Metodos bayesianos - Interrogación 2 - parte II

Sebastian Baeza

Interrogación 2 - parte II

Pregunta 1)

Parte a).

Para que la muestra sea representativa se debe suponer que la probabilidad de que un participante vote a favor de la propuesta (o en contra) es independiente de si responde la encuesta telefónica o la rechaza. Además, es necesario suponer que las observaciones son independientes entre sí, en el sentido clásico, mientras que desde el enfoque bayesiano suponemos que son permutables.

Parte b).

Desde el punto de vista clásico, proponemos que cada observación de la muestra distribuye $x_i \overset{iid}{\sim} \text{Bern}(p)$ (verosimilitud), donde p es la probabilidad/proporción de individuos de Coralville a favor de la propuesta.

Luego, se quiere probar $H_0 : p \geq 0.5$ versus $H_1 : p < 0.5$. Luego para un modelo Bernoulli, podemos usar el test Binomial: rechazar H_0 SSI $\sum_{i=1}^{327} x_i \leq \text{Bin}_{(\alpha)}(327, 0.5)$. Entonces, rechazamos cuando $131 \leq \text{Bin}_{(\alpha)}(327, 0.5)$ (redondeamos $327 \cdot 0.4$), de aquí se sigue que el valor-p para este test es 0.00019.

```
round(327*0.4)
```

```
[1] 131
```

```
pbinom(round(327*0.4), 327, 0.5)
```

```
[1] 0.0001927988
```

Otro test que se puede realizar, es el test Z (aproximación), ya que $n = 327$ es suficientemente grande. Luego, rechazamos H_0 SSI $\frac{137-327\cdot 0.5}{\sqrt{327\cdot 0.5(1-0.5)}} \leq Z_{(\alpha)}$. De aquí se sigue que el valor-p de este test es 0.0017.

```
pnorm((137-327*0.5)/sqrt(327*0.5*0.5))
```

```
[1] 0.001689871
```

Luego, se sigue que en ambos test el valor-p es extremadamente bajo. En consecuencia, rechazamos $H_0 : p \geq 0.5$ (para una significancia $\alpha = 0.05$), es decir, se concluye que la proporción de personas que aprobará la medida es menor al 50%.

Parte c).

Desde el enfoque bayesiano, suponiendo que no tenemos conocimiento previo, lo conveniente es plantear una priori no informativa, supondremos una uniforme plana $p \sim U(0,1)$ (ya que la proporción p solo puede estar entre 0 y 1). Sabemos que la verosimilitud es proveniente de una bernoulli $x_i|p \overset{iid}{\sim} \text{Bern}(p); i = 1, ..., 327$. Luego, este modelo es conocido y la posteriori es $p|x_i \sim \text{Beta}(1 + \sum_{i=1}^{n} x_i, 1 + n - \sum_{i=1}^{n} x_i)$.

Parte d).

Consideremos el test de hipótesis $H_0 : p \geq 0.5$ versus $H_1 : p < 0.5$ y a la función de pérdida generalizada 0-1:

$$l(a_i, w_i) = \begin{cases} 0 & , a_0 = w_0; a_1 = w_1 \\ c_1 & , a_1 = w_0 \\ c_2 & , a_0 = w_1 \end{cases}$$

Con $c_1 = 2$ y $c_2 = 1$, ya que se considera más grave errar aceptando H_1, o sea considerar que la mayoría de la población no está a favor de la medida, cuando en realidad si lo está. En este caso se perderá mucho (daños materiales, vidas, molestia, etc.) si hay una inundación en la que la población no tuvo un respaldo de los impuestos siendo que querían tener dicho respaldo.

Entonces $\frac{c_2}{c_1+c_2} = \frac{1}{3}$, y por otro lado $P(p \geq 0.5|\sum_{i=1}^{327} x_i = 131) = 0.00016$. Luego como esta probabilidad es menor a $\frac{c_2}{c_1+c_2}$, se sigue que la decisión óptima es aceptar H_1.

```r
sumx <- round(327*0.4)
n <- 327
1 - pbeta(0.5, 1 + sumx, 1 + n - sumx)
```

```
[1] 0.0001590998
```

Luego, lo concluido es análogo desde ambos enfoques. Incluso desde el enfoque clásico el valor-p es tan bajo que para toda significancia razonable rechazaríamos $p \geq 0.5$. Mientras que desde el enfoque bayesiano la probabilidad de $p \geq 0.5$ es tan pequeña que aceptaríamos $p < 0.5$ inclusive ante valores de c_1 muy altos (o c_2 bajos).

Pregunta 2

Primero, notemos que las hipótesis a evaluar son $H_0 : \theta = 0$ (primer modelo, donde la pendiente no es significativa) versus $H_1 : \theta \neq 0$ (segundo modelo, con pendiente significativa). Entonces, el Factor de Bayes para este caso, hipótesis simple versus compuesta, es:

$$BF = \frac{p(y|M_0)}{p(y|M_1)} = \frac{p(y|\theta = 0)}{\int_{\mathbb{R}} p(y|\theta, x)p(\theta)d\theta}$$

Luego, del enunciado $(y|\theta = 0 \sim N(0, \sigma^2))$, se sigue que $p(y|\theta = 0) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2\right)\right\}$.

Por otro lado, para el denominador usaremos la expresión ofrecida por Tierney y Kadane, a fin de tener una expresión analítica. Entonces notemos que en el segundo modelo:

$p(y|\theta, x)p(\theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \theta x_i)^2\right\} \cdot (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{\theta^2}{2\tau^2}\right\}$

$= (2\pi\sigma^2)^{-\frac{n}{2}}(2\pi\tau^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n x_i y_i + \theta^2(\sum_{i=1}^n x_i^2)) - \frac{\theta^2}{2\tau^2}\right\}$

$= C \cdot (2\pi\tau^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(-2\theta\sum_{i=1}^n x_i y_i + \theta^2(\sum_{i=1}^n x_i^2)) - \frac{\theta^2}{2\tau^2}\right\}$

Donde $C = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}$ es una constante que no depende de θ, y coincide con el valor anteriormente mostrado del numerador del Factor de Bayes.

$= C \cdot (2\pi\tau^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2\tau^2}(-2\tau^2\theta\sum_{i=1}^n x_i y_i + \tau^2\theta^2(\sum_{i=1}^n x_i^2) + \sigma^2\theta^2)\right\}$

$= C \cdot (2\pi\tau^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2\tau^2}(-2\tau^2\theta\sum_{i=1}^n x_i y_i + \theta^2(\tau^2\sum_{i=1}^n x_i^2 + \sigma^2))\right\}$

$= C \cdot (2\pi\tau^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{\tau^2\sum_{i=1}^n x_i^2 + \sigma^2}{2\sigma^2\tau^2}(-2\theta(\tau^2\sum_{i=1}^n x_i^2 + \sigma^2)^{-1}\tau^2\sum_{i=1}^n x_i y_i + \theta^2)\right\}$

Luego, podemos ver que la expresión anterior es proporcional (con respecto a θ) a: $\exp\left\{-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-1}(-2\theta(\tau^2\sum_{i=1}^n x_i^2+\sigma^2)^{-1}\tau^2\sum_{i=1}^n x_iy_i+\theta^2)\right\}$, lo que corresponde al kernel de una normal. Luego, $\theta|y,x \sim N\left(\frac{\tau^2\sum_{i=1}^n x_iy_i}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}, \left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\right) \equiv N\left(\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\frac{\sum_{i=1}^n x_iy_i}{\sigma^2}, \left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\right)$. Esto nos servirá más adelante ya que la moda y media coinciden en la normal.

Luego, para usar la aproximación, necesitamos determinar $\frac{\partial^2\log(p(y|\theta,x)p(\theta))}{\partial\theta^2}$. Entonces, se sigue que $\frac{\partial^2\log(p(y|\theta,x)p(\theta))}{\partial\theta^2} =$

$$\frac{\partial^2}{\partial\theta^2}\left(\log(C)-\frac{1}{2}\log(2\pi\tau^2)-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-1}(-2\theta(\tau^2\sum_{i=1}^n x_i^2+\sigma^2)^{-1}\tau^2\sum_{i=1}^n x_iy_i+\theta^2)\right)$$

$$=\frac{\partial}{\partial\theta}\left(-\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-1}(-(\tau^2\sum_{i=1}^n x_i^2+\sigma^2)^{-1}\tau^2\sum_{i=1}^n x_iy_i+\theta)\right)$$

$$=-\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-1}$$

Entonces $\Sigma = \left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)$ (esto es esperable, la varianza exacta a posteriori). Luego, tenemos todo lo necesario para estimar la integral que corresponde a $p(y|M_1) = \int_{\mathbb{R}} p(y|\theta,x)p(\theta)d\theta \approx$

$$\left(\frac{2\pi\cdot\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{1/2}\cdot C\cdot(2\pi\tau^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-1}(-2\bar{\theta}(\tau^2\sum_{i=1}^n x_i^2+\sigma^2)^{-1}\tau^2\sum_{i=1}^n x_iy_i+\bar{\theta}^2)\right\}$$

Usando que la posteriori es conocida, y que su media es igual a la moda, se tiene que:

$$p(y|M_1)\approx\left(\frac{\sigma^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{1/2}\cdot C\cdot\exp\left\{\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\left(\frac{\sum_{i=1}^n x_iy_i}{\sigma^2}\right)^2-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\left(\frac{\sum_{i=1}^n x_iy_i}{\sigma^2}\right)^2\right\}$$

Entonces el Factor de Bayes se puede aproximar por:

$$BF\approx\left(\frac{\sigma^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)^{-\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}\left(\frac{\sigma^2\tau^2}{\tau^2\sum_{i=1}^n x_i^2+\sigma^2}\right)\left(\frac{\sum_{i=1}^n x_iy_i}{\sigma^2}\right)^2\right\}$$

Pregunta 3

Parte a).

Cargamos los datos solicitados:

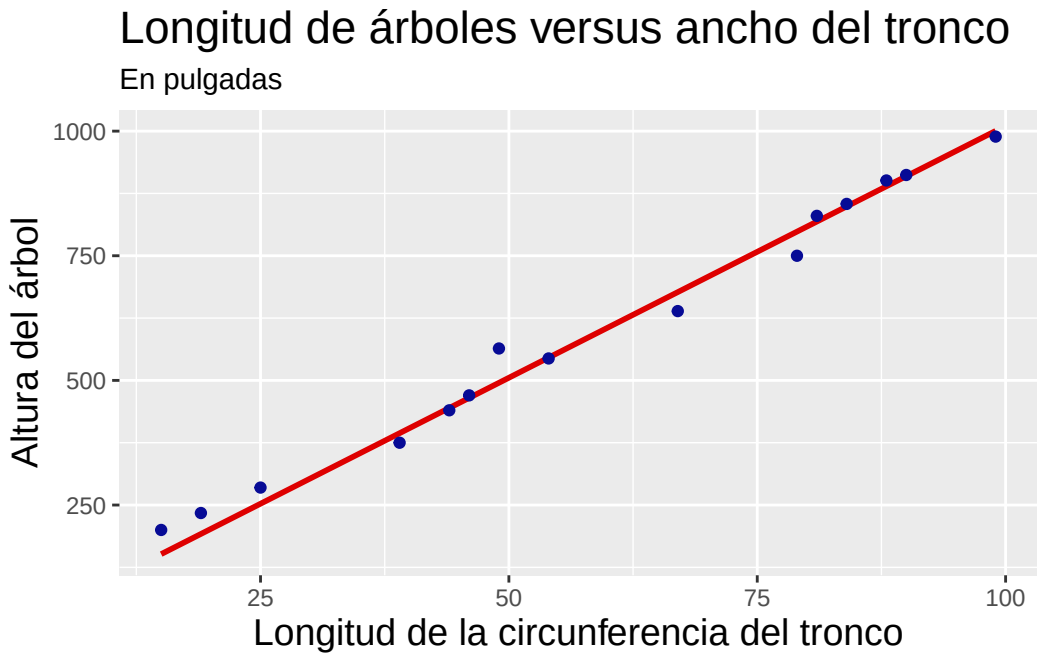
```
library(tidyverse)
arboles <- read_delim("arboles.csv")
```

Ahora ajustamos el modelo sin considerar intercepto:

```
modelo <- lm(altura ~ circunferencia - 1, arboles)
arboles$ajuste <- modelo$fitted.values
```

Luego, graficamos lo pedido, el gráfico de dispersión y la recta del modelo de regresión ajustado:

```
arboles %>%
  ggplot(aes(x = circunferencia, y = altura)) +
  geom_line(aes(y = ajuste), colour = "#DE0000", lwd = 1) +
  geom_point(colour = "#070B96") +
  labs(title = "Longitud de árboles versus ancho del tronco",
       y = "Altura del árbol",
       x = "Longitud de la circunferencia del tronco",
       subtitle = "En pulgadas") +
  theme(plot.title = element_text(size = 17),
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14))
```



Luego, para verificar la significancia del predictor (circunferencia) podemos recurrir al resumen del modelo:

```
summary(modelo)$coefficients
```

```
              Estimate Std. Error  t value     Pr(>|t|)
circunferencia 10.10574    0.12927 78.17546 6.831493e-20
```

Luego, al obtener el valor-p del test t, notamos que es muy bajo, del orden de 10^{-20}. Entonces, para una significancia de $\alpha = 0.05$ rechazamos la hipótesis nula de que el coeficiente asociado a la variable circunferencia es 0. Este regresor es significativo.

Parte b).

Desde el punto de vista bayesiano, consideremos el test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, que podemos evaluar mediante el Factor de Bayes estimado en 2.:

```
x <- arboles$circunferencia
y <- arboles$altura
sigma2 <- 250^2
tau2 <- 100^2
varpost <- sigma2 * tau2 / (tau2 * sum(x) + sigma2)
(BF <- (varpost/tau2)^(-1/2) * exp(-varpost * (sum(x*y)/sigma2)^2 / 2))
```

```
[1] 0
```

```
log10(BF)
```

```
[1] -Inf
```

```
-log10(BF)
```

```
[1] Inf
```

Luego, hay una pérdida de cifras significativas en el cálculo, pues $BF \approx 0$. Sin embargo, esto es suficiente para notar que $-\log_{10} BF > 2$ (ver desarrollo alternativo en script), por tanto, usando el criterio de Kass y Raftery la evidencia es decisiva a favor de H_1. Por ende, desde ambos enfoques la conclusión es la misma, que la circunferencia es una variable significativa para la regresión y la fuerza con que se concluye esto es muy alta sin importar el enfoque usado.