Assignment 3 - STA2201H Applied Statistics II

Luis Correia - Student No. 1006508566

March 04th 2020

Question 1 - IQ

Scoring on IQ tests is designed to produce a normal distribution with a mean of 100 and a standard deviation of 15 when applied to the general population. Now suppose we are to sample n individuals from a particular town and then estimate μ, the town-specific mean IQ score, based on the sample of size n. Let Y_i denote the IQ score for the i-th person in the town of interest, and assume

$$Y_1, Y_2, ..., Y_n | \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right) \tag{1}$$

For this question, will assume that the standard deviation of the IQ scores in the town is equal to 15, then mean is equal to 113 and the number of observations is equal to 10. Additionally, for Bayesian inference, the following prior will be used:

$$\mu \sim N\left(\mu_0, \sigma_0^2\right) \tag{2}$$

with $\mu_0 = 100$ and $\sigma_{\mu_0}^2 = 15$.

(a) Write down the posterior distribution of μ based on the information above. Give the Bayesian point estimate and a 95

{*Answer.*}

As seen in class slides, from week-6, the likelihood of the data, given by $Y_1, Y_2, ..., Y_n | \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$ is given by:

$$\begin{aligned} p(\mathbf{y}|\mu, \sigma^2) &= \prod_{i=1}^{n} p(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

$$\implies p(\mathbf{y}|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) \tag{3}$$

Considering that the prior on μ can be expressed by

$$\mu \sim N\left(\mu_0, \sigma_0^2\right) \tag{4}$$

We have also seen that by (3) and (4), the posterior is normal distributed as:

$$\mu|\mathbf{y}, \sigma^2 \sim N\left(\frac{\mu_0/\sigma^2_{\mu_0} + n \times \bar{y}/\sigma^2}{1/\sigma^2_{\mu_0} + n/\sigma^2}, \frac{1}{1/\sigma^2_{\mu_0} + n/\sigma^2}\right) \tag{5}$$

As we have been given that:

- $\mu_0 = 100$
- $\sigma_{\mu_0} = 15$
- $n = 10$
- $\bar{\mathbf{y}} = 113$
- $\sigma = 15$

Then, substituting these values on (5) we have the posterior as:

$$\mu|\mathbf{y}, \sigma^2 \sim N\left(\frac{100/15^2 + 10 \times 113/15^2}{1/15^2 + 10/15^2}, \frac{1}{1/15^2 + 10/15^2}\right)$$

$$\implies \mu|\mathbf{y}, \sigma^2 \sim N\left(\frac{1230}{11}, \frac{225}{11}\right) \tag{6}$$

From (6) follows that the Bayesian point estimate is given by:

$$\hat{\mu}_{Bayes} = E(\mu|\mathbf{y}, \sigma^2) = \frac{1230}{11} = 111.82 \tag{7}$$

In order to obtain the 95% credible interval, we need just to calculate the `qnorm` probability of distribution given by (6) for the desired level of confidence.

By doing so, we obtained the following result:

Table 1: 95% C.I. for Bayesian point estimation

Lower	Upper
102.9539	120.6825

(b) Suppose that (unknown to us) the true mean IQ score is μ^*. To evaluate how close an estimator is to the truth, we might want to use the mean squared error (MSE) $MSE[\hat{\mu}|\mu^*] = E[(\hat{\mu} - \mu^*)^2|\mu^*]$. Show the MSE is equal to the variance of the estimator plus the bias of the estimator squared, i.e. $MSE[\hat{\mu}|\mu^*] = Var[\hat{\mu}|\mu^*] + Bias(\hat{\mu}|\mu^*)^2$.

{*Answer.*}

From the definition of MSE, i.e., starting from $MSE[\hat{\mu}|\mu^*] = E[(\hat{\mu} - \mu^*)^2|\mu^*]$ and adding $E[(\hat{\mu}|\mu^*]$ in the left side of the equation we have:

$$\begin{aligned}
MSE[\hat{\mu}|\mu^*] &= E[(\hat{\mu}-\mu^*)^2|\mu^*] \\
&= E[(\hat{\mu}-E[\hat{\mu}|\mu^*]+E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*] \\
&= E[(\hat{\mu}-E[\hat{\mu}|\mu^*])^2+2[(\hat{\mu}-E[\hat{\mu}|\mu^*])(E[\hat{\mu}|\mu^*]-\mu^*)]+(E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*] \\
&= E\{(\hat{\mu}-E[\hat{\mu}|\mu^*])^2|\mu^*\}+E\{2[(\hat{\mu}-E[\hat{\mu}|\mu^*])(E[\hat{\mu}|\mu^*]-\mu^*)]|\mu^*\}+E\{(E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*\} \\
&= Var[\hat{\mu}|\mu^*]+E\{2[(\hat{\mu}-E[\hat{\mu}|\mu^*])(E[\hat{\mu}|\mu^*]-\mu^*)]|\mu^*\}+E\{(E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*\} \\
&= Var[\hat{\mu}|\mu^*]+2E\{(\hat{\mu}-E[\hat{\mu}|\mu^*])(E[\hat{\mu}|\mu^*]-\mu^*)|\mu^*\}+E\{(E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*\}
\end{aligned}$$

Note that the 1st term after $Var[\hat{\mu}|\mu^*]$ is:

$$\begin{aligned}
E\{(\hat{\mu}-E[\hat{\mu}|\mu^*])(E[\hat{\mu}|\mu^*]-\mu^*)|\mu^*\} &= \\
E\{\hat{\mu}E[\hat{\mu}|\mu^*]-\hat{\mu}\mu^*-(E[\hat{\mu}|\mu^*])^2+\mu^*E[\hat{\mu}|\mu^*]|\mu^*\} &= \\
\hat{\mu}E\{E[\hat{\mu}|\mu^*]|\mu^*\}-\mu^*E[\hat{\mu}|\mu^*]-E\{(E[\hat{\mu}|\mu^*])^2|\mu^*\}+\mu^*E\{E[\hat{\mu}|\mu^*]|\mu^*\} &= \\
\hat{\mu}E[\hat{\mu}|\mu^*]-\mu^*E[\hat{\mu}|\mu^*]-E[\hat{\mu}|\mu^*]^2+\mu^*E[\hat{\mu}|\mu^*] &= \\
\hat{\mu}E[\hat{\mu}|\mu^*]-E[\hat{\mu}|\mu^*]^2 &= \\
E[\hat{\mu}|\mu^*](\hat{\mu}-E[\hat{\mu}|\mu^*]) &= \\
E[\hat{\mu}|\mu^*](E[\hat{\mu}|\mu^*]-E[\hat{\mu}|\mu^*]) &= \\
0
\end{aligned}$$

and considering that $Bias[\hat{\mu}|\mu^*]=E[\hat{\mu}|\mu^*]-\mu^*$, we have the last part of expression:

$$\begin{aligned}
E\{(E[\hat{\mu}|\mu^*]-\mu^*)^2|\mu^*\} &= \\
E\{(Bias[\hat{\mu}|\mu^*])^2|\mu^*\} &= \\
(Bias[\hat{\mu}|\mu^*])^2
\end{aligned}$$

Then

$$\implies MSE[\hat{\mu}|\mu^*]=Var[\hat{\mu}|\mu^*]+(Bias[\hat{\mu}|\mu^*])^2. \tag{8}$$

(c) Suppose that the true mean IQ score is 112. Calculate the bias, variance and MSE of the Bayes and ML estimators. Which estimator has a larger bias? Which estimator has a larger MSE?

{*Answer.*}

First, considering that MLE of a normal distribution is the sample mean, we have that:

$$\hat{\mu}_{ML}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{9}$$

From (9), calculating the bias for we have that:

$$\begin{aligned}
Bias[\hat{\mu}_{ML}|\mu^*] &= E[\hat{\mu}_{ML}|\mu^*]-\mu^* \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}y_i\middle|\mu^*\right]-\mu^* \\
&= \frac{1}{n}\sum_{i=1}^{n}E[y_i|\mu^*]-\mu^* \\
&= \frac{1}{n}n\mu^*-\mu^*=0
\end{aligned}$$

$$\implies Bias[\hat{\mu}_{ML}|\mu^*] = 0. \tag{10}$$

In this case, from (8) we have:

$$\begin{aligned} MSE[\hat{\mu}_{ML}|\mu^*] &= Var[\hat{\mu}_{ML}|\mu^*] + (Bias[\hat{\mu}_{ML}|\mu^*])^2 \\ &= Var[\hat{\mu}_{ML}|\mu^*] + 0 \\ &= Var[\hat{\mu}_{ML}|\mu^*] \\ &= Var\left[\frac{1}{n}\sum_{i=1}^{n} y_i \middle| \mu^*\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[y_i|\mu^*] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Assuming the parameters provided to item (a), i.e., $\sigma = 15$ and $n = 10$, we have:

$$MSE[\hat{\mu}_{ML}|\mu^*] = Var[\hat{\mu}_{ML}|\mu^*] = \frac{\sigma^2}{n} = \frac{15^2}{10} = 22.5. \tag{11}$$

From item(a), in (7) we calculated the Bayes estimator, so the Bias for Bayes estimator is given by:

$$Bias[\hat{\mu}_{Bayes}|\mu^*] = \frac{1230}{11} - 112 = 111.82 - 112 = -0.18. \tag{12}$$

In item(a), in (6) we calculated the Variance for Bayes estimator, which is

$$Var[\hat{\mu}_{Bayes}|\mu^*] = \frac{225}{11} = 20.45. \tag{13}$$

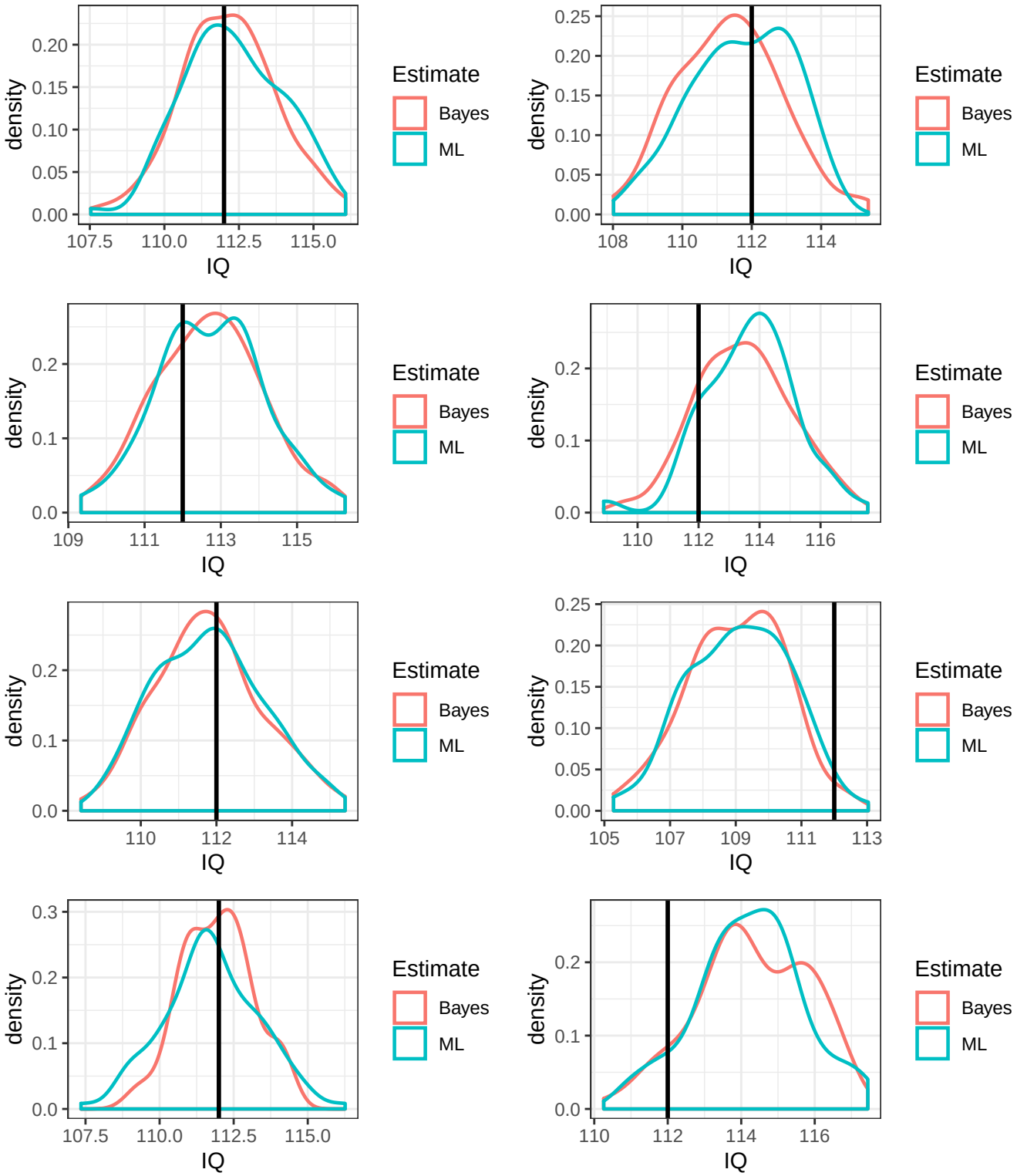
Then, from (12) and (13) we have the MSE for Bayes estimator equals to:

$$\begin{aligned} MSE[\hat{\mu}_{Bayes}|\mu^*] &= Var[\hat{\mu}_{Bayes}|\mu^*] + Bias[\hat{\mu}_{Bayes}|\mu^*]^2 \\ &= \frac{225}{11} + \left(\frac{1230}{11} - 112\right)^2 \\ &= 20.45 + (-0.18)^2 \\ &= 20.49. \end{aligned}$$

In this sense, we can conclude that $\hat{\mu}_{ML}$ **has the larger MSE** and $\hat{\mu}_{Bayes}$ **has the larger bias**.

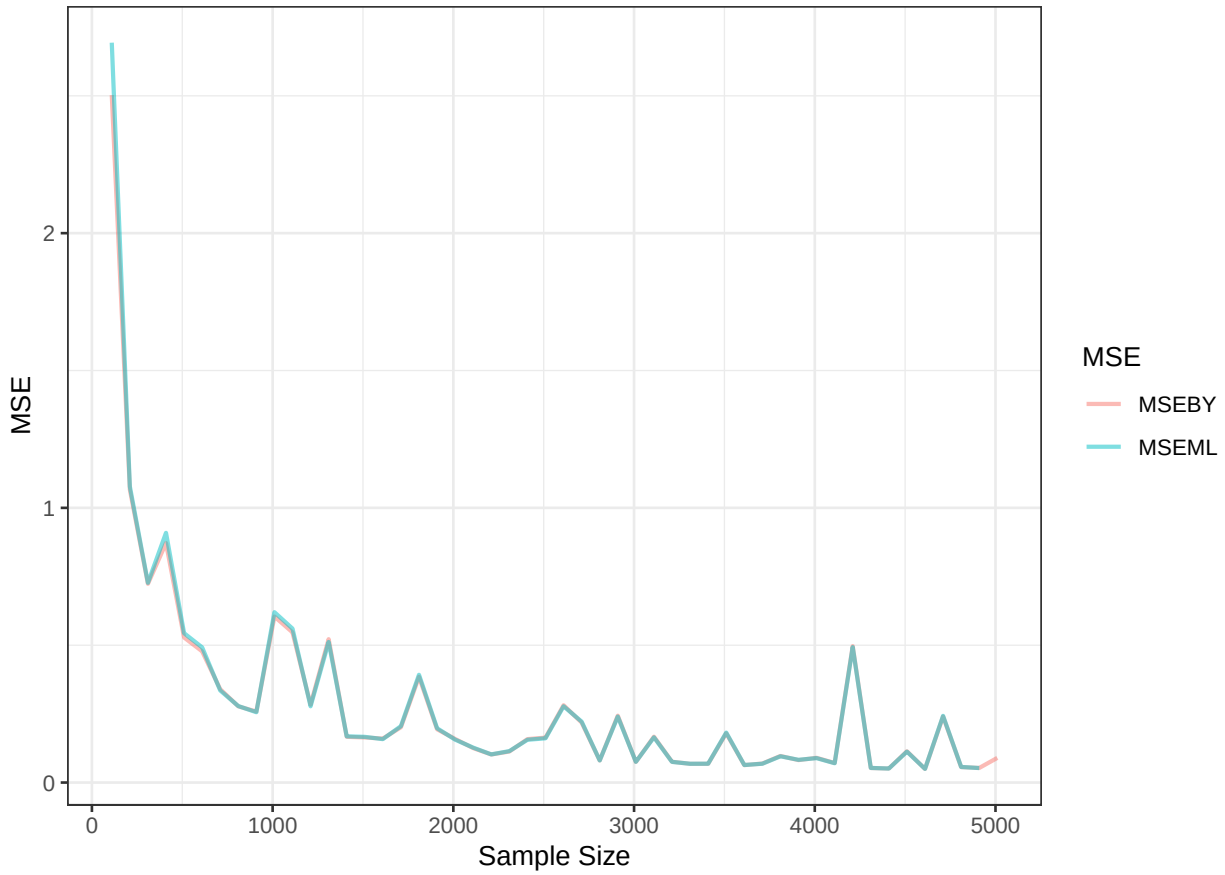
(d) Write down the sampling distributions for the ML and Bayes estimates, again assuming $\mu^* = 112$ and $\sigma = 15$. Plot the two distributions on the one graph. Summarize your understanding of the differences in bias, variance and MSE of the two estimators by describing how these differences relate to differences in the sampling distributions as plotted. To further illustrate the point, obtain the Bayes and ML MSEs for increasing sample sizes and plot the ratio (Bayes MSE)/(ML MSE) against sample size.

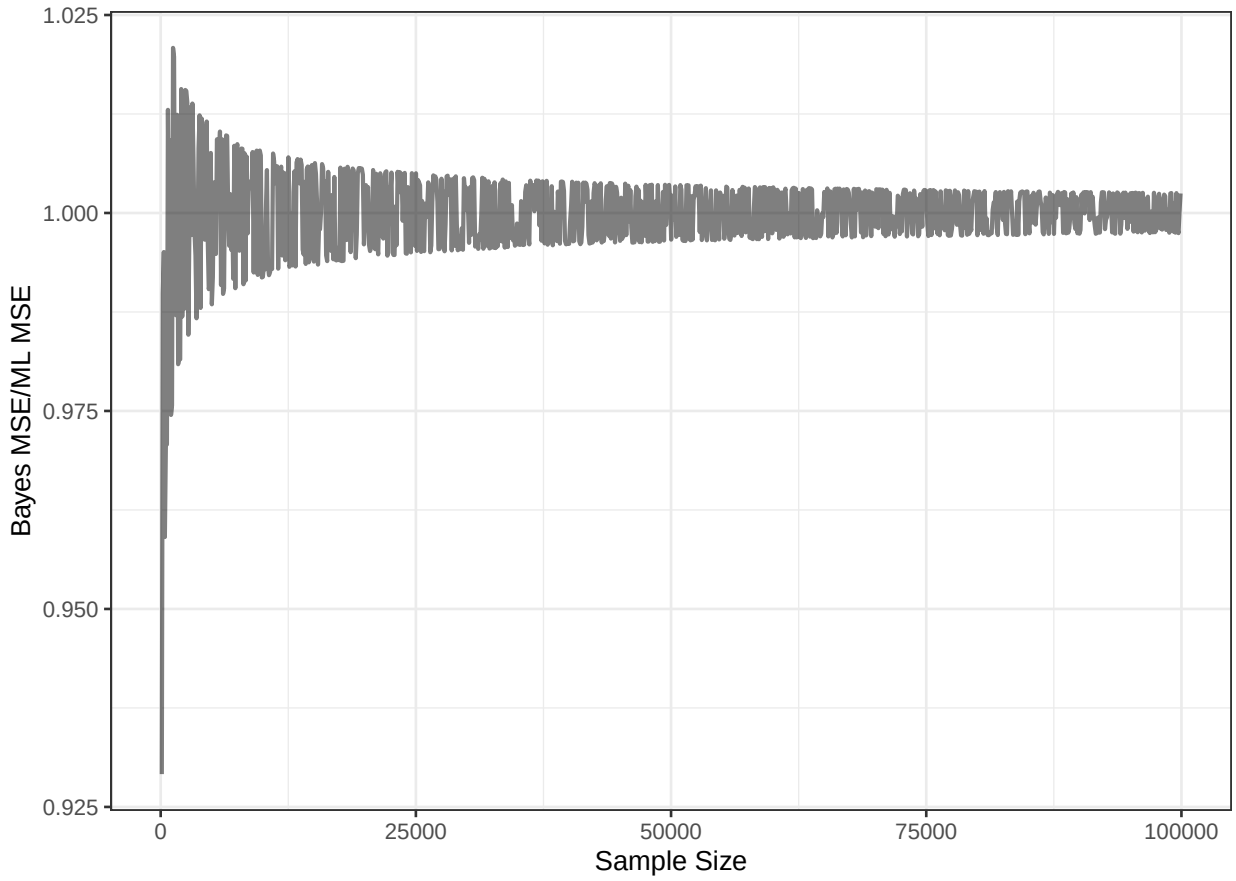
Let's take a look at some samples comparing both estimates - these charts were generated using the prior assumption $N(\mu_0, \sigma^2_{\mu_0})$, $\mu^* = 112$ (true mean) and known σ^2. We can see in some of them, the distribution of probability mass for Bayes estimator smaller when compared with ML estimates.



Now we generated 10.000 simulations using a crescent sample size in the interval $[100, 100000]$ and plot the

curves of MSE for both estimators and the ratio (Bayes MSE)/(ML MSE) against sample size.





Comments:

- The first obvious conclusion about Bias is that Bayes estimates have larger bias when compared with the ML estimate, which is unbiased. As a consequence, the probability mass is more distributes around the true mean, in most of the cases, for ML sampling than in Bayes sampling, which presents some deviance from real mean. On the other hand, ML estimate is exactly distributed around the true mean;
- The variance of Bayes estimates seems to be smaller than ML estimates. On the other hand, most of the time, the Bias of Bayes estimator is greater than in ML reflecting the unbiased characteristic of MLE;
- The ratio between Bayes MSE / ML MSE shows a convergence to 1, as the sample size increases. Asymptotically we can expect both MSE's to be equivalents.

Question 2 - Gibbs Sampling

(a) Suppose the parameter vector of interest θ has been divided into d components $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$. In Gibbs Sampling, each θ_j^s at iteration s is sampled from the conditional distribution given all other components of θ, i.e., $p(\theta_j|\theta_{-j}^{s-1}, y)$, where θ_{-j}^{s-1} is all the components of θ except for all j at their current values:

$$\theta_{-j}^{s-1} = (\theta_1^s, \ldots, \theta_{j-1}^{s-1}, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1}) \tag{14}$$

Write down an expression for the proposal distribution of $J(\theta^*|\theta^{s-1}$ for the Gibbs sampler and show that Gibbs sampling is a special case of the Metropolis-Hastings algorithm with $r = 1$.

{*Answer.*}

Let's consider the components for the full conditional distributions for **Gibbs Sampling** given by:

$$\theta_{-j}^{s-1} = (\theta_1^s, \ldots, \theta_{j-1}^{s-1}, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1}) \tag{15}$$

Let's also consider the **Metropolis-Hastings** algorithm which proposes to extract samples of the new θ^* from a distribution given by $J(\theta^*|\theta^{s-1})$ which is given by:

$$J(\theta^*|\theta^{s-1}) = J(\theta_1^s, \ldots, \theta_{j-1}^{s-1}, \theta_j^*, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1} \Big| \theta_1^s, \ldots, \theta_{j-1}^{s-1}, \theta_j^{s-1}, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1}) \tag{16}$$

Using (15) we can rewrite $J(\theta^*|\theta^{s-1})$ as follows:

$$\begin{aligned} J(\theta^*|\theta^{s-1}) &= J(\theta_j^*, \theta_{-j}^{s-1} \Big| \theta_j^{s-1}, \theta_{-j}^{s-1}) \\ &= p(\theta_j^* \Big| \theta_1^s, \ldots, \theta_{j-1}^{s-1}, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1}, y) \\ &= p(\theta_j^* \Big| \theta_j^{s-1}, \theta_{-j}^{s-1}, y) \end{aligned}$$

Then, for the new θ^* its full conditional probability is given by:

$$\implies J(\theta^*|\theta^{s-1}) = p(\theta_j^* \Big| \theta_j^{s-1}, \theta_{-j}^{s-1}, y) \tag{17}$$

From the definition of MH algorithm, we use the ratio of conditional probability of the new θ^* and the conditional probability of previous sample to decide to include or not the new estimate. In other words, we look at the ratio r given by:

$$r = \frac{p\Big(\theta^* \Big| y\Big)}{p\Big(\theta^{s-1} \Big| y\Big)} \tag{18}$$

By using the (17) and (18) we can rewrite the ratio r in the following way:

$$
\begin{aligned}
r &= \frac{p\big(\theta^*\big|y\big)/J_s(\theta^*|\theta^{s-1})}{p\big(\theta^{s-1}\big|y\big)/J_s(\theta^{s-1}|\theta^*)} \\
&= \frac{p(\theta_j^*, \theta_{-j}^{s-1}|y)/J_s(\theta_j^*, \theta_{-j}^{s-1}|\theta_j^{s-1}, \theta_{-j}^{s-1})}{p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)/J_s(\theta_j^{s-1}, \theta_{-j}^{s-1}|\theta_j^*, \theta_{-j}^{s-1})} \\
&= \frac{p(\theta_j^*, \theta_{-j}^{s-1}|y)/p\big(\theta_j^*\big|\theta_{-j}^{s-1}, y\big)}{p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)/p\big(\theta_j^{s-1}\big|\theta_{-j}^{s-1}, y\big)} \\
&= \frac{p(\theta_j^*, \theta_{-j}^{s-1}|y)/p\big(\theta_j^*\big|\theta_{-j}^{s-1}, y\big)}{p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)/p\big(\theta_j^{s-1}\big|\theta_{-j}^{s-1}, y\big)} \\
&= \frac{p(\theta_j^*, \theta_{-j}^{s-1}|y)p\big(\theta_j^{s-1}\big|\theta_{-j}^{s-1}, y\big)}{p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)p\big(\theta_j^*\big|\theta_{-j}^{s-1}, y\big)} \\
&= \frac{p(\theta_j^*, \theta_{-j}^{s-1}|y)p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)p\big(\theta_{-j}^{s-1}\big|y\big)}{p\big(\theta_j^{s-1}, \theta_{-j}^{s-1}\big|y\big)p\big(\theta_j^*, \theta_{-j}^{s-1}\big|y\big)p\big(\theta_{-j}^{s-1}\big|y\big)} \\
&= 1.
\end{aligned}
$$

So, we can conclude that, **a MH algorithm with $r = 1$ is a Gibbs Sampler**.

(b) This question relates to IQ example above. Let's make things a bit more realistic and assume the observed standard deviation is 13. Assume the observed sample mean is still 113. We will use the following priors to estimate both μ and σ in a Bayesian model:

$$\mu \sim N(\mu_0, \sigma_{\mu_0}^2) 1/\sigma^2 \sim Gamma\big(\nu_0/2, \nu_0/2 \times \sigma_0^2\big) \tag{19}$$

Let's set $\mu_0 = 100$, $\sigma_{\mu_0} = \sigma_0 = 15$ and $\nu_0 = 1$. Use Gibbs sampling in R to obtain posterior samples for μ and σ. Notes:

- you don't need to derive the full conditionals if you don't want to, can just use the expressions in lecture notes
- use sample mean and precision for initial values
- obtain 1000 samples
- code should be well commented so it's clear what is going on

Output required:

- trace plots for μ and σ
- histogram of posterior samples for μ and σ
- point estimates and 95% CI for μ and σ

{*Answer.*}

In this item it will be used the lecture results from Week-6 (Metropolis Hastings), slides #19-21, i.e., as seen in class, the full conditionals for μ and σ are:

$$\mu\big|\mathbf{y}, \sigma^2 \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2} + n \times \frac{\bar{y}}{\sigma^2}}{\frac{1}{\sigma_{\mu_0}^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_{\mu_0}^2} + \frac{n}{\sigma^2}}\right) \tag{20}$$

and

$$\frac{1}{\sigma^2}\Big|\mathbf{y}, \mu \sim Gamma\Big(\frac{\nu_n}{2}, \frac{\nu_n}{2} \times \sigma_n^2\Big) \tag{21}$$

with the following known parameters: $\mu_0 = 100$, $\sigma_{\mu_0} = \sigma_0 = 15$ and $\nu_0 = 1$, and for initial values $\mu^{(1)} = 113$ and $\sigma^{(1)} = 13$.

To implement Gibbs-Sampler I wrote 02 functions representing the full conditionals from μ and σ accepting the suggestion of using the lecture materials. These functions were used interactively within a loop, retro-feeding the estimate of the new parameter using those generated in previous step, implementing the loop where the samples were generated - the details of implementation can be verified in the `.rmd` file provided.

The simulation of samples for μ and σ resulted in the following graphs:

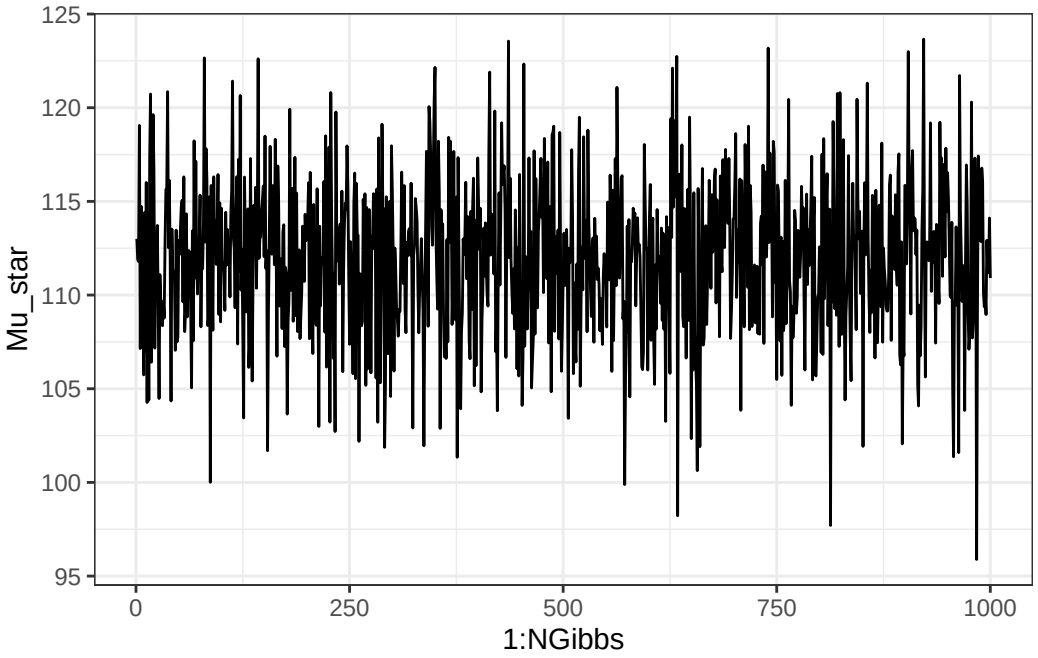


Figure 1: Mean - Gibbs sampling

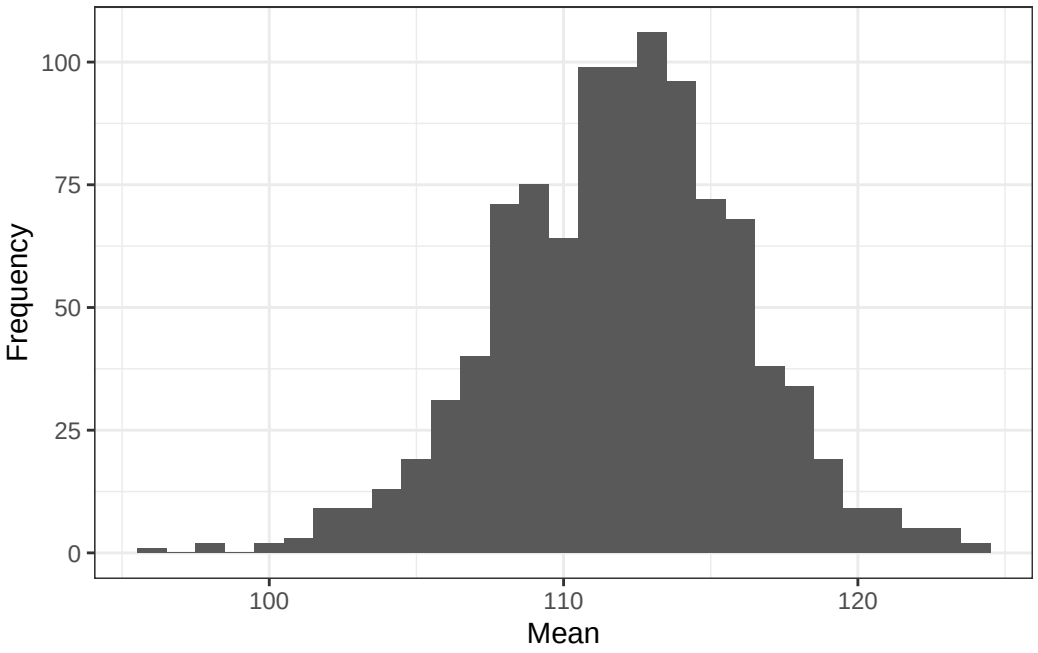


Figure 2: Histogram Mean - Gibbs sampling

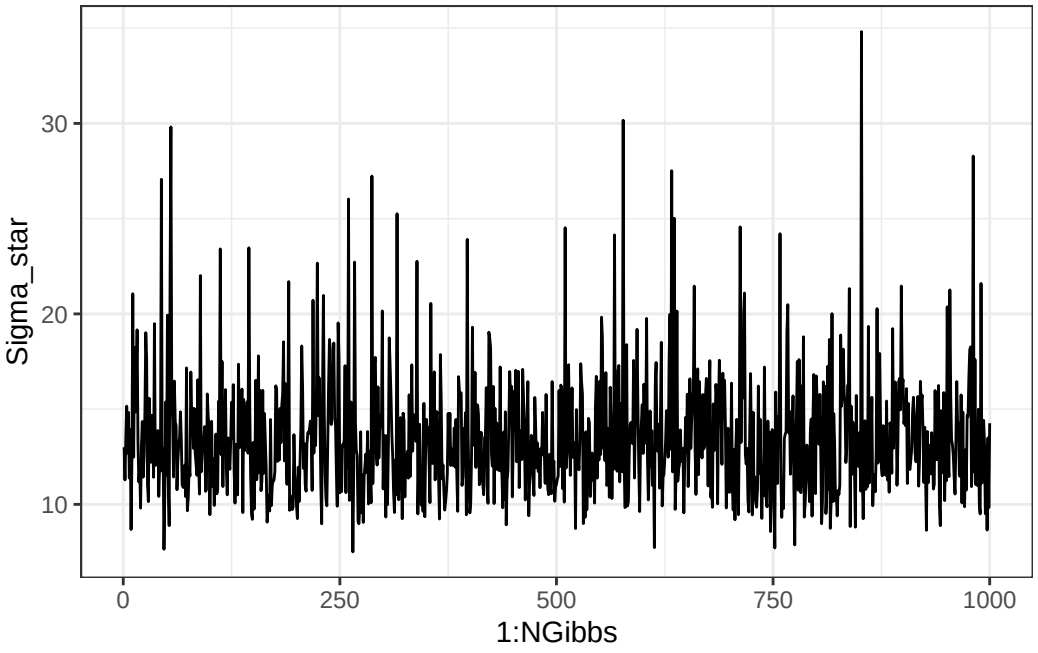


Figure 3: Sigma - Gibbs sampling

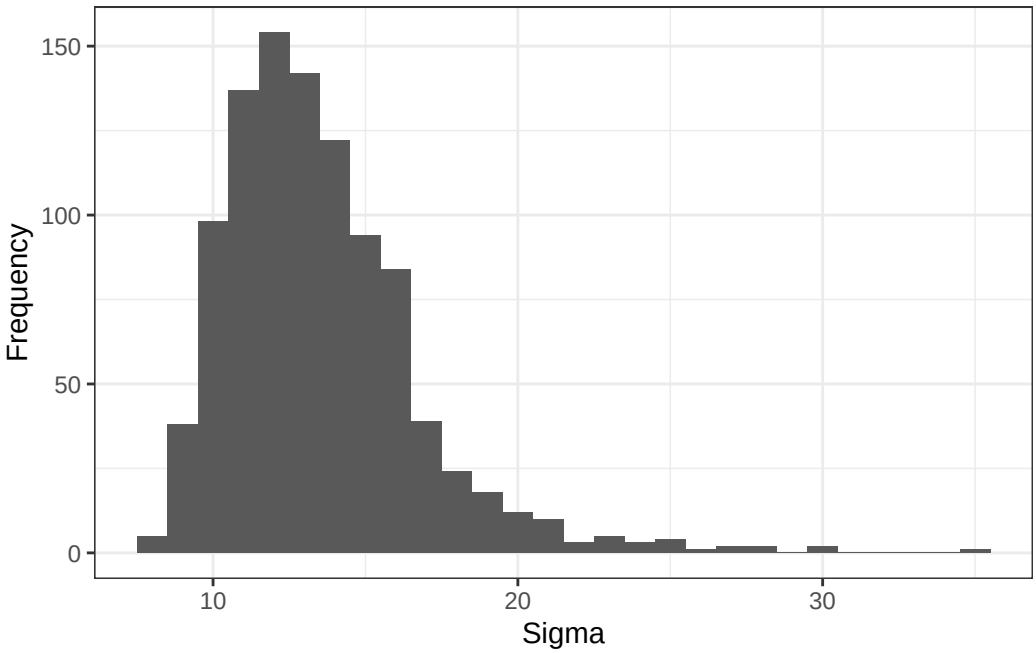


Figure 4: Histogram Sigma - Gibbs sampling

Estimates of both plots seems to converge to their true values of μ and σ and histograms represent quite well the sample distributions, as expected. The 95% C.I for both μ and σ are given below, besides the sample statistics, $\hat{\mu}$ and $\hat{\sigma}$:

Table 2: 95% C.I. for Mu and Sigma - Gibbs Sampling

Sample	Lower	Upper
111.95888	103.893509	120.02424
13.47727	7.205962	19.74857

Question 3 - Wells

This question uses data looking at the decision of households in Bangladesh to switch drinking water wells in response to their well being marked as unsafe or not. A full description from the Gelman Hill text book (page 87):

"*Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure. Any locality can include wells with a range of arsenic levels. The bad news is that even if your neighbor's well is safe, it does not mean that yours is safe. However, the corresponding good news is that, if your well has a high arsenic level, you can probably find a safe well nearby to get your water from—if you are willing to walk the distance and your neighbor is willing to share. [In an area of Bangladesh, a research team] measured all the wells and labeled them with their arsenic level as well as a characterization as "safe" (below 0.5 in units of hundreds of micro grams per liter, the Bangladesh standard for arsenic in drinking water) or "unsafe" (above 0.5). People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. A few years later, the researchers returned to find out who had switched wells.*"

The outcome of interest is whether or not household i switched wells:

The data we are using for this question are here: http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat and you can load them in directly using `d <- read.table(url("the_url_above"))`

The variables of interest for this questions are:

- switch, which is y_i above
- arsenic, the level of arsenic of the respondent's well
- dist, the distance (in meters) of the closest known safe well

(a) Do an exploratory data analysis illustrating the relationship between well-switching, distance and arsenic. Think about different ways of effectively illustrating the relationships given the binary outcome. As usual, a good EDA includes well-thought-out descriptions and analysis of any graphs and tables provided, well-labelled axes, titles etc.

{*Answer.*}

The database contains data for households considered *unsafe*, i.e., the `arsenic` level is greater than 0.5 units of hundreds of micro grams per liter. First we will investigate some relations between variables to identify particular behaviors, possible correlations that can provide useful insights when studying the target variable, i.e., `switch` which measures if household switched to well a nearest safe well, given the unsafe condition of current well providing water to him/her and his/her families.

Table 3: General Statistics

switch	arsenic	dist	assoc	educ
Min. :0.0000	Min. :0.510	Min. : 0.387	Min. :0.0000	Min. : 0.000
1st Qu.:0.0000	1st Qu.:0.820	1st Qu.: 21.117	1st Qu.:0.0000	1st Qu.: 0.000
Median :1.0000	Median :1.300	Median : 36.761	Median :0.0000	Median : 5.000
Mean :0.5752	Mean :1.657	Mean : 48.332	Mean :0.4228	Mean : 4.828
3rd Qu.:1.0000	3rd Qu.:2.200	3rd Qu.: 64.041	3rd Qu.:1.0000	3rd Qu.: 8.000
Max. :1.0000	Max. :9.650	Max. :339.531	Max. :1.0000	Max. :17.000

Now we will analyse general aspects for all variables.

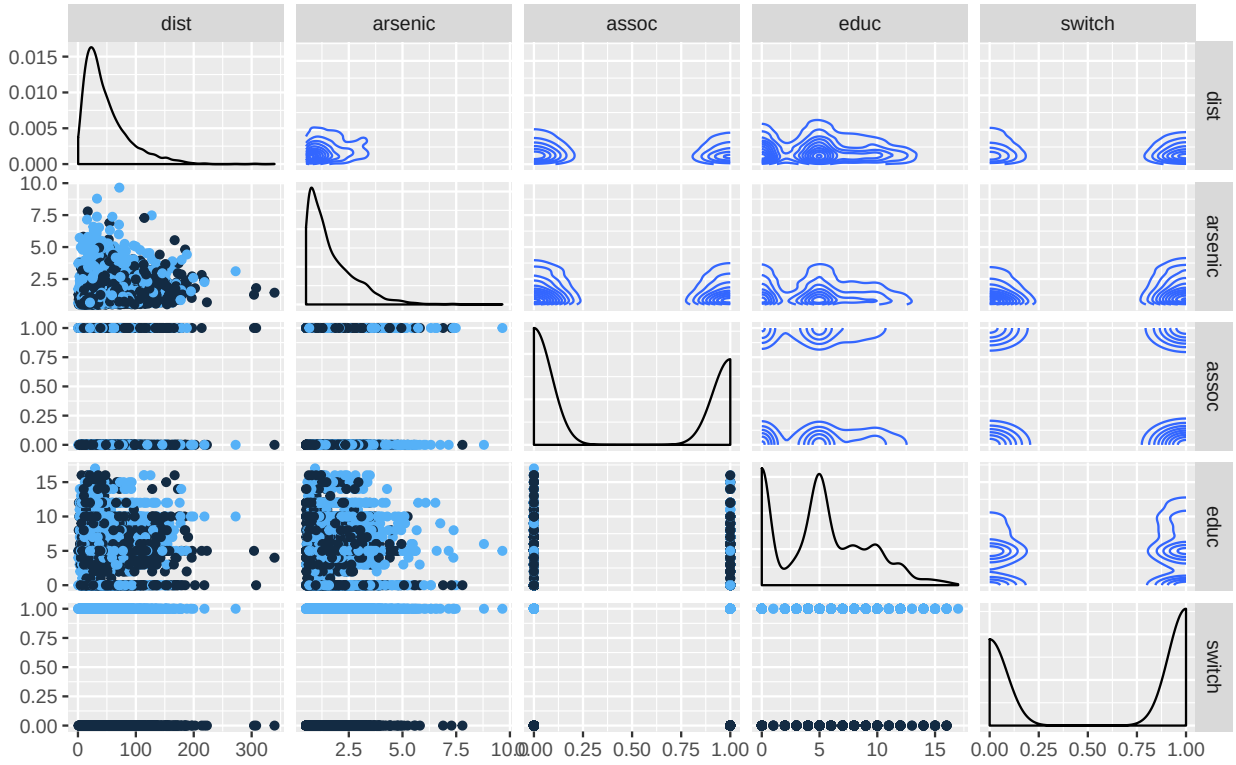


Figure 5: Cross-Correlation between variables

From the pairs of variables, we can see the sample distribution of `dist` and `arsenic` seems to be more correlated. An interesting aspect is that `educ` has a mixed distribution, bi-modal on `zero` and around `fifth` which can represent different behaviors or perception of risk in continuous use of a contaminated well. It can be further investigated in future studies/model adjustments.

The heat-maps below will provide an additional view of correlation between covariates.

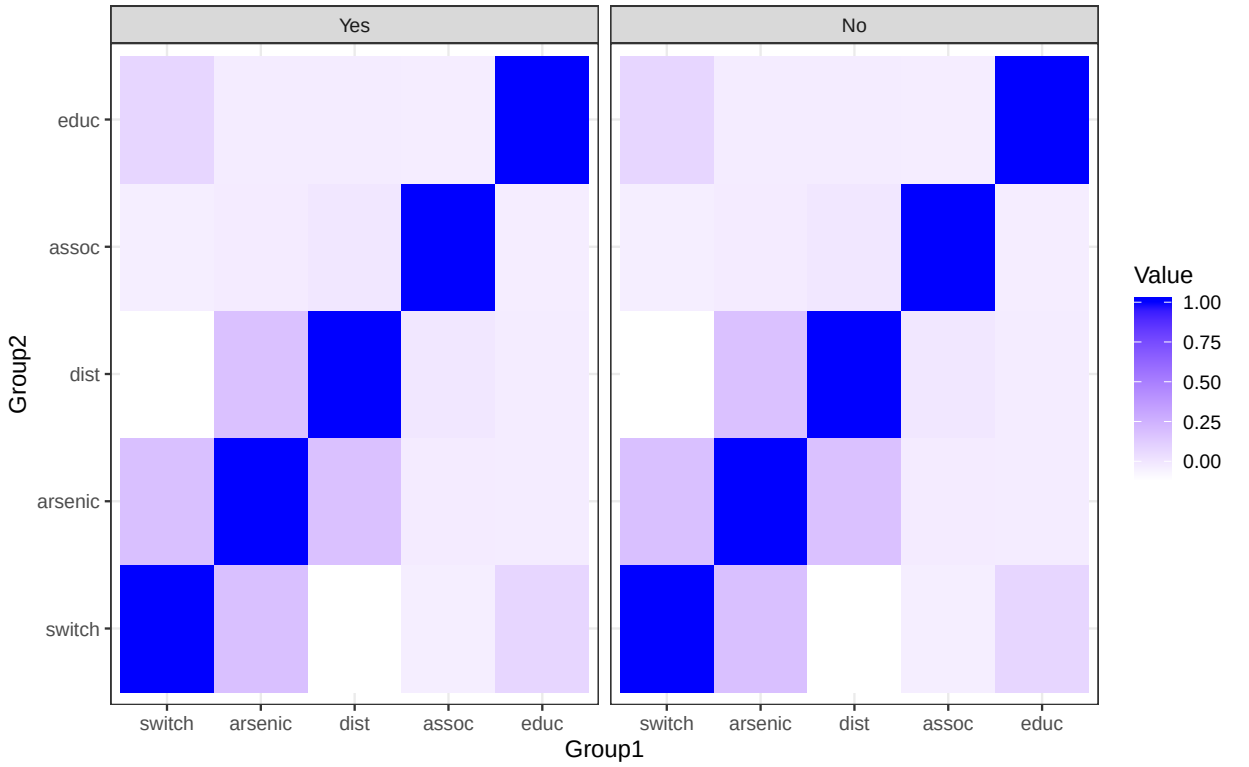


Figure 6: Correlation of variables depending on Switch = Yes/No

As expected, the decision to switch is highly correlated with variables `arsenic` and `dist`, with no relevant association with others. A special mention to `educ` which apparently the third greater correlation with the decision of switch/no-switch well can be considered in further studies. This makes sense because, depending of the level of education, the perception of risk in continuing using unsafe well may influence the decision to switch to another safe well.

More attention will be now spent on variables `arsenic` and `dist`, as they appear to be more correlated with the variable of interest and can explain what influences the decision to switch well.

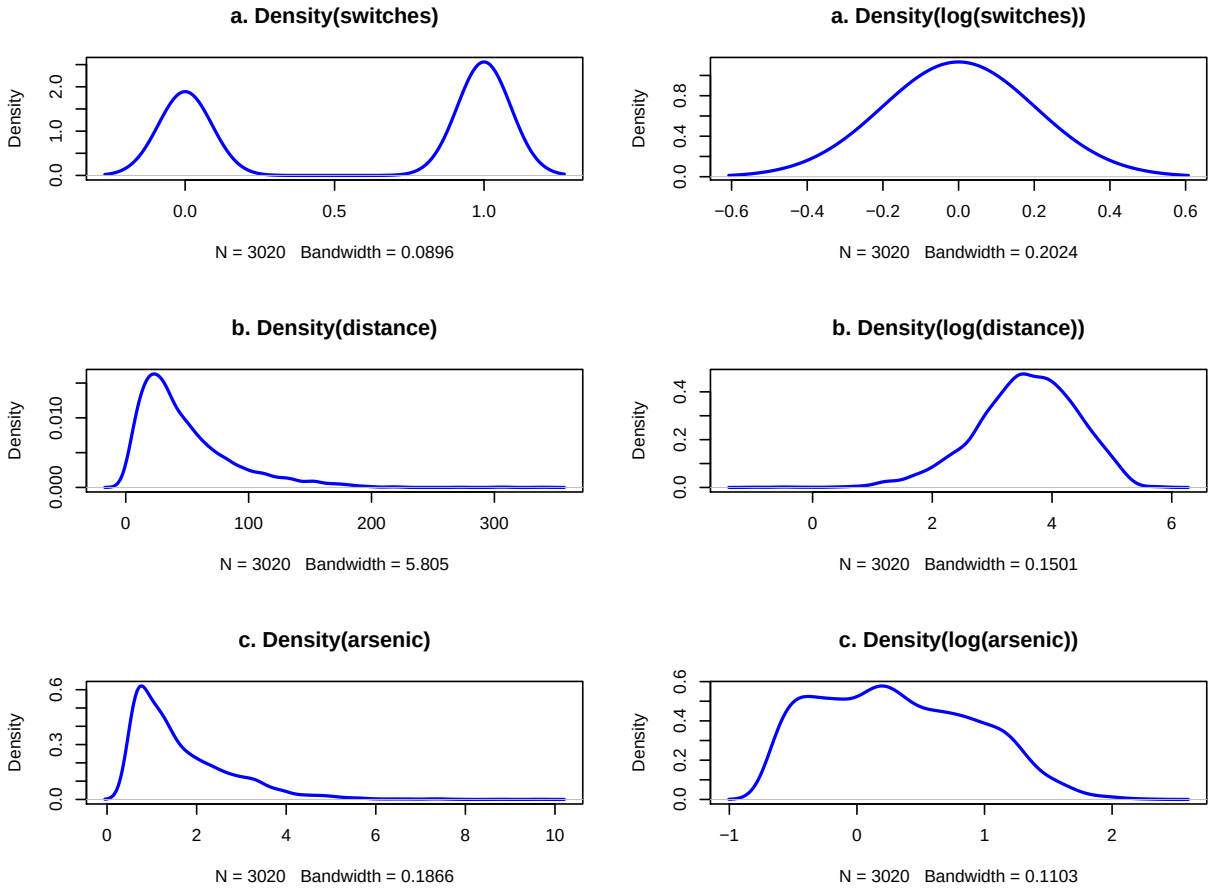


Figure 7: Densities of Switch, Arsenic and Distance

From the graphs above we can verify the distribution of the variable of interest is normal for each condition of switch/no-switch choice, i.e., "yes" or "no" for switched well. As it has a binary output, the distribution of variable `switch` is either of the closely related logistic or `probit` regression models may be used to model it, using *Bernoulli* distribution with `logit` link function map linear predictions in $]-\infty,+\infty[$ into a probability values in the interval $[0,1]$.

Variables `arsenic` and `dist` have both similarities with with Poisson Process or Gamma distributions then `log` transformation might help normalize these covariates.

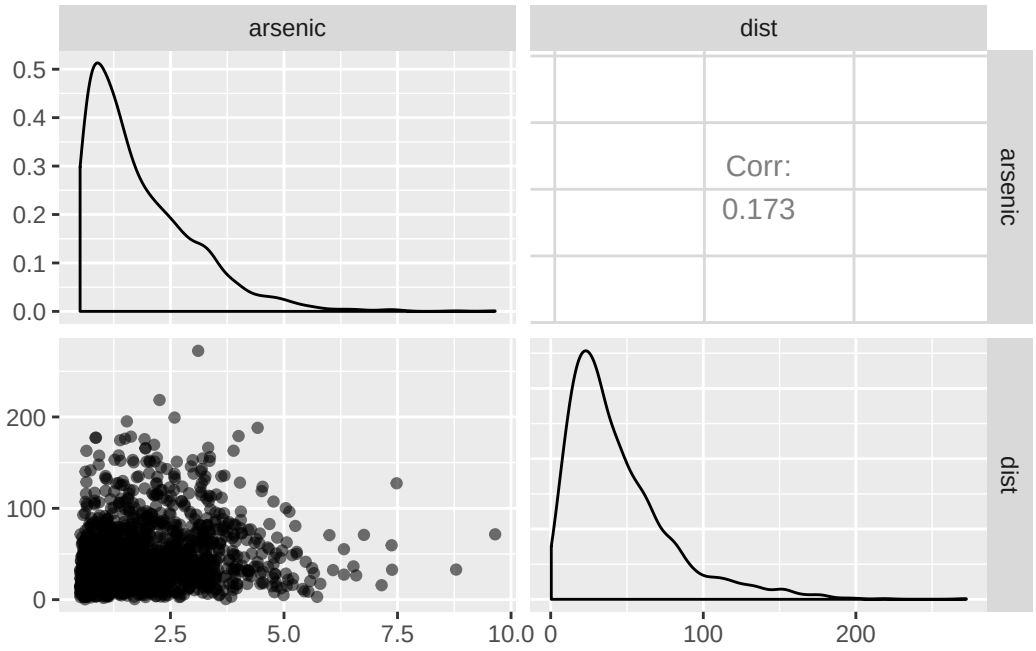


Figure 8: Scatterplots of variables & switch = Yes

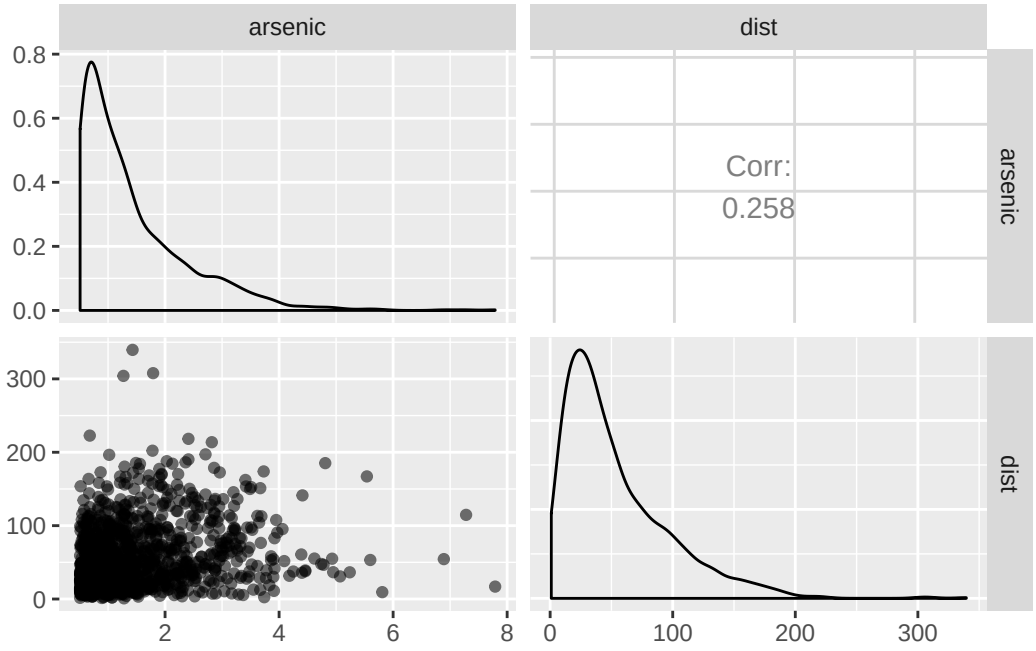


Figure 9: Scatterplots of variables & switch = No

Now, analyzing both groups (switches/no-switched), we can observe some association between `arsenic` and `dist` have greater intensity on group "no-switched". As well, the amplitude of nominal measurements of both variables are greater on group "no-switched", suggesting that higher distances and higher concentration of arsenic might be linked with the choice of no switch well.

In the next questions we will adjust 02 (two) STAN models to analyse in deep these relationships.

Assume $y_i \sim \ Bern(p_i)$, where p_i refers to the probability of switching. Consider two candidate models.

- Model 1:

$$logit(p_i) = \beta_0 + \beta_1 \times (d_i - \overline{d}) + \beta_2 \times (a_i - \overline{a}) + \beta_3 \times (d_i - \overline{d})(a_i - \overline{a})$$

- Model 2:

$$logit(p_i) = \beta_0 + \beta_1 \times (d_i - \overline{d}) + \beta_2 \times (log(a_i) - \overline{log(a)}) + \beta_3 \times (d_i - \overline{d})(log(a_i) - \overline{log(a)})$$

where d_i is `distance` and a_i is `arsenic` level.

(b) Fit both of these models using Stan. Put N(0, 1) priors on all the βs. You should generate point-wise log likelihood estimates (to be used in later questions), and also samples from the posterior predictive distribution (unless you'd prefer to do it in R later on). For model 1, interpret each coefficient.

{*Answer.*}

Interpretation of coefficients for Model 1

The coefficients for Model 1 are as follows:

Table 4: Coefficients for Model 1

beta0	0.3521339
beta1	-0.0087493
beta2	0.4693929
beta3	-0.0017790

Some aspects we can identify from Model 1:

```
##
## Intercept: (propension to switch)  0.587135
```

```
##
## Cofficient for distance (probability to switch due to distance above avg level):  0.372198
```

```
##
##  -> var probability of switch of each 100m in distance, controlling for arsenic:  -0.2149369
```

```
##
## Coefficient arsenic - probability of switch for each 1 unit in arsenic level:  0.6945603
```

```
##
##  -> var probability of switch for each 1 arsenic, controlling for distance:  0.1074254
```

```
##
## Coefficient interaction - variation on probability of switch for each 100m in distance and
##          arsenic in 1 unit above the average level, due just for interaction:  -0.04422185
```

```
##
## Coefficient interaction - probability of switch for each 1 unit in arsenic and
##          distance above in 100m of average level, due just for interaction:  0.4424298
```

We can summarize the main conclusions as follows:

- *Constant term*: $logit^{-1}(0.3521) = 0.5871$ is the estimated probability of switching, i.e., the householders have the natural propensity to switch well, if the distance to the nearest safe well and the arsenic level in the current well are both near the average in the whole data (which means both measurements $d_i - \overline{d}$ and $a_i - \overline{a}$ equals to zero) is about 58,7%. Depending of the other coefficients/variables, the total probability may increase our decrease;
- *Coefficient for distance*: as we are working with centered data, each 100 meters of this variable represents a difference of 100 meters from the current well to the average nearest wells - we may understand this well is " farther" than the others in relation of the mean distance. This is the coefficient for distance (on the `logit` scale), if arsenic level is at its average distance of nearest safe well and will represent a decrease in probability of switching (because its sign is negative) to a level of $logit^{-1}(0.3521 - 0.00875 * 100) = 0.3721$. In other words, this coefficient leads to the conclusion that each 100 meters of distance corresponds to an approximate 21.49% negative difference in probability of switching;
- *Coefficient for arsenic*: in the same way, now comparing two wells with difference of 1 in arsenic level, if the distance to the nearest safe well is at its average value for both, we have that the probability of switching is $logit^{-1}(0.3521 + 0.46939 * 1) = 0.6945$, i.e., representing that an increase of each additional unit of arsenic corresponds to an approximate 10.7% positive increase in probability of switching.
- *Coefficient to interaction Distance-Arsenic*: the interaction between distance and arsenic, in the light of the coefficient which is -0.001779, means that an increase of 1 above the average level of in arsenic will add -0.001779 to the coefficient for distance which leads to an increase of importance of distance for households with higher arsenic levels. On the other way, when comparing the probability of switch, the effect of interaction is negative, i.e., for each 100m added to distance from the average distance to nearest well and 1 unit to arsenic from average level decreases the probability of switch in -0.04422 when compared with the effect without the interaction which may lead to the conclusion of wells with higher distances from the average level and higher levels of arsenic reduces the probability of switch. The overall probability of switch for this case (w/ interaction) is 14.5% lower than if there was no interaction.

(c) Let $t(\mathbf{y}) = \sum_{i=1}^{n} 1(y_i = 1, a_i < 0.82) / \sum_{i=1}^{n} 1(a_i < 0.82)$ i.e. the proportion of households that switch with arsenic level less than 0.82. Calculate $t(\mathbf{y}^{rep})$ for each replicated data-set for each model, plot the resulting histogram for each model and compare to the observed value of $t(\mathbf{y})$. Calculate $P(t(\mathbf{y}^{rep}) < t(\mathbf{y}))$ for each model. Interpret your findings.

{*Answer.*}

```
##
## Level of Arsenic:  0.82
```

```
##
## Observed Probability of switch well, given arsenic <  0.82  :  0.4338
```

For each model we calculated the $t(\mathbf{y}^{rep})$ obtaining the following results:

```
##
## Probability of t(y^rep) is less than t(y^obs) in Model 1:  0.005
```

```
##
## Probability of t(y^rep) is less than t(y^obs) in Model 2:  0.28
```

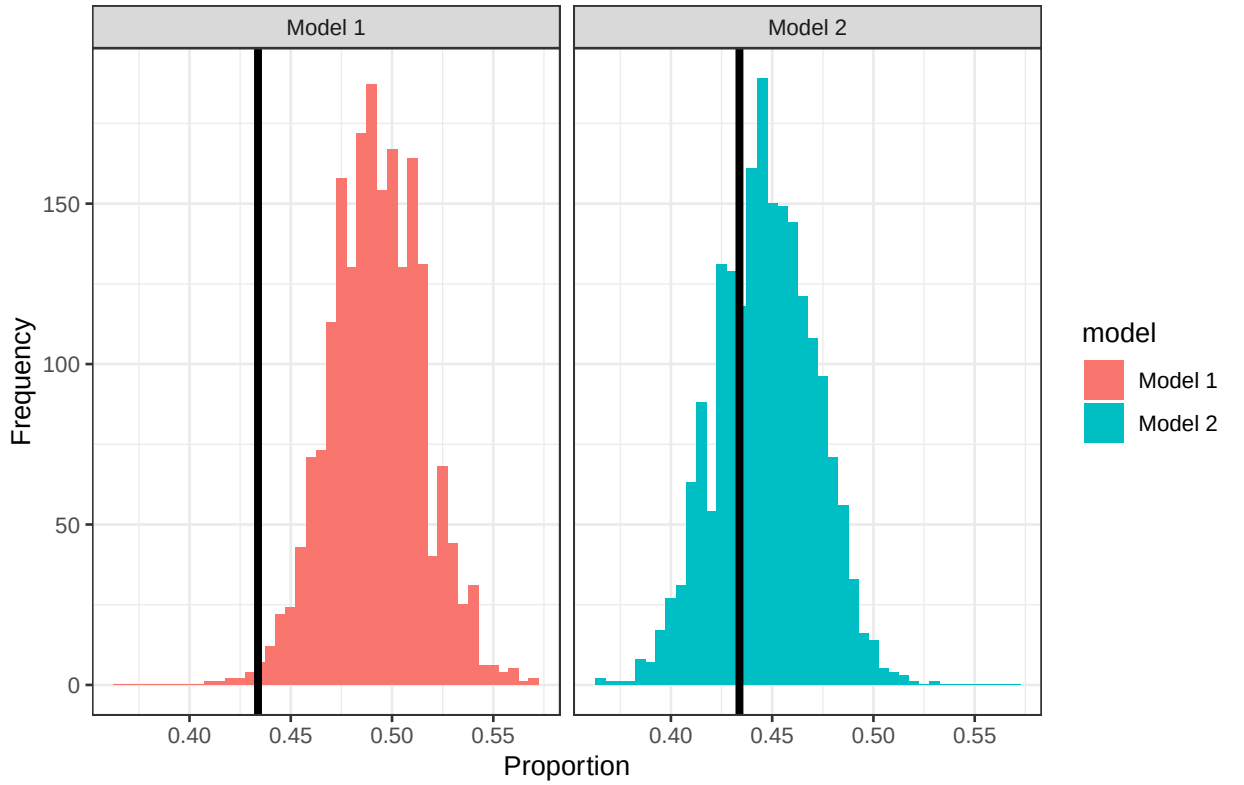


Figure 10: Model 1 vs. Model 2 - Probability of Switch = Yes, given Arsenic < 0.82

When comparing both histograms, this scenario points out that the number of households who decided to switch to the nearest safe well, having arsenic levels less than the limit of 0.82 is *greater in Model 2 than in Model 1.* This can be summarized by the frequency of times the probability of a household switch well, given its `arsenic` level is less than 0.82 (i.e., $P_1(Y_i = 1|a_i < 0.82)$) is less than the observed probability. In our case, Model 2 presented a higher frequency than Model 1, as it could be seen on histograms above.

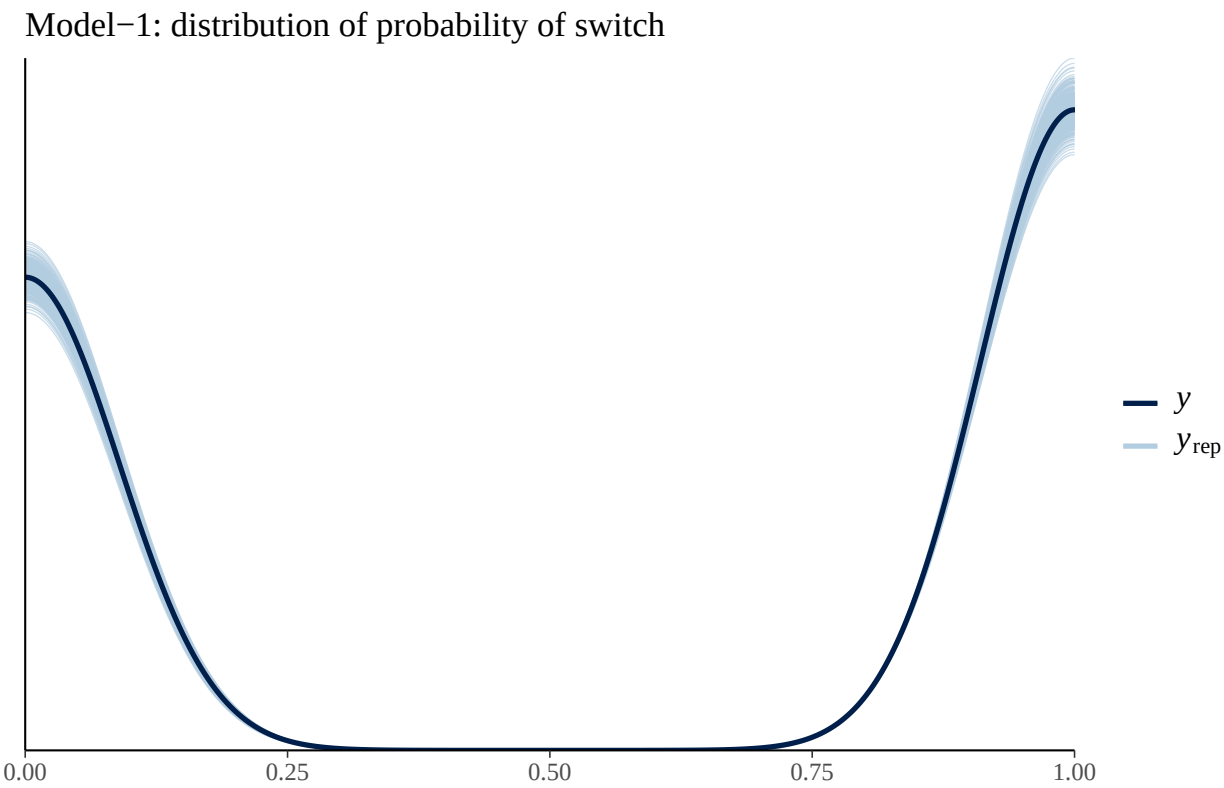
In this sense and based on each model, the simulated probability of switch, given `arsenic` is less than 0.82 is less than observed same probability is:

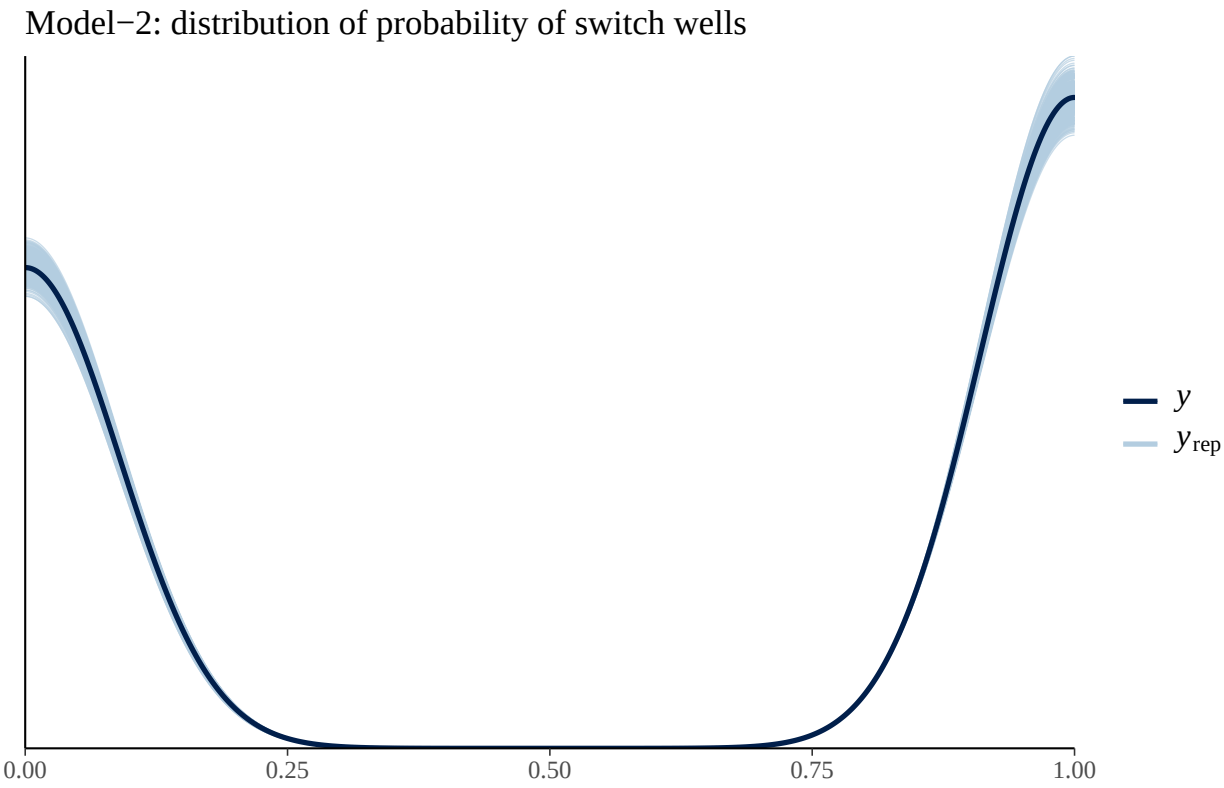
- $P_1(t(\mathbf{y}^{rep}) < t(\mathbf{y}^{obs})) = 0.005$ (Model 1)
- $P_2(t(\mathbf{y}^{rep}) < t(\mathbf{y}^{obs})) = 0.28$ (Model 2)

(d) Use the `loo` package to get estimates of the expected `log`-point-wise predictive density for each point, $ELPD_i$. Based on $\sum_i ELPD_i$, which model is preferred?

{*Answer.*}

Let's first verify if the densities, by sampling 20% of simulations generated and plot it against the observed distribution of y_i.





The graphs seems to be quite good, adherent to the sample distribution of data.

Now, by simulating LOO process to calculate the ELPD for both models we have the following results:

```
## Warning: Relative effective sample sizes ('r_eff' argument) not specified.
## For models fit with MCMC, the reported PSIS effective sample sizes and
## MCSE estimates will be over-optimistic.

##
## Computed from 2000 by 3020 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1968.0 15.9
## p_loo         4.3  0.3
## looic      3935.9 31.7
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.

## Warning: Relative effective sample sizes ('r_eff' argument) not specified.
## For models fit with MCMC, the reported PSIS effective sample sizes and
## MCSE estimates will be over-optimistic.

##
## Computed from 2000 by 3020 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1952.4 16.3
## p_loo         4.0  0.1
## looic      3904.8 32.7
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.

##        elpd_diff se_diff
## model1   0.0       0.0
## model2 -15.6       4.4
```

In our case, **Model 2** has the lowest LOO-IC, so we can consider it the preferred model.

(e) Create a scatter plot of the ELPDi's for Model 2 versus the ELPDi's for Model 1. Create another scatter plot of the difference in ELPDi's between the models versus `log(arsenic)`. In both cases, color the dots based on the value of y_i. Interpret both plots.

{*Answer.*}

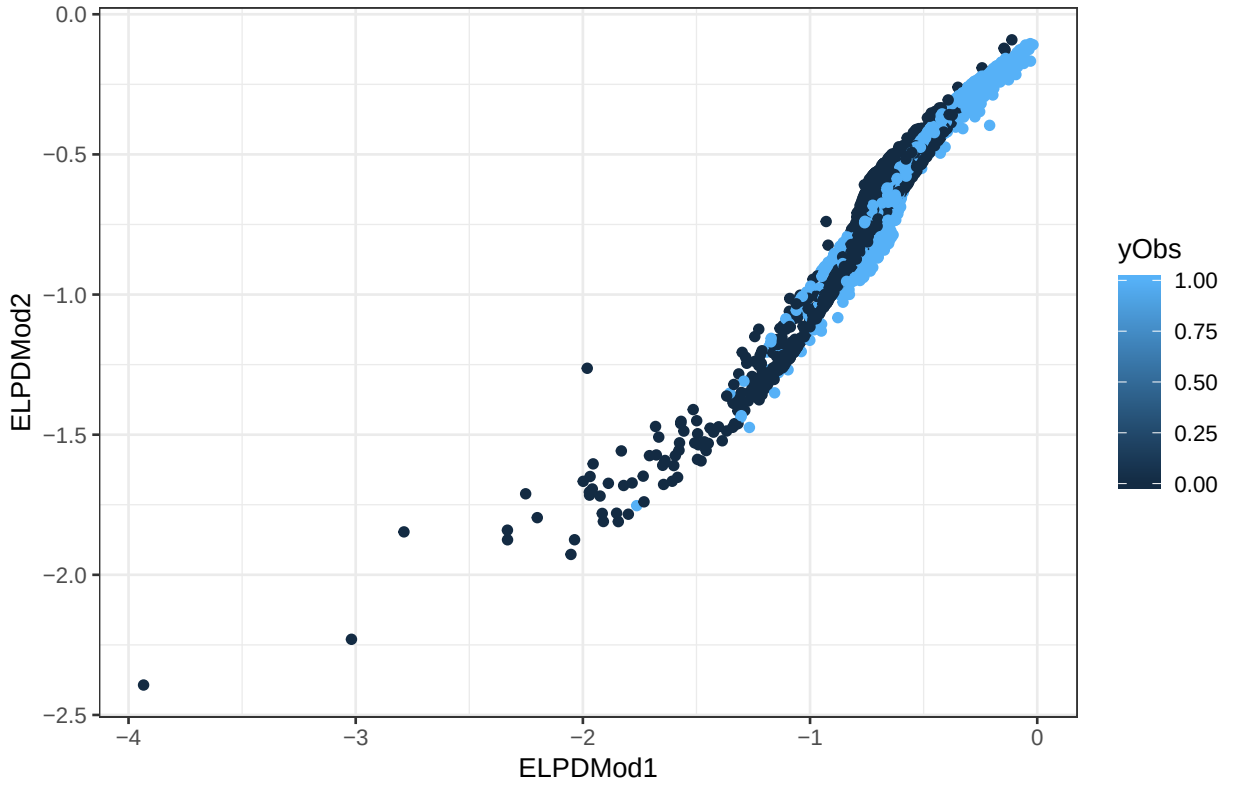


Figure 11: Model 1 vs. Model 2 - Scatterplots for ELPDs

The scatterplot between the ELPD's shows a slight S-shape with concentration of blue-dots in, which represents households who switched wells, in the upper-right of graph and in the opposite, black-dots in the low-left side, representing those households who didn't changed wells. We can see that ELPD for Model 2 are smaller than ELPD's from model 1, representing that Model 2 is the preferred model when compared with Model 1.

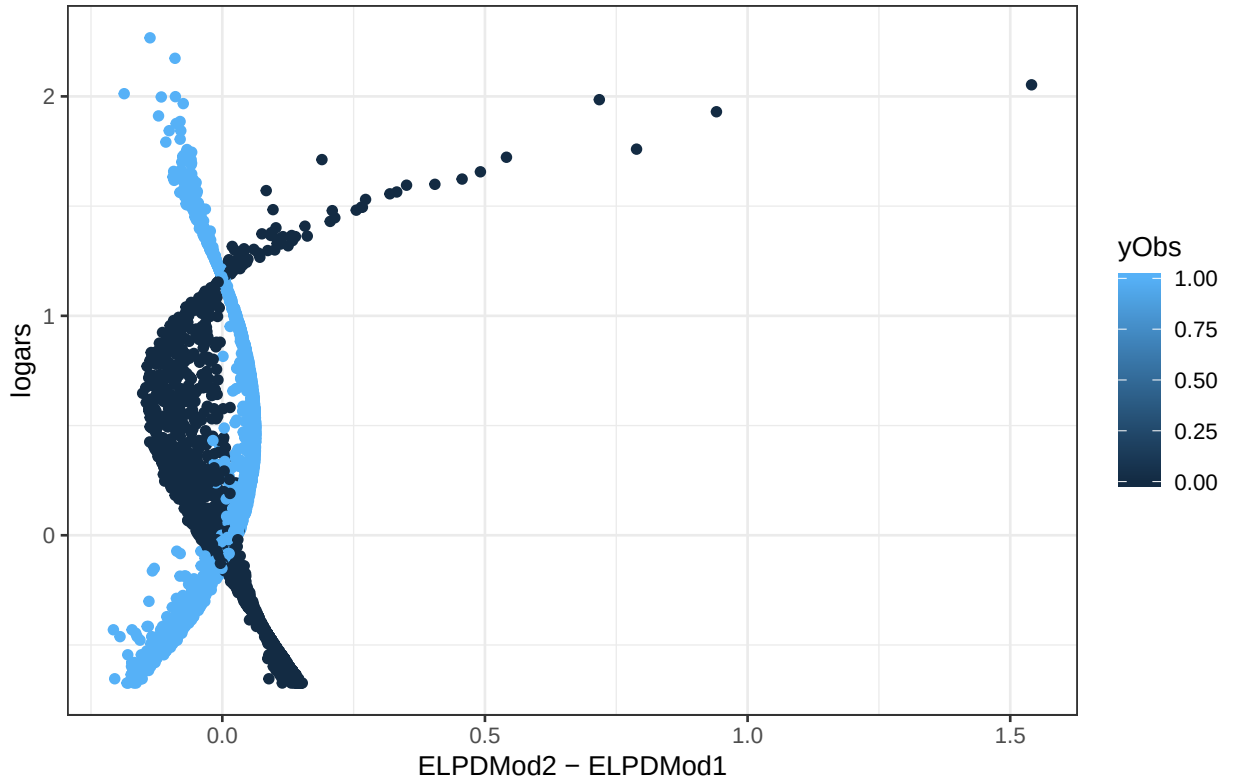


Figure 12: Difference ELPD for Model 1 vs. Model 2

This graph shows some interesting aspects:

- the points to the right to the `zero` in x-axis shows ELPDs from Model 2 greater than in Model 1, and vice-versa. It shows the majority of concentration of points in left side, which is where Model 2 is preferred against Model 1;
- in the y-axis, it shows the concentration of points above `zero` which represents those households with difference of level of arsenic greater than 1.0 unit, and below those with less concentration of arsenic. We can also see that most of the points are located above the 0-level which are the portion of households with greater levels of arsenic on their wells which means households who "should more" change wells than the rest;
- looking jointly at both axis, if we divide the graph in 4 quadrants both in relation to `zero`, we see that that in the upper left quadrant, Model 2 is better to predict points above `arsenic>=2.7` because it shows better the discrimination between people who switched than those who doesn't representing "good classification"; in upper right quadrant, shows the reverse: poor classification for those households with higher `arsenic` levels; the other quadrants shows the reverse, but in both cases, it reinforces Model 2 as preferred model than Model 1.

(f) Given the outcome in this case is discrete, we can directly interpret the $ELPD_i$'s. In particular, what is $exp(ELPD_i)$?

{*Answer.*}

From the definition of Expected Log Point-wise Predictive Density (ELPD), we have:

$$ELPD = \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i}) \tag{22}$$

When applying $\exp ELPD$ we have:

$$\begin{aligned}\exp(ELPD) &= \exp\Big(\sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i})\Big) \\ &= \prod_{i=1}^{n} p(y_i|\mathbf{y}_{-i}) \\ &= p(\mathbf{y}).\end{aligned}$$

In this sense, $\exp ELPD_i$ is the **marginal probability of y**.

Applying $\exp(ELPD)$ and plotting it for each model, against the observed `switch` we can see the division around 0.5 in probability representing the S-curve in action for each case - if the expected probability of switch is greater than 0.5, it is more likely people switch to a nearest well than for those whose probability is less than 0.5. This can be seen n next graph.

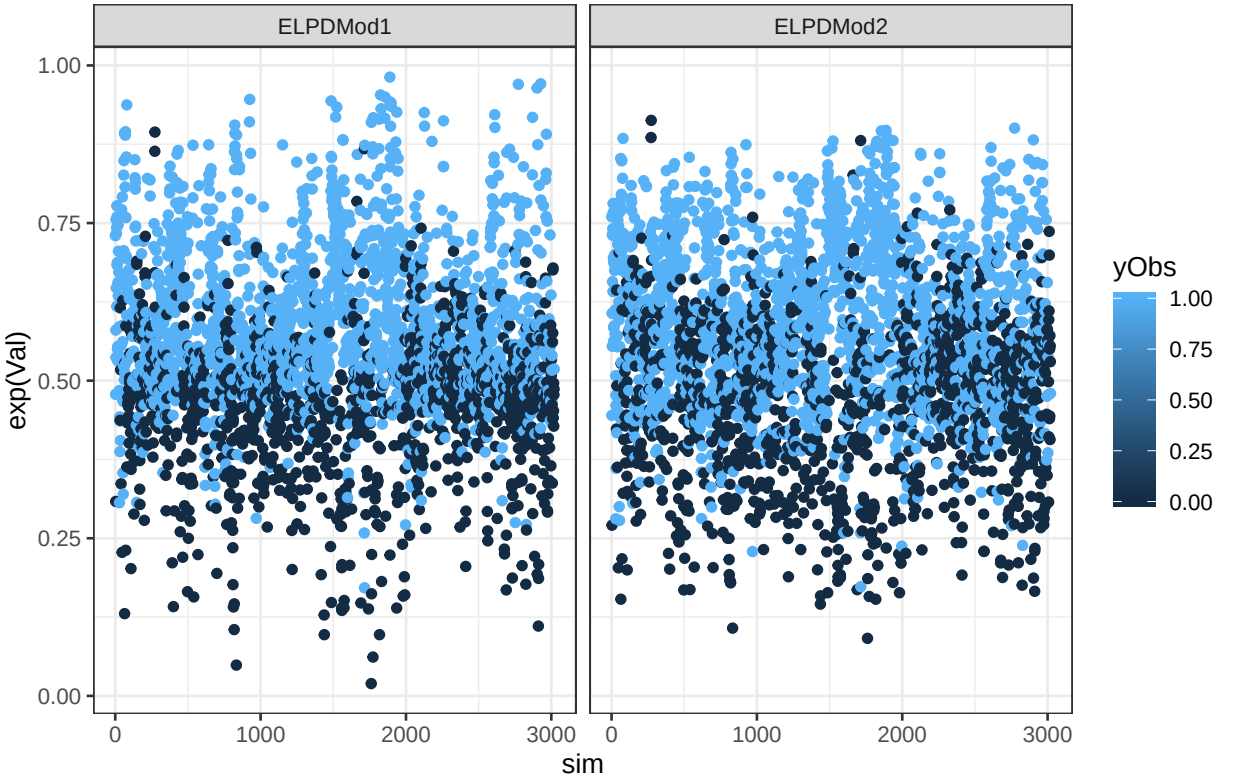


Figure 13: Calculation of exp(ELPDs) for each model vs. Switch

From the two graphs, we see that Model 2 have less dispersion around $p = 0.5$ possibly indicating it is slightly better than Model 1 when predicting the probability of switch.

(g) For each model recode the $ELPD_i$'s to get $\hat{y}_i = E(Y_i|\mathbf{y}_{-i})$. Create a binned residual plot, looking at the average residual $y_i - \hat{y}_i$ by `arsenic` for Model 1 and by `log(arsenic)` for Model 2. Split the data

such that there are 40 bins. On your plots, the average residual should be shown with a dot for each bin. In addition, add in a line to represent +/- 2 standard errors for each bin. Interpret the plots for both models.

{*Answer.*}

For this question we will use as reference the book *Data Analysis Using Regression and Multilevel/Hierarchical Models* from Gelman.

In logistic models the residuals can be defined as observed $y_i $minus expected values $\hat{y}_i$ as follows :

$$\begin{aligned} Resid(y_i) &= y_i - \hat{y}_i \\ &= y_i - [y_i \times \exp ELPD_i + (1 - y_i) \times \exp ELPD_i] \end{aligned}$$

As the data are discrete and so are the residuals, the binned residuals plot is useful to verify visually the model adjustment. The bins are created by dividing the data into categories (bins) based on their fitted values and then plotting the average residual versus the average fitted value for each bin.

As requested by the question, we will plot additionally 02 dotted lines representing the +/- 2 standard errors for each bin.

For comparison, as Gelman did on page#98, we will plot the built-in function `binnedplot` from package `arm` but, in order avoid conflict with packages `dplyr` and `tidyr`, we will call only the function and not loading the entire library.

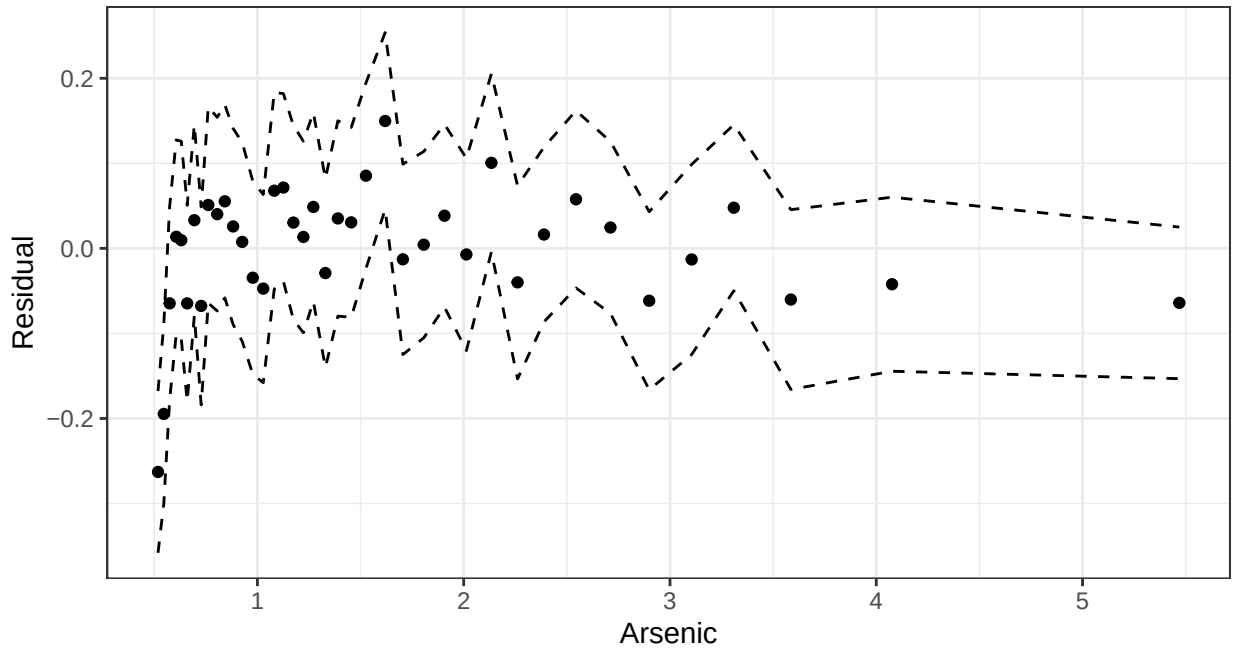
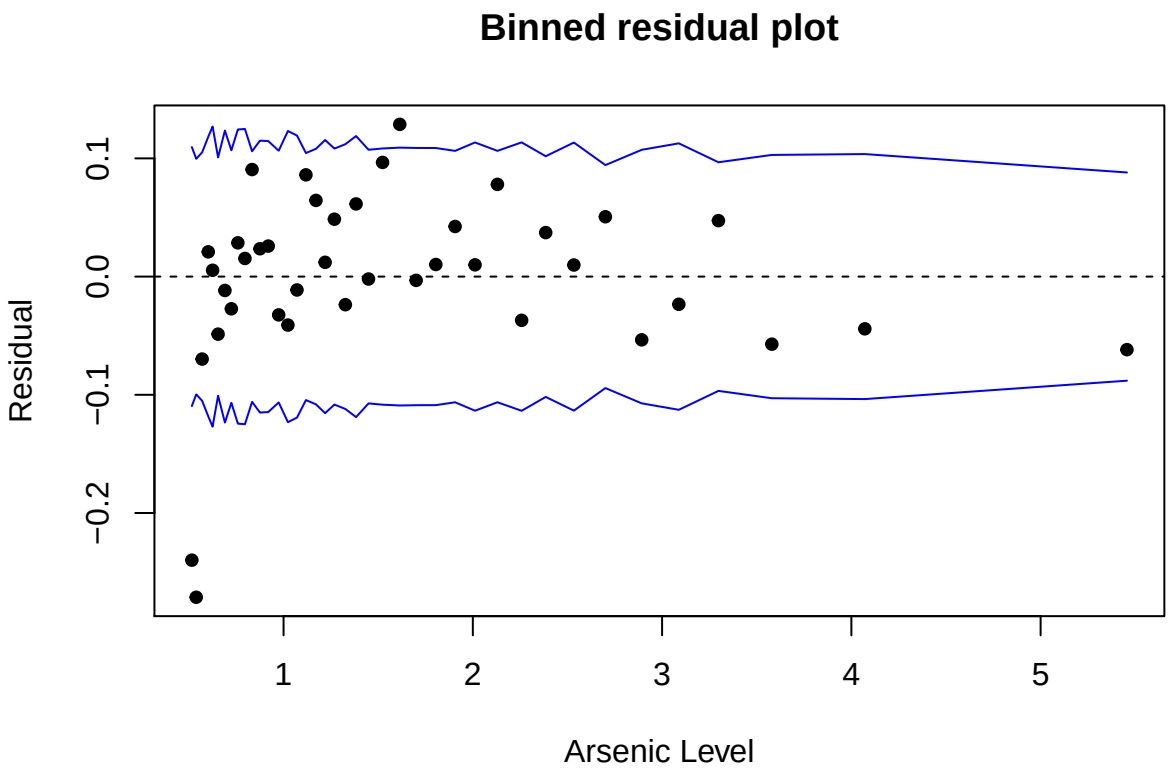


Figure 14: Binned residual plot for Models 1



For Model 2 we have the following result:

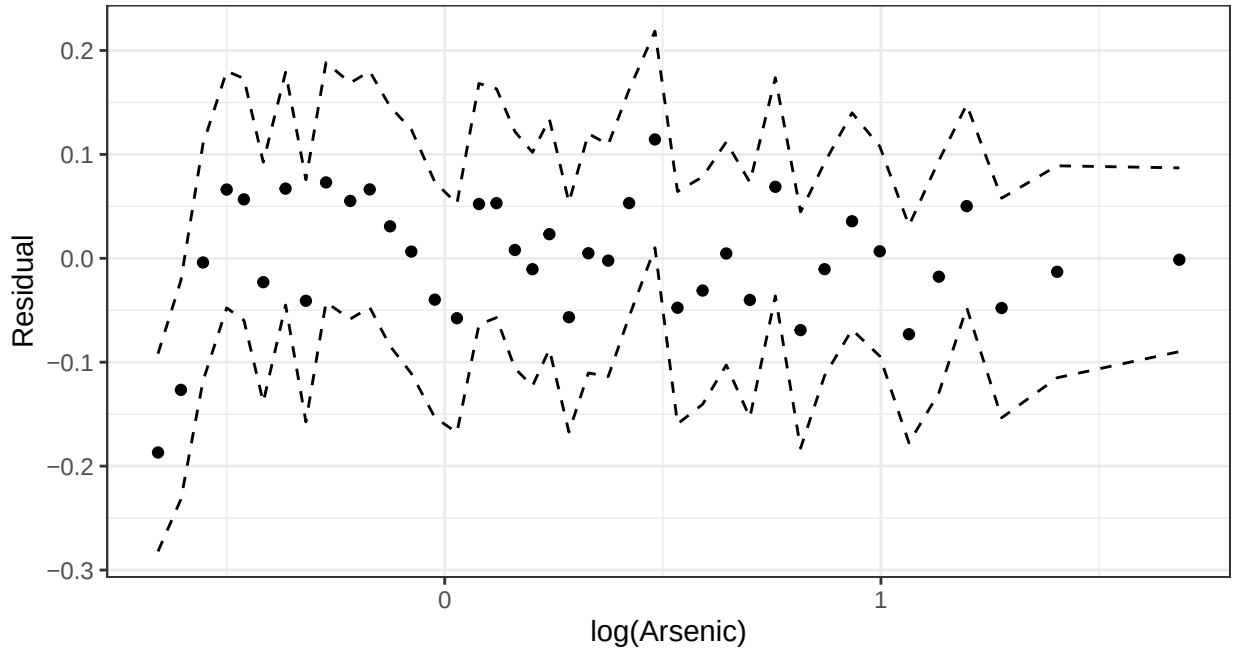
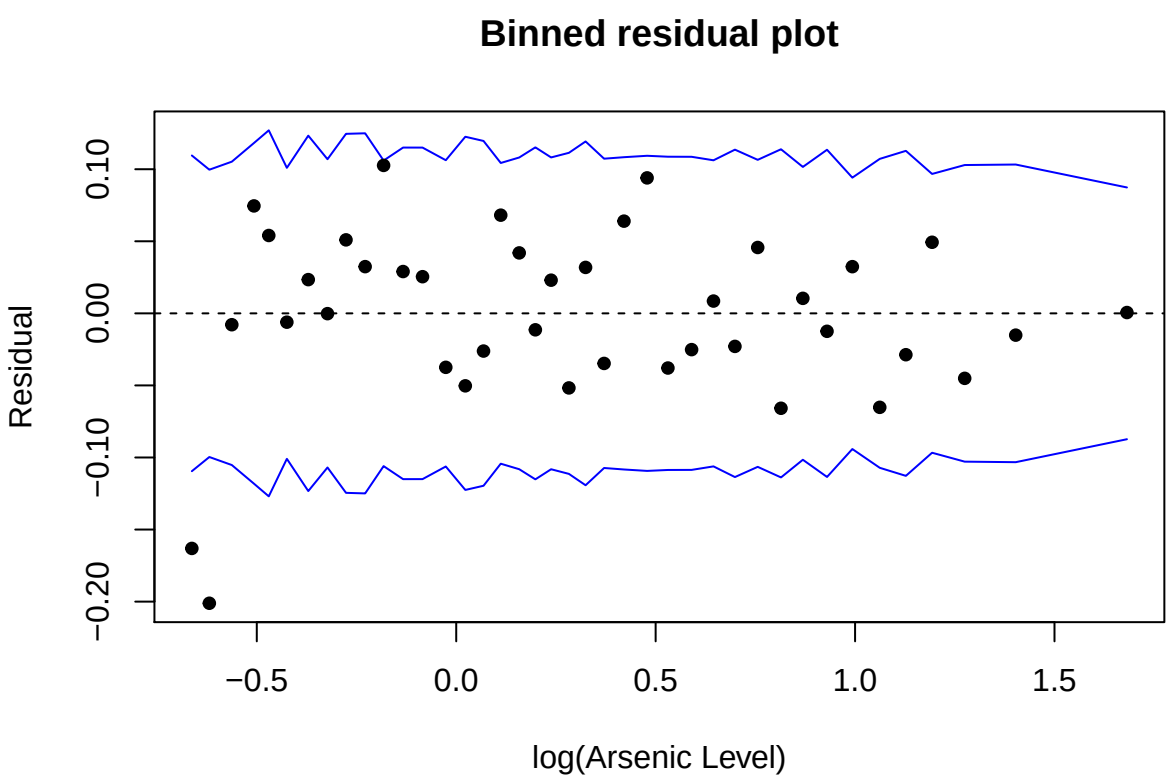


Figure 15: Binned residual plot for Models 2



Comparing the binned plots we can see that Model 2 has better fit:

- the residuals seems to be more uniformly distributed in Model 2 than in Model 1, which presents a high concentration of bins up to level 1.0 of `arsenic` concentration;
- the amplitude of residuals is smaller in Model 2 when compared to Model 1, which suggests it is a best predictor for y;
- both built-in `binnedplot` shows similar the distribution of points on graphs for Models 1 and 2 and included the theoretical 95% error bounds indicated by the *blue* line.
- In all graphs we can see that at lower levels of arsenic, the residuals shows *two* points with highly negative residuals which would indicate those individuals are in discordance in $\sim 20\%$ with the probability of switch than indicated by the model.

In general these plots confirms what we have seen on previous questions, which indicates Model 2 as the preferred model.