.

Inferencia en los coeficientes de regresión

Igual que en el caso de la regresión lineal, pero evaluando si una variable tiene un efecto significativo una vez que controlamos por las demás variables en el modelo:

H0: βj=0, controlando por las demás variables H1: βj≠0, controlando por las demás variables

Distribución muestral t con n-(k+1) grados de libertad

Donde n= tamaño de la muestra k= número de variables independientes en el modelo

Supuestos del modelo de regresión

Ausencia de casos atípicos

Corresponden a observaciones individuales que no siguen el patrón de relación de los demás casos. Es decir, son casos que el modelo predice menos bien que a los demás casos.

- Estos casos pueden influenciar resultados, en particular si la muestra es chica.
- Se recomienda correr el modelo excluyendo casos atípicos para ver si cambian los estadísticos (pendiente y constante).

Relación lineal

- Y y X se relacionan por medio de una ecuación lineal (gráficamente, la relación forma una recta).
- Si la relación es aproximadamente lineal, tiene sentido usar modelos de regresión lineal.
- Se puede evaluar en base a diagramas de dispersión (si el n es pequeño)
- Una correlación alta entre las variables es indicación de que la relación es lineal.
- Si la relación no es lineal, considerar utilizar otro tipo de regresión o transformar variables (por ejemplo, ver efectos de variables al cuadrado).

Ausencia de multicolinealidad entre las variables dependientes

Cuando dos o más variables independientes están altamente correlacionadas.

- En estas situaciones resulta difícil estimar cuál de las dos variables es la que explica la variable dependiente, generando errores estándar altos y baja precisión de los coeficientes calculados.
- Para identificar esta situación, hay que revisar la matriz de correlación entre las variables y detectar correlaciones de 0,8 o más. Si este es el caso, es recomendable eliminar una de las dos variables del modelo.

Ausencia de multicolinealidad entre las variables dependientes

- También existen estadísticos que miden la presencia de multicolinealidad al correr el análisis de regresión. En particular:
- **Factor de inflación de la varianza, VIF**: indicador de cuánto aumenta el error estándar debido a problemas de multicolinealidad.
- Sacamos la raíz cuadrada e interpretamos el valor resultante como en cuantas veces mayor es el error estándar debido a problemas de multicolinealidad. Por ejemplo, un VIF de 4 significa que el error estándar es 2 veces mayor de lo que sería si las variables no estuvieran correlacionadas.
- Un VIF mayor a 2.5 es considerado como indicando problemas de multicolinealidad.

Homocedasticidad de los errores

Los residuos tienen una varianza constante a lo largo de los distintos valores de Y

- Residuos: diferencia entre el valor estimado por el modelo y el valor observado que tiene un caso en la muestra
- Esto quiere decir que el modelo es igualmente apropiado para predecir valores bajos (por ejemplo, bajos ingresos) y altos de la variable dependiente (por ejemplo, altos ingresos).
- El contrario es heterocedasticidad
  Solución: Modelos con errores estándares robustos

Normalidad en la distribución de los residuos

Los residuos en torno a los valores estimados de Y se distribuyen normalmente.

- Si los residuos se distribuyen normalmente, quiere decir que la mayor parte de los residuos se encuentran en torno a 0 (es decir, son valores que se alejan poco del valor observado).
- A su vez, son cada vez menos los residuos a medida que estos valores son mayores en términos absolutos.

Análisis de regresión lineal múltiple

Sesión N° 4
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López

Universidad Diego Portales