Regresión lineal

Parcial 1

Dr. José Juan Castro Alva

UDLAP

30 de octubre de 2022

Correlación y Regresión

1. Correlación

2. Introducción

Datos bivariados

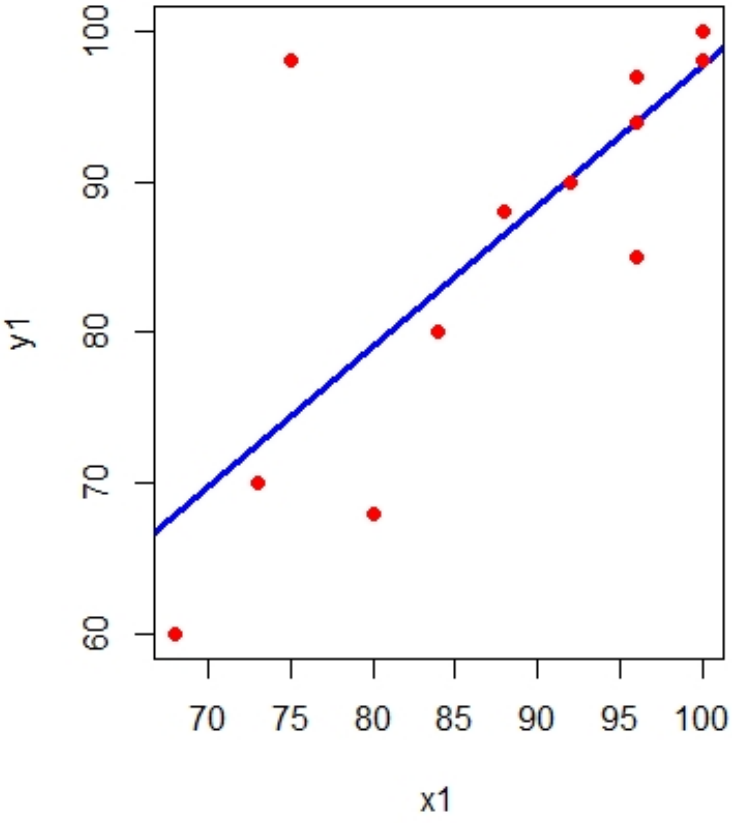
Cuando dos variables se miden en una sola unidad experimental, los datos resultantes se denominan **datos bivariados**.
Muchas veces es interesante explorar la relación entre las dos variables en caso de que exista alguna.

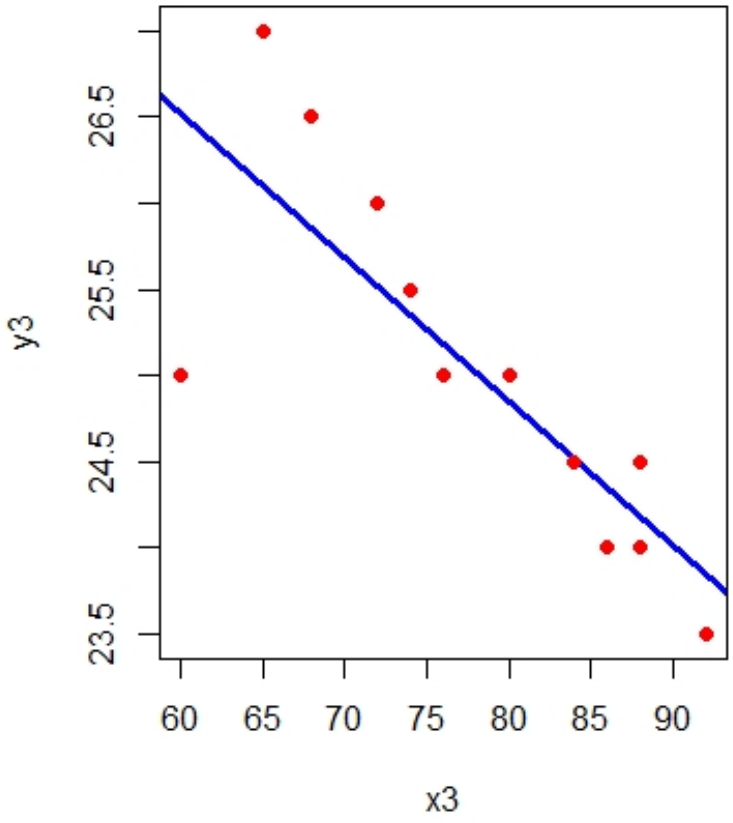
Método gráfico

- Análisis univariado.
  - Barplot
  - Histogram
  - Boxplot
  - qqplot
- Análisis biivariado. Scatter Plot

Ventajas de usar Scaterplot

- Permite observar si existe relación entre dos variables, x y y, de tipo lineal, no lineal o sin relación.
- Tendencia constante hacia arriba o hacia abajo que siga un modelo en líneal.
- Observar si la mayoría de los puntos siguen una tendencia lineal.
- Detectar Outliers


y1
x1
60
70
80
90
100
70
75
80
85
90
95
100


y3
x3
23.5
24.5
25.5
26.5
60
65
70
75
80
85
90

Introducción

El modelo de regresión lineal es un modelo estadístico que se utiliza para predecir el comportamiento (valor) de un una variable de respuesta cuantitativa, en términos de una variable independiente llamada predictor.

Puntos importantes a considera

- ¿Existe algún tipo de relación entre la variable respuesta y la o las variables predictoras?
- en caso de exister relación entre la variable respuesta y predictora. ¿Qué tan fuerte es la relación?
- ¿Cómo contribuye cada una de las variables predictoras?
- ¿Cómo estimar el efecto de las variables predictoras?
- ¿Cómo pronosticar el comportamiento de las variable respuesta?
- ¿Qué forma tienen la relación?

Modelo probabilistico de linea recta

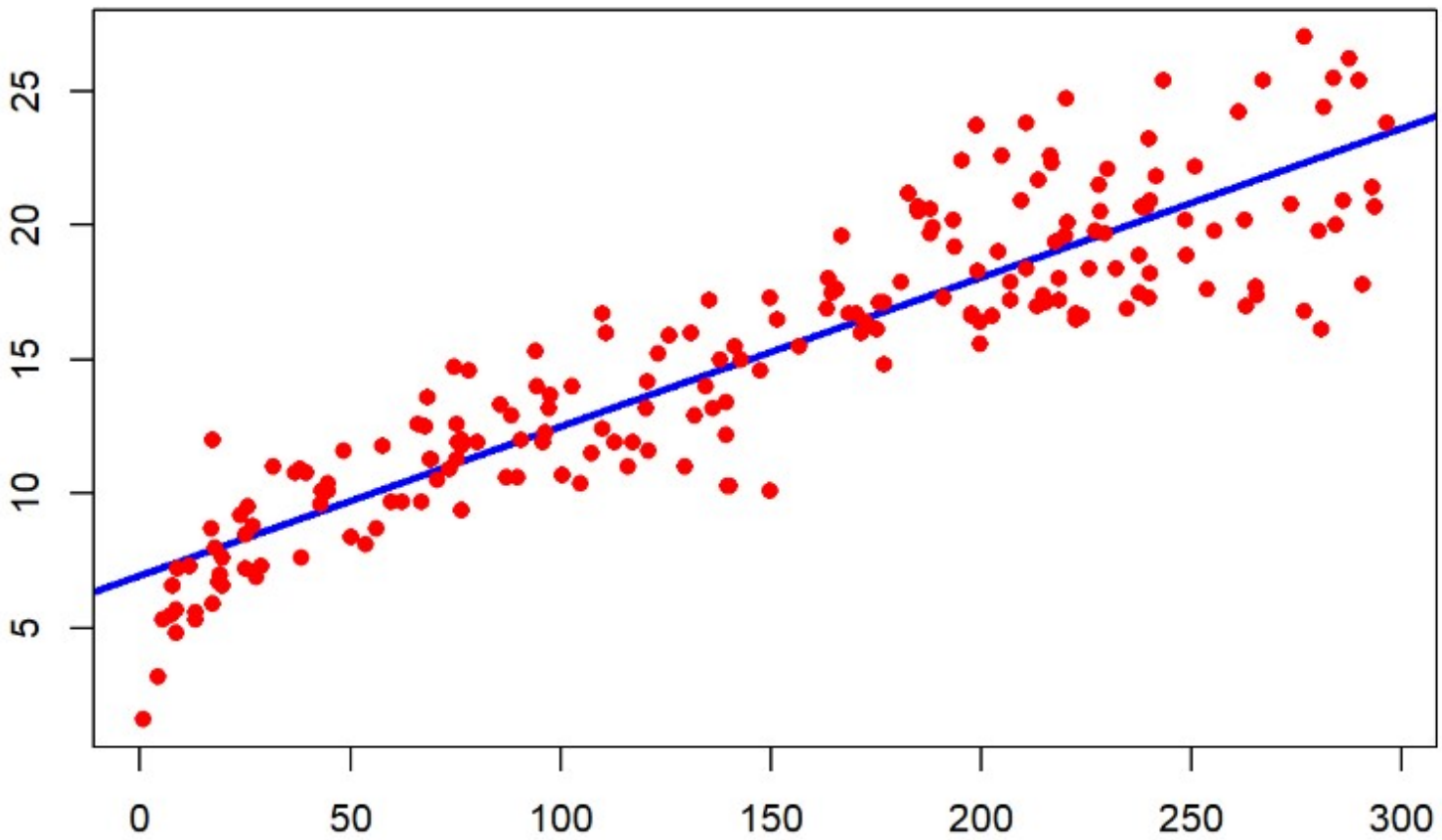
El modelo lineal de primer orden es el modelo de probabilidad más simple, el cual se representa de la siguiente manera.

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde

- y: Denota la variable dependiente o variable de respuesta
- x: Variable independiente o variable predictora.
- ϵ: La componente de error aleatorio
- β_0: El y-interceptor de la recta
- β_1: La pendiente de la linea recta
- $E[y] = \beta_0 + \beta_1 x$: La componente determinista

Representación Gráfica



Ajuste del modelo: Método de mínimos cuadrados

Definición (Erorr)

Se le llama error a la diferencia entre el valor observado y el valor pronosticado por el modelo de linea recta.

La forma en que se mide el error es considerando la distancia de la liner recta vertical que va del punto observado a la linea recta. Generalmente los errores se consideran como una componente aleatoria.

- SE: suma de los errores
- SSE: Suma de el cuadrado de los errores

Método de mínimos cuadrados

La recta que minimiza la suma de cuadrados de los errores, se le llama recta del mejor ajuste, **recta de mínimos cuadrados** o recta de regresión.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son las estimaciones de los parámetros β_0 y β_1 respectivamente.

Para reducir al mínimo las distancias desde los puntos a la recta ajustada, se utiliza **el principio de mínimos cuadrados**.

Principio de mínimos cuadrados

La suma del cuadrado de las desviaciones por lo general se denomina suma de cuadrados de error (SSE) y se define como:

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x\right)^2$$

Estimadores de mínimos cuadrados de $\hat{\beta}_0$ y $\hat{\beta}_1$

Los estimadores son

$$\hat{\beta}_1 = \frac{SC_{xy}}{SC_x}$$

y

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

donde las cantidades SC_{xy}, SC_x y SC_y están definidas como:

$$\begin{aligned} SC_{xy} &= \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = \sum x_i y_i - n\bar{x}\bar{y} \\ SC_x &= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n(\bar{x})^2 \\ SC_y &= \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = \sum y_i^2 - n(\bar{y})^2 \end{aligned}$$

Se tiene que SC_x y SC_y siempre deben ser positivas.

Ejercicio

Se muestran las calificaciones del examen de matemáticas de 10 estudiantes de primer año de universidad, junto con sus calificaciones finales en cálculo. Encuentre la recta de predicción de mínimos cuadrados para estos datos.

Calificación Matemáticas	Calificación Cálculo
39	65
43	78
21	52
64	82
57	92
47	89
28	73
75	98
34	56
52	75

	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
	39	65	1521	2535	4225
	43	78	1849	3354	6084
	21	52	441	1092	2704
	64	82	4096	5248	6724
	57	92	3249	5244	8464
	47	89	2209	4183	7921
	28	73	784	2044	5329
	75	98	5625	7350	9604
	34	56	1156	1904	3136
	52	75	2704	3900	5625
suma	460	760	23634	36854	59816

Así,

$$
\begin{aligned}
SC_{xy} &= \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 36854 - \frac{(460)(760)}{10} = 1894 \\
SC_x &= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 23634 - \frac{(460)^2}{10} = 2474 \\
SC_y &= \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 59816 - \frac{(760)^2}{10} = 2056 \\
\bar{x} &= \frac{\sum x_i}{n} = \frac{460}{10} = 46 \\
\bar{y} &= \frac{\sum y_i}{n} = \frac{760}{10} = 76
\end{aligned}
$$

Entonces:

$$b = \frac{SC_{xy}}{SC_x} = \frac{1894}{2474} = 0.7655618432$$

y

$$a = \bar{y} - b\bar{x} = 76 - 0.7655618432(46) = 40.78415521$$

La recta de regresión es entonces:

$$\hat{y} = a + bx = 40.78415521 + 0.7655618432x$$

Recta de regresión

La recta de regresión se puede usar para predecir y para un valor determinado de x sustituyendo el valor apropiado de x en la ecuación. Por ejemplo, si se obtuvo una calificación de $x = 60$ en matemáticas, la calificación que se espera obtener para cálculo es

$$\hat{y} = a + b(60) = 40.78415521 + 0.7655618432(60) = 86.7178658$$

Solución usando R-studio

```
mate<-c(39,43,21,64,57,47,28,75,34,52)
calculo <- c(65,78,52,82,92,89,73,98,56,75)
plot(mate,calculo, col= "red",pch=16,lwd=3,
abline(lm(calculo~mate),col="blue",lwd=3))

model <- lm(calculo~mate)
model
summary(model)
```

Solución usando R-studio

```
Call:
lm(formula = calculo ~ mate)

Coefficients:
(Intercept)         mate
40.7842       0.7656
```

Solución usando R-studio

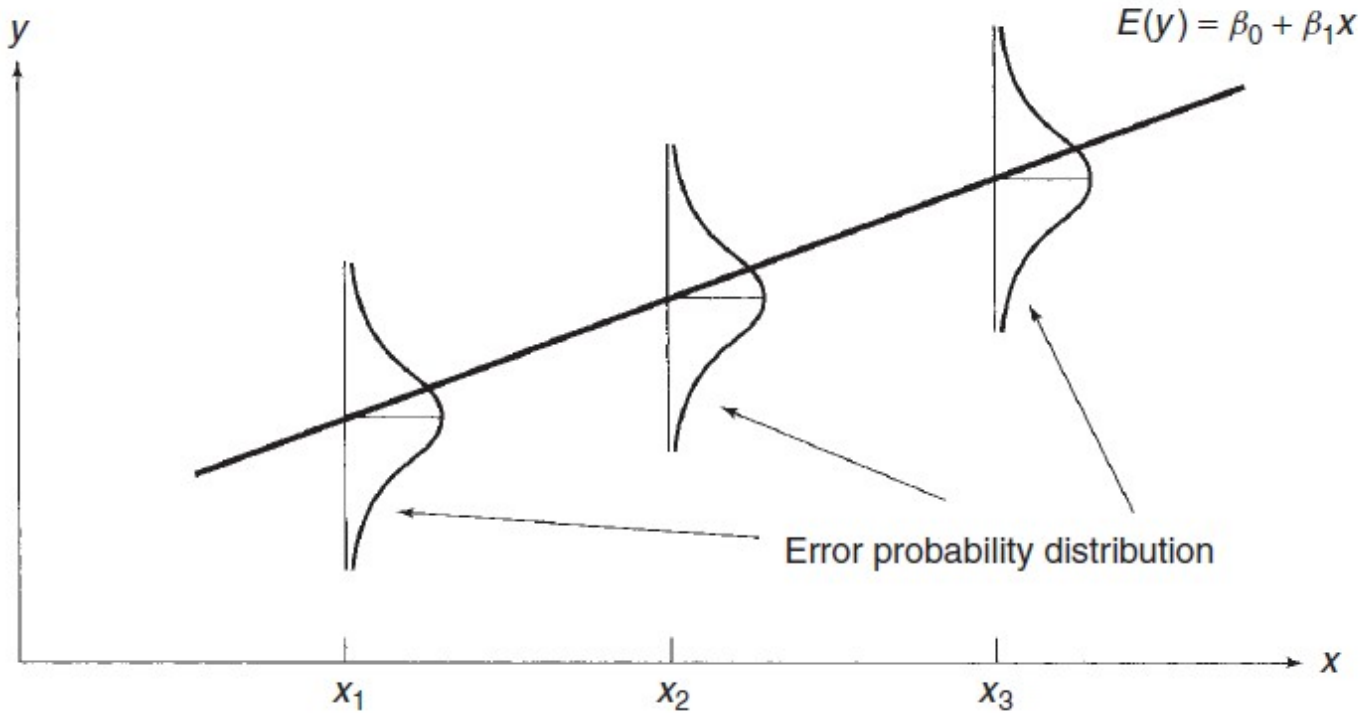
```
Call:
lm(formula = calculo ~ mate)

Residuals:
Min       1Q  Median      3Q     Max
-10.813  -5.629  -2.531   6.758  12.234
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.7842     8.5069   4.794  0.00137 **
mate          0.7656     0.1750   4.375  0.00236 **
---
Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 8.704 on 8 degrees of freedom
Multiple R-squared:  0.7052,Adjusted R-squared:  0.6684
F-statistic: 19.14 on 1 and 8 DF,  p-value: 0.002365
```

Problema de aplicación

Considerando la base de datos Advertising muestran las ventas (en miles de unidades) de un producto en particular en función de los presupuestos publicitarios (en miles de dólares) para TV, radio y newspaper media.
Se le pide sugerir, un plan de marketing para el próximo año que resultará en altas ventas de productos.
¿Qué información sería útil para ofrecer tal recomendación?

Supuestos de la distribución de ϵ

- La media de la distribución de probabilidad de ϵ es cero, ie $E[e_i] = 0$ para todo i. Esto implica que $E[y] = \beta_0 + \beta_1 X$.
- La varianza de la distribución de probabilidad de ϵ es constante i.e, $V[e_i = \sigma^2]$ para todo i.
- La distribución de probabilidad de ϵ es normal, i.e $e_i \sim N(0, \sigma^2)$.
- Los errores asociados con dos observaciones diferentes son independiente


y
$E(y) = \beta_0 + \beta_1 x$
Error probability distribution
x
x_1
x_2
x_3

Estimación de σ^2

En la práctica generalmente σ^2 es desconocido y se debe estimar usando los datos observados, es conocido que el estimador de σ^2 es S^2

Estimación de σ^2 para el modelo de regresión lineal

$$S^2 = \frac{SSE}{\text{df for error}} = \frac{SSE}{n-2}$$

donde

$SSE = \sum_i (y_i - \hat{y}_i)^2 = SC_y - \hat{\beta}_1 SC_{xy}$

$SC_y = \sum y_i^2 - n(\bar{y})^2$

$SC_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$

Así S^2 estima la varianza de ϵ y $S = \sqrt{S^2}$ estima la desviación estándar de ϵ.

Regla empírica: Se espera que aproximadamente el 95 % de los datos observados (valores en eje Y) se encuentren a dos desviaciones estándar con respecto al valor pronosticado por la recta de mínimos cuadrados.

Coeficiente de variación

El coeficiente de variación es la razón de la desviación estándar estimada de ϵ entre la media muestral de la variable dependiente $\bar{y}$, este coeficiente se mide en porcentaje

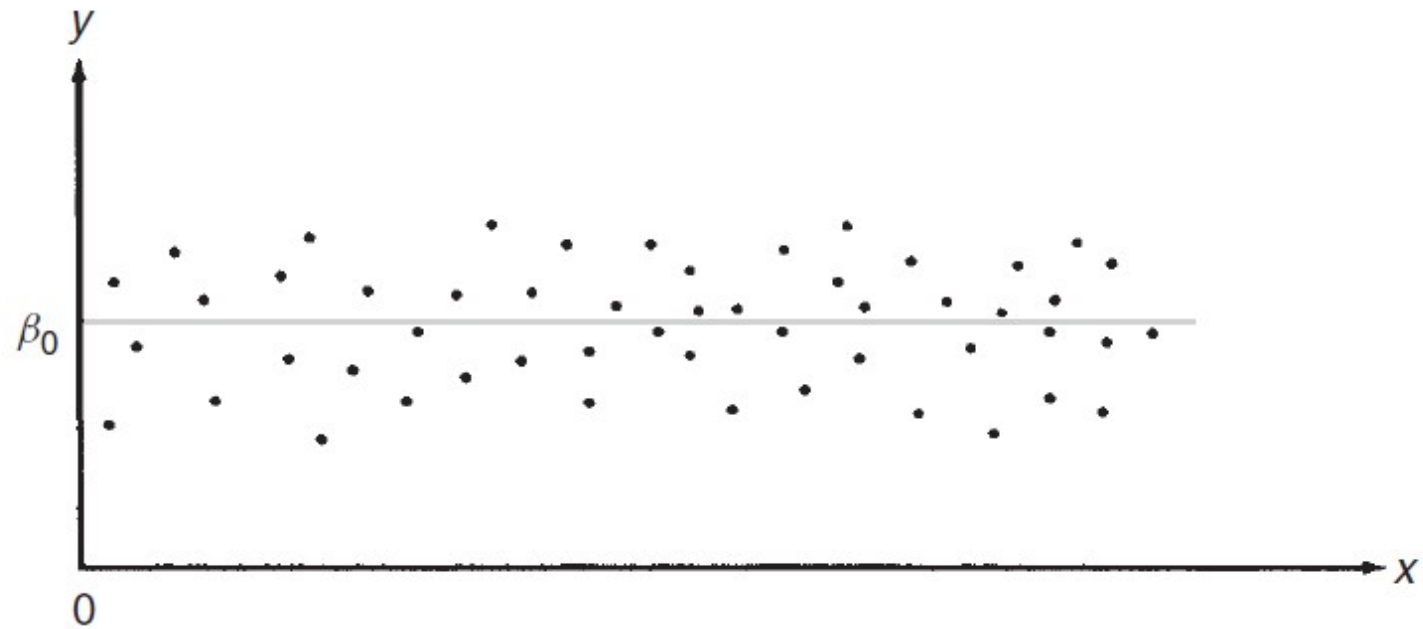
$$C.V = 100\frac{S}{\bar{y}}$$

En un modelo con buen desempeño se espera que el el C.V sea menor o igual al 10 %, esto garantizará mejores aproximaciones en los valores de predicción.

Inferencia sobre β_1

Considere el modelo probabilístico $y = \beta_0 + \beta_1 X + \epsilon$.
¿Qué pasaría si la información de la variable (predictora) x no contribuye en absoluto a la predicción del valor de y?

Si x contribuye en el modelo probabilistico, entonces, la parte determinista del modelo $E[y] = \beta_0 + \beta_1 X$ no se vería afectada por los cambios en x, por lo cual la predicción simpre sería la misma para cualquier valor de x.



Prueba de Hipótesisi para β_1

La hipótesis nula establece que x no contribuye información y la hipótesis alternativa establece que esas variables están linealmente correlacionadas con una pendiente diferente de cero.

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

Si se cumplen los supuestos sobre ϵ mencionados anteriormente, entonces la distribución muestral de $\hat{\beta}_1$ es normal con media β_1 y desviación estándar

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SCx}}$$

Prueba de hipótesis para $\hat{\beta}_1$

Test statistic: $t = \hat{\beta}_1 / s_{\hat{\beta}_1} = \dfrac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$

	ONE-TAILED TESTS		TWO-TAILED TEST
	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$
	H_a: $\beta_1 < 0$	H_a: $\beta_1 > 0$	H_a: $\beta_1 \neq 0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve
			$2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha >$ p-value, or, if test statistic falls in rejection region

where $P(t > t_\alpha) = \alpha$, $P(t > t_{\alpha/2}) = \alpha/2$, t_c = calculated value of the test statistic, the t-distribution is based on $(n-2)$ df and $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ true})$.

Assumptions: The four assumptions about ε listed in Section 3.4.

Intervalo de confianza para β_1

El intervalo de de confianza para β_1 a un nivel de significancia de $100(1-\alpha)$ está dado por

$$\hat{\beta}_1 \pm (t_{\alpha/2}) S_{\hat{\beta}_1}, \tag{1}$$

donde $S_{\hat{\beta}_1} = \dfrac{S}{SC_x}$ y $t_{\alpha/2}$ está basado en $n-2$ grados de libertad.

Coeficiente de correlación

El coeficiente de correlación de Pearson r mide la relación lineal que existe entre dos variables, el valor que toma este coeficiente está entre -1 y 1. este coeficiente se calcula de la siguiente forma

$$r = \frac{SC_{xy}}{\sqrt{SC_x SC_y}}$$

Observación

Una correlación fuerte no implica causalidad, es decir, si hay una alta correlación no implica que cambios de x causen los cambios en y. La conclusión correcta sería que existe una tendencia de tipo lineal entre la variable respuesta y la predictora.

Coeficiente de determinación

El coeficiente de determinación se puede interpretar como la proporción de la variabilidad muestral total de los valores de y explicada por la relación lineal entre y y x. Este coeficiente se calcula de la sigueinte forma.

$$r^2 = \frac{SC_y - SSE}{SC_y} = 1 - \frac{SSE}{SC_y}$$

De esta forma el $100r^2\,\%$ de la variación muestral en y es explicada por el modelo de línea recta.

Error estándar de $\hat{y}$

La desviación estándar de la distribución muestral del estimador $\hat{y}$ del valor medio de y en un valor particular de x, digamos, x_p, es

$$\sigma_{\hat{y}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

donde σ es la desviación estándar del error aleatorio ϵ y se estima σ por S.

Error estándar de la predicción

La desviación estándar del error de predicción para el predictor $\hat{y}$ de un valor de y para $x = p$ es

$$\sigma_{y-\hat{y}} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Intervalos de confianza

Intervalo de confianza para el valor medio de y para $x = x_p$

$$\hat{y} \pm (t_{\alpha/2}) S \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Intervalo de predicción para un valor individual de y para $x = x_p$

$$\hat{y} \pm (t_{\alpha/2}) S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

donde $t_{\alpha/2}$ está basado en $(n-2)$ grados de libertad.

Bibliografía

Casella, G. and R. L. Berger (2021).
Statistical inference.
Cengage Learning.

Lilja, D. J. (2016).
Linear Regression Using R: An Introduction to Data Modeling.
University of Minnesota Libraries Publishing.

Mendenhall, W., T. Sincich, and N. S. Boudreau (2003).
A second course in statistics: regression analysis, Volume 6.
Prentice Hall New York.

Montgomery, D., E. A. Peck, and G. G. Vining (2006).
Introducción al análisis de regresión lineal.
México: Limusa Wiley.