STAT/BIOSTAT 534 Statistical Computing
Homework 5

Adrian Dobra
adobra@uw.edu

1 Background: Bayesian Inference in Normal Linear Regression

Let $Y = X_1$ be a continuous response variable and $X_{(2:p)} = (X_2, \ldots, X_p)$ be the vector of explanatory variables. Denote by D_1 the first column of the $n \times p$ data matrix D, by $D_{(2:p)}$ the columns $(2:p) = \{2, \ldots, p\}$ of D. The linear regression containing all the explanatory variables X_{-1} is denoted by $[1|(2:p)]$ and has coefficients $\beta_{(2:p)} = (\beta_2, \ldots, \beta_p)$. We center and scale the observed variables such that their sample means are zero and their sample standard deviations are one. As such, there is no need for an intercept parameter.

The full regression $[1|(2:p)]$ is given by

$$p(Y|X_{(2:p)} = x_{(2:p)}, \beta_{(2:p)}) = N\left(\sum_{j=2}^{p} \beta_j x_j, \sigma^2\right). \quad (1)$$

Now assume that only some explanatory variables X_A, $A \subset (2:p)$, are present in the linear regression (1). That is, we set to zero the regression coefficients associated with the rest of the explanatory variables:

$$\beta_j = 0,\ j \in (2:p) \subset A.$$

We denote with $[1|A]$ the regression that involves only the explanatory variables X_A. Moreover, we denote by $|A|$ the number of elements of A and by D_A the columns of D indexed by A. The equation associated with regression $[1|A]$ is

$$p(Y|X_A = x_A, \beta_A) = N\left(\sum_{j \in A} \beta_j x_j, \sigma^2\right). \quad (2)$$

We consider the following Bayesian specification of the regression model (2). The prior for σ^2 is $p(\sigma^2) = \text{inverse-Gamma}((|A| + 2)/2, 1/2)$ and, conditional on σ^2, the regression

coefficients β_A have independent priors $p(\beta_j) = N(0, \sigma^2)$, $j \in A$. The corresponding posterior distributions are

$$\begin{aligned} p(\sigma^2|D_1, D_A) &= \text{inverse-Gamma}\left((n + |A| + 2)/2, \left(1 + D_1^T D_1 - D_1^T D_A M_A^{-1} D_A^T D_1\right)\right), \\ p(\beta_A|\sigma^2, D_1, D_A) &= N_{|A|}\left(M_A^{-1} D_A^T D_1, \sigma^2 M_A^{-1}\right), \end{aligned}$$

where $M_A = I_{|A|} + D_A^T D_A$. The marginal likelihood of $[1|A]$ therefore given by

$$p(D_1, D_A|[1|A]) = \frac{\Gamma\left((n + |A| + 2)/2\right)}{\Gamma\left((|A| + 2)/2\right)} \left(\det M_A\right)^{-1/2} \left(1 + D_1^T D_1 - D_1^T D_A M_A^{-1} D_A^T D_1\right)^{-(n+|A|+2)/2} \quad (3)$$

2 Homework Problems

The data file "erdata.txt" has $n = 158$ samples (rows) and $p = 51$ variables (columns). The first variable (column 1) is your response variable (levels of a probe corresponding with the estrogen receptor transcription factor). The remaining 50 columns represent the expression levels of 50 genes that are highly correlated with the response. All these variables are continuous.

2.1 First problem (50 points)

In this problem you will use C to produce a function that calculates the marginal likelihood (3) associated with a linear regression. In Problem 2 Homework 1 I asked you to write the same function in R. The file "main.cpp" you will use is:

```
#include "matrices.h"

int main()
{
  int n = 158; //sample size
  int p = 51; //number of variables
  int i;

  int A[] = {2,5,10};//indices of the variables present in the regression
  int lenA = 3; //number of indices
  char datafilename[] = "erdata.txt";

  //allocate the data matrix
  double** data = allocmatrix(n,p);

  //read the data
  readmatrix(datafilename,n,p,data);
```

```
  printf("Marginal likelihood of regression [1|%d",A[0]);
  for(i=1;i<lenA;i++)
  {
    printf(",%d",A[i]);
  }
  printf("] = %.5lf\n",marglik(n,p,data,lenA,A));

  //free memory
  freematrix(n,data);

  return(1);
}
```

You task is to write the function

```
double marglik(int n,int p,double** data,int lenA,int* A);
```

If everything goes well, your program should run like this:

```
stu5:~/534> ./matrices
Marginal likelihood of regression [1|2,5,10] = -59.97893
```

2.2 Second problem (50 points)

In this problem you will use GSL to produce a C function that calculates the marginal likelihood (3) associated with a linear regression. In Problem 2 Homework 1 I asked you to write the same function in R, while in the first problem of this homework I asked you to write the same function using some numerical functions we wrote together in class. Now all the relevant numerical functions should come from GSL, that is, you need to use the functions from the file "matrices.cpp" located in

`/VectorsAndMatrices/MatricesGSL/`

The file "main.cpp" you will use is:

```
#include "matrices.h"

int main()
{
  int n = 158; //sample size
  int p = 51; //number of variables
  int i;

  int A[] = {2,5,10};//indices of the variables present in the regression
  int lenA = 3; //number of indices
  char datafilename[] = "erdata.txt";

  //allocate the data matrix
  gsl_matrix* data = gsl_matrix_alloc(n,p);

  //read the data
  FILE* datafile = fopen(datafilename,"r");
  if(NULL==datafile)
  {
    fprintf(stderr,"Cannot open data file [%s]\n",datafilename);
    return(0);
  }
  if(0!=gsl_matrix_fscanf(datafile,data))
  {
    fprintf(stderr,"File [%s] does not have the required format.\n",datafilename);
    return(0);
  }
  fclose(datafile);

  printf("Marginal likelihood of regression [1|%d",A[0]);
  for(i=1;i<lenA;i++)
  {
    printf(",%d",A[i]);
  }
  printf("] = %.5lf\n",marglik(data,lenA,A));

  //free memory
  gsl_matrix_free(data);

  return(1);
}
```

Your task is to write the function

```
double marglik(gsl_matrix* data,int lenA,int* A);
```

If everything goes well, your program should run like this:

```
stu5:~/534> ./matrices
Marginal likelihood of regression [1|2,5,10] = -59.97893
```

Points will be deducted if your version of the function uses numerical routines not defined in the library "GSL". It is okay to use usual mathematical functions defined in "math.h", e.g. "log" or "lgamma".