447 TEXTOS DOCENTS

REGRESIÓN LINEAL SIMPLE EN BIOESTADÍSTICA

Santiago Ríos
Marta Cubedo

Departamento de Genética,
Microbiología y Estadística
Universidad de Barcelona

UNIVERSITAT DE BARCELONA

REGRESIÓN LINEAL SIMPLE EN BIOESTADÍSTICA

Santiago Ríos
Marta Cubedo

Departamento de Genética,
Microbiología y Estadística
Universidad de Barcelona

447 TEXTOS DOCENTS

Universitat de Barcelona
Edicions

Índice

PRÓLOGO

Muchos científicos y estudiantes que se dedican al estudio de las Ciencias de la Vida y Farmacología consideran la Estadística como una disciplina fundamental en su trabajo cotidiano. Sin embargo, muchos de ellos manifiestan cierta inseguridad en el tratamiento estadístico de sus resultados, lo que los lleva a realizar consultas sobre Estadística a los profesionales de Análisis de Datos. Todas estas razones nos han conducido a elaborar esta colección de problemas de regresión lineal simple, basados en la experiencia. Los datos utilizados corresponden a situaciones reales, lo que hace que el tema resulte más motivante y comprensible para el lector. Para seguir los temas tratados solo se necesita un conocimiento muy básico de Estadística, aunque diseñamos y analizamos los problemas de una manera formal y rigurosa. En cada uno de los problemas resueltos, los cálculos se realizan con el software libre R.

Capítulo 1

REGRESIÓN Y CORRELACIÓN

1.1. Regresión lineal entre dos variables aleatorias

En muchos casos, dadas dos variables aleatorias, X, Y, resulta interesante relacionar linealmente la Y y la variable X. Para ello, debemos encontrar la combinación lineal de X, de modo que se ajuste de la mejor forma posible a Y, $\widetilde{Y} = \alpha + \beta X$. El criterio para obtener la combinación lineal será el de los mínimos cuadrados:

$$F(\alpha,\beta) = E(Y-\alpha-\beta X)^2 = \text{mín.}$$

Con lo que se obtiene

$$\beta = \frac{cov(X,Y)}{var(X)} \qquad \alpha = E(Y)-\beta E(X).$$

1.2. Coeficiente de correlación de Pearson

El grado de relación lineal entre X, Y se cuantifica por el coeficiente de correlación ρ cuyo valor viene dado por la siguiente expresión:

$$\rho^2 = \frac{cov^2(X,Y)}{var(X)var(Y)}.$$

Se considera que es el porcentaje de variabilidad de la Y, que depende de la X.
Las propiedades de este parámetro son:

(a) $-1 \le \rho \le +1$.

(b) $\rho^2 = 1 \Rightarrow Y = \alpha + \beta X$.

(c) Si $\rho^2 = 0$, se dice que las variables están incorrelacionadas. En particular, si son independientes, $\rho = 0$. El recíproco, en general, no es cierto.

(d) El coeficiente de correlación es invariante por transformaciones lineales de las variables, es decir, $U = \mu + \lambda X$, $V = \mu' + \lambda' Y \Rightarrow \rho' = corr(U,V) = \rho = corr(X,Y)$.

1.3. Varianza residual

La varianza residual $\tilde{\sigma}^2$ es el residuo del modelo, es decir, la varianza de la variable aleatoria $Y-\widetilde{Y}$, diferencia entre variable dependiente y la determinada por la recta de regresión. Su valor se puede calcular a partir de la expresión $\tilde{\sigma}^2 = \sigma_Y^2(1-\rho^2)$, $\sigma_Y^2 = var(Y)$. Es la varianza de la variable aleatoria $(Y/X = x)$. La raíz cuadrada de la varianza residual se conoce como error típico.

Capítulo 2

INFERENCIA SOBRE LA REGRESIÓN

En la práctica, para analizar la relación lineal entre la X y la Y, que supondremos normales y bivariantes, partimos de una muestra de n pares de observaciones de la variable (X, Y) (Tabla 1).

Tabla 1

Observaciones	X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

2.1. Estimación de parámetros

Tanto los parámetros de la regresión, coeficientes α, β, como el coeficiente de correlación ρ y varianza residual $\tilde{\sigma}^2$, deben estimarse a partir de la muestra. Sus estimaciones son a, b, r y $\tilde{s}^2$, respectivamente.

$$b = \frac{s_{xy}}{s_x^2},\ a = \bar{y} - b\,\bar{x},\ r = \frac{s_{xy}}{s_x\, s_y},\ \tilde{s}^2 = s_y^2(1-r^2),$$

donde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,\ s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2,\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

2.2. Contrastes en regresión

En general, interesa conocer si las variables analizadas están incorrelacionadas; para ello, nos planteamos realizar el siguiente contraste de hipótesis:

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0. \tag{1}$$

Si H_0 es cierta, el estadístico

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

sigue una distribución T de Student con $n-2$ grados de libertad, donde n es el tamaño muestral.

Rechazamos H_0 a nivel de significación ϵ si $|t| > k$, siendo k un valor real que cumple $p(T > k) = \frac{\epsilon}{2}$.

El contraste de hipótesis (1) es equivalente al contraste: $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, siendo β la pendiente de la recta de regresión.

2.3. Intervalos de confianza para la predicción y para la media

El intervalo de confianza para el verdadero valor de la predicción de la variable dependiente Y, para un valor determinado de la variable aleatoria $X = x_0$, con un nivel de confianza $1-\epsilon$, viene dado por:

$$y_0 \pm t(\epsilon)\sqrt{\frac{n}{n-2}}\,\tilde{s}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{n s_x^2}}.$$

El intervalo de confianza para el verdadero valor de la media de la variable $Y/X = x_0$, con un nivel de confianza $1-\epsilon$, viene dado por:

$$y_0 \pm t(\epsilon)\sqrt{\frac{n}{n-2}}\,\tilde{s}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{n s_x^2}},$$

donde $y_0 = a + b\,x_0$ y $p(T > t(\epsilon)) = \frac{\epsilon}{2}$.

Los estimadores a, b de α, β tienen por varianza estimada, respectivamente:

$$\widehat{var(b)} = \frac{\tilde{s}^2}{\sum_{i=1}^n (x_i-\bar{x})};\quad \widehat{var(a)} = \frac{\tilde{s}^2\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2};\quad \widehat{cov(a,b)} = \frac{-\tilde{s}^2\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}.$$

Los intervalos de confianza para el verdadero valor de β y α son, respectivamente:

$$b \pm \frac{t(\epsilon)\tilde{s}}{s_x\sqrt{n-2}} \qquad a \pm t(\epsilon)\tilde{s}\frac{\sqrt{\sum_{i=1}^n x_i^2}}{s_x\sqrt{n(n-2)}}.$$

Capítulo 3

MODELO LINEAL

La regresión de una variable Y sobre otra X puede ser desarrollada como modelo lineal, donde Y es una variable aleatoria observable y X es una variable controlada por el experimentador, la designaremos como x y no le exigimos normalidad. La mayoría de los resultados en los puntos anteriores son válidos también para este caso y los resultados que vamos a exponer a continuación son asimismo válidos siempre y cuando x sea una variable controlada.

Capítulo 4

COEFICIENTE DE CORRELACIÓN DE SPEARMAN

El coeficiente de correlación de Spearman es una medida estadística de la relación monótona entre datos apareados.

Su interpretación es similar a la de Pearson. Si r_s es el coeficiente de correlación de Spearman, $0 \leq |r_s| \leq 1$. Cuanto más cerca esté el valor absoluto a 1, más fuerte será la relación monótona.

El cálculo del coeficiente de correlación de Spearman y la significación posterior requiere que se cumplan los siguientes supuestos sobre los datos observados:

1. Los datos deben ser ordinales.
2. A diferencia de la correlación de Pearson, no existe el requisito de normalidad bivariante de (X, Y), por lo tanto es una estadística no paramétrica.

El coeficiente de correlación de rango de Spearman, denotado por r_s, se puede calcular aplicando la siguiente fórmula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

donde d_i es la diferencia en los rangos de los dos valores asociados de las dos variables asociadas.

En el caso del par observado $(x_i(j), y_i(k))$ que ocupan los lugares j y k respectivamente, $d_i = j - k$, $i = 1, \ldots, n$.

Capítulo 5

PROBLEMAS

Problema 1

Portmán es una localidad de la Región de Murcia situada a orillas del mar Mediterráneo donde se explotaron minas de plomo. Los residuos de la explotación minera se vertieron al mar en su bahía, lo que produjo su contaminación. Un laboratorio ha investigado recientemente el pH de sus aguas y su relación con la distancia en kilómetros desde los puntos donde se obtuvo la muestra de agua a la costa. Los resultados aparecen en la siguiente tabla:

km	4,5	15,5	31,2	39,6	54,1	56,1	66,7	81.0	82,1
pH	3,20	3,60	4,09	4,18	5,57	5,70	6,04	7,37	7,89

1. ¿Cuál es la variable dependiente?
2. Representar los datos de la tabla en un sistema de coordenadas cartesianas. Analizar las características del gráfico que se consideren más interesantes.
3. Calcular el coeficiente de correlación muestral. Analizar, a un nivel de significación del 0,1%, si existe relación lineal entre la distancia y el *pH*. Explicar los valores obtenidos.
4. Estimar la recta de regresión que relaciona la distancia con el *pH*.
5. Obtener una estimación del *pH* en los siguientes lugares: a distancias de 0 km, 50 km y 100 km de la costa. Comentar los resultados.
6. Obtener un intervalo de la predicción realizada para 50 km al 95% de confianza.
7. Obtener un intervalo de la media de los valores del *pH* del agua a 50 km al 95% de confianza.
8. ¿Entre qué valores se encuentra el *pH* a 50 km de la costa con probabilidad 0,95?

Solución:

1. La variable dependiente es *pH*.

   Representamos en el eje de abscisas la distancia (D) en km y en el eje de ordenadas el *pH*. A medida que la distancia aumenta, el *pH* también lo hace (figura 1).

2. Representación gráfica de los datos.

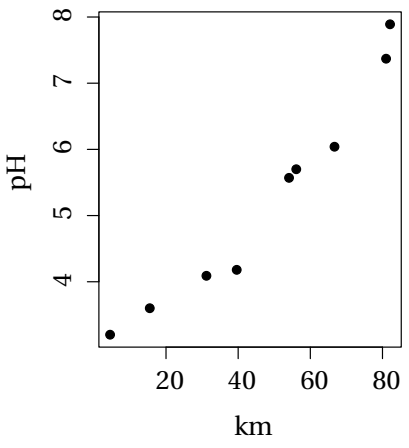

Figura 1. Relación pH-distancia.

3.

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i\,y_i-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}^2}}=0{,}972$$

$$\sum_{i=1}^{n}x_i\,y_i=2632{,}05;\sum_{i=1}^{n}x_i=430{,}8;\sum_{i=1}^{n}y_i=47{,}64;\sum_{i=1}^{n}x_i^2=26626{,}42;\sum_{i=1}^{n}y_i^2=273{,}966;n=9$$

$$t=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}=10{,}975.$$

Al consultar la tabla de la T de Student con 7 g.l., se tiene: $p(|T|>5{,}408)=0{,}001$. Puesto que $|10{,}9745|>5{,}408$, podemos aceptar que a un nivel de significación del 0,1%, hay relación lineal entre la distancia y el *pH*.

El coeficiente de correlación es positivo, lo que confirma que, a medida que la distancia aumenta, el *pH* también lo hace. El valor absoluto de t es muy superior al obtenido en las tablas, por lo que podemos considerar que la relación lineal es fuerte.

4.

$$b=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i\,y_i-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}=0{,}0585604,\ \ a=\bar{y}-b\,\bar{x}=2{,}49024$$

$$pH=2{,}490+0{,}059\text{ km}$$

5. A 0 km: $2{,}490+0{,}059*0=2{,}490$. A 50 km: $2{,}490+0{,}059*50=5{,}418$. A 200 km: $2{,}490+0{,}059*200=14{,}202$.

   El resultado para 200 km no tiene sentido, por lo que no se debe considerar.

6. El intervalo de confianza para el verdadero valor de la estimación de 50 km es:

$$5{,}419=\pm t(0{,}05)\sqrt{\frac{9}{7}}\,\tilde{s}\sqrt{1+\frac{1}{9}+\frac{(50-47{,}86666667)^2}{9*667{,}2733333}}.$$

   Como $t(0{,}05)=2{,}306$ y la estimación de la varianza residual $\tilde{\sigma}^2$ es $\tilde{s}^2=s^2(1-r^2)$. $\tilde{s}=\sqrt{2{,}421288889*(1-0{,}972148)}$, el intervalo de confianza es:

$$I=(5{,}419\pm1{,}0311)=(4{,}388\div6{,}450).$$

7. El intervalo de confianza para el verdadero valor de la media del *pH* del agua a 50 km es:

$$5{,}419\pm t(0{,}05)\sqrt{\frac{9}{7}}\,\tilde{s}\sqrt{\frac{1}{9}+\frac{(50-47{,}86666667)^2}{9*667{,}2733333}}.$$

   Como $t(0{,}05)=2{,}306$ y la estimación de la varianza residual $\tilde{\sigma}^2$, es $\tilde{s}^2=s^2(1-r^2)$.$\tilde{s}=\sqrt{2{,}421288889*(1-0{,}972148)}$, el intervalo de confianza es:

$$I=(5{,}419\pm0{,}328)=(5{,}091\div5{,}747).$$

8. La variable aleatoria $H = (pH/D = 50)$ se distribuye según una Normal de media, la estimación obtenida en el apartado 4 y varianza, la residual; $H \sim N(\mu = 5{,}418; \sigma = 0{,}260)$. Por consiguiente:

$$p(5{,}418 - 1{,}96 * 0{,}260 \leq H \leq 5{,}418 + 1{,}96 * 0{,}260) = 0{,}95$$

El intervalo buscado es $I = (4{,}9077 \div 5{,}9283)$.

El significado de los intervalos obtenidos en los apartados 6 y 8 es distinto. En el apartado 8 se indica aproximadamente, entre qué valores se encuentra la variable *pH* condicionada a $D = 50$, con probabilidad 0,95 y en el apartado 6 se obtiene un intervalo de confianza para la verdadera predicción de *pH* al 0,95 de confianza. La longitud del intervalo obtenido en 6 tiende al obtenido en 8, cuando el tamaño muestral tiende a infinito.

Resolución en «R»

Entrada de datos

```
Dist <−c(4.5, 15.5, 31.2, 39.6, 54.1, 56.1, 66.7, 81.0, 82.10)
pH <−c(3.20, 3.60, 4.09, 4.18, 5.57, 5.70, 6.04, 7.37, 7.89)
```

Significación del modelo lineal

```
reg1 <−lm(pH ~ Dist)
summary(reg1)
```

Representación gráfica

```
plot(Dist,pH,xlab="Km.",ylab="pH",pch=19 )
```

Cálculo de la correlación Pearson y nivel de significación

```
modelo <−lm(pH ~ Dist)
cor(Dist, pH)
cor.test(Dist, pH)
```

Cálculo de la región crítica para el coeficiente de correlación a nivel de significación del 0,001

```
qt(0.9995, 7, lower.tail=TRUE)
qt(0.0005, 7, lower.tail=TRUE)
```

Intervalos de confianza

```
confint(reg1,level=0.95)
newdata <−data.frame(Dist=c(0,50,100,200))
```

Intervalos de confianza para la predicción

```
predict(reg1, newdata, interval='prediction',level = 0.95)
```

Intervalos de confianza para la media

```
predict(reg1, newdata, interval='confidence',level = 0.95)
```

Contraste para la pendiente (tabla ANOVA)

```
reg1.aov <−anova(reg1)
reg1.aov
```

Predicción del pH a 50 km

```
new <−data.frame(Dist=50)
mu <−predict(reg1,new)
sd <−sqrt(reg1.aov[2,3])
```

Intervalo de confianza de la predicción del pH a 50 km a nivel del 0,95

```
LI <−qnorm(0.25,mu,sd)
LS <−qnorm(0.975,mu,sd)
c(LI,LS)
```

Problema 2

Se supone que la relación entre dos variables x e Y es lineal, donde Y toma un valor constante e igual a cero cuando $x = 0$. Es decir $E(Y/x) = \beta x$. Determinar β por el método de mínimos cuadrados. Para analizar el efecto que sobre la temperatura corporal tiene una determinada manta térmica usada en los rescates de montaña, se midió el incremento de la temperatura de la zona central del cuerpo y la de la piel en 10 personas a -10 °C con una velocidad del viento de 2,7 ms^{-1}, durante 25 minutos cada 5 minutos. Los promedios de los resultados aparecen en la siguiente tabla:

Tiempo	5 m	10 m	15 m	20 m	25 m	30 m
Zona central	0,07	−0,07	−0,13	−0,17	−0,27	−0,38
Piel	−2,81	−3,53	−4,08	−4,53	−4,92	−5,17

1. Representar los datos de la temperatura en un sistema de coordenadas cartesianas.
2. Proponer un modelo para los datos, explicando las razones de la elección.
3. Estimar los parámetros del modelo escogido.
4. Calcular los incrementos de temperatura predichos para los seis tiempos utilizados en el estudio.

Solución:

Sean las n observaciones:

$x:$	$x_1, x_2, \ldots, x_n$
$y:$	$y_1, y_2, \ldots, y_n$

El valor de β será el que minimice la expresión:

$$\Phi(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2,$$

por lo tanto, se ha de cumplir que:

$$\frac{\partial \Phi}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \beta x_i) x_i = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

1. Representación gráfica de las temperaturas. En el eje de abscisas representamos el incremento en la piel, y en el eje de ordenadas, el incremento en el centro (figura 2).

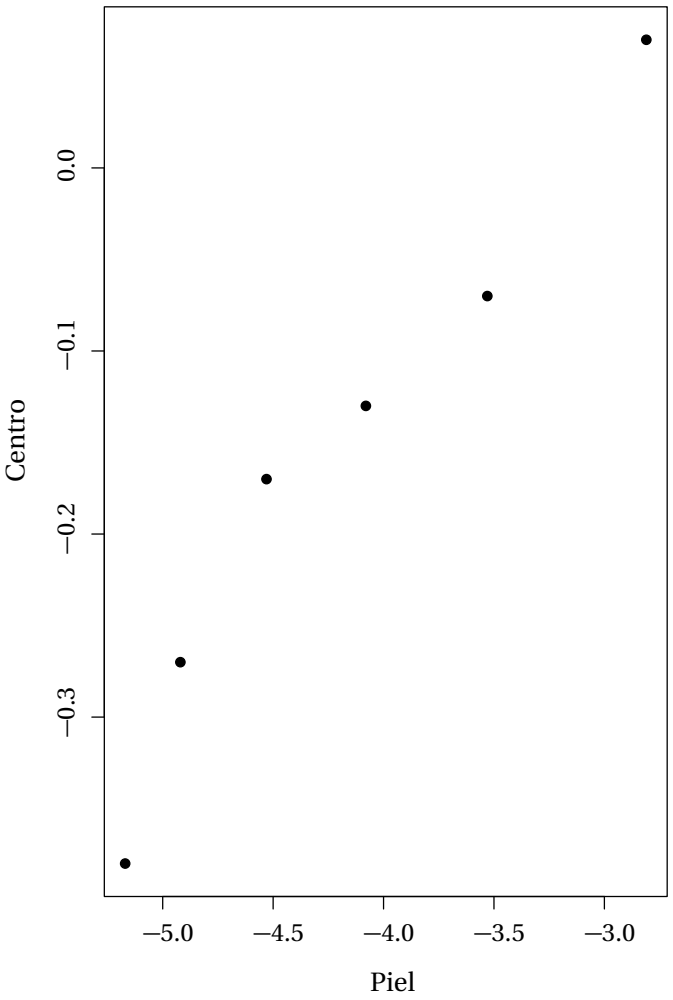

Figura 2. Relación incremento piel-incremento centro.

2. A medida que aumenta la temperatura en la piel es mayor, en el centro también lo es. Cuando el incremento en la piel es cero, en el centro también lo es (figura 2).

   Se puede proponer el modelo:

$$\Delta T(\text{centro}) = \beta \Delta T(\text{piel}).$$

3.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = 0{,}04282.$$

4. A incremento en piel $-2.81°$: $0.04282 * (-2.81) = -0.1203154$. A $-3.53°$: -0.1511435. A $-4.08°$: -0.1746928. A $-4.53°$: -0.1939604. A $-4.92°$: -0.2106590. A $-5.17°$: -0.2213632.

Zona central observada	0,07	−0,07	−0,13	−0,17	−0,27	−0,38
Zona central predicha	−0,1203	−0,1511	−0,1747	−0,1940	−0,2107	−0,2213
Piel	−2,81	−3,53	−4,08	−4,53	−4,92	−5,17