

Capítulo 8. Contraste de hipótesis y selección de modelos

Índice

1. Modelos anidados y no anidados	1
2. Razón de verosimilitudes	2
3. Selección de modelos no anidados	9
4. Ejercicios	11

1. Modelos anidados y no anidados

En este capítulo veremos varios métodos para comparar el ajuste de dos o más modelos a unos mismos datos. Dos modelos estadísticos se dice que están *anidados* cuando uno es un caso particular del otro. En concreto, un modelo anidado tiene la misma forma matemática que otro general, pero con algunos parámetros del modelo general fijados a valores constantes. Por ejemplo, si en la distribución gamma fijamos $\alpha = 1$ se obtiene la función de densidad

$$f(x) = \beta \exp(-\beta x)$$

que resulta ser la misma que una distribución exponencial. Entonces la distribución exponencial es un caso particular de la gamma.

Supongamos un segundo ejemplo. Tomamos dos muestras de dos poblaciones y asumimos que cada una de ellas sigue la distribución normal. El primer modelo estadístico asume que ambas poblaciones tienen igual varianza pero la media puede ser diferente, es decir tendríamos las distribuciones $\text{normal}(\mu_1, \sigma)$ y $\text{normal}(\mu_2, \sigma)$ para cada una de las poblaciones. El segundo modelo asume que $\mu_1 = \mu_2$. El segundo modelo está anidado en el primero porque matemáticamente son iguales pero se ha impuesto una restricción en los parámetros. De la comparación entre estos dos modelos puede concluirse si la hipótesis de igualdad de medias puede mantenerse o no. Estos dos modelos son precisamente los que se están comparando cuando se realiza un contraste de igualdad de medias con una prueba como la T o el ANOVA.

Modelos no anidados son aquellos que tienen distinta forma matemática. Supongamos por ejemplo que ajustamos dos distribuciones a unos mismos datos, la normal y la exponencial. Queremos saber cual sería la distribución más adecuada, por lo que el objetivo del análisis es comparar las distribuciones y seleccionar una de ellas. En este análisis no hay una hipótesis involucrada, simplemente se trata de seleccionar la distribución más adecuada para unos datos. Los modelos normal y exponencial no están anidados porque su forma matemática es diferente y no es posible obtener uno de ellos imponiendo restricciones en los parámetros del otro.

Los métodos estadísticos empleados en la comparación de modelos depende de si están anidados o no. Cuando se comparan modelos anidados realizamos un contraste de bondad de ajuste mediante el estadístico chi-cuadrado de la razón de verosimilitudes. Al comparar modelos no anidados se utilizan medidas para calcular el grado de discrepancia entre cada modelo y los datos.

2. Razón de verosimilitudes

La razón de verosimilitudes es un método general para diseñar contrastes de hipótesis. Al igual que la estimación máximo-verosímil es una guía de actuación para estimar parámetros que puede aplicarse a cualquier distribución, con la razón de verosimilitudes ocurre igual. Es un método que puede adaptarse a innumerables casos concretos, tales como contrastar hipótesis acerca de un sólo parámetro, hipótesis sobre conjuntos de parámetros, contrastes de bondad de ajuste en los que se compran un modelo estadístico con un modelo base o modelo *saturado*, contrastes en los que se comparan modelos que siguen distintas funciones de probabilidad, etc. Desde la prueba T del contraste de medias, pasando por el estadístico F del ANOVA hasta los contrastes de bondad de ajuste en todo tipo de modelos (tablas de contingencia, log-lineal, análisis factorial, etc.), todos ellos son razones de verosimilitudes.

El test de la razón de verosimilitudes tiene buenas propiedades estadísticas porque, al igual que la estimación por máxima-verosimilitud, aprovecha toda la información contenida en la muestra. Por ello es posible demostrar, aunque no lo haremos en este curso, que el contraste de la razón de verosimilitudes es el contraste que alcanza mayor potencia estadística. De modo informal, podemos entender que de todos los procedimientos que podríamos crear para contrastar una hipótesis estadística, el test de la razón de verosimilitudes es el que tiene mayor capacidad de detectar cuando una determinada hipótesis es falsa.

El contraste de la razón de verosimilitudes generalizado se utiliza para comparar dos modelos anidados, M_1 y M_2 , donde M_1 es un caso particular de M_2 . Además asume que el modelo mas general, M_2 , es correcto. El resultado carece de sentido si M_2 no ajusta a los datos. Además M_1 incluye un vector de parámetros $\theta = (\theta_1, \dots, \theta_j)$. Los parámetros del modelo M_2 son $\omega = (\omega_1, \dots, \omega_k)$. Cómo M_1 es un caso particular de M_2 , incluye un número menor de parámetros ($j < k$).

La lógica del contraste es asumir que el modelo general es correcto en la población, y valoramos si al imponer restricciones en los parámetros el modelo resultante deja de ajustar. En concreto, la hipótesis nula, H_0 , afirma que el modelo restringido, M_1 , ajusta a los datos. La hipótesis alternativa, H_1 , afirma que M_1 no ajusta.

Los modelos se comparan utilizando su función de verosimilitud. En concreto, si el valor de la función de verosimilitud de M_1 y M_2 es similar podemos mantener la hipótesis de que M_1 ajusta a los datos, dado que partíamos de asumir que M_2 ajusta. Si la función de verosimilitud de M_1 es significativamente menor que la de M_2 rechazamos la hipótesis de que M_1 ajusta. La forma de comparar las funciones de verosimilitud es calcular su cociente

$$\frac{f(\mathbf{x}; M_1)}{f(\mathbf{x}; M_2)}.$$

donde $f(\mathbf{x}; M_1)$ es el valor de la función de verosimilitud para el primer modelo cuando esta se evalúa en su estimador máximo-verosímil, y $f(\mathbf{x}; M_2)$ es el máximo de la función de verosimilitud para el segundo modelo. El cociente toma valores en el intervalo $[0, 1]$. El valor 1 indica que ambos modelos son igualmente probables; cuanto menor sea el cociente menor será la verosimilitud de M_1 en comparación con la de M_2 .

Ejemplo 1. Supongamos que se desea contrastar la hipótesis $H_0 : \mu = 100$ frente a $H_1 : \mu \neq 100$ en una variable distribuida normalmente con desviación típica desconocida. Los dos modelos a comparar son, M_1 es el modelo normal($\mu = 100, \sigma = 15$), en el que μ se fija al valor a contrastar y σ se estima a partir de los datos. El modelo general, M_2 , es la normal(μ, σ), en el que ambos parámetros son libres y se estiman a partir de la muestra. El siguiente código R calcula la razón de verosimilitudes a partir de la muestra $\mathbf{x} = (116, 85, 90, 124)$.

```
x <- c(116, 85, 90, 125)
lk <- 0
likelihood <- function(mu, sigma){
  lk <- 1
  for(i in 1:length(x)){
    lk <- lk * dnorm(x[i], mu, sigma)
```

```

}
  return(lk)
}

media <- mean(x)
desviacion <- sd(x)

M1 <- likelihood(100, desviacion)
M2 <- likelihood(media, desviacion)
razon <- M1/M2

cat(
  sprintf('media = %5.2f, desviacion = %5.2f\nlk(M1) = %14.12f, lk(M2) = %14.12f, Razon = %6.4f',
    media, desviacion, M1, M2, razon)
)

## media = 104.00, desviacion = 19.51
## lk(M1) = 0.000000035859, lk(M2) = 0.000000039004, Razon = 0.9194

```

Matemáticamente la razón de verosimilitudes nunca puede ser mayor de 1 porque M_1 está incluido en M_2 , por lo que M_1 no puede ajustar a los datos mejor que M_2 . Sin embargo, es necesario tener un criterio que nos permita juzgar el valor de la razón de verosimilitudes para tomar una decisión estadística a partir del mismo.

El criterio estadístico para tomar una decisión sobre H_0 a partir de la razón de verosimilitudes consiste en calcular el estadístico

$$G^2 = -2 \log \frac{f(\mathbf{x}; M_1)}{f(\mathbf{x}; M_2)} .$$

La distribución asintótica de G^2 es chi-cuadrado con $k - j$ grados de libertad (es decir los grados de libertad son la diferencia en el número de parámetros estimados con ambos modelos). Por tanto, la zona crítica se sitúa en la parte derecha de la distribución chi-cuadrado y se define por $G^2 \geq_{1-\alpha} \chi^2_{k-j}$. Al estar la zona crítica en la cola derecha, el nivel crítico o p -valor se calcula como la probabilidad hacia la derecha del valor observado de G^2 .

Desarrollando el logaritmo, el estadístico G^2 toma la siguiente forma que es más adecuada para hacer los cálculos cuando se programa la estimación máximo-verosímil en R.

$$G^2 = -2 \log f(\mathbf{x}; M_1) + 2 \log f(\mathbf{x}; M_2) .$$

El término $-2 \log l(\hat{\omega})$ se denomina en estadística la *desvianza*, donde $l(\hat{\omega})$ es el valor máximo de la función log-verosimilitud. La desvianza mide la distancia del modelo a los datos, a menor desvianza menor distancia. Si los datos son muy probables de acuerdo con un modelo, $l(\hat{\omega})$ será alto y la desvianza baja. El estadístico G^2 es una diferencia de desvianzas, y otros estadísticos de selección de modelos como el AIC también se basan en la desvianza.

Ejemplo 2. Continuando con el ejemplo 1, queremos tomar una decisión sobre H_0 utilizando un nivel de significación $\alpha = 0,05$. Como la muestra es pequeña no tenemos mucha confianza en que la distribución de G^2 se aproxime a chi-cuadrado, pero aún así haremos el cálculo. El siguiente código R, que es continuación del visto en el ejemplo 1, calcula G^2 y su nivel crítico.

```

alpha <- 0.05
G2 = -2*log(razon)
gl <- 1
p_valor <- pchisq(G2, gl, lower.tail=F)

```

```
decision <- ifelse(p_valor <= alpha, "Rechazar H0", "Mantener H0")
cat(sprintf('G2 = %5.2f, g1 = %1.0f p = %5.3f. Decision: %s', G2, g1, p_valor, decision))
```

```
## G2 = 0.17, g1 = 1 p = 0.682. Decision: Mantener H0
```

Ejemplo 3. Tenemos dos muestras de datos dicotómicos procedentes de distintas poblaciones. Los datos son

$$\mathbf{x}_1 = (1, 1, 1, 0, 0, 1, 1, 1, 0, 0)$$

$$\mathbf{x}_2 = (0, 0, 0, 1, 1, 1, 0, 0)$$

Asumimos que la distribución es binomial y queremos contrastar la hipótesis de que el parámetro π es igual en ambas poblaciones, es decir

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 \neq \pi_2$$

Podemos utilizar el contraste de la razón de verosimilitudes dado que el modelo en H_0 es un caso particular del modelo en H_1 . En concreto, bajo H_0 la función de verosimilitud es el producto de dos distribuciones binomiales con igual π :

$$f(\mathbf{x}_1, \mathbf{x}_2; H_0) = f(\mathbf{x}_1, \pi)f(\mathbf{x}_2, \pi)$$

donde $f(\mathbf{x}_1, \pi)$ y $f(\mathbf{x}_2, \pi)$ son distribuciones binomiales. Por tanto la función de log-verosimilitud de M_1 es la siguiente

$$l_1(\pi) = \log f(\mathbf{x}_1, \pi) + \log f(\mathbf{x}_2, \pi)$$

Bajo H_1 , la función de probabilidad de los datos observados es el producto de dos distribuciones binomiales con diferente π , por lo que la función de log-verosimilitud de M_2 es

$$l_2(\pi_1, \pi_2) = \log f(\mathbf{x}_1, \pi_1) + \log f(\mathbf{x}_2, \pi_2)$$

El siguiente código R maximiza las funciones de log-verosimilitud de los dos modelos para obtener los estimadores. Además calcula el estadístico G^2 y su nivel crítico.

```
x1 = c(1, 1, 1, 0, 0, 1, 1, 1, 0, 0)
x2 = c(0, 0, 0, 1, 1, 1, 0, 0)

s1 = sum(x1)
s2 = sum(x2)

n1 = length(x1)
n2 = length(x2)

M1 <- function(p) { log(dbinom(s1, n1, p)) + log(dbinom(s2, n2, p)) }
M2 <- function(q) { log(dbinom(s1, n1, q[1])) + log(dbinom(s2, n2, q[2])) }

fit1 <- optim(c(0.5), f=M1, method="Brent", lower=0, upper=1, control=list(fnscale=-1))
fit2 <- optim(c(0.5, 0.5), f=M2, control=list(fnscale=-1))
```

```
G2 = -2*fit1$value + 2*fit2$value
gl <- 1
p_valor <- pchisq(G2, gl, lower.tail=F)

cat(sprintf('pi = %5.3f, pi2(1) = %5.3f, pi2(2) = %5.3f, G2 = %6.3f, p-valor = %5.3f',
            fit1$par, fit2$par[1], fit2$par[2], G2, p_valor))

## pi = 0.500, pi2(1) = 0.600, pi2(2) = 0.375, G2 = 0.908, p-valor = 0.341
```

Por tanto, podemos mantener la H_0 y no podemos descartar que ambas poblaciones tengan el mismo valor de π .

Ejemplo 4. Supongamos que un investigador en psicología cognitiva toma datos de tiempo de reacción en segundos al realizar una determinada tarea y encuentra el resultado

$$\mathbf{x} = (5, 3, 7, 9, 1)$$

Vamos a realizar la estimación de la distribución gamma y a contrastar la hipótesis $H_0 : \mu = 4$.

El siguiente código R realiza la estimación máximo-verosímil incluyendo los errores típicos

```
x = c(5, 3, 7, 9, 1)

logLikelihood <- function(p){
  lk <- 0
  for(i in 1:length(x)){
    lk <- lk + log(dgamma(x[i], p[1], p[2]))
  }
  return(lk)
}

fit_M2 <- optim(c(1,1), f=logLikelihood, method="L-BFGS-B", control=list(fnscale=-1),
               lower=0, hessian=TRUE)

H <- fit_M2$hessian
Informacion <- - H
VarCovar <- solve(Informacion)

cat(sprintf("Metodo de maxima-verosimilitud, a = %5.2f(%4.2f), b = %5.2f(%4.2f)",
            fit_M2$par[1], sqrt(VarCovar[1,1]), fit_M2$par[2], sqrt(VarCovar[2,2])))

## Metodo de maxima-verosimilitud, a = 2.24(1.33), b = 0.45(0.30)
```

De acuerdo con $H_0 : \mu = 4$, y teniendo en cuenta que en una distribución gamma $\mu = \alpha/\beta$, la hipótesis nula equivale a aplicar la siguiente restricción en la distribución gamma: $\alpha = 4\beta$. Con el siguiente código R estimamos el modelo que implementa esta restricción

```
logLikelihood_R <- function(beta){
  alpha <- 4*beta
  lk <- 0
  for(i in 1:length(x)){
    lk <- lk + log(dgamma(x[i], alpha, beta))
  }
  return(lk)
}
```

```
fit_M1 <- optim(c(1), f=logLikelihood_R, method="Brent", lower=0, upper=1,
               control=list(fnscale=-1), hessian=TRUE)

Informacion <- - fit_M1$hessian
Se <- sqrt(1/Informacion)

cat(sprintf("Metodo de maxima-verosimilitud, modelo restringido, a = %5.2f(%4.2f), b = %5.2f(%4.2f)",
            4*fit_M1$par, 4*Se, fit_M1$par, Se))
```

```
## Metodo de maxima-verosimilitud, modelo restringido, a = 2.03(1.19), b = 0.51(0.30)
```

Al imponer una restricción en el modelo se reduce el número de parámetros estimados, lo que redundará en unos menores errores típicos. Para comparar el modelo general con el restringido y tomar una decisión sobre H_0 utilizamos la razón de verosimilitudes.

```
G2 = -2*fit_M1$value + 2*fit_M2$value
gl <- 1
p_valor <- pchisq(G2, gl, lower.tail=F)
decision <- ifelse(p_valor <= 0.05, "Rechazar H0", "Mantener H0")
cat(sprintf('G2 = %5.2f, gl = %1.0f p = %5.3f. Decision: %s', G2, gl, p_valor, decision))
```

```
## G2 = 0.57, gl = 1 p = 0.449. Decision: Mantener H0
```

Ejemplo 5. En este ejemplo veremos un caso en el que la función de densidad de la muestra no forma parte de las incluidas en el paquete base de R sino que la tenemos que programar. Sea $X > 0$ una variable aleatoria con función de densidad

$$f(x) = \frac{x^4}{24\beta^5} \exp\left(-\frac{x}{\beta}\right).$$

donde el espacio paramétrico es $\beta > 0$. Vamos a realizar un contraste con $\alpha = 0,05$ y las hipótesis

$$H_0 : \beta = 1$$

$$H_1 : \beta \neq 1$$

a partir de la muestra $\mathbf{x} = (5, 2, 4, 3, 1)$.

El primer paso es definir la función de verosimilitud, que tiene la siguiente forma bajo el supuesto de muestreo aleatorio simple

$$\begin{aligned} f(\mathbf{x}) &= f(x_1)f(x_2) \times \cdots \times f(x_n) \\ &= \frac{\prod_{i=1}^n x_i^4}{24^n \beta^{5n}} \exp\left(-\frac{n\bar{X}}{\beta}\right) \end{aligned}$$

Por tanto, la función de log-verosimilitud es

$$l(\beta) = 4 \sum_{i=1}^n \log x_i - n \log 24 - 5n \log \beta - \frac{n\bar{X}}{\beta}.$$

Para realizar la razón de verosimilitudes tenemos que comparar dos modelos. El modelo general, M_2 , no impone ninguna restricción en β y su función de log-verosimilitud es la que se acaba de indicar. El modelo restringido, M_1 , impone $\beta = 1$ y la log-verosimilitud es

$$l_0 = 4 \sum_{i=1}^n \log x_i - n \log 24 - n \bar{X}.$$

donde l_0 indica que este es el modelo derivado de H_0 y en el desarrollo de esta fórmula se ha tenido en cuenta que $\log 1 = 0$.

Necesitamos entonces calcular ambas funciones, $l(\beta)$ y l_0 , y a partir de ellas obtener G^2 y el p -valor. El código R es el siguiente, en la programación de la log-verosimilitud se han eliminado todas las constantes que no son relevantes para el cálculo de los estimadores ni de G^2 :

```
x <- c(5, 2, 4, 3, 1)
n <- length(x)
media <- mean(x)

l <- function(beta) -5*n*log(beta) - n*media/beta

fit_M2 <- optim(c(1), f=l, method="Brent", lower=0, upper=20,
               control=list(fnscale=-1), hessian=TRUE)

fit_M1 <- l(1)

G2 = -2*fit_M1 + 2*fit_M2$value
gl <- 1
p_valor <- pchisq(G2, gl, lower.tail=F)
decision <- ifelse(p_valor <= 0.05, "Rechazar H0", "Mantener H0")
cat(sprintf(
  'beta estimada = %5.2f, Se = %5.3f\nG2 = %5.2f, gl = %1.0f p = %5.3f, Decision: %s',
  fit_M2$par, sqrt(-1/fit_M2$hessian), G2, gl, p_valor, decision))

## beta estimada = 0.60, Se = 0.120
## G2 = 5.54, gl = 1 p = 0.019, Decision: Rechazar H0
```

Entonces rechazamos H_0 y concluimos que el valor $\beta = 1$ no es compatible con los datos muestrales.

Ejemplo 6. La regresión beta se utiliza con una variable que represente una proporción o una calificación acotada entre 0 y 1. Por ejemplo, supongamos que en un contexto educativo se ha recogido datos acerca del aprovechamiento de las clases de informática. La variable P indica la proporción del tiempo de clase que los niños pierden utilizando las redes sociales; la variable X indica el número de alumnos presentes en el aula. Los datos disponibles son

$$\begin{aligned} \mathbf{p} &= (0, 5, 0, 2, 0, 1, 0, 6, 0, 3, 0, 7, 0, 6, 0, 6, 0, 9, 0, 4) \\ \mathbf{x} &= (10, 4, 2, 15, 8, 20, 16, 18, 19, 12) \end{aligned}$$

Aplicaremos una regresión beta de P sobre X con las siguientes características. Este modelo asume que P sigue una distribución beta con media y varianza

$$\begin{aligned} E(P) &= \mu \\ \text{Var}(P) &= \frac{\mu(1-\mu)}{1+\phi} \end{aligned}$$

Además μ y ϕ los calculamos del siguiente modo

$$\mu = \frac{\exp(a + bx)}{1 + \exp(a + bx)}$$

$$\phi = \exp(-v)$$

donde v es un parámetro de variabilidad, a mayor v mayor varianza de P . El intercepto y la pendiente de la regresión son a y b . Estos tres son los parámetros a estimar en la regresión beta.

La función de densidad de P es $\text{beta}(\alpha, \beta)$, donde

$$\alpha = \mu\phi$$

$$\beta = (1 - \mu)\phi$$

El objetivo del análisis estadístico es contrastar la hipótesis de que la pendiente de la regresión es cero en la población. Para ello estimaremos dos modelos, un modelo restringido en el que se fija $b = 0$ y un modelo general en el que b lo estimamos a partir de los datos. Compararemos el ajuste de ambos modelos mediante la razón de verosimilitudes para tomar una decisión acerca de la hipótesis nula.

El siguiente código R realiza los cálculos. Se ha programado por separado la función de verosimilitud de los dos modelos. Cada función de verosimilitud recibe como argumento un vector llamado *parametros*, cuyos elementos son (V, a, b) para el modelo 1 y (V, a) para el modelo 2 (dado que en este modelo $b = 0$). La función de verosimilitud se ha calculado sumando la contribución de cada una de las observaciones de la muestra, es decir $l(\text{parametros}) = \sum_{i=1}^n \log f_i(p_i)$, donde $f_i(p_i)$ es la función de densidad de la observación p_i calculada mediante una distribución beta.

El código R también muestra el valor estimado de los parámetros para el modelo 2, obtenido a partir del Hessiano de la función de log-verosimilitud. Finalmente, el código muestra el diagrama de dispersión con el pronóstico de la regresión (μ) superpuesto.

```
p <- c(0.5, 0.2, 0.1, 0.6, 0.3, 0.7, 0.6, 0.6, 0.9, 0.4)
x <- c(10, 4, 2, 15, 8, 20, 16, 18, 19, 12)

logLikelihood_1 <- function(parametros){
  lk <- 0
  for(i in 1:length(x)){
    phi <- exp(-parametros[1])
    mu <- exp(parametros[2]) / (1+exp(parametros[2]))
    alpha <- mu * phi
    beta <- (1-mu) * phi
    lk <- lk + log(dbeta(p[i], alpha, beta))
  }
  return(lk)
}

logLikelihood_2 <- function(parametros){
  lk <- 0
  for(i in 1:length(x)){
    phi <- exp(-parametros[1])
    mu <- exp(parametros[2] + parametros[3]*x[i]) / (1+exp(parametros[2] + parametros[3]*x[i]))
    alpha <- mu * phi
    beta <- (1-mu) * phi
    lk <- lk + log(dbeta(p[i], alpha, beta))
  }
  return(lk)
}
```



```

fit_M1 <- optim(c(0,0), f=logLikelihood_1, control=list(fnscale=-1), hessian=TRUE)
fit_M2 <- optim(c(0,0,0), f=logLikelihood_2, control=list(fnscale=-1), hessian=TRUE)

G2 = -2*fit_M1$value + 2*fit_M2$value
gl <- 1
p_valor <- pchisq(G2, gl, lower.tail=F)
decision <- ifelse(p_valor <= 0.05, "Rechazar H0", "Mantener H0")
Se <- sqrt(diag(solve(-fit_M2$hessian)))

cat(sprintf("G2 = %5.2f, gl = %1.0f p = %5.3f, Decision: %s
Parametros estimados. v = %5.2f(%4.2f), a = %5.2f(%4.2f) b = %5.2f(%4.2f)",
G2, gl, p_valor, decision, fit_M2$par[1], Se[1], fit_M2$par[2], Se[2], fit_M2$par[3], Se[3]))

## G2 = 19.56, gl = 1 p = 0.000, Decision: Rechazar H0
## Parametros estimados. v = -3.31(0.44), a = -2.13(0.33) b = 0.17(0.02)

xaxis <- seq(2, 20, by=0.01)
mu <- exp(fit_M2$par[2] + fit_M2$par[3]*xaxis) / (1+exp(fit_M2$par[2] + fit_M2$par[3]*xaxis))
plot(x,p)
lines(xaxis,mu, lwd=2, col="navyblue")

```

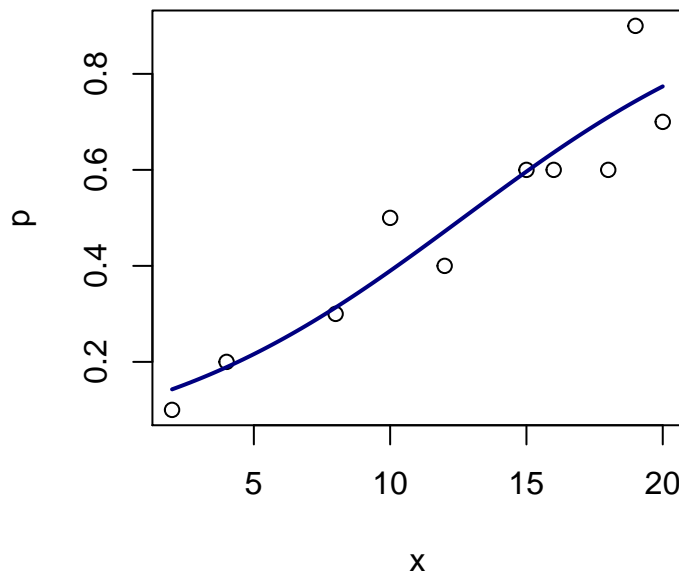


Figura 1: Regresión beta

3. Selección de modelos no anidados

El estadístico de la razón de verosimilitudes solo puede emplearse cuando tenemos modelos anidados. Una situación muy habitual es la comparación de modelos que no están anidados, para lo cual tenemos que recurrir

a otros estadísticos cómo el criterio de información de Akaike (AIC).

El AIC no es un estadístico de contraste. Al utilizarlo no hay ninguna hipótesis que mantener o rechazar. Es un estadístico que proporciona información descriptiva útil para comparar la distancia entre varios modelos y los datos. Con esta información podemos seleccionar el modelo que más se aproxime a los datos y descartar los demás, pero esto no significa nada en términos de ajuste o no ajuste. Podría darse el caso de que ningún modelo ajuste, que lo hagan todos o cualquier situación intermedia entre estas. El AIC simplemente mide la distancia a los datos.

El estadístico AIC es sencillo de calcular pero se basa en un desarrollo bastante sofisticado utilizando la teoría de la información. AIC pondera dos aspectos, la distancia los datos y la parsimonia del modelo. La distancia a los datos se evalúa mediante la función log-verosimilitud, cuanto mayor sea la log-verosimilitud de un modelo más probables son los datos de acuerdo con ese modelo. La parsimonia indica si el modelo es excesivamente complejo; la log-verosimilitud de cualquier modelo siempre puede mejorar añadiendo nuevos parámetros, sin embargo estos parámetros también pueden producir mejoras insignificantes en la log-verosimilitud. La parsimonia evalúa que el modelo no tiene parámetros innecesarios. Con esto se pretende evitar el *sobre-ajuste*, que es el fenómeno que se produce cuando un modelo tiene demasiados parámetros y algunos de ellos son irrelevantes.

El cálculo del AIC es

$$AIC = -2l(\hat{\theta}) + 2k,$$

donde $l(\hat{\theta})$ es el valor máximo de la log-verosimilitud y k es el número de parámetros del modelo. Cómo el AIC es una medida de distancia a los datos, valores altos de AIC indican mayor distancia. En la comparación de modelos, el escogido será el que obtenga menor AIC.

Ejemplo 6. Disponemos de una m.a.s. aleatoria de diez valores

$$\mathbf{x} = (7, 3, 3, 1, 20, 4, 12, 2, 0, 2).$$

Queremos comparar dos modelos para estos datos, la distribución normal y la exponencial. Como no son modelos anidados, utilizaremos el AIC para escoger el más adecuado. Con respecto a la distribución normal, sus parámetros son μ y σ , y los estimadores máximo-verosímiles son la media y la desviación típica de la muestra:

$$\begin{aligned}\hat{\mu} &= 5,4 \\ \hat{\sigma} &= 5,87\end{aligned}$$

El estimador máximo-verosímil de la distribución exponencial es

$$\hat{\omega} = \frac{1}{\bar{X}} \approx 0,19.$$

El siguiente código R implementa los cálculos. El cálculo de la varianza de X se ha corregido porque el estimador máximo-verosímil de σ^2 es la varianza sesgada mientras que la función `var` de R devuelve la varianza sesgada.

```
x = c(7, 3, 3, 1, 20, 4, 12, 2, 0, 2)
n <- length(x)

media <- mean(x)
desviacion <- sqrt((n-1)*var(x)/n)
omega <- 1/media
```

```

l_normal <- sum(log(dnorm(x, media, desviacion)))
l_exponencial <- sum(log(dexp(x, omega)))

AIC_normal <- -2*l_normal + 4
AIC_exponencial <- -2*l_exponencial + 2

cat(sprintf("AIC normal = %6.2f. AIC exponencial = %6.2f\nmedia %5.2f Sd = %5.2f omega = %5.2f",
            AIC_normal, AIC_exponencial, media, desviacion, omega))

## AIC normal = 67.77. AIC exponencial = 55.73
## media 5.40 Sd = 5.87 omega = 0.19

```

Por tanto la distribución exponencial tiene una menor discrepancia con los datos que la exponencial.

4. Ejercicios

Ejercicio 1. Se ha tomado la siguiente muestra de una distribución exponencial:

$$\mathbf{x} = (2, 6, 5, 3) .$$

Obtenga el valor estimado de ω , el error típico y contraste la hipótesis $H_0 : \omega = 0,1$ mediante el estadístico G^2 en R.

Ejercicio 2. Se ha tomado la siguiente muestra de una distribución de Poisson:

$$\mathbf{x} = (8, 7, 11, 2) .$$

Estime λ , obtenga el error típico y contraste la hipótesis $H_0 : \lambda = 4$ mediante el estadístico G^2

Ejercicio 3. Supongamos que se toma la siguiente muestra de una distribución normal con el objetivo de contrastar la hipótesis $H_0 : \mu = 20, \sigma^2 = 9$:

$$\mathbf{x} = (30, 33, 26, 29) .$$

Realice el contraste utilizando el estadístico G^2 en R.

Ejercicio 4. Una variable sigue la distribución de Poisson en dos poblaciones diferentes. Tomamos una muestra de cada población y se encuentra el resultado

$$\begin{aligned}\mathbf{x}_1 &= (4, 3, 6, 7) \\ \mathbf{x}_2 &= (15, 12, 14, 7)\end{aligned}$$

Contraste la hipótesis $H_0 : \lambda_1 = \lambda_2$ frente a $H_1 : \lambda_1 \neq \lambda_2$ utilizando el estadístico G^2 en R.

Ejercicio 5. Un psicólogo está estudiando el tiempo en segundos en que se tarda en resolver una tarea, Y , en función del número de operaciones mentales involucradas en ella, X . Los datos de que dispone son

$$\begin{aligned}\mathbf{x} &= (1, 2, 3, 4, 5) \\ \mathbf{y} &= (2, 4, 5, 7, 6)\end{aligned}$$

Cómo la variable dependiente es el tiempo de respuesta, va a aplicar la regresión gamma. La media y varianza de Y se calculan del modo

$$\mu = \exp(a + bx)$$

$$\sigma^2 = \exp(v)$$

Utilice R para contrastar la hipótesis $H_0 : b = 0$ y realice el diagrama de dispersión con la línea de regresión superpuesta.