CARRERA DE INGENIERÍA DE SISTEMAS

Análisis de Regresión

Regresión Lineal Simple

ESTADÍSTICA APLICADA
PROF. MSC. VILMA S. ROMERO ROMERO
5TO CICLO | 16-06-2022

UNIVERSIDAD DE LIMA

Introducción

- La regresión lineal es una técnica estadística muy poderosa.
- El análisis de regresión permite desarrollar una ecuación (modelo de regresión) que muestra la relación de las variables, a partir de los datos de una muestra.
- Los objetivos de un análisis de regresión son:
  - Identificar **variables explicativas (X)** relacionadas con la **variable respuesta (Y)**.
  - Describir la forma de la relación entre las variables explicativas y la variable respuesta.
  - Proporcionar una ecuación de predicción de la variable respuesta en base a las variables explicativas.
- Existen abundantes opciones para ajustar modelos de regresión en R. Vito Ricci, en el 2005, creó una lista de alrededor de 205 funciones que se pueden utilizar. (https://cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf)

UNIVERSIDAD DE LIMA

Introducción

- Algunos ejemplos:
  - Estimar el precio de una vivienda en función de su superficie.
  - Estimar el tiempo de ejecución de un programa en base a la velocidad del procesador.
  - Estimar la nota obtenida en el curso de Estadística según el número de horas de estudio semanal.
  - Análisis de la efectividad del marketing, los precios y las promociones en las ventas de un producto.

Tipo de Regresión	Uso Típico
Lineal Simple	Predecir una variable respuesta cuantitativa en base a una variable explicativa cuantitativa.
Lineal Múltiple	Predecir una variable respuesta cuantitativa en base a una o varias variables explicativas.
Polinomial	Predecir una variable respuesta cuantitativa en base a una variable explicativa cuantitativa donde la relación considerada es un polinomio de orden n.
Robusta	Predecir una variable respuesta cuantitativa en base a una o varias variables explicativas considerando una metodología resistente al efecto de observaciones influyentes.
Logística	Predecir una variable categórica en base a una o varias variables explicativas.
Multivariada	Predecir más de una variable respuesta en base a una o más variables explicativas.
Poisson	Predecir una variable respuesta que representa conteos en base a una o más variables explicativas.
Multilevel	Predecir una variable respuesta en base a datos que poseen una estructura jerárquica (por ejemplo, estudiantes dentro de salones dentro de un colegio). Conocidos también como Modelos Mixtos o Jerárquicos.
No Lineal	Predecir una variable respuesta cuantitativa en base a una o varias variables explicativas, donde la forma del modelo es no lineal.
No Paramétrica	Predecir una variable respuesta cuantitativa en base a una o varias variables explicativas, donde la forma del modelo se deriva de los datos y no es especificado apriori.
Series de Tiempo	Modelar datos de series de tiempo con errores correlacionados.
Cox Proportional Hazards	Predecir el tiempo a la ocurrencia de un evento (muerte, falla, reincidencia) en base a una o más variables explicativas.

Tabla tomada de Kabacoff (2015)

¿Qué es Regresión Lineal Simple?

Predecir una **variable respuesta cuantitiva (Y)** desde **una única variable predictora (X)**.

La relación que se establece entre dichas variables es una línea recta (conocida como **línea de regresión**).

Nota:
"Simple" = Evaluación de la relación existente entre dos variables.

Modelo de Regresión Lineal Simple

$$Y_i = \boxed{\beta_0 + \beta_1 X_i} + \varepsilon_i \qquad (1)$$

- β_0 y β_1 son los **coeficientes de regresión**.
  - β_0 = Intercepto con el eje Y poblacional
  - β_1 = Pendiente poblacional.
- ε_i = Error aleatorio que representa la variabilidad no explicada por el modelo lineal (errores de muestreo, otras variables no consideradas).
- $\boxed{\beta_0 + \beta_1 X_i}$ = Componente Lineal

Variable Y	**Variable X**
Variable de interés	Variable Explicativa
Variable Dependiente	Variable Independiente
Variable Respuesta	Covariable

- Los parámetros a estimar son: β_0, β_1 y σ

UNIVERSIDAD DE LIMA

Ecuación de Regresión Lineal Simple

La **ecuación de regresión lineal simple estimada** o **línea de predicción** proporciona una **estimación** de la regresión lineal poblacional.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{2}$$

- $\hat{\beta}_0$ = Estimación del intercepto de regresión.
- $\hat{\beta}_1$ = Estimación de la pendiente de regresión.
- $\hat{Y}_i$ = Valor estimado (o predicho) de Y para la observación i.
- X_i = Valor de X para la observación i.

UNIVERSIDAD DE LIMA

Supuestos

El modelo debe satisfacer los siguientes supuestos:

1) **Linealidad**, la relación entre X e Y es lineal.
2) **Homogeneidad**, el valor esperado de los errores es cero. Hay errores por exceso y por defecto que en promedio se anulan.

$$E(\varepsilon_i) = 0$$

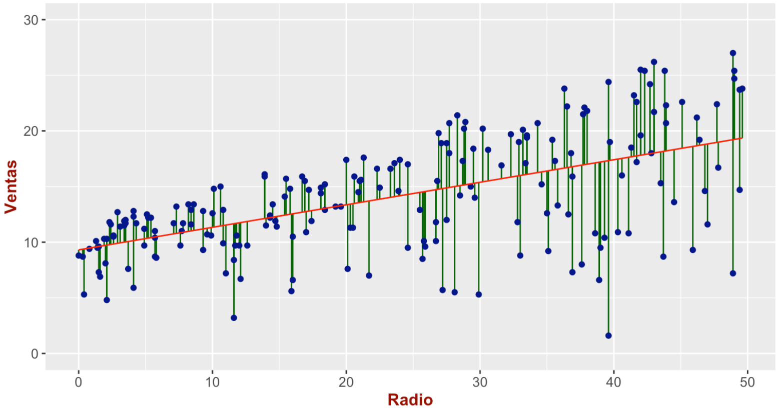
3) **Homocedasticidad**, la varianza de los errores es constante.

$$V(\varepsilon_i) = \sigma^2$$

4) **Independencia**, los errores son independientes. Esto implica que las observaciones son independientes.
5) **Normalidad**, los errores siguen una distribución normal.

$$\varepsilon_i \sim N(0, \sigma^2)$$

Estimación - Idea inicial



- **Residual**: Diferencia entre el valor observado y su valor estimado asociado. Indica que tan lejos está la predicción del modelo en ese punto.

$$\boxed{e_i = y_i - \hat{y}_i}$$

UNIVERSIDAD DE LIMA

Método de Mínimos Cuadrados Ordinarios

Encontrar el valor de los coeficientes de regresión de tal modo que la suma de los cuadrados de las diferencias entre los valores observados y la línea de regresión sea mínima.

Matemáticamente, minimizar la suma de cuadrados de los residuales:

$$Q = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2$$

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3}$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{4}$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (x_i y_i - x_i \hat{\beta}_0 - \hat{\beta}_1 x_i^2) \tag{5}$$

UNIVERSIDAD DE LIMA

Método de Mínimos Cuadrados Ordinarios

Igualando a cero las expresiones (4) y (5), se obtienen las "Ecuaciones normales"

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{6}$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{7}$$

Despejando los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

UNIVERSIDAD DE LIMA

Condiciones para la recta de mínimos cuadrados

1) Linealidad, los datos deben mostrar una tendencia lineal.
2) Normalidad, generalmente los residuales deben ser casi normales.
   - Si la condición no se cumple $\Rightarrow$ Presencia de outliers u observaciones influyentes.
3) Variabilidad constante, la variabilidad de las observaciones alrededor de la línea de mínimos cuadrados permanece aproximadamente constante.
4) Independencia
   - Prestar atención a datos de series de tiempo, datos secuenciales que tienen una estructura subyacente que debe considerarse en el modelo.

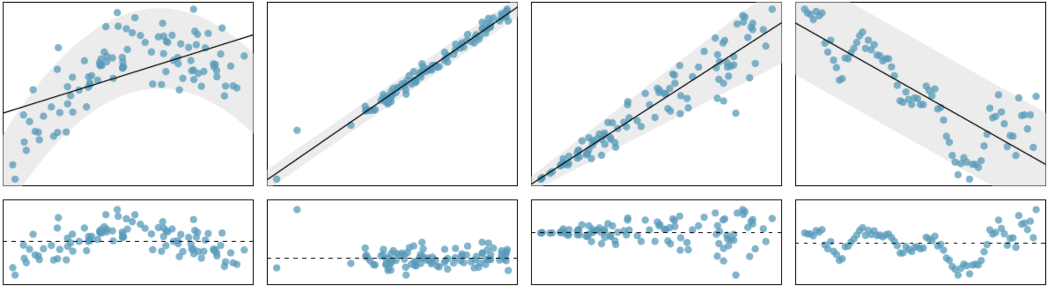


Fig. 1: Cuatro ejemplos que muestran que los métodos de regresión lineal son insuficientes para aplicar a los datos. **Panel 1**, una línea recta no se ajusta a los datos. **Panel 2**, hay dos valores atípicos. **Panel 3**, la variabilidad de los datos alrededor de la línea aumenta con valores mayores de X. **Panel 4**, se muestra un conjunto de datos de series temporales, donde las observaciones sucesivas están altamente correlacionadas. Imagen tomada de Diez et al. (2015)

Interpretación de los coeficientes estimados

Interpretación del Intercepto $\hat{\beta}_0$
Es el valor esperado de Y cuando X toma el valor de 0.

Interpretación de la pendiente o coeficiente de regresión $\hat{\beta}_1$
Es el cambio promedio en Y producido por el cambio en una unidad de X. Además, si:

- $\hat{\beta}_1 > 0$: Tendencia lineal creciente.
- $\hat{\beta}_1 < 0$: Tendencia lineal decreciente.
- $\hat{\beta}_1 = 0$: $y = \hat{\beta}_0$, no existe regresión e Y permanece constante para cualquier valor de X.

UNIVERSIDAD DE LIMA

Estimación de la varianza del error

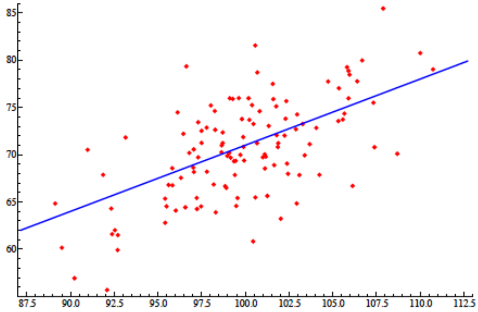
Un estimador insesgado de σ^2 es:

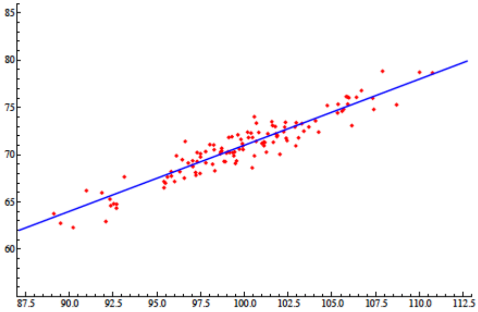
$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i}{n-2}$$

- s_e^2 también se denomina cuadrado medio del error (MSE) o varianza residual.
- La desviación estándar de s_e^2 se denomina error estándar de estimación.

Interpretación de la varianza del error σ^2

Describe qué tan grandes son los errores en promedio.





- Valor pequeño = Los datos se encuentran más cercanos a la línea de regresión.
- Determinan el ancho de los intervalos predictivos.

UNIVERSIDAD DE LIMA

Inferencia sobre el Modelo de Regresión

- Intervalos de confianza, para obtener una medida de precisión de los coeficientes estimados.
- Pruebas de hipótesis, para evaluar si un valor determinado puede ser el verdadero valor del parámetro.

Intervalo de confianza para los parámetros

El intervalo de confianza del $(1-\alpha)100\%$ para estimar el **intercepto** está dado por:

$$\boxed{\beta_0 \in \left\langle \hat{\beta}_0 \pm t_{\left(n-2;1-\frac{\alpha}{2}\right)} es(\hat{\beta}_0) \right\rangle}$$

Donde $es(\hat{\beta}_0)$ es el error estándar de $\hat{\beta}_0$:

$$es(\hat{\beta}_0) = \sqrt{s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}$$

La longitud del intervalo disminuirá si:

- El tamaño de la muestra aumenta.
- La varianza de las x_i aumenta.
- La varianza residual disminuye.
- La media de las x_i disminuye.

UNIVERSIDAD DE LIMA

Intervalo de confianza para los parámetros

El intervalo de confianza del $(1-\alpha)100\%$ para estimar la **pendiente** está dado por:

$$\boxed{\beta_1 \in \left\langle \hat{\beta}_1 \pm t_{\left(n-2;1-\frac{\alpha}{2}\right)} es(\hat{\beta}_1) \right\rangle}$$

Donde $es(\hat{\beta}_1)$ es el error estándar de $\hat{\beta}_1$:

$$es(\hat{\beta}_1) = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \sqrt{\frac{s_e^2}{(n-1)s_x^2}}$$

La longitud del intervalo disminuirá si:

- El tamaño de la muestra aumenta.
- La varianza de las x_i aumenta.
- La varianza residual disminuye.

Prueba de Hipótesis para la pendiente

1) Hipótesis:
   $H_0 : \beta_1 = 0$ (la variable X no es significativa en el modelo)
   $H_1 : \beta_1 \neq 0$ (la variable X es significativa en el modelo)

2) Especificar el nivel de significación α.

3) Calcular el valor del estadístico de prueba: $t_0 = \dfrac{\hat{\beta}_1}{es(\hat{\beta}_1)} \sim t_{(n-2)}$

4) Región crítica y regla de decisión:

$$RC = \left\langle -\infty; t_{(n-2,\alpha/2)} \right\rangle \cup \left\langle t_{(n-2,1-\alpha/2)}; \infty \right\rangle$$

$\Rightarrow$ **Rechazar** H_0 si $t_0 < t_{(n-2,\alpha/2)}$ o $t_0 > t_{(n-2,1-\alpha/2)}$

UNIVERSIDAD DE LIMA

Tabla de Análisis de Varianza (ANOVA)

La variación total de la variable respuesta con respecto a su media se puede descomponer en:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Variación Total	=	Variación explicada	+	Variación no explicada
Suma de Cuadrados Total (SCT)	=	Suma de Cuadrados de la Regresión (SCR)	+	Suma de Cuadrados del Error (SCE)

Tabla de Análisis de Varianza (ANOVA)

Fuente de variación	**Grados de libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\mathbf{F_0}$
Regresión	1	SCR	$CMR = \frac{SCR}{1}$	$F_0 = \frac{CMR}{CME}$
Error	$n-2$	SCE	$CME = \frac{SCE}{n-2}$	
Total	$n-1$	SCT		

Coeficiente de Determinación

Medida de bondad de ajuste que indica la proporción de varianza en Y que puede ser explicada por medio de X.

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$.
  - $\mathbf{R^2 = 1}$: Todos los valores de los datos caen en la línea de regresión, correlación perfecta entre las variables X e Y.
  - $\mathbf{R^2 = 0}$: Todos los valores ajustados son iguales a una misma constante, no existe correlación entre las variables.
- Cuanto mayor sea el valor de R^2, mejor será el ajuste de la línea de regresión a los datos.

UNIVERSIDAD DE LIMA

Recursos Adicionales I

Devore, J. (2019). *Introducción a la probabilidad y estadística para ingeniería y ciencias*. Cengage, 1 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Johnson, R. A. (2012). *Probabilidad y estadística para ingenieros*. Pearson Educación, 8 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

Kokoska, S. (2015). *Introductory Statistics*. W. H. Freeman and Company, 2 edition.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2015). *Introducción a la probabilidad y estadística*. Cengage, 14 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

UNIVERSIDAD DE LIMA

Recursos Adicionales II

Millones, R., Barreno, E., Vásquez, F., and Castillo, C. (2017). *Estadística Descriptiva y Probabilidades: Aplicaciones en la ingeniería y los negocios*. Lima: Fondo Editorial de la Universidad de Lima, 1 edition. Código Biblioteca U.Lima: 519.53 E.

Triola, M. (2018). *Estadística*. Pearson Educación, 12 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

UNIVERSIDAD DE LIMA