Advanced Economics Statistics

FALL 2010

Seventh-Assignment Answer Sheet

by Freddy Rojas Cama

1. The answer to exercise 6.15 (b) of CB states "$g(\overline{X}, S^2)$ has zero expectation so $(\overline{X}, S^2)$ not complete." We give an example of $P(g(T) = 0) \neq 1$. By definition 6.2.21 of Casella and Berger (2002), let $f(t|\theta)$ be a family of pdf's or pmf's for a statistic $T(x)$. The family of probability distributions is called "*complete*" if $E_\theta g(T) = 0$ for all θ implies $P_\theta[(g(T)) = 0] = 1$ for all θ..Equivalently, θ is named a complete statistic. In solution of 6.15 b in Casella and Berger (2001) we have that $g(T)$ is;

$$g(T) = \frac{n}{a+n}\overline{X}^2 - \frac{S^2}{a}$$

which has expectation equal to zero. We claim that the related family of probability distributions is not complete in fact; we begin with the following expression in order to prove this;

$$\frac{n}{a+n}\overline{X}^2 - \frac{S^2}{a} = c$$

where c is a constant. In terms of matrices

$$\frac{n}{a+n}\frac{(X'i)(i'X)}{n^2} - \frac{1}{a}\frac{X'AX}{n-1} = c$$

where A is an idempotent matrix equal to $I - \frac{1}{n}ii'$. we can rewrite the above expression as following;

$$\frac{1}{a+n}\frac{(X'i)(i'X)}{n} - \frac{1}{a}\frac{X'AX}{n-1} = c$$

we name $\left(\frac{1}{a+n}\right)\frac{1}{n} = b$ and $\left(\frac{1}{a}\right)\frac{1}{n-1} = d$, thus we have

$$b \cdot (X'i)(i'X) - d \cdot X'AX = c \tag{1}$$

above expression is a polinomial where b, d and c are chosen in order to fulfill the above equality, and one of the solutions is taking $c = 0$. Thus, $P_\theta[(g(T)) = 0] \neq 1$ because there are another solutions $(g(T) \neq 0)$ where the expression (1) is hold.

2. Considering an example in genetic modeling which is a genetic linkage multinomial model we observe the multinomial vector (x_1, x_2, x_3, x_4) with the probabilities $\frac{1}{2}+\frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta)$ and $\frac{1}{4}(\theta)$ respectively.

   (a). We show that this is a curved exponential family. Thus,

$$f(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}P_1^{x_1}P_2^{x_2}P_3^{x_3}P_4^{x_4}$$

according to chapter 3, a family of pdf's or pmf's is called an exponential family if it can be expressed as;

$$f(x|\lambda) = h(x)\, c(\lambda) \exp\left(\sum\nolimits_i w_i(\lambda) \cdot t_i(x)\right)$$

In this case;

$$\begin{aligned} h(x) &= \frac{n!}{x_1!x_2!x_3!x_4!} \\ c(\lambda) &= n! \\ w_i(\lambda) &= \ln P_i \\ t_i &= x_i \end{aligned}$$

(b). We find a sufficient statistic for θ. In this case;

$$\begin{aligned} \frac{p(x|\theta)}{q(T(x)|\theta)} &= \left(\frac{n!}{x_1!x_2!x_3!x_4!}\right)^{-1} \frac{P_1^{x_1}P_2^{x_2}P_3^{x_3}P_4^{x_4}}{P_1^{x_1}P_2^{x_2}P_3^{x_3}P_4^{x_4}} \\ \frac{p(x|\theta)}{q(T(x)|\theta)} &= \left(\frac{n!}{x_1!x_2!x_3!x_4!}\right)^{-1} \end{aligned}$$

thus, $\frac{p(x|\theta)}{q(T(x)|\theta)}$ does not depend of θ, therefore the category countings are the sufficient statistics to estimate multinomial parameters.

(c). According to wikipedia[1], A sufficient statistic is minimal sufficient if it can be represented as a function of any other sufficient statistic. In other words, $S(X)$ is minimal sufficient if and only if $S(X)$ is sufficient, and if $T(X)$ is sufficient, then there exists a function f such that $S(X) = f(T(X))$. A useful characterization of minimal sufficiency is that when the density f_θ exists, $S(X)$ is minimal sufficient if and only if $\frac{f_\theta(x)}{f_\theta(y)}$ is independent of $\theta \Leftrightarrow S(x) = S(y)$ Our proposal is $T(x')$ which states that sufficient statistics are $t_i = x_i$ for $i = 1,2$ and 3 and $t_4 = n - \sum_{j=1}^{3} x_j$.

$$f(x_1, x_2, x_3, x_4) = \frac{n!}{x_1'!x_2'!x_3'!(n - x_1' - x_2' - x_3')!} P_1^{x_1'} P_2^{x_2'} P_3^{x_3'} \left(1 - \sum\nolimits_{j=1}^{3} P_j\right)^{n - x_1' - x_2' - x_3'}$$

according to chapter 3, a family of pdf's or pmf's is called an exponential family if it can be expressed as

$$f(x|\lambda) = h(x)\, c(\lambda) \exp\left(\sum\nolimits_i w_i(\lambda) \cdot t_i(x)\right)$$

In this case;

$$\begin{aligned} h(x) &= \frac{1}{x_1'!x_2'!x_3'!(n - x_1' - x_2' - x_3')!} \\ c(\lambda) &= n! \\ w_i &= \ln P_i \text{ for i=1,2 and 3} \\ t_i &= x_i' \text{ for i=1,2 and 3} \\ w_4 &= \ln\left(1 - \sum\nolimits_{j=1}^{3} P_j\right) \\ t_4 &= n - x_1' - x_2' - x_3' \end{aligned}$$

[1] See http://en.wikipedia.org/wiki/Sufficient_statistic

We checked out the sufficiency of our proposal

$$\frac{p(x'|\theta)}{q(T(x')|\theta)} = \left(\frac{n!}{x_1'!x_2'!x_3'!(n-x_1'-x_2'-x_3')!}\right)^{-1} \frac{\left(\prod_j^3 P_j^{x_j'}\right)\left(1-\sum_{j=1}^3 P_j\right)^{n-x_1'-x_2'-x_3'}}{P_1^{x_1'}P_2^{x_2'}P_3^{x_3'}\left(1-\sum_{j=1}^3 P_j\right)^{n-x_1'-x_2'-x_3'}}$$

$$\frac{p(x'|\theta)}{q(T(x')|\theta)} = \left(\frac{n!}{x_1'!x_2'!x_3'!(n-x_1'-x_2'-x_3')!}\right)^{-1}$$

we know that our proposal is a function of statistics $t_i = x_i'$ for all i in order to compute the probabilities of ocurrency in the multinomial model. We check the minimal sufficiency;

$$\frac{f_\theta(x')}{f_\theta(x)} = \frac{\frac{n!}{x_1!x_2!x_3!(n-x_1-x_2-x_3)!}P_1^{x_1}P_2^{x_2}P_3^{x_3}\left(1-\sum_{j=1}^3 P_j\right)^{n-x_1-x_2-x_3}}{\frac{n!}{x_1!x_2!x_3!x_4!}P_1^{x_1}P_2^{x_2}P_3^{x_3}P_4^{x_4}}$$

if $x_1 = x_1'$ we have

$$\frac{f_\theta(x')}{f_\theta(x)} = \frac{x_4!\left(1-\sum_{j=1}^3 P_j\right)^{n-x_1'-x_2'-x_3'}}{(n-x_1'-x_2'-x_3')!P_4^{x_4}} \tag{2}$$

the expression (2) is dependent of parameters of the probability function; as we know $T(x')$ is a function of $T(x)$, and $T(x')$ is a sufficient statistic (see above). Then $T(x')$ is a minimal sufficient statistic.

3. This exercise it is an application of *generalized-method-of-moments* methodology;

(i). We find the method of moment estimates of θ; the data are;

$$\begin{aligned} x_1 &= 125 \\ x_2 &= 18 \\ x_3 &= 20 \\ x_4 &= 34 \end{aligned}$$

The moments to match;

$$P_i = \widehat{P}_i$$

In this case we have

$$\begin{aligned} \frac{x_1}{\sum_i^4 x_i} &= \frac{1}{2}+\frac{\theta}{4} \\ \frac{x_2}{\sum_i^4 x_i} &= \frac{1}{4}(1-\theta) \\ \frac{x_3}{\sum_i^4 x_i} &= \frac{1}{4}(1-\theta) \\ \frac{x_1}{\sum_i^4 x_i} &= \frac{\theta}{4} \end{aligned}$$

for each above equation there is a θ which solves those equations;

$$\begin{aligned}
\widehat{\theta} &= 4\frac{x_1}{\sum_i^4 x_i} - 2 = 0.5381 \\
\widehat{\theta} &= -4\frac{x_2}{\sum_i^4 x_i} + 1 = 0.6345 \\
\widehat{\theta} &= -4\frac{x3}{\sum_i^4 x_i} + 1 = 0.5939 \\
\widehat{\theta} &= 4\frac{x_4}{\sum_i^4 x_i} = 0.6904
\end{aligned}$$

(ii). We use GMM estimates in order to get an estimate of θ;

$$\min \left(P - \widehat{P}\right)' A \left(P - \widehat{P}\right)$$

where $\widehat{P} = \{\frac{x_1}{\sum_i^4 x_i}, \frac{x_2}{\sum_i^4 x_i}, \frac{x_3}{\sum_i^4 x_i}, \frac{x4}{\sum_i^4 x_i}\}'$, $P = \{\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\}'$ and A is a identity matrix. Thus, we minimize the following;

$$\begin{aligned}
S(\theta) &= \left(\frac{x_1}{\sum_i^4 x_i} - \frac{1}{2} - \frac{\theta}{4}\right)^2 + \left(\frac{x_2}{\sum_i^4 x_i} - \frac{1}{4}(1-\theta)\right)^2 + \left(\frac{x_3}{\sum_i^4 x_i} - \frac{1}{4}(1-\theta)\right)^2 \ldots \\
&+ \left(\frac{x4}{\sum_i^4 x_i} - \frac{1}{4}\theta\right)^2
\end{aligned}$$

we minimize the above expression;

$$\begin{aligned}
\frac{\partial S(\theta)}{\partial \theta} &= 2\left(\frac{x_1}{\sum_i^4 x_i} - \frac{1}{2} - \frac{\theta}{4}\right)\frac{1}{4} + 2\left(\frac{x_2}{\sum_i^4 x_i} - \frac{1}{4}(1-\theta)\right)\frac{1}{4} + 2\left(\frac{x_3}{\sum_i^4 x_i} - \frac{1}{4}(1-\theta)\right)\frac{1}{4} \ldots \\
&2\left(\frac{x4}{\sum_i^4 x_i} - \frac{1}{4}\theta\right)\frac{1}{4} = 0
\end{aligned}$$

re-arranging terms

$$\frac{\partial S(\theta)}{\partial \theta} = \left(-\frac{x_1}{2\sum_i^4 x_i} + \frac{x_2}{2\sum_i^4 x_i} + \frac{x_3}{2\sum_i^4 x_i} - \frac{x_4}{2\sum_i^4 x_i}\right) + \frac{\theta}{2} = 0$$

replacing data into above identity

$$\begin{aligned}
-\frac{\theta}{2} &= -\frac{125}{2 \cdot 197} + \frac{18}{2 \cdot 197} + \frac{20}{2 \cdot 197} - \frac{34}{2 \cdot 197} \\
-\frac{\theta}{2} &= -0.30711 \\
\widehat{\theta} &= 0.61421
\end{aligned}$$

4. This question continues the discussion in previous question:

(i). There are 197 events classified into one of 4 possible categories $X = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$.The probability of ocurrency is $\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta)$ and $\frac{1}{4}(\theta)$ respectively. Now, the log-likelihood function is;

$$\log L(\theta) = x_1 \ln\left(\frac{1}{2} + \frac{\theta}{4}\right) + x_2 \ln\left(\frac{1}{4}(1-\theta)\right) + x_3 \ln\left(\frac{1}{4}(1-\theta)\right) + x_4 \ln\left(\frac{1}{4}\theta\right)$$

maximizing the log-likelihood;

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{x_1}{2+\theta} - \frac{x_2 + x_3}{1-\theta} + \frac{x_4}{\theta} = 0 \quad (3)$$

so,

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{125}{2+\theta} - \frac{38}{1-\theta} + \frac{34}{\theta} = 0$$
$$\theta = 0.6270$$

See Pedersen (2001) which show the same result. We verified by the second order condition that the solution maximizes the likelihood function (we reach a global solution), in this case we take the expression (3) and perform another derivative with respect to parameter θ

$$\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta} = -\frac{x_1}{(2+\theta)^2} - \frac{x_2 + x_3}{(1-\theta)^2} - \frac{x_4}{(\theta)^2} < 0$$

Thus, the solution is a global maximum (a negative value of the second derivative ensures that we reach a global solution).

(ii). We wrote a GAUSS procedure for plotting the grid search, we used a grid of 1000 points between the space 0.1 and 0.9. It was too easy to do this search because the space for seeking the solution is bounded and known. The figure 4.1 shows the

search

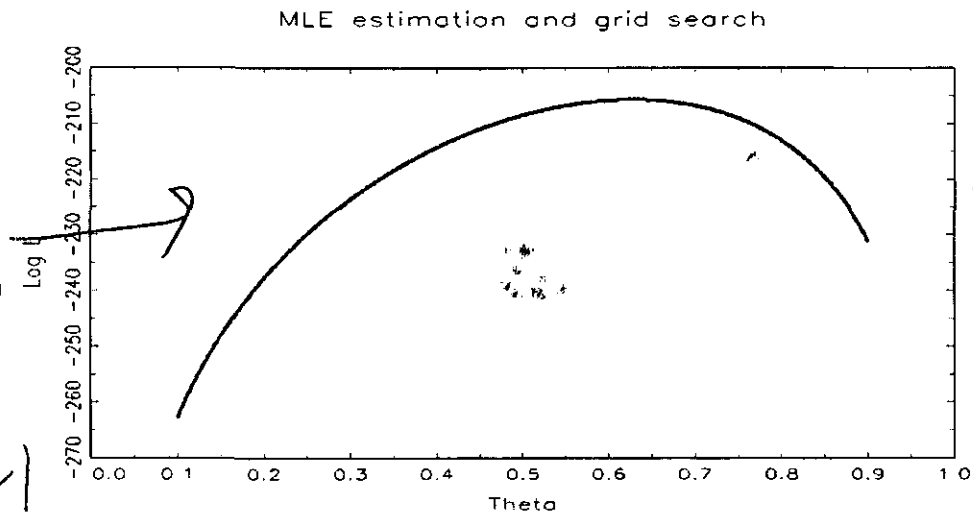

Figure 4.1

We identify the value of `theta` which maximizes the log-likelihood using the following command in GAUSS

```
theta[maxindc(lnL)];
```

where `theta` is the vector where we look for the solution, the solution belongs to [0.9 0.1] interval, and `lnL` is the loolikelihood, as a result we have $\theta_{grid} = 0.6269$. Anyway, we also use the maxlik and we set up the following procedure:

```
proc(1) = max_b(p0,data);
local x,L;
x=data;
L=(0.5+p0/4)^x[1]*((1-p0)/4)^x[2]*((1-p0)/4)^x[3]*(p0/4)^x[4];
retp(ln(L));
endp;
```

please see appendix for the GAUSS code. In this case we got $\theta_{maxlik} = 0.6268$. In anycase, we got the same results as in *4.i.*

References

[1] Casella, G and R. Berger. 2002. Statistical Inference. Second Edition, Duxbury Advanced Studies.

[2] Mendenhall and Scheaffer. 1973. Mathematical Statistics with Applications. Duxbury Press. North Scituate, Massachusetts.

[3] `Mathworld website. http://mathworld.wolfram.com/`

[4] GAUSS kernel density library. GAUSS.

[5] GAUSS maxlik library. GAUSS.

[6] M.P Wand & M.C Jones. 1995. Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman & Hall, 1995.

[7] Pedersen, Ted (2001). A gentle Introduction to EM algorithm. Slides in www.d.umn.edu/~tpederse/Docs/emnlp01-em-slides.ppt

[8] Minimal Sufficient statistic (Wikipedia). http://en.wikipedia.org/wiki/Sufficient_statistic.

1 Appendix

1.1 Question 4.ii: GAUSS code

```
new;
  cls;
  /*=============================================================================================
  /* Code by Freddy Rojas Cama */
  // Last update November 15th 2010
  // Rutgers University - Phd program in Economics
  /*=============================================================================================
  library pgraph kernel maxlik;
  pqgwin many;
  print "Assignment 7th";
  print "";
  /*******************************************************
  @Question 4.ii@
  *******************************************************/
  x=125|18|20|34;
  // Likelihood
  grid=(0.90-0.1)/(1000-1);
  theta=seqa(0.1,grid,1000);
  L=(0.5+theta./4).^x[1].*((1-theta)./4).^x[2].*((1-theta)./4).^x[3].*(theta./4).^x[4];
  lnL=ln(L);
  // Graph
  graphset;
   _pcolor = { 9 }; /* Colors for series */
   _pmcolor = { 1, 8, 2, 8, 8, 8, 8, 8, 15 };
   /*Colors for axes, title, x and y labels, date, box, and background */
   _plwidth={12}; /*Controls line thickness for main curves*/
   // _paxht=0.20; /*Controls size of axes labels*/
   _ptitlht = 0.18; /*Controls main title size */
  _pltype={6};
  _plegctl = { 1 5 0.04 13};
  _plegstr="MLE estimation and grid search";
  title("MLE estimation and grid search");
  ylabel("Log L");
  xlabel("Theta");
  xy(theta,lnL);
  //maxlik
  { x,f,g,cov,retcode } = MAXLIK(x,0,&max_b,0.3);
   proc(1)=max_b(p0,data);
   local x,L;
   x=data;
   L=(0.5+p0/4)^x[1]*((1-p0)/4)^x[2]*((1-p0)/4)^x[3]*(p0/4)^x[4];
   retp(ln(L));
```

```
 endp;
" ";
" ";
"Argument which maximixes the objective funtion (using maxlik)";; x;
"Argument which maximixes the objective funtion (using grid-search)";; theta[maxindc(lnL)];
"maxlikelihood";; f;
"gradient";; g;
```