La recta de regresión

Inferencia sobre la recta de regresión

Grado en Biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

17 de diciembre de 2017

El problema

Disponemos de una muestra de (pares de) valores

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

Podemos

- calculamos una recta cuyos coeficientes minimizan el error cuadrático
- Usar la identidad Anova y r^2 para cuantificar qué parte de la variabilidad de Y se debe al azar y qué parte a la existencia de un modelo con
- Calcular el coeficiente de correlación lineal de Pearson para medir la tendencia de los pares de puntos a estar alineados.

TODO DEPENDE DE (y se refiere a) LA MUESTRA CONCRETA CON LA QUE TRABAJAMOS

El problema

Con cada muestra $(x_1, y_1), \cdots, (x_n, y_n)$ la recta de regresión "muestral"

$$y(x) = b_0 + b_1 x$$

es una "manifestación" de la recta teórica

$$y(x) = \beta_0 + \beta_1 x$$

Para cada muestra los coeficientes de *esa* recta de regresión

$$b_0 \qquad b_1$$

son **estimadores** de los coeficientes de la recta teórica

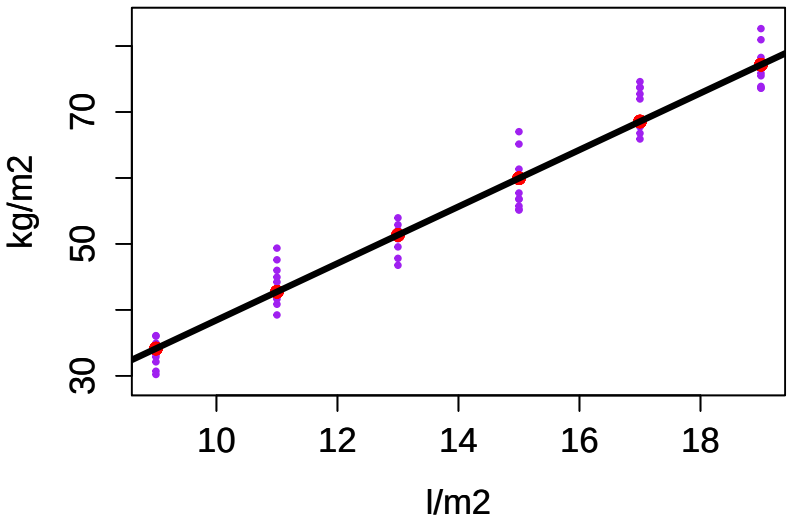
$$\beta_0 \qquad \beta_1$$

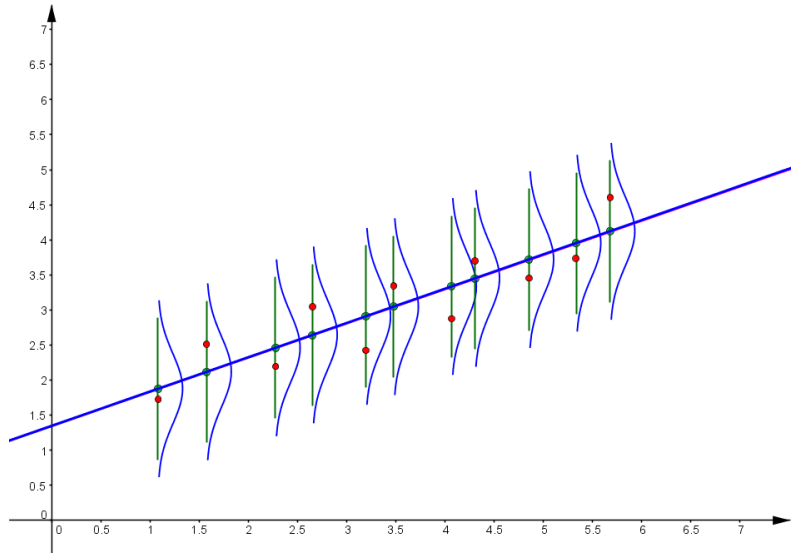
Buscamos una distribución de probabilidad para hacer inferencia sobre esos estimadores

- Intervalo de confianza para el verdadero valor de β_0 y β_1
- Contraste de hipótesis sobre β_0 y β_1

Modelo de regresión lineal simple: hipótesis

Ejemplo: Se estudia la relación entre las variables X intensidad de regadío (l/m^2) e Y productividad de cierto cultivo (kg/m^2)





Para cada valor x_* de X las medidas de Y se distribuyen como una normal

- media $\mu_{Y_{x_*}} = \beta_0 + \beta_1 x_*$
- la misma desviación típica σ para todos los x_*

Modelo de regresión lineal simple: hipótesis

Para cada valor x_* fijo de la variable X resulta que

$$Y_{x_*} \sim N(\beta_0 + \beta_1 x_*, \sigma),$$

En otras palabras

$$y = \beta_0 + \beta_1 \cdot x + \epsilon, \qquad \text{siendo } \epsilon \sim N(0, \sigma).$$

lo que equivale a que para todos los resíduos e_j

$$e_j \sim N(0, \sigma)$$

1. Los residuos son normales.
2. Las varianzas de los residuos son iguales.
3. Además, independencia de las muestras.

En la práctica, se combinan

- Gráficos de diagnóstico: QQ-plot y dispersión de los resíduos
- Contrastes paramétricos/no paramétricos

Estadístico para β_1

Si se cumplen las hipótesis del modelo de regresión lineal simple

Estadístico para β_1, la pendiente de la recta teórica de regresión

El estadístico

$$\Xi = \frac{b_1 - \beta_1}{\sqrt{\frac{ECM}{(n-2)s^2(x)}}} \sim t_{n-2} \tag{1}$$

sigue una distribución t de Student con $n-2$ grados de libertad.

Con esto se puede

1. Calcular intervalo de confianza para β_1
2. Contrastar hipótesis sobre β_1

Ejemplos

Ejemplo: Con los datos del ejemplo del nivel de irrigación y la productividad del cultivo, $\alpha = 0{,}05$ se tiene

```
datos = read.table(file = "datos_ej_res_normales.csv",
sep = ";", header = T)
colnames(datos) = c("Riego", "Productividad")
modelo = lm(datos$Productividad ~datos$Riego)
confint(modelo, level=0.95)
```

	2.5 %	97.5 %
(Intercept)	-2.1931112	4.645111
datos$Riego	0.7287371	1.203263

Contraste sobre β_1

Ejemplo: Con los datos del ejemplo del nivel de irrigación y la productividad del cultivo, $\alpha = 0{,}05$, para contrastar

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0,$$

se tiene

$$\Xi = \frac{b_1 - \beta_1}{\sqrt{\frac{ECM}{(n-2)s^2(x)}}} \approx \frac{0.92 - 0}{\sqrt{\frac{20.74}{(60-2)11.2}}} = 5.303$$

de donde

$$p - valor = 2P(t_{58} > 5.303) =$$

```
2*pt(5.303, df=58, lower.tail = FALSE) =1.853406e-06
```

Por tanto, se rechaza H_0.

Más información en sección 10.4.4 del libro Ejemplo punto influyente