

Quant II - Problem Set I - Estimators

Solution

Kenneth Benoit

1. R revision

(a) `library(foreign)`
`d <- read.dta("dail2002.dta")`

```
summary(d$votes1st)
summary(d$spend_total)
summary(d$incumb)
```

(b) `d$incumb <- as.factor(d$incumb)`

(c) `table(d$incumb, d$wonseat)`
`chisq.test(d$incumb, d$wonseat)`

We reject the null hypothesis that there is no relationship between being an incumbent and having won a seat, since it is very unlikely to obtain a Chi-square statistic of this magnitude in the sample if the null is true.

(d) `t.test(d$spend_total ~ d$incumb)`

The mean for total spending is 11080 Euros for non-incumbents, and 21854 Euros for incumbents. The null hypothesis that the means in the two groups are equal is rejected, since the probability of getting a t-statistic of size -12.87 or more extreme is very low under the assumption that the means in the population are equal.

(e) `reg <- lm(d$spend_total ~ d$incumb + d$senator + d$councillor)`
`summary(reg)`

The intercept reflects total spending if all independent variables are zero. A candidate who is neither an incumbent TD nor an incumbent senator nor an incumbent councillor spends on average 10183 Euros. Candidates who are incumbent councillors spend around 6061 Euros more than those who are not. Being an incumbent senator is reflected in a total spending figure that lies about 6128 Euros above the one for candidates who are not senators. And incumbent TDs spend *ceteris paribus* about 11583 Euros more than other candidates. All these differences are statistically significant at least at the 5% level. The R-squared amounts to .29, and the F-test, which tests whether all the coefficients are jointly zero, is statistically significant as well.

2. Bias, efficiency, and MSE.

- (a) A C B from left to right
- (b) B, the one on the right
- (c) C in the middle or B on the right, hard to decide by eye-balling
- (d) probably A on the left (there are dots outside the circle which are hard to see), it might also be B on the right
- (e) A on the left

3. Estimating the sample proportion

- (a)

$$MSE(P) = Var(P) + (Bias(P))^2 \quad (1)$$

$$= \frac{\pi(1-\pi)}{N} + (E(P) - \pi)^2 \quad (2)$$

$$= \frac{\pi(1-\pi)}{N} + (\pi - \pi)^2 \quad (3)$$

$$= \frac{\pi(1-\pi)}{N} \quad (4)$$

P is consistent.

(b)

$$MSE(P^*) = Var(P^*) + (Bias(P^*))^2 \quad (5)$$

$$= Var(P^*) + (E(P^*) - \pi)^2 \quad (6)$$

$$= \left(\frac{N}{N+2}\right)^2 \frac{\pi(1-\pi)}{N} \quad (7)$$

$$+ \left[\frac{N}{N+2}\pi + \frac{1}{N+2} - \pi \right]^2$$

$$= \left(\frac{N}{N+2}\right)^2 \frac{\pi(1-\pi)}{N} \quad (8)$$

$$+ \left[\left(\frac{N}{N+2} - 1\right)\pi + \frac{1}{N+2} \right]^2$$

$$= \left(\frac{N}{N+2}\right)^2 \frac{\pi(1-\pi)}{N} \quad (9)$$

$$+ \left[\left(\frac{-2}{N+2}\right)\pi + \frac{1}{N+2} \right]^2$$

$$= \left(\frac{N}{N+2}\right)^2 \frac{\pi(1-\pi)}{N} \quad (10)$$

$$+ \left[\left(\frac{1}{N+2}\right)(1-2\pi) \right]^2$$

$$= \frac{N^2}{(N+2)^2} \frac{\pi(1-\pi)}{N} + \frac{1}{(N+2)^2} (1-2\pi)^2 \quad (11)$$

$$= \left(\frac{1}{N+2}\right)^2 (N\pi - N\pi^2 + 1 - 4\pi + 4\pi^2) \quad (12)$$

$$= \frac{\pi^2(4-N) + \pi(N-4) + 1}{(N+2)^2} \quad (13)$$

$$= \frac{\pi(1-\pi)(N-4) + 1}{(N+2)^2} \quad (14)$$

P^* is consistent.

(c) If we conceive of efficiency as relative efficiency in terms of the MSE:

$$\frac{MSE(P)}{MSE(P^*)} = \frac{\frac{\pi(1-\pi)}{N}}{\frac{\pi(1-\pi)(N-4)+1}{(N+2)^2}} \quad (15)$$

For $N = 10$ we get:

$$14.4 \frac{\pi(1 - \pi)}{6\pi(1 - \pi) + 1} \quad (16)$$

The result is (rounded figures):

π	rel. eff.
0	0
.1	.84
.2	1.18
.3	1.34
.4	1.42
.5	1.44
.6	1.42
.7	1.34
.8	1.18
.9	.84
1	0

In R, this may be calculated as follows:

```
f <- function(p,n) {  
  ((p*(1-p))/n)/((p*(1-p)*(n-4)+1)/((n+2)^2))  
}  
  
pseq <- seq(0,1,by =.1)  
releff <- f(pseq,10)  
  
overv <- cbind(pseq,releff)  
overv
```

Note: One may interpret relative efficiency also in terms of the variance instead in terms of the MSE.

- (d) A short answer: If π is around .5.
A long answer: We may state that we prefer P^* if its MSE is smaller than the one of P . Then:

$$MSE(P) > MSE(P^*) \quad (17)$$

$$\frac{\pi(1-\pi)}{N} > \frac{\pi(1-\pi)(N-4)+1}{(N+2)^2} \quad (18)$$

Multiplying by $N(N+2)^2 > 0$ yields:

$$(\pi - \pi^2)(N+2)^2 > N(\pi - \pi^2)(N-4) + N \quad (19)$$

Dividing by $\pi - \pi^2 > 0$ yields:

$$(N+2)^2 > N^2 - 4N + \frac{N}{\pi - \pi^2} \quad (20)$$

$$8N + 4 > \frac{N}{\pi - \pi^2} \quad (21)$$

Multiplying by $\frac{\pi - \pi^2}{8N+4} > 0$ and re-arranging yields:

$$-\pi^2 + \pi - \frac{N}{8N+4} > 0 \quad (22)$$

Setting the equation equal to zero yields:

$$\pi_{1/2} = \frac{-1 \pm \sqrt{1 - \frac{N}{2N+1}}}{-2} \quad (23)$$

The first derivative of the left part of equation 22 is $-2\pi + 1$, so $\pi = .5$ is a maximum. Then $MSE(P^*) < MSE(P)$ if $\pi_2 < \pi < \pi_1$.

4. Various things can be done here. For example, a Monte Carlo simulation that compares P and P^* may look as follows:

```
###small simulation

#model for data generation, define function

simsucc <- function(nsim,size,prob){
rbinom(n=nsim,size=size,prob=prob)
}

#use this to generate data by calling function in two steps
```

```

#first, set values for simulation here
nsim <- 10000
size <- 10
prob <- .5

#second, actual call
gensucc <- simsucc(nsim=nsim, size=size, prob=prob)

#estimate with different estimators

estp <- gensucc/size
estpstar <- (gensucc/size)*(size/(size+2))+(1/(size+2))
biasestp <- mean(estp)-prob
varestp <- var(estp)
mseestp <- varestp + biasestp^2
biasestpstar <- mean(estpstar)-prob
varestpstar <- var(estpstar)
mseestpstar <- varestpstar + biasestpstar^2

#obtain sampling distributions, bias, variance and MSE
summary(estp)
biasestp
varestp
mseestp
summary(estpstar)
biasestpstar
varestpstar
mseestpstar

```