M5MS04 - Bayesian Statistics Project 2

Thomas Ng

February 18, 2013

1 Question 1

1.1 1a

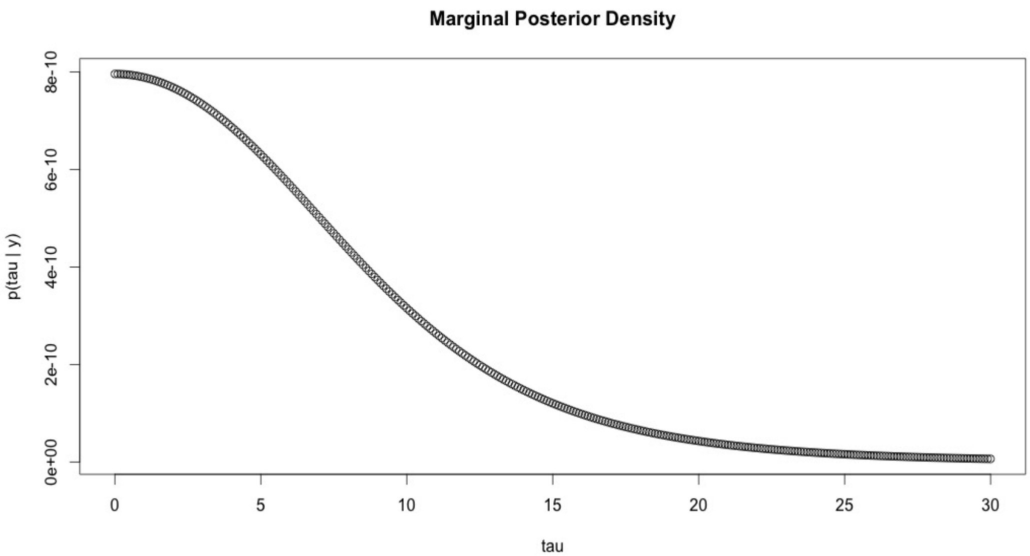
handwritten

1.2 1b

handwritten

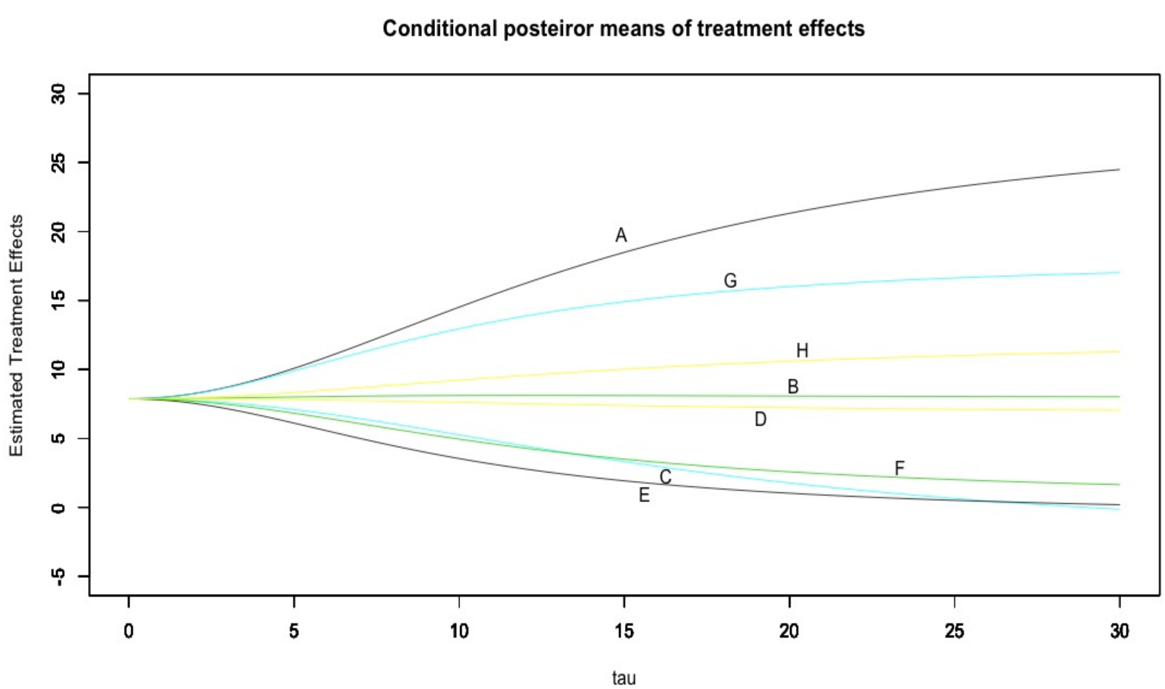
1.3 1c - Reproduce computations for the educational testing example

First, we reproduce the *marginal posterior density* $p(\tau|y)$. Basically, we just plot the marginal poserior density $p(\tau|y)$ against τ .



Second, we plot the *Conditional posterior means of treatment effects*, $E(\theta_j|\tau, y)$. Essentially, we plot $E(\theta_j|\tau, y)$ against τ. Notice the line for school C crosses the lines for E and F because C has a higher measurement error . This can be seen in the dataframe and its estimate is hence shrunk more strongly toward the overall mean in the Bayesian analysis.
We compute the conditional mean $E[\theta_j|\tau, y]$ by using tower property. The detailed calculations is written on the following page:

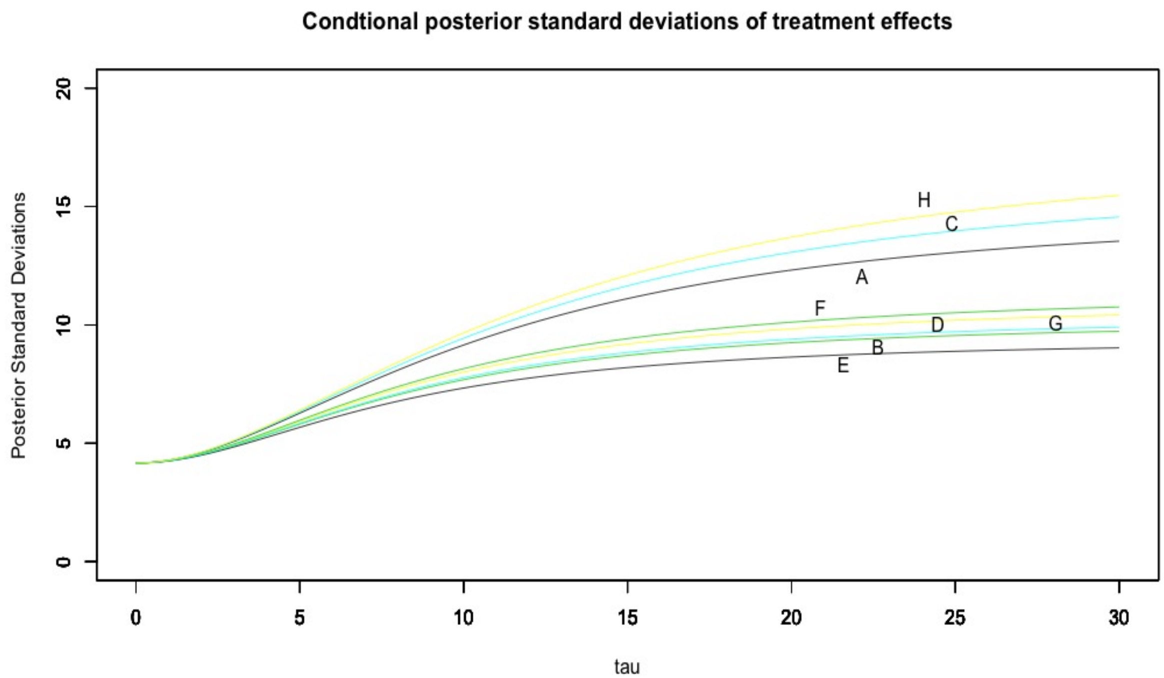
$$E[\theta_j|\tau, y] = E[E[\theta_j|\mu, \tau, y]|\tau, y]$$



Third, we reproduce the plot *Conditional posterior standard deviations of treatment effects*, $sd(\theta_j|\tau, y)$. Here, we use the following (detailed calculations on next page):

$$Var(\theta_j|\tau, y) = E[Var(\theta_j|\mu, \tau, y)|\tau, y] + Var(E[\theta_j|\mu, \tau, y]|\tau, y)$$

We obtain the above and square root it to get $sd(\theta_j|\tau, y)$.



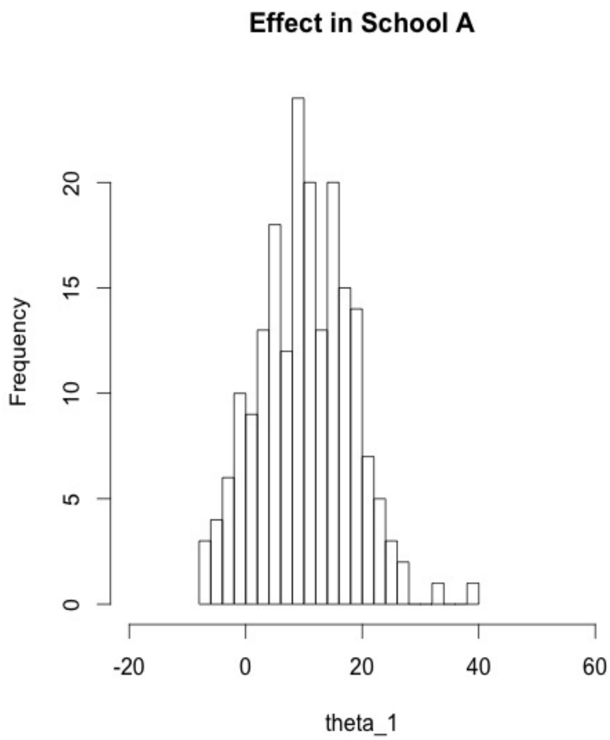
Fourth, to reproduce the summary table of 200 simulations of treatment effects in the eight schools, we adopt the following theoretical algorithm:

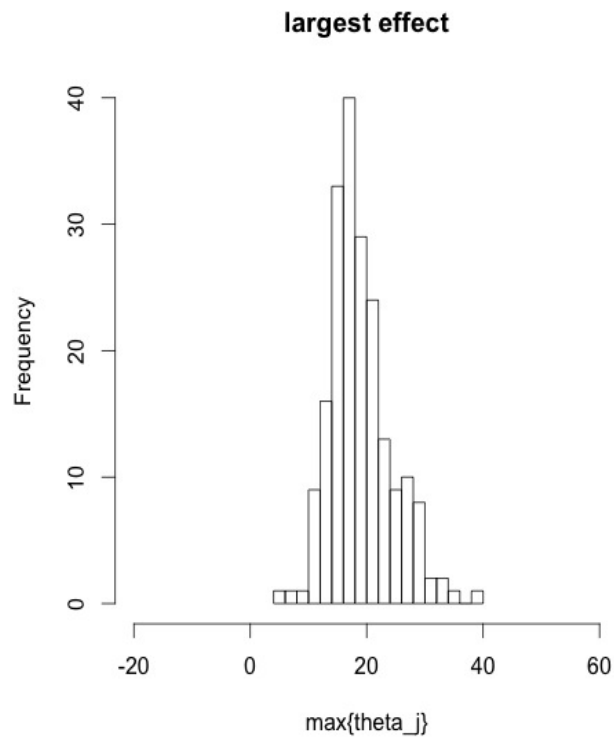
- Sample $(\tau^2)^{(t)} simp(\tau^2|y)$ using the grid method
- Sample $\mu^{(t)} \sim p(\mu|(\tau^2)^{(t)}, y)$
- Sample $\theta_j^{(t)} \sim p(\theta_j|\mu^{(t)}, (\tau^2)^{(t)}, y)$ indepdently for $j = 1, ..., J$

After we sample the θ_j s, we can compute their posterior quantiles. In my case, we set the number of simulations to be 200.

```
School      Posterior quantiles
----------------------------------------
      2.5% 25% 50% 75% 97.5%
  A     -4   5  11  16    27
  B     -8   0   6  13    21
  C    -10  -1   3   9    20
  D     -7   1   6  12    23
  E    -10  -1   3   8    19
  F    -11  -1   4   9    17
  G     -3   6  11  16    25
  H    -10   1   6  12    22
```

Fifth, we reproduce histograms of 2 quantities of interest computed from the 200 simulation draws: (a) the effect in school A, θ_j; (b) the largest effect, $max\{\theta_j\}$.





The second histogram on the right depicts the effect of the most successful of the 8 coaching programs at each simulation .

1.4 1d

To compute the probability that its coaching program is the best is equivalent to traversing the vector of simulations of the treatment effects of each school / coaching program and count the number of θ_j that is equal to $max\{\theta_j\}$.
For example, for school A, we traverse the vector of θ_A s and count the number of elements in the vector θ_A that is equal to $max\{\theta_j\}$. Thus,

$$Pr(\theta_i = max\{\theta_j\}) = Pr(school\ i\ is\ the\ best) = \frac{num\ of\ \{\theta_i = max\{\theta_j\}\}}{num\ of\ simulations}$$

```
School  1 prob of being best coaching program:
 ---------------------------------------
0.26 or  52  /  200

School  2 prob of being best coaching program:
 ---------------------------------------
0.105 or  21  /  200

School  3 prob of being best coaching program:
 ---------------------------------------
0.07 or  14  /  200

School  4 prob of being best coaching program:
 ---------------------------------------
0.135 or  27  /  200
```

```
School  5 prob of being best coaching program:
 ---------------------------------------
0.035 or  7  /  200

School  6 prob of being best coaching program:
 ---------------------------------------
0.06 or  12  /  200

School  7 prob of being best coaching program:
 ---------------------------------------
0.21 or  42  /  200

School  8 prob of being best coaching program:
 ---------------------------------------
0.125 or  25  /  200
```

For each pair of schools , j and k, we compute the probability that the coaching program for school j is better than that for school k again by traversing the vectors of θ_j and θ_k and then compare them elements by elements to count how many elements of θ_j is greater than θ_k $\forall\ k \neq j$. The matrix below with row j and columns k illustrates this probability. For instance, the first row shows $Pr(\theta_1 > \theta_k)\ \ \forall\ \ k \neq 1$

$$Pr(\theta_j > \theta_k) =$$

```
     [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]   
[1,] "\\"    "0.645" "0.75"  "0.675" "0.77"  "0.75"  "0.525" "0.625"
[2,] "0.355" "\\"    "0.61"  "0.51"  "0.6"   "0.555" "0.32"  "0.5"  
[3,] "0.25"  "0.39"  "\\"    "0.415" "0.515" "0.515" "0.275" "0.38" 
[4,] "0.325" "0.49"  "0.585" "\\"    "0.62"  "0.605" "0.355" "0.495"
[5,] "0.23"  "0.4"   "0.485" "0.38"  "\\"    "0.485" "0.255" "0.385"
[6,] "0.25"  "0.445" "0.485" "0.395" "0.515" "\\"    "0.265" "0.405"
[7,] "0.475" "0.68"  "0.725" "0.645" "0.745" "0.735" "\\"    "0.66" 
[8,] "0.375" "0.5"   "0.62"  "0.505" "0.615" "0.595" "0.34"  "\\"   
```

1.5 1e - Suppose $\tau^2 = \infty$

When $\tau^2 = \infty$, we can directly sample θ_j from $\theta_j|\mu, \tau^2, y \sim N(y_j, \sigma_j^2)$, since the distribution of θ_j doesnt depend on mu anymore. This means different schools / coaching programs havev different means y_j ,i.e. group means have no pattern and nothing in common.

$$E[\theta_j|Y, \mu, \sigma^2, \tau^2] = y_j$$

Therefore, individual coaching programs have their own individual means.
Under this assumption, if we repeat the computation of the probabilities in part(d), we obtain:

$$Pr(\theta_i = max\{\theta_j\}) = Pr(school\ i\ is\ the\ best) = \frac{num\ of\ \{\theta_i = max\{\theta_j\}\}}{num\ of\ simulations}$$

```
School  1 prob of being best coaching program:
 ---------------------------------------
0.505 or  101  /  200

School  2 prob of being best coaching program:
 ---------------------------------------
0.035 or  7  /  200

School  3 prob of being best coaching program:
 ---------------------------------------
```

```
0.015 or  3  /  200

School  4 prob of being best coaching program:
 ---------------------------------------
0.025 or  5  /  200

School  5 prob of being best coaching program:
 ---------------------------------------
0 or  0  /  200

School  6 prob of being best coaching program:
 ---------------------------------------
0.005 or  1  /  200

School  7 prob of being best coaching program:
 ---------------------------------------
0.205 or  41  /  200

School  8 prob of being best coaching program:
 ---------------------------------------
0.21 or  42  /  200
```

$$Pr(\theta_j > \theta_k) =$$

```
     [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
[1,] "\\"    "0.88"  "0.935" "0.88"  "0.96"  "0.96"  "0.695" "0.7"
[2,] "0.12"  "\\"    "0.745" "0.555" "0.745" "0.66"  "0.25"  "0.42"
[3,] "0.065" "0.255" "\\"    "0.28"  "0.415" "0.355" "0.115" "0.22"
[4,] "0.12"  "0.445" "0.72"  "\\"    "0.66"  "0.585" "0.205" "0.35"
[5,] "0.04"  "0.255" "0.585" "0.34"  "\\"    "0.48"  "0.065" "0.245"
[6,] "0.04"  "0.34"  "0.645" "0.415" "0.52"  "\\"    "0.135" "0.27"
[7,] "0.305" "0.75"  "0.885" "0.795" "0.935" "0.865" "\\"    "0.605"
[8,] "0.3"   "0.58"  "0.78"  "0.65"  "0.755" "0.73"  "0.395" "\\"
```

1.6 1f - Suppose $\tau^2 = 0$

In this case, we have

$$\mu|\tau^2, y \sim N\left(\frac{\sum_j \frac{y_j}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}}, \frac{1}{\sum_j \frac{1}{\sigma_j^2}}\right)$$

and

$$\theta_j|\mu, \tau^2, y \sim N(\mu, 0)$$

This essentially means $\theta_j = \mu$. Thus, we can sample μ from $\mu|\tau^2, y$ and then set $\theta_j = \mu$.
The implication here is that individual coaching programs have the same mean, i.e. group means are all equal. So there is no difference between groups. Thus.

$$E[\theta_j|Y, \mu, \tau^2, \sigma^2] = \mu$$

We shall expect the probability of school j being the best coaching program should be the same across all schools. We shoud also expect the probability that the coaching program for school j is better than that for school k to be roughly the same (i.e. equally likely) for all j, because individual schools have same means,

i.e. no difference between them.
Thus, we obtain:

$$Pr(\theta_i = max\{\theta_j\}) = Pr(school\ i\ is\ the\ best) = \frac{num\ of\ \{\theta_i = max\{\theta_j\}\}}{num\ of\ simulations}$$

```
School  1 prob of being best coaching program:
 ---------------------------------------
0.145 or  29  /  200

School  2 prob of being best coaching program:
 ---------------------------------------
0.11 or  22  /  200

School  3 prob of being best coaching program:
 ---------------------------------------
0.12 or  24  /  200

School  4 prob of being best coaching program:
 ---------------------------------------
0.11 or  22  /  200

School  5 prob of being best coaching program:
 ---------------------------------------
0.075 or  15  /  200

School  6 prob of being best coaching program:
 ---------------------------------------
0.14 or  28  /  200

School  7 prob of being best coaching program:
 ---------------------------------------
0.145 or  29  /  200

School  8 prob of being best coaching program:
 ---------------------------------------
0.155 or  31  /  200
```

$$Pr(\theta_j > \theta_k) =$$

```
     [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
[1,] "\\"    "0.505" "0.455" "0.51"  "0.545" "0.48"  "0.52"  "0.47"
[2,] "0.495" "\\"    "0.42"  "0.49"  "0.495" "0.42"  "0.475" "0.47"
[3,] "0.545" "0.58"  "\\"    "0.53"  "0.555" "0.48"  "0.505" "0.485"
[4,] "0.49"  "0.51"  "0.47"  "\\"    "0.53"  "0.485" "0.495" "0.46"
[5,] "0.455" "0.505" "0.445" "0.47"  "\\"    "0.45"  "0.46"  "0.45"
[6,] "0.52"  "0.58"  "0.52"  "0.515" "0.55"  "\\"    "0.5"   "0.535"
[7,] "0.48"  "0.525" "0.495" "0.505" "0.54"  "0.5"   "\\"    "0.49"
[8,] "0.53"  "0.53"  "0.515" "0.54"  "0.55"  "0.465" "0.51"  "\\"
```

1.7 1g - Discuss how answers in parts d - f differ

In part d, we have the bayesian continuum. Part e and f represent the limit case, where we get the classical dichotomy, while part d repsents the bayesian continuum , i.e. a weighted average of the 2 extremes.
In part e, when $\tau^2 \to \infty$, this means group means have no pattern, nothing in common. The estimated

treatment effect is essentially the observed treatment effect: $\hat{\theta}_j = y_j$. This basically reiterates the results of data and hence, uninformative, as we can simply observe from the observed treatment effect that school A has the best coaching program.
In part f, when $\tau^2 \to 0$, this means group means are all equal, i.e. $\hat{\theta_j} = E[\mu|y, \sigma^2, \tau^2] = \bar{y}..$. In this case, all coaching programs are equally as good. Each school can be the best coaching program with probability $\frac{1}{8}$. This is uninformative either, since setting $\tau^2 = 0$ assume there is no difference between groups and thus obscures what the data is trying to tell us.
Relatively speaking, Part d is an informative and optimal case, as it is the weighted average of both extremes in part e and f. It constitutes a comparison with the classical dichotomy through **the Shrinkage Parameter**:
The *Shrinkage Parameter for school(subgroup j)* is:

$$B_j = \frac{1/\tau^2}{1/\sigma_j^2 + 1/\tau^2} = \frac{\sigma_j^2}{\tau^2 + \sigma_j^2}$$

Our estimate of $(\theta_j - \mu)$ is $(y_j - \mu)$ "shrunk by B_j", *i.e.shrink group means towards global means*

$$E[\theta_j|Y, \mu, \sigma^2, \tau^2] - \mu = (1 - B_j)(y_j - \mu)$$