STA2201H Winter 2023 Assignment 2

Due: 11:59pm ET, March 17

What to hand in: .Rmd or .qmd file and the compiled pdf, and any stan files

How to hand in: Submit files via Quercus

1. IQ

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

$$Y_1, Y_2, \ldots, Y_n | \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$$

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

$$\mu \sim N\left(\mu_0 = 100, \sigma^2_{\mu_0} = 15^2\right)$$

a) Write down the posterior distribution of μ based on the information above. Give the Bayesian point estimate and a 95% credible interval of μ, $\hat{\mu}_{Bayes} = E(\mu|\boldsymbol{y})$.

Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_{10})$.

The likelihood function is

$$f(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{10} (2\pi\sigma^2)^{-1/2} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^5} e^{-\frac{\sum_{i=1}^{10}(y_i-\mu)^2}{2\sigma^2}}$$

The prior function is

$$p(\mu) = \frac{1}{15\sqrt{2\pi}} e^{\frac{-(\mu-100)^2}{2(15^2)}}$$

The posterior distribution of μ is

$$p(\mu|\mathbf{y},\sigma^2) = f(\mathbf{y}|\mu,\sigma^2)p(\mu) \propto e^{-\frac{1}{2}\left[\left(\frac{10}{\sigma^2}+\frac{1}{15^2}\right)\left[\mu-\left(\frac{10\overline{y}}{\sigma^2}+\frac{100}{15}\right)\right]^2\right]} \text{ (Lecture 4 page 13)}$$

i.e. $\mu|\mathbf{y},\sigma^2 \sim N\left(\frac{\frac{100}{15^2}+10\frac{\overline{y}}{\sigma^2}}{\frac{1}{15^2}+\frac{10}{\sigma^2}}, \frac{1}{\frac{1}{15^2}+\frac{10}{\sigma^2}}\right) \equiv N\left(\frac{\frac{100}{15^2}+10\frac{113}{15^2}}{\frac{1}{15^2}+\frac{10}{15^2}}, \frac{1}{\frac{1}{15^2}+\frac{10}{15^2}}\right) \equiv N\left(\frac{1230}{11}, \frac{225}{11}\right)$

The Bayesian point estimate is $\hat{\mu}_{Bayes} = E(\mu|\boldsymbol{y}) = \frac{1230}{11} \approx 111.81$

A 95% credible interval of μ is $E(\mu|\mathbf{y}) \pm 1.96\sqrt{Var(\mu|\mathbf{y})} = \frac{1230}{11} \pm 1.96\sqrt{\frac{225}{11}} \approx (102.954, 120.683)$

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) $\text{MSE}\left[\hat{\mu}|\mu^*\right] = E\left[\left(\hat{\mu}-\mu^*\right)^2|\mu^*\right]$. Show the MSE is equal to the variance of the estimator plus the bias of the estimator squared, i.e.

$$\text{MSE}\left[\hat{\mu}|\mu^*\right] = \text{Var}\left[\hat{\mu}|\mu^*\right] + \text{Bias}\left(\hat{\mu}|\mu^*\right)^2$$

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

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

Let $\mu^* = 112$.

Bias:

$Bias(\hat{\mu}_{Bayes} \mid \mu^* = 112) = E(\hat{\mu}_{Bayes} \mid \mu^*) - 112 = \frac{1230}{11} - 112 = \frac{2}{11} \approx 0.182$

$Bias(\hat{\mu}_{MLE} \mid \mu^* = 112) = E(\hat{\mu}_{MLE} \mid \mu^*) - 112 = 113 - 112 = 1$

Variance:

$Var(\hat{\mu}_{Bayes} \mid \mu^* = 112) = \frac{225}{11} \approx 20.455$

$Var(\hat{\mu}_{MLE} \mid \mu^* = 112) = \frac{15^2}{10} = \frac{45}{2} = 22.5$

MSE:

$MSE(\hat{\mu}_{Bayes} \mid \mu^* = 112) = Var(\hat{\mu}_{Bayes} \mid \mu^* = 112) + \big[Bias(\hat{\mu}_{Bayes} \mid \mu^* = 112)\big]^2 = \frac{2479}{121} \approx 20.488$

$MSE(\hat{\mu}_{MLE} \mid \mu^* = 112) = Var(\hat{\mu}_{MLE} \mid \mu^* = 112) + \big[(\hat{\mu}_{MLE} \mid \mu^* = 112)\big]^2 = 22.5 + 1^2 = 23.5$

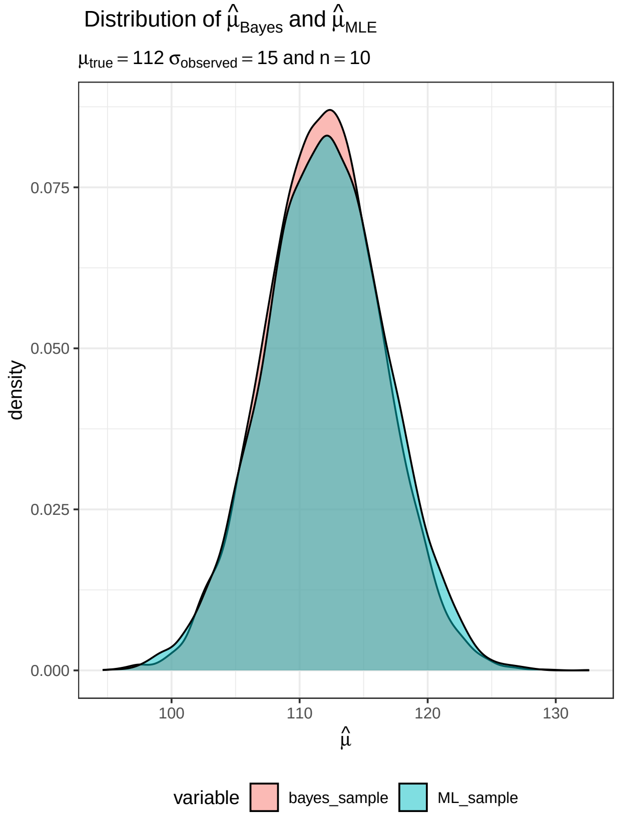
The ML estimator resulted in higher bias, variance, and MSE than those produced by the Bayes estimator.

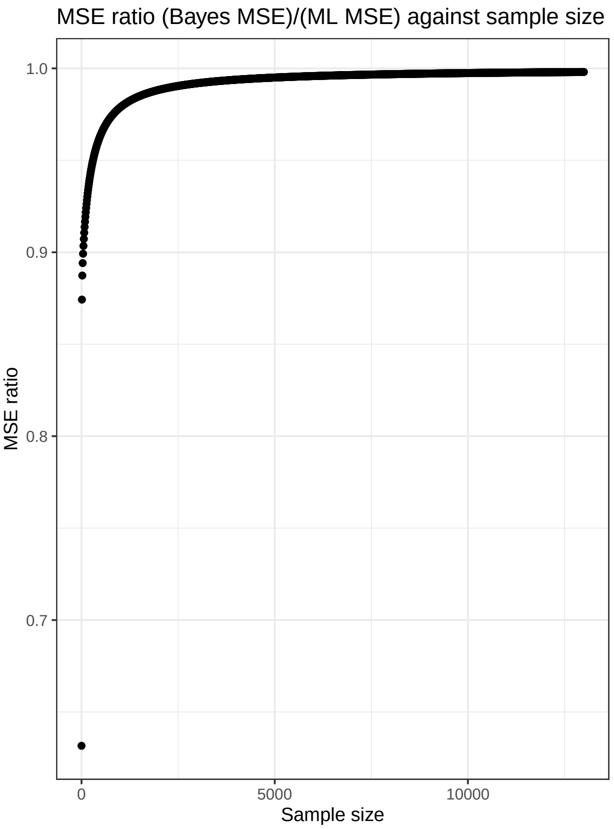
d) Write down the sampling distributions for the ML and Bayes estimates, again assuming $\mu^* = 112$ and $\sigma = 15$. Plot the two distributions on the one graph. Summarize your understanding of the differences in bias, variance and MSE of the two estimators by describing how these differences relate 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.

Sampling distributions for the ML and Bayes estimates (n = 10):

$$\hat{\mu}_{Bayes} \sim N\Big(\frac{1230}{11}, \frac{225}{11}\Big) \text{ and } \hat{\mu}_{MLE} \sim N\Big(112, \frac{15^2}{10}\Big) \equiv N(112, 22.5)$$

In the frequentist approach, we use a point estimate $\hat{\mu}_{MLE}$ to estimate the true value of μ^*. In the Bayesian approach, we get a distribution of the estimates of μ^*, say $\hat{\mu}_{Bayes}$, under a prior opinion, and estimate μ^* by the expected value of $\hat{\mu}_{Bayes}$. Bias is the difference between the expected values of either estimator and μ^*. Moreover, the ML estimator is unbiased (i.e. $\mu^* = E(\hat{\mu}_{MLE})$), and we see that the MSE ratio converges to 1 (i.e. $MSE_{Bayes} \to MSE_{MLE}$) as n increases, implying the Bayes estimator $\hat{\mu}_{Bayes}$ is asymptotically unbiased. For a small sample size, it is hard to choose between the MLE and Bayes estimate since $\hat{\mu}_{MLE}$ is unbiased but has a larger variance than that of $\hat{\mu}_{Bayes}$, so we need the mean square error (MSE) criteria, which takes into account concerns about both bias and variance of estimators, describing the difference between the estimate and the actual parameter. Here, the difference in the sampling distributions of the two estimators results from the difference in how their estimate diverges from the true parameter.





2. Gompertz

Gompertz hazards are of the form

$$\mu_x = \alpha e^{\beta x}$$

for $x \in [0, \infty)$ with $\alpha, \beta > 0$. It is named after Benjamin Gompertz, who suggested a similar form to capture a 'law of human mortality' in 1825.

This question uses data on deaths by age in Sweden over time. The data are in the `sweden` file in the class repo. I grabbed the data from the Human Mortality Database. We will assume that the deaths we observe in a particular age group are Poisson distributed with a rate equal to the mortality rate multiplied by the population, i.e.

$$D_x \sim \text{Poisson}(\mu_x P_x)$$

where x refers to age. In this question we will be estimating mortality rates using the Gompertz model as described above.

a) Describe, with the aid of a couple of graphs, some key observations of how mortality above age 50 in Sweden has changed over time.

Figure 1

Starting levels of mortality rates in Sweden from 1990 to 2020

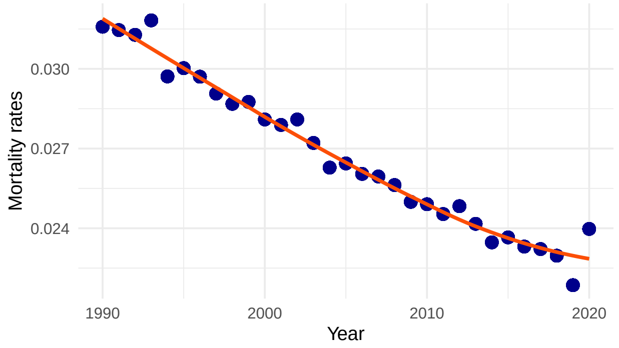


Figure 2

Mortality rates across ages from 1990 to 2020

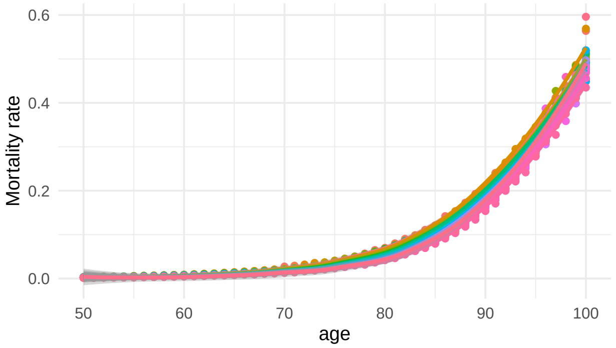


Figure 3

Observed death tolls in Sweden from 1990 to 2020

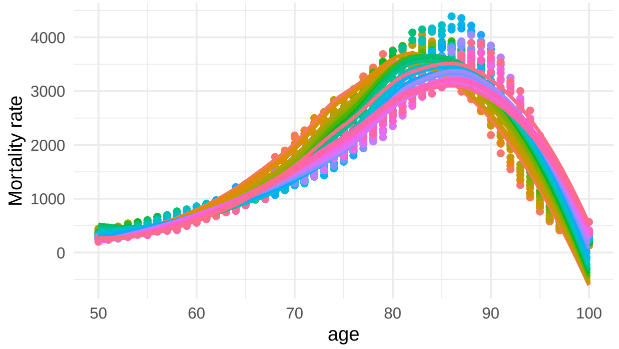
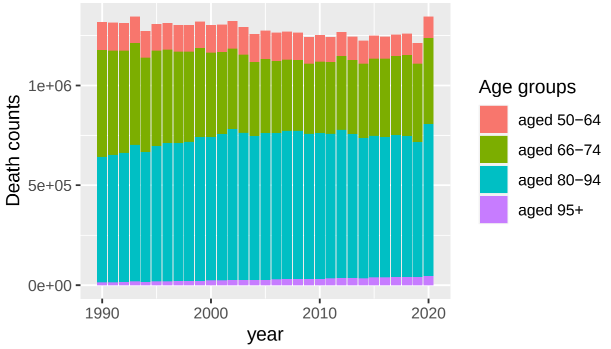


Figure 4

Observed death tolls by age in Sweden from 1990 to 2020



The starting levels of mortality rates in Sweden appear to have declined sharply from 1990 to 2020, except for 2020, when we observed the first spike in the mortality rate over the 31-year period, which we could attribute to the emergence of COVID-19. Also, mortality rates of older people were higher and varied in greater variation between years than those of their respective younger demographic groups. We also note that most Swedish residents passed away between 75 and 85 years old, as indicated by the steep climb in mortality rate in Figure 3, and people aged 80+ were the most vulnerable to the pandemic, as evidenced by the drastic shift in death tolls between 2019 and 2020 in Figure 4.

b) Carry out prior predictive checks for α and β, based on populations by age in Sweden in 2020. Summarize what you find and what you decide to be weakly informative priors for these parameters.

In the Gompertz model, parameter α characterizes how the starting level of mortality rate (i.e. mortality rate at age 50) changes over time while the parameter β captures the increase in mortality rate over age. Graphically, we see the starting level of mortality rate decreases over time (Figure 1), so we place a Half-normal(0.5, 0.002^2) prior on parameter α, i.e. we believe the mortality rate of Swedish residents aged 50 decreased by about 50% over the period between 1990 and 2020. Also, as seen in Figure 2, we believe that the rate of mortality increases over age by 3% for every unit increment in age and put Half-normal(0.03, 0.0055^2) prior on the parameter β. Prior predictive checks justified that the two weakly informative priors on α and β induce relatively similar behavior in simulated data to collected data (i.e. the spike in death counts among populations aged between 75 and 85, followed by a drop in death counts among the 85+ residents.)

Figure 5

Simulated and Observed death tolls (purple)

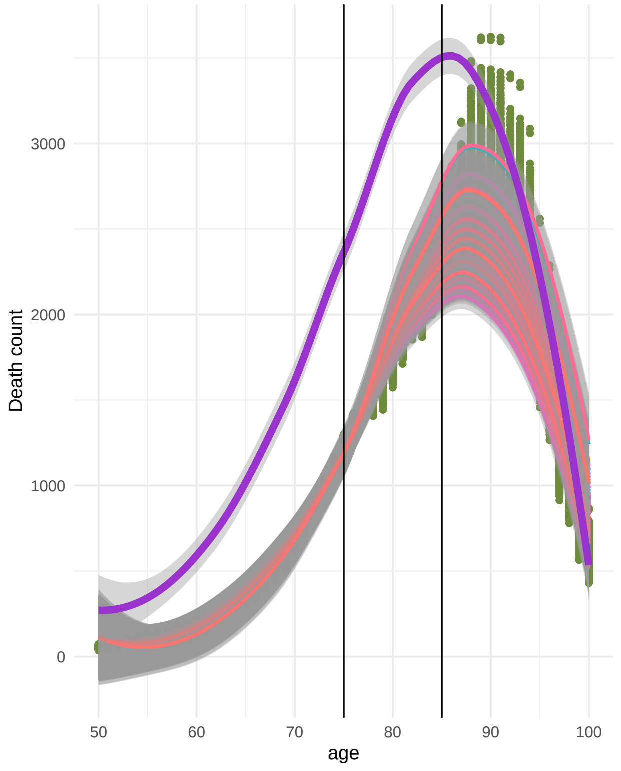
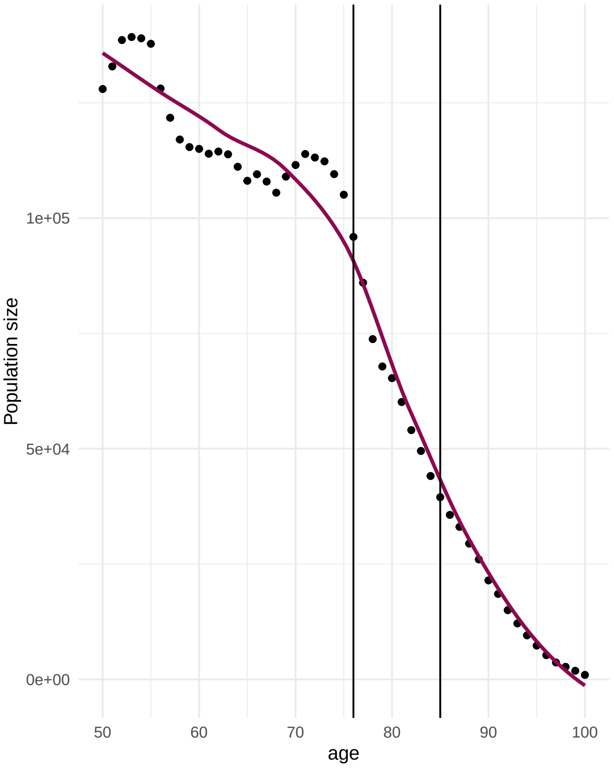


Figure 6

Population sizes across age



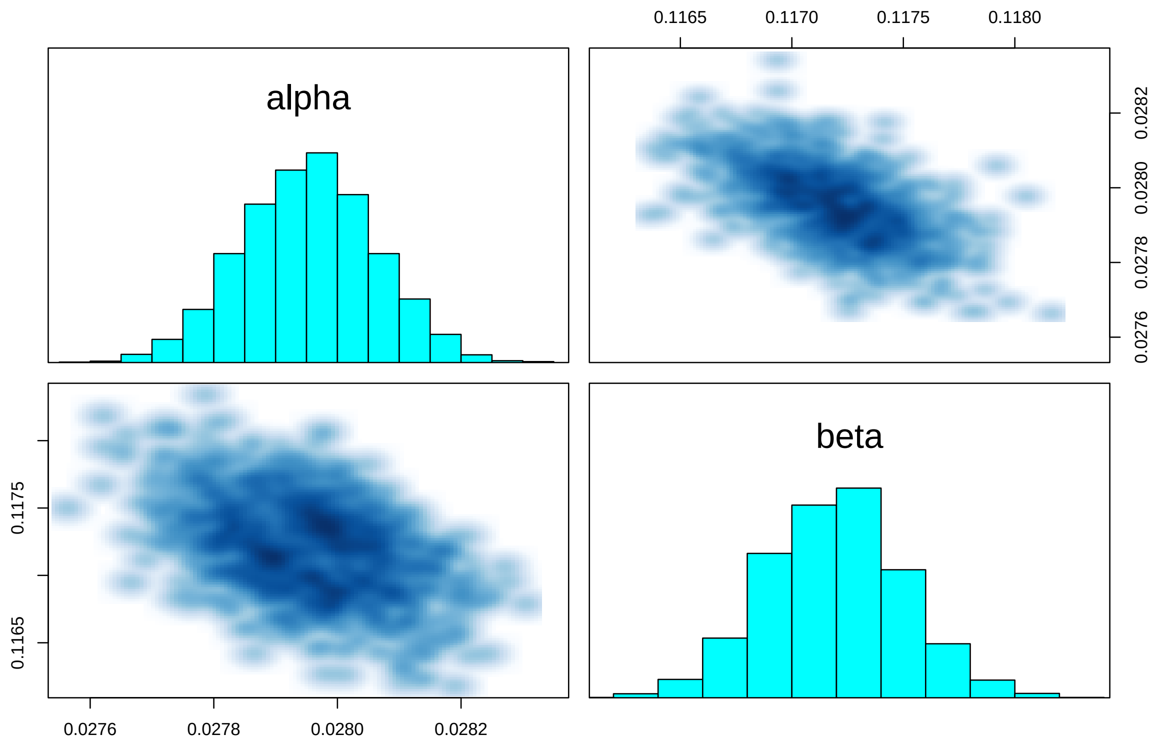
c) Fit a model in Stan to estimate α and β for the year 2020. Note that it may be easier to specify the likelihood on the log scale (you can do this in Stan using the `poisson_log` function). Priors should be informed by your prior predictive checks and any other information available. Ensure that the model has converged and other diagnostics are good. Interpret your estimates for α and β.

Likelihood: Y $\sim$ Poisson(η) where $\eta = \alpha P_x e^{\beta x} = e^{\beta x + log(\alpha P_x)}$ and P_x is population aged x

Priors: $\alpha \sim N^+(0.5, 0.002^2)$ and $\beta \sim N^+(0.03, 0.0055^2)$

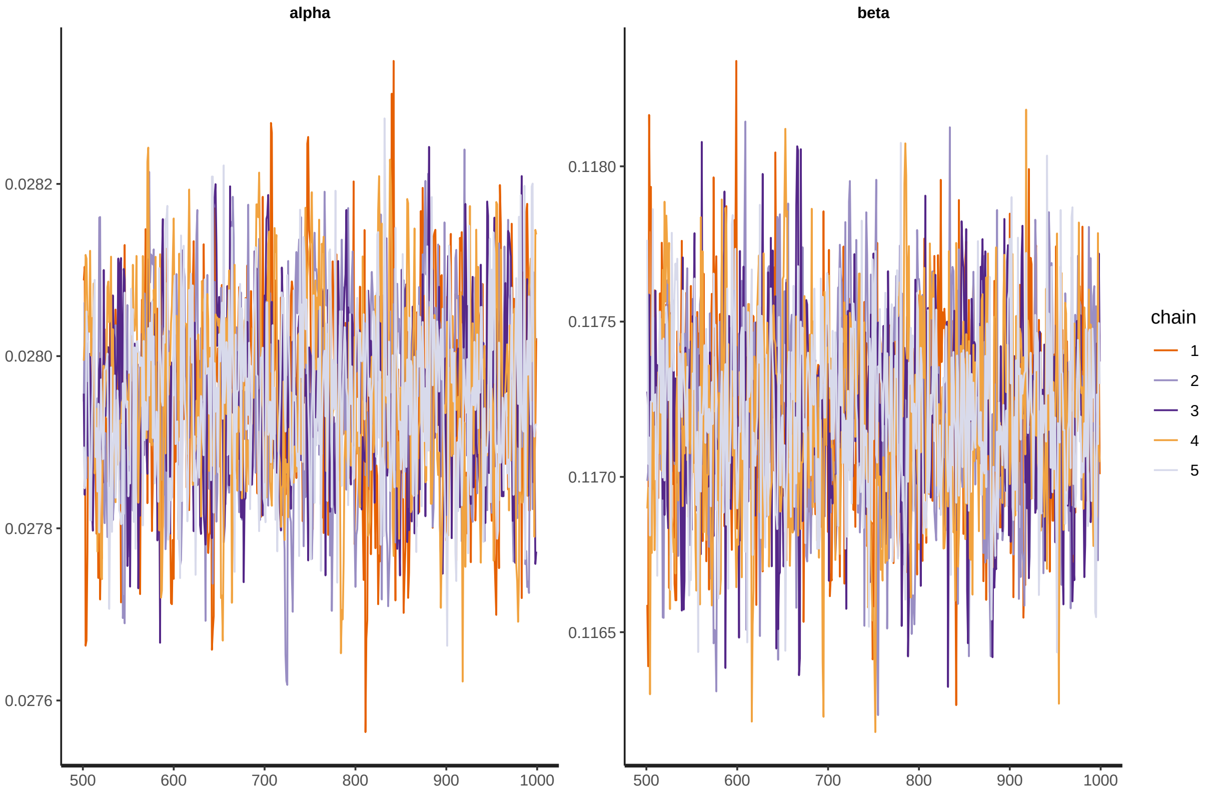
```
##             mean      se_mean           sd       2.5%        25%       50%
## alpha 0.02795425 3.321918e-06 0.0001080160 0.02774274 0.02787942 0.0279548
## beta  0.11719250 8.754139e-06 0.0003105289 0.11658774 0.11697685 0.1171972
##              75%      97.5%    n_eff     Rhat
## alpha 0.02802842 0.02816609 1057.301 1.001929
## beta  0.11739573 0.11781304 1258.280 1.001221
```

Since $\hat{R} < 1.05$ and effective sample size $n_{eff} > 100$ for the estimates of α and β, the model converged. Here, we found that centering the `age` covariate reduced correlation between chains and improved the efficiency of MCMC sampling (i.e. the model converged relatively fast). Model convergence can also be shown using the `pairs` plot.



The Gompertz model with Half-normal priors on α and β estimates that the starting level of mortality rate decreased by about 2.79% from 2019 to 2020 (i.e. the proportion of Swedish aged 75 passed away in 2020 is 2.79% lower than that in 2019), and the mortality rate increased by roughly

$e^{0.11739472} - 1 \approx 12.45\%$ for every unit increment in age in 2020. Performing visual inspection of the trace plots of α and β, we observed random scatter around mean values, suggesting that the chains mixed well and converged toward the target distributions.

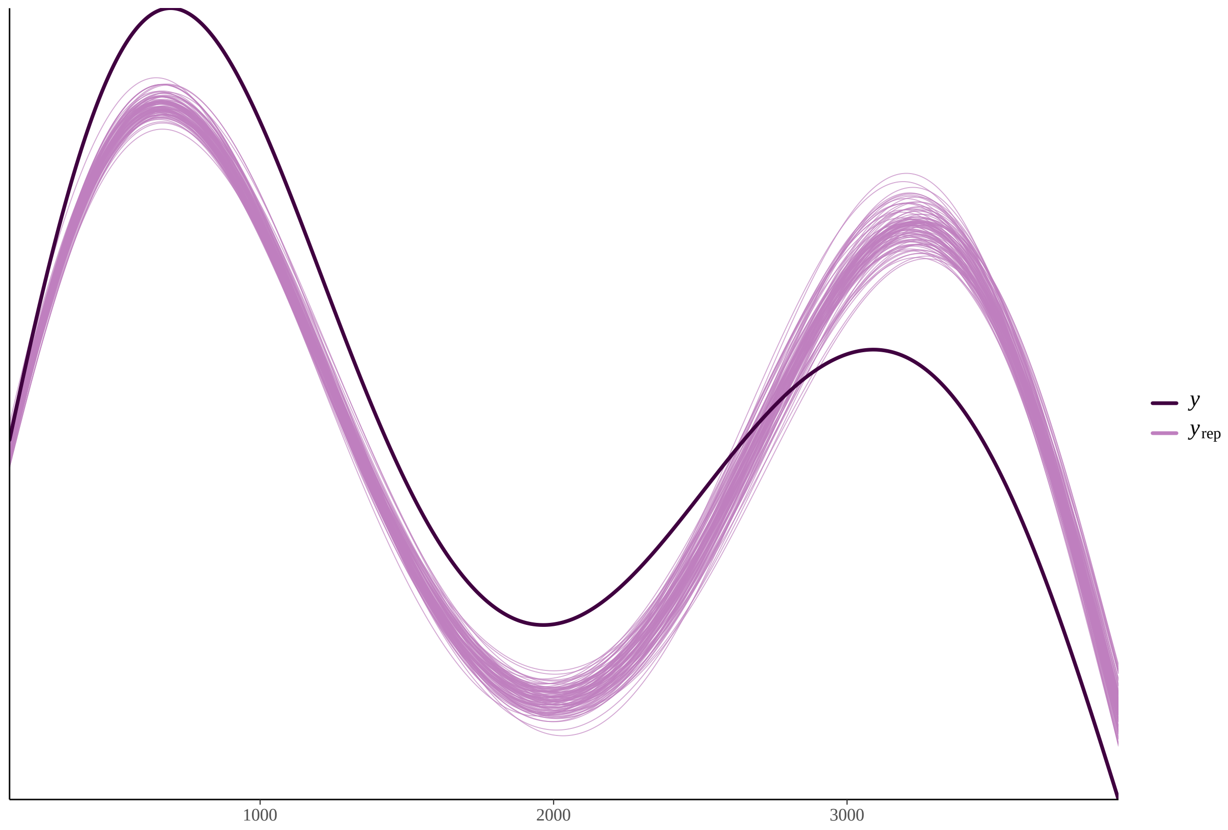


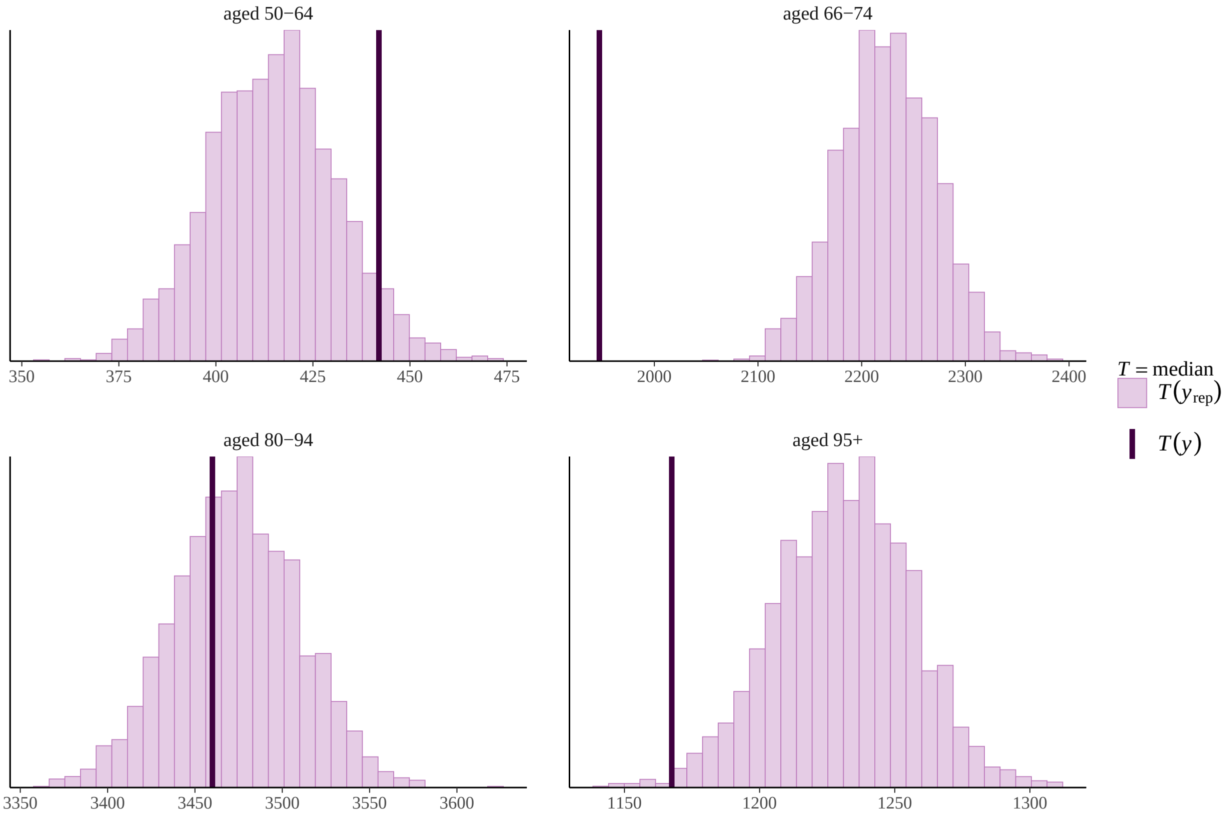
d) Carry out some posterior predictive checks to assess model performance.

Graphical posterior predictive checks (PPCs):

In the plot below, the dark line is the distribution of the observed death tolls, and each of the 100 lighter purple lines is the kernel density estimate of one of the replications of death tolls from the posterior predictive distribution. It is easy to see that the model underestimates death tolls of less than about 2600 and overestimates the counterpart, i.e. the model fails to account for variations in death tolls and population sizes between different age groups. To be more specific, the mortality rate that we are trying to model is the fraction of the death count to the population size at a point in time of a particular age, and here the spike in death tolls, alongside the plunge in size between populations aged less than 65 and those aged between 66 to 85 (Figures 5 and 6), was not fully captured by the Gompertz hazards.

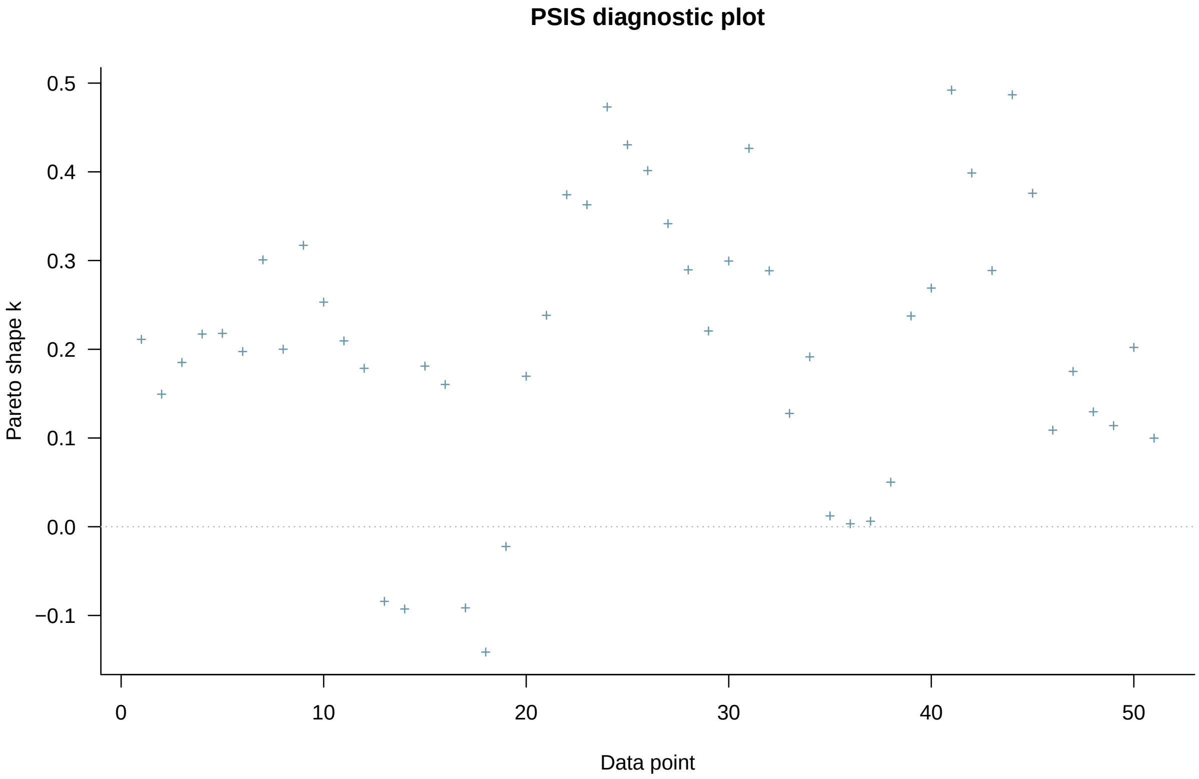
Another way to see this is to look at the distributions of the median of the replicated data sets over four age groups from the posterior predictive distribution and compare them with that of the observed death tolls. The histograms below indicate that the predicted median death tolls, compared to the observed median death tolls of the respective age groups, are higher for the population aged above 66 and lower for the counterpart.


y
y_{rep}
1000
2000
3000


aged 50−64
350
375
400
425
450
475
aged 66−74
2000
2100
2200
2300
2400
aged 80−94
3350
3400
3450
3500
3550
3600
aged 95+
1150
1200
1250
1300
T = median
$T(y_{\text{rep}})$
$T(y)$

Leave-One-Out Cross-Validation (LOO-CV):

The PSIS diagnostic plot indicates no influential points as none of the Pareto k estimates exceeds 0.7, suggesting the Monte Carlo sampling yields stable estimates for Swedish death tolls in 2020 with relatively small variance and bias, thus the LOO predictive distribution for point i does not deviate from the full predictive distribution.



Leave-One-Out Probability Integral Transform (LOO-PIT):

Examining the probabilities that the posterior predicted data $\tilde{y}$ has a lower value than the observed data y_i when we remove the i observation, say $p_i = p(\tilde{y} \leq y_i | y_{-i})$, we see the model is not very well calibrated for the sample of the year 2020 as there is more observed data than expected for both low and high values. This might be due to a small sample size of the data set. The QQ-plot of p_i has few data points at both ends deviating from the straight line while its center aligns with the straight line, making p_i's approximately normally distributed. Figure 9 and 10 also illustrate that the model death toll estimates for residents aged 66 to 85 are flawed as the observed data point is not contained inside or located close by either endpoints of the predictive interval for those ages.

Figure 7: LOO−PIT

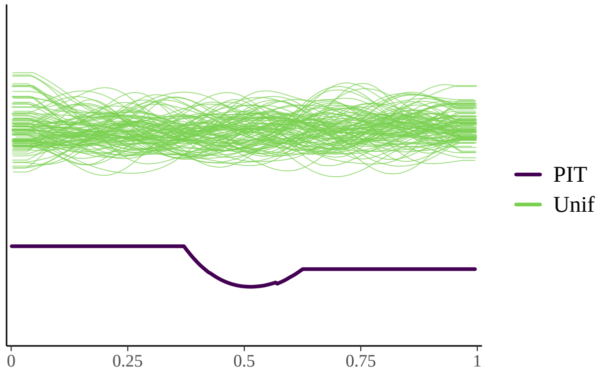


Figure 8 Quantile – Quantile plot of $p_i = p(|(\tilde{y} \le y_i, y_{-i}$

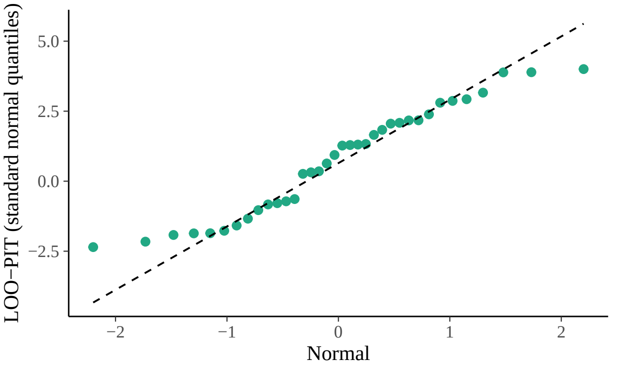


Figure 9: LOO predictive intervals vs observations

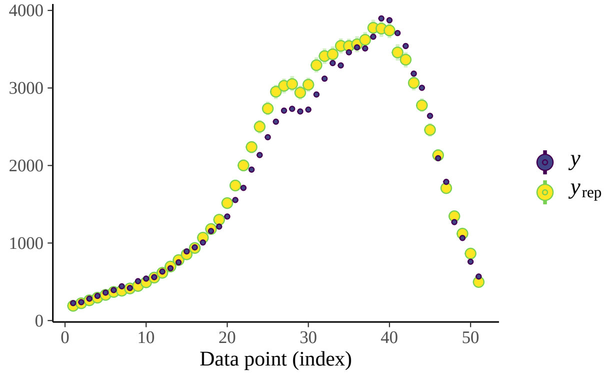
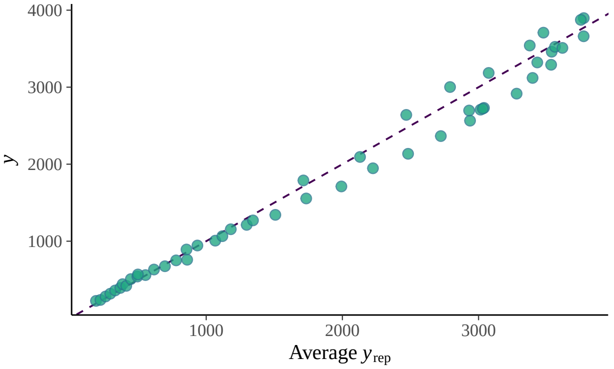
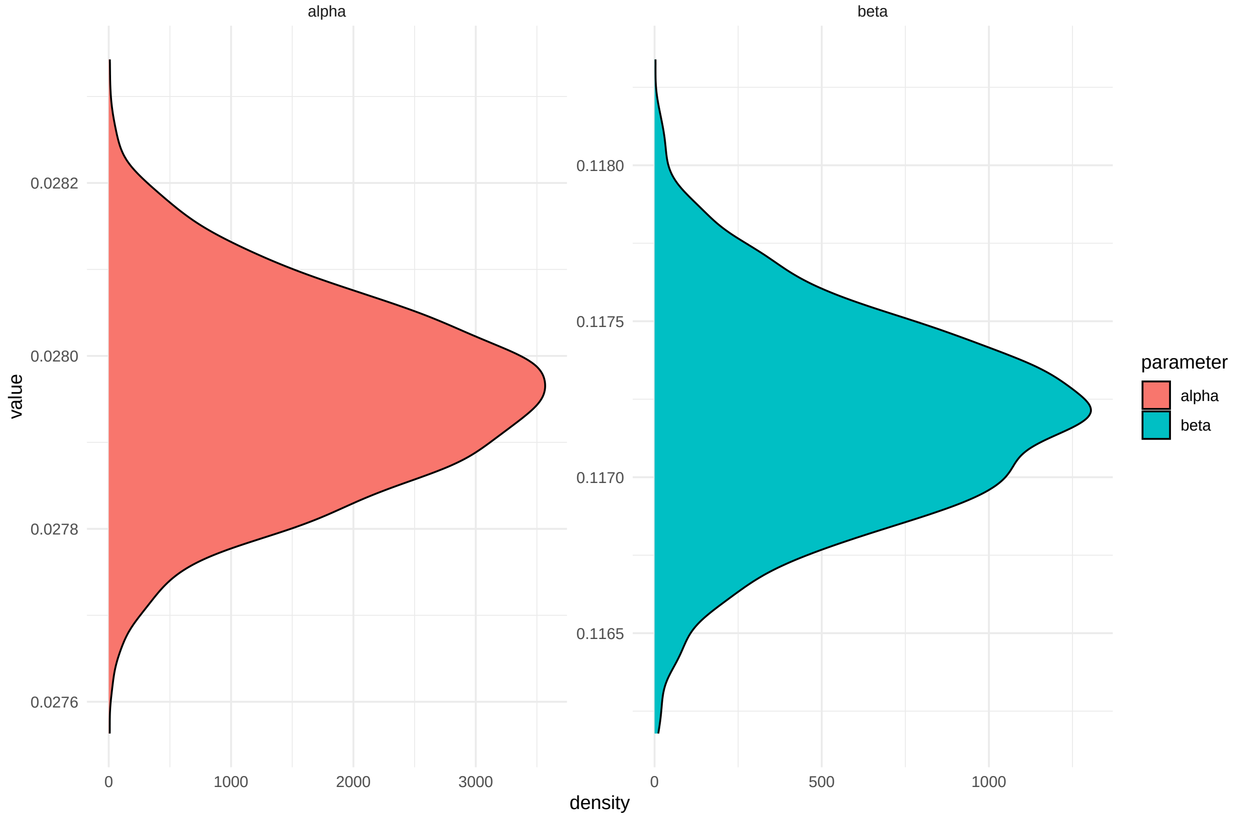


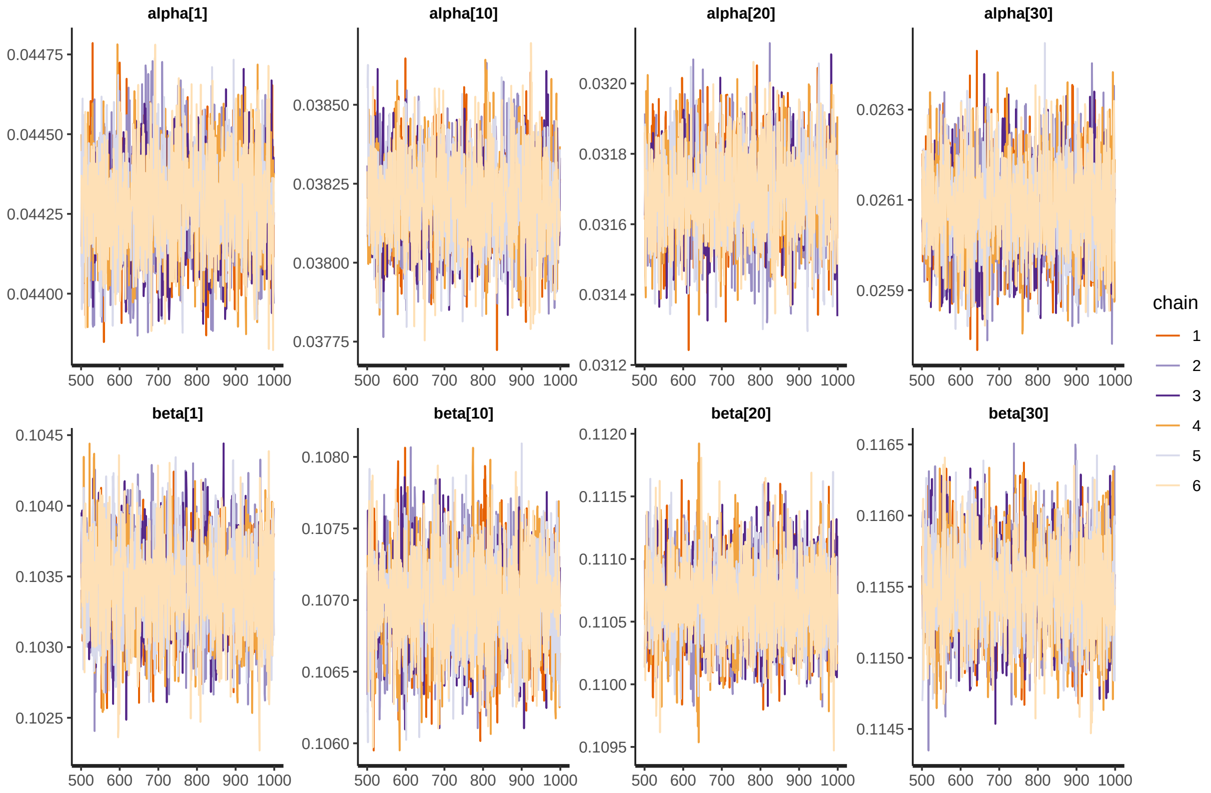
Figure 10: Prediction per datapoint vs the Observed valu



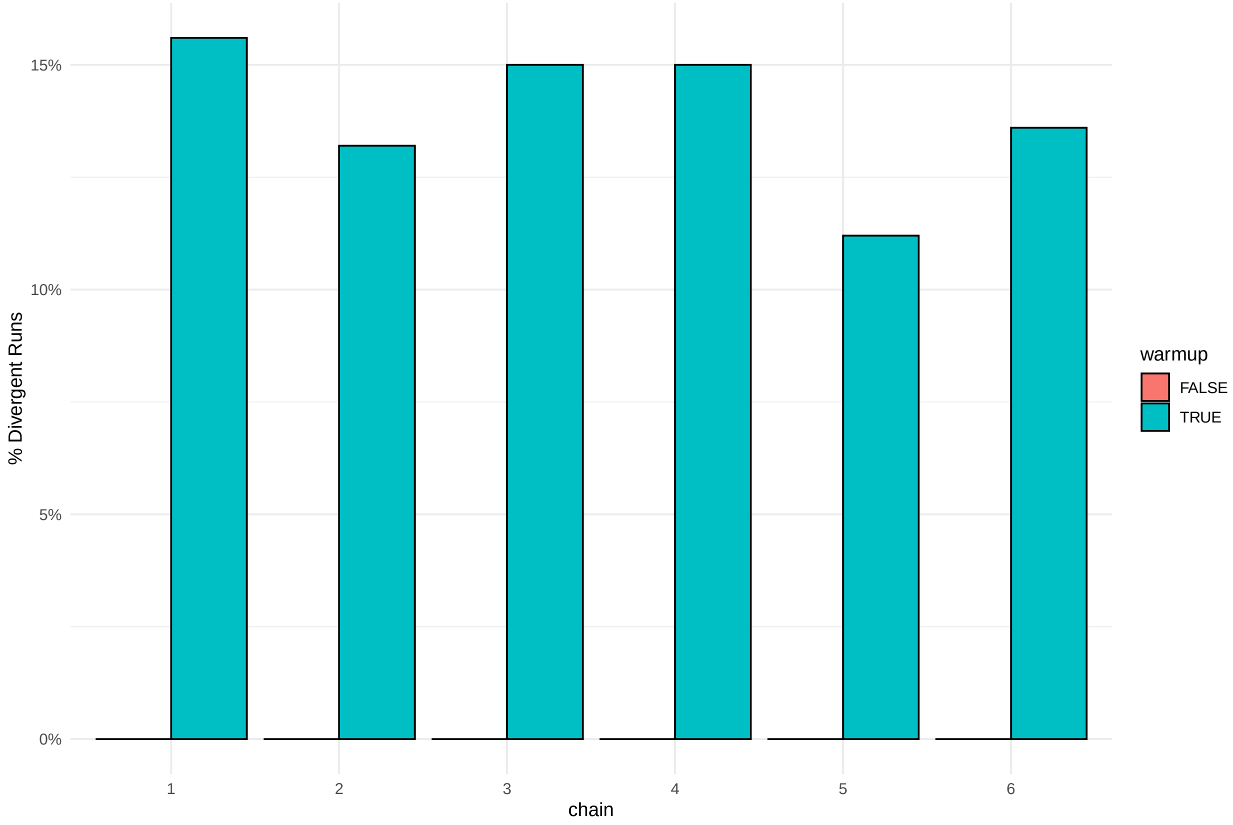


e) Now extend your model to estimate α and β in every year over the interval 1990-2020. Plot the resulting point estimates and 95% credible intervals for your estimates of α and β over time. Comment briefly on what you observe.

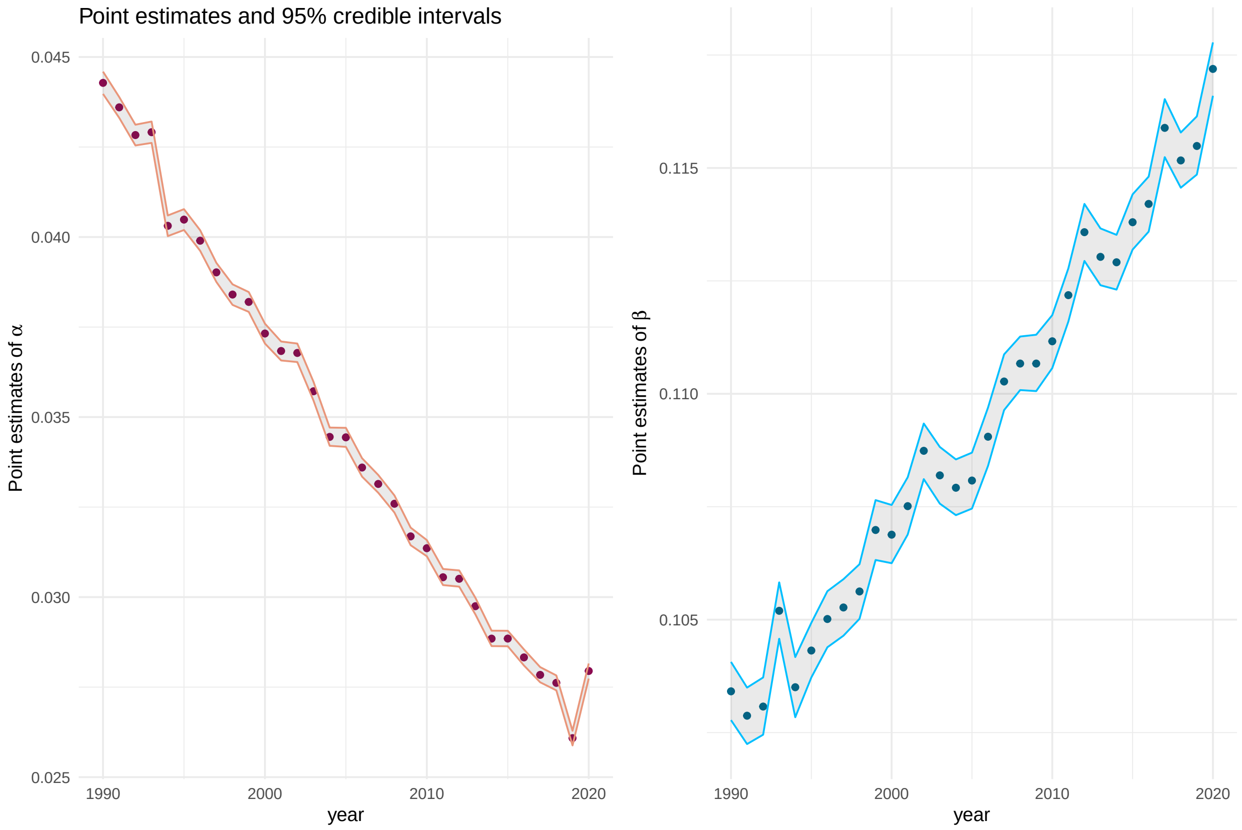
```
##                   mean      se_mean           sd       2.5%        25%        50%
## alpha[1]  0.04428164 2.726460e-06 0.0001567764 0.04397415 0.04417251 0.04428316
## alpha[2]  0.04360171 3.148199e-06 0.0001508027 0.04331043 0.04349598 0.04360136
## alpha[3]  0.04283214 2.793791e-06 0.0001488186 0.04254353 0.04273084 0.04283130
## alpha[4]  0.04291145 2.546880e-06 0.0001518423 0.04261404 0.04281232 0.04290912
## alpha[5]  0.04031151 2.483833e-06 0.0001459151 0.04002525 0.04021257 0.04030974
## alpha[6]  0.04048394 2.781082e-06 0.0001472321 0.04019536 0.04038463 0.04048289
## alpha[7]  0.03989772 3.418449e-06 0.0001434728 0.03961719 0.03980043 0.03989483
## alpha[8]  0.03901881 2.497533e-06 0.0001367159 0.03875327 0.03892393 0.03902095
## alpha[9]  0.03840277 2.618878e-06 0.0001439433 0.03811482 0.03830788 0.03840267
## alpha[10] 0.03819810 2.711468e-06 0.0001412781 0.03792170 0.03809972 0.03819547
## alpha[11] 0.03732028 3.672412e-06 0.0001372431 0.03704278 0.03723096 0.03731979
## alpha[12] 0.03683547 2.335645e-06 0.0001344811 0.03657226 0.03674712 0.03683268
## alpha[13] 0.03678114 2.203604e-06 0.0001328039 0.03652811 0.03669426 0.03677831
## alpha[14] 0.03571638 2.169517e-06 0.0001320373 0.03545963 0.03562754 0.03571730
## alpha[15] 0.03444947 2.856923e-06 0.0001285797 0.03420225 0.03436134 0.03445039
## alpha[16] 0.03443739 2.286880e-06 0.0001315937 0.03417270 0.03435449 0.03443724
## alpha[17] 0.03359880 2.561882e-06 0.0001332859 0.03334594 0.03350751 0.03359980
## alpha[18] 0.03314284 2.011423e-06 0.0001269323 0.03289802 0.03305755 0.03314116
## alpha[19] 0.03259271 2.043113e-06 0.0001229735 0.03234741 0.03250832 0.03259266
## alpha[20] 0.03168673 2.268535e-06 0.0001232150 0.03144169 0.03160484 0.03168727
## alpha[21] 0.03135829 2.535273e-06 0.0001167286 0.03113439 0.03127819 0.03136014
## alpha[22] 0.03055168 1.834561e-06 0.0001150923 0.03033399 0.03047379 0.03054966
## alpha[23] 0.03050932 2.046467e-06 0.0001154734 0.03029131 0.03043176 0.03050845
## alpha[24] 0.02974771 2.030895e-06 0.0001167685 0.02951984 0.02967104 0.02974685
## alpha[25] 0.02884720 1.954691e-06 0.0001115224 0.02863946 0.02876995 0.02884575
## alpha[26] 0.02884880 1.826619e-06 0.0001115888 0.02863539 0.02877311 0.02884861
## alpha[27] 0.02832563 2.012555e-06 0.0001124911 0.02810050 0.02824878 0.02832456
## alpha[28] 0.02784054 3.486201e-06 0.0001074581 0.02763487 0.02776724 0.02783918
## alpha[29] 0.02761961 1.807015e-06 0.0001045588 0.02741020 0.02754968 0.02761917
## alpha[30] 0.02608370 1.871679e-06 0.0001020187 0.02588092 0.02601524 0.02608279
## alpha[31] 0.02794885 2.480725e-06 0.0001063227 0.02773811 0.02787836 0.02794996
## beta[1]   0.10341324 6.741201e-06 0.0003309888 0.10277575 0.10318529 0.10340765
## beta[2]   0.10287304 6.713511e-06 0.0003157636 0.10224589 0.10266079 0.10287976
## beta[3]   0.10307595 6.292336e-06 0.0003284379 0.10245268 0.10285045 0.10307600
## beta[4]   0.10519603 4.916334e-06 0.0003162568 0.10457695 0.10498350 0.10518680
## beta[5]   0.10350406 5.417685e-06 0.0003365774 0.10283974 0.10328073 0.10349838
## beta[6]   0.10431626 6.364689e-06 0.0003179269 0.10371915 0.10408711 0.10432363
## beta[7]   0.10501428 5.657244e-06 0.0003203646 0.10438873 0.10479285 0.10501648
## beta[8]   0.10526954 5.087550e-06 0.0003233596 0.10464482 0.10504535 0.10525855
## beta[9]   0.10562407 4.598245e-06 0.0003146928 0.10501843 0.10541440 0.10562553
## beta[10]  0.10698494 5.456755e-06 0.0003343241 0.10632106 0.10676160 0.10698682
## beta[11]  0.10688123 5.145894e-06 0.0003253067 0.10624993 0.10666809 0.10688108
## beta[12]  0.10751278 4.865072e-06 0.0003200013 0.10688493 0.10730337 0.10751204
## beta[13]  0.10873820 4.445358e-06 0.0003133372 0.10811133 0.10852579 0.10873791
## beta[14]  0.10819462 4.777257e-06 0.0003239990 0.10756810 0.10797464 0.10818502
## beta[15]  0.10792040 5.190738e-06 0.0003158425 0.10731332 0.10769937 0.10791768
## beta[16]  0.10807975 4.840721e-06 0.0003130825 0.10745894 0.10787031 0.10807088
## beta[17]  0.10904991 4.932776e-06 0.0003309869 0.10840411 0.10882282 0.10905536
## beta[18]  0.11027237 4.337971e-06 0.0003111845 0.10963930 0.11006293 0.11028178
## beta[19]  0.11067003 4.696619e-06 0.0003112967 0.11008270 0.11045269 0.11067172
## beta[20]  0.11066970 4.899024e-06 0.0003219360 0.11006019 0.11044581 0.11067152
## beta[21]  0.11116256 5.075574e-06 0.0003062761 0.11056953 0.11095383 0.11116315
## beta[22]  0.11218420 4.447940e-06 0.0003040916 0.11159852 0.11198207 0.11218516
## beta[23]  0.11357844 5.151444e-06 0.0003256686 0.11294188 0.11335671 0.11357907
## beta[24]  0.11303201 4.679878e-06 0.0003183683 0.11240360 0.11281102 0.11303105
## beta[25]  0.11291207 4.643680e-06 0.0003145229 0.11230722 0.11269932 0.11291172
## beta[26]  0.11379888 4.470418e-06 0.0003121451 0.11319314 0.11358332 0.11379416
## beta[27]  0.11420368 4.471995e-06 0.0003120718 0.11359051 0.11399343 0.11421002
## beta[28]  0.11588747 6.138040e-06 0.0003184833 0.11523977 0.11567675 0.11588276
## beta[29]  0.11516698 4.311784e-06 0.0003128053 0.11456272 0.11495343 0.11516428
## beta[30]  0.11548674 4.931909e-06 0.0003262210 0.11485408 0.11527049 0.11548088
## beta[31]  0.11719290 5.358232e-06 0.0003071685 0.11659762 0.11698904 0.11719017
##                  75%      97.5%     n_eff      Rhat
## alpha[1]  0.04438802 0.04459294 3306.4585 1.0003917
## alpha[2]  0.04370636 0.04388745 2294.5290 0.9992501
## alpha[3]  0.04293552 0.04311940 2837.4389 1.0002934
## alpha[4]  0.04301436 0.04320695 3554.4184 0.9987741
## alpha[5]  0.04040681 0.04059895 3451.0852 1.0009251
## alpha[6]  0.04058387 0.04077400 2802.7053 0.9999328
## alpha[7]  0.03999082 0.04018864 1761.4937 0.9999991
## alpha[8]  0.03911223 0.03928282 2996.5107 1.0010757
## alpha[9]  0.03849575 0.03868649 3021.0124 1.0003732
## alpha[10] 0.03829467 0.03847252 2714.8170 1.0001365
## alpha[11] 0.03741236 0.03759520 1396.6202 1.0034674
## alpha[12] 0.03692792 0.03709841 3315.1938 0.9992977
## alpha[13] 0.03687260 0.03704218 3632.0744 0.9994540
## alpha[14] 0.03580586 0.03598005 3703.9655 0.9998282
## alpha[15] 0.03453667 0.03470869 2025.5738 1.0088383
## alpha[16] 0.03452266 0.03470133 3311.1852 0.9988292
## alpha[17] 0.03369129 0.03385656 2706.7623 0.9997307
## alpha[18] 0.03323115 0.03339152 3982.3340 0.9996889
## alpha[19] 0.03267624 0.03282370 3622.7504 0.9986567
## alpha[20] 0.03176851 0.03192983 2950.0962 1.0011781
## alpha[21] 0.03143903 0.03158392 2119.8488 1.0003460
## alpha[22] 0.03062988 0.03078101 3935.7538 0.9992166
## alpha[23] 0.03058572 0.03074166 3183.8643 0.9996617
## alpha[24] 0.02982299 0.02998186 3305.7961 0.9992283
## alpha[25] 0.02892154 0.02906765 3255.1243 0.9989813
## alpha[26] 0.02892274 0.02906562 3732.0353 0.9989707
## alpha[27] 0.02840085 0.02854767 3124.2164 0.9994015
## alpha[28] 0.02791286 0.02805664  950.1088 1.0033945
## alpha[29] 0.02768963 0.02782509 3348.0949 0.9999131
## alpha[30] 0.02615327 0.02628859 2970.9592 0.9994179
## alpha[31] 0.02801692 0.02815663 1836.9375 1.0015947
## beta[1]   0.10363928 0.10406395 2410.7486 1.0019275
## beta[2]   0.10308412 0.10349599 2212.2019 1.0002961
## beta[3]   0.10329558 0.10371964 2724.4734 0.9999613
## beta[4]   0.10541258 0.10582562 4138.0628 0.9994233
## beta[5]   0.10373874 0.10417136 3859.6009 1.0009188
## beta[6]   0.10452526 0.10492945 2495.1700 0.9990761
## beta[7]   0.10523074 0.10562988 3206.8551 0.9991736
## beta[8]   0.10549028 0.10589588 4039.7468 1.0002227
## beta[9]   0.10583272 0.10622707 4683.7012 0.9985645
## beta[10]  0.10720580 0.10764665 3753.7598 0.9981462
## beta[11]  0.10709051 0.10753885 3996.3573 1.0004151
## beta[12]  0.10773087 0.10815243 4326.3828 0.9990547
## beta[13]  0.10895191 0.10934256 4968.3321 0.9986482
## beta[14]  0.10841270 0.10882279 4599.7068 0.9994181
## beta[15]  0.10813569 0.10854751 3702.3960 1.0006078
## beta[16]  0.10829144 0.10869719 4183.0914 0.9990221
## beta[17]  0.10927221 0.10969528 4502.3459 1.0001364
## beta[18]  0.11048475 0.11087280 5145.9136 0.9995415
## beta[19]  0.11088396 0.11126822 4393.1732 0.9985447
## beta[20]  0.11088434 0.11130763 4318.3715 0.9992683
## beta[21]  0.11137603 0.11174228 3641.2964 0.9998200
## beta[22]  0.11239040 0.11277686 4674.0240 0.9990981
## beta[23]  0.11380046 0.11420457 3996.6293 0.9997144
## beta[24]  0.11325169 0.11366090 4627.9715 0.9995936
## beta[25]  0.11311884 0.11351946 4587.5425 1.0002837
## beta[26]  0.11400407 0.11441545 4875.4735 0.9990154
## beta[27]  0.11442416 0.11480916 4869.7480 0.9987383
## beta[28]  0.11610488 0.11652426 2692.2413 0.9998582
## beta[29]  0.11538191 0.11578473 5263.0090 0.9988841
## beta[30]  0.11569561 0.11614256 4375.1559 0.9989495
## beta[31]  0.11740180 0.11778016 3286.3245 1.0003189
```



Examining trace plots of some representative point estimates of α and β, we see all chains mixed well, and there are no divergences across all chains during the sampling period.



The point estimates of α declined over time, similar to the trend of starting levels of mortality rates in Sweden in Figure 1, whereas the point estimates of β increased by about 13% between 1990 and 2020, suggesting the rate at which mortality rate increased given one unit increment in age increased over the period of 1990-2020. The 95% credible intervals for all point estimates of α and β interpreted as the interval for which there is a 95% probability the true values of α and β lie, are small, indicating it is with low uncertainty that the point estimates yielded by MCMC are flawed.



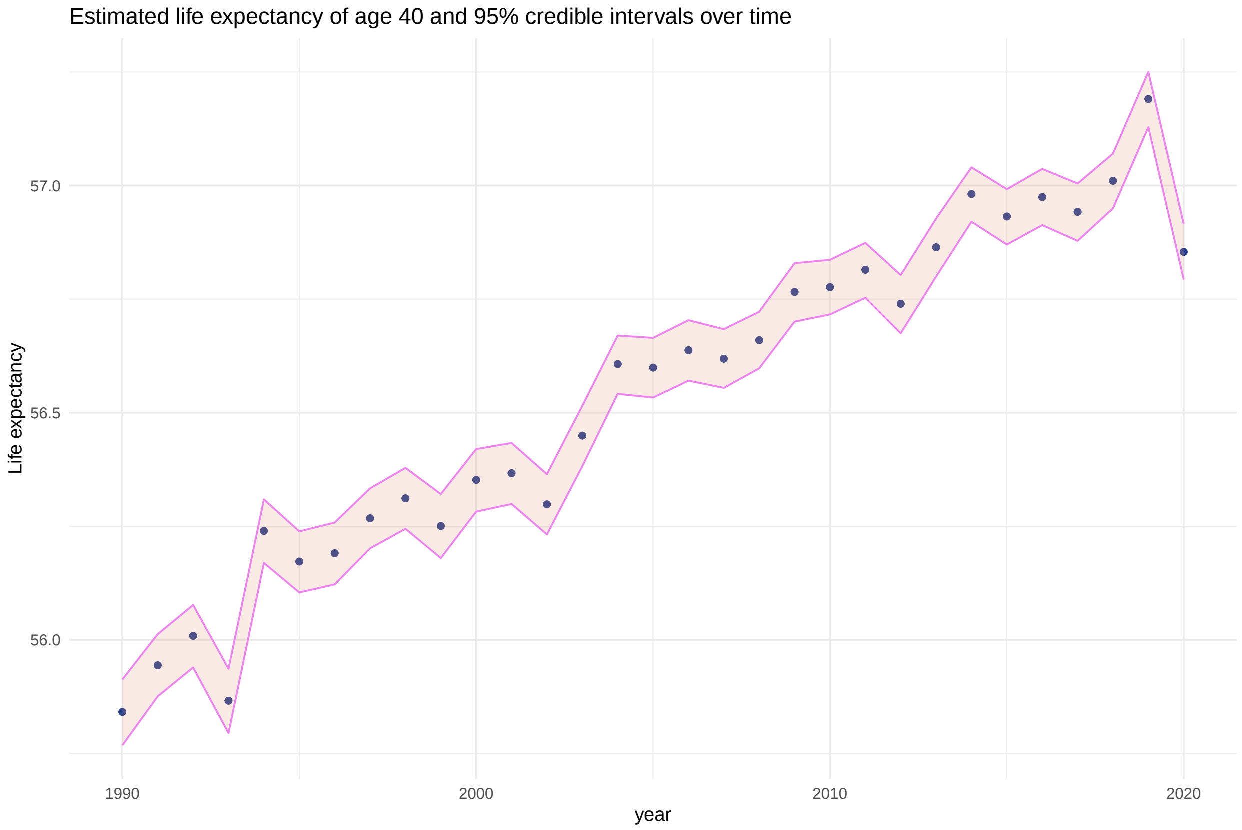
Point estimates and 95% credible intervals

f) Life expectancy at age x is defined as

$$\int_x^\omega e^{-\mu_a} da$$

where ω is the oldest age group (you may assume this is age 100). Life expectancy is the expected number of years of life left at age x. The integral can be approximated by summing over discrete age groups. Based on your estimates in the previous question, estimate life expectancy at age 40 (note starting age!) for every year from 1990-2020. Plot your resulting point estimates and 95% credible intervals over time and comment briefly.

The overall life expectancy of Swedish residents aged 40 is estimated to have increased by 1.5 years between 1990 and 2019, which tailed off for about 0.3 years in 2020 due to the pandemic.


Estimated life expectancy of age 40 and 95% credible intervals over time
Life expectancy
57.0
56.5
56.0
1990
2000
2010
2020
year

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 neighbors 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 micrograms 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:

$$y_i = \begin{cases} 1 & \text{if household } i \text{ switched to a new well} \\ 0 & \text{if household } i \text{ continued using its own well.} \end{cases}$$

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 `read_table`.

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 metres) 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.

```
##            arsentic_safety
## switch_well    unsafe
##          no  0.4248344
##          yes 0.5751656
```

First, we note that the distributions of `dist` and `arsenic` are skewed, so we log transformation might be necessary to reduce the skewness of measurement variables. Performing EDA, we noticed that, given all wells in the study are unsafe, the longer the distance Bengalis had to commute to the safer well, the less willing they were to switch. Also, on average, the average distance households that didn't switch to safer wells would have to commute is 9.1793 meters longer, while the arsenic level in wells used by them is only 0.411839 units of hundreds of micrograms per litre lower. The finding is that long-distance commutes and seemingly lower arsenic levels compared to that in wells used by other households discouraged about 42.5% of Bengali households from switching to safer wells, even though long-term exposure to arsenic from drinking water can lead to accumulation of toxicity. It is also noteworthy that, from Figure 6, we observed the households that used the most arsenic-contaminated wells were also the farthest from a safe well.

Figure 1

Distribution of distance

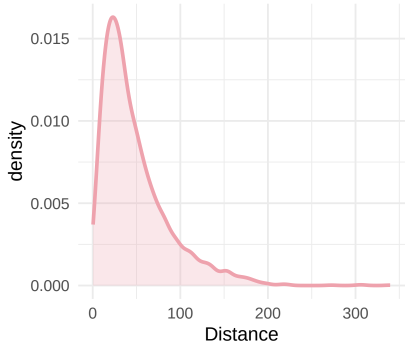


Figure 2

Distribution of arsenic level

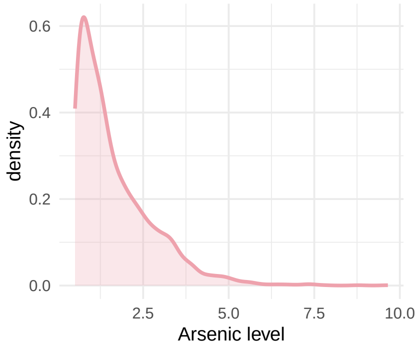


Figure 3

Households

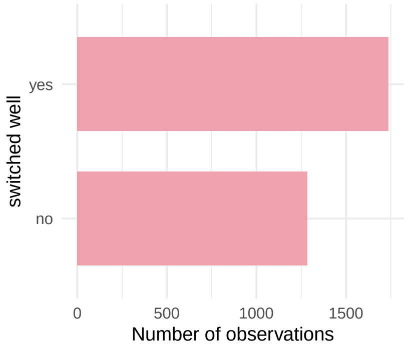


Figure 4

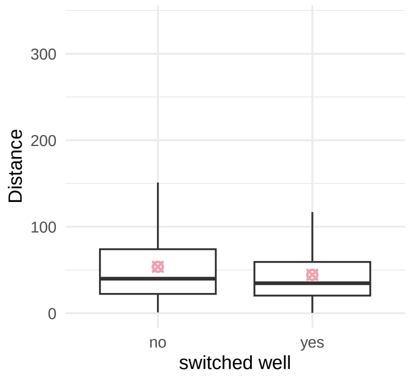


Figure 5

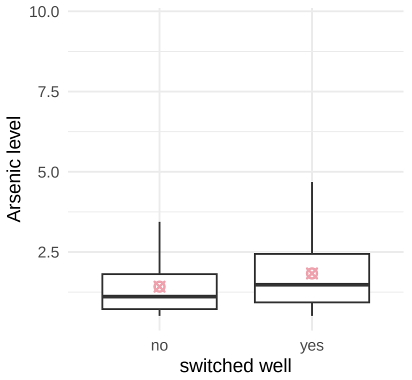
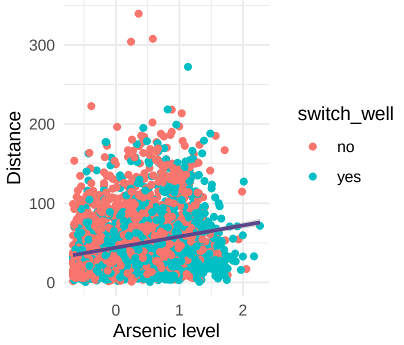


Figure 6

Distance by Arsenic level



b) Fit both of these models using Stan. Put $N(0,1)$ priors on all the βs. You should generate pointwise 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.

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

Model 1:

$$\operatorname{logit}(p_i) = \beta_0 + \beta_1 \cdot \left(d_i - \bar{d}\right) + \beta_2 \cdot (a_i - \bar{a}) + \beta_3 \cdot \left(d_i - \bar{d}\right)(a_i - \bar{a})$$

```
##                    mean      se_mean            sd          2.5%           25%
## beta[1]  0.350539835 1.198775e-03 0.038497942  0.274988226  0.324239321
## beta[2] -0.008791021 1.877680e-05 0.001025688 -0.010737683 -0.009481738
## beta[3]  0.469978800 1.437647e-03 0.041187975  0.392492952  0.441127527
## beta[4] -0.001739962 1.960718e-05 0.001013028 -0.003622189 -0.002469569
##                  50%          75%         97.5%     n_eff      Rhat
## beta[1]  0.350616589  0.376385843  0.4269102472 1031.3348 1.0023888
## beta[2] -0.008817922 -0.008084694 -0.0067569234 2983.9227 0.9994960
## beta[3]  0.469224167  0.496962890  0.5522033618  820.7982 1.0039903
## beta[4] -0.001761819 -0.001048877  0.0002195051 2669.3960 0.9987015
```

Interpretation:

- $\hat{\beta}_0 \approx 0.35$: Any households whose well was contaminated by an average arsenic level (i.e. $a_i - \overline{a} = 0$) and lived an average distance to a safe well (i.e. $d_i - \overline{d} = 0$) had a $\frac{e^{0.35}}{1+e^{0.35}} \approx 58\%$ probability of switching, or equivalently, had the odds of switching of such households was about 1.42.
- $\hat{\beta}_1 \approx -0.0088$: For any households whose well was contaminated by the average arsenic level (i.e. $a_i - \overline{a} = 0$), for a one-meter increase (or decrease) in its distance to the safe, we expect to see about $e^{0.0088} - 1 \approx 0.88\%$ decrease (or increase) in the odds of switching.
- $\hat{\beta}_2 \approx 0.4699$ For any households that lived an average distance to a safe well (i.e. $d_i - \overline{d} = 0$), for a one-unit increase (or decrease) in arsenic level, we expect to see about $e^{0.4699} - 1 \approx 60\%$ increase (or decrease) in the odds of switching.
- $\hat{\beta}_3 \approx -0.0017$: The interaction term reflects the effect of distance on the odds, or probability of switching, as arsenic levels vary. Here, for each arsenic level, below average (0.71), average (1.65693), and above average (3.28), the effect of distance on the odds, or probability of switching, decreases as distance increases.

```
##                    values distance arsenic
## 1 below average distance 16.80000 0.71000
## 2                        16.80000 1.65693
## 3                        16.80000 3.28000
## 4       average distance 48.33186 0.71000
## 5                        48.33186 1.65693
## 6                        48.33186 3.28000
## 7 above average distance 80.70000 0.71000
## 8                        80.70000 1.65693
## 9                        80.70000 3.28000
```

Model 2:

$$\begin{aligned}\text{logit}\left(p_i\right) =& \beta_0 + \beta_1 \cdot \left(d_i - \bar{d}\right) + \beta_2 \cdot \left(\log\left(a_i\right) - \overline{\log(a)}\right) \\ &+ \beta_3 \cdot \left(d_i - \bar{d}\right)\left(\log\left(a_i\right) - \overline{\log(a)}\right)\end{aligned}$$

where d_i is distance and a_i is arsenic level.

```
##                  mean      se_mean          sd         2.5%          25%
## beta[1]  0.342770565 1.213353e-03 0.039579854  0.266172441  0.315887817
## beta[2] -0.009449160 1.961613e-05 0.001063770 -0.011524649 -0.010146078
## beta[3]  0.867489333 2.186559e-03 0.067490596  0.734572582  0.821698049
## beta[4] -0.002284441 3.754476e-05 0.001824817 -0.005937678 -0.003527038
##                   50%          75%        97.5%     n_eff     Rhat
## beta[1]  0.342053689  0.369190367  0.420054139 1064.0789 1.001162
## beta[2] -0.009455151 -0.008754823 -0.007322791 2940.8205 1.000941
## beta[3]  0.867347754  0.913595278  0.998591664  952.7178 1.002678
## beta[4] -0.002280913 -0.001067791  0.001383599 2362.3274 0.999878
```

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

The predicted proportion of households whose wells were contaminated by arsenic levels less than 0.82 switched to safer wells yielded by Model 1 is very close to 0, whereas about 28% by Model 2, i.e. the predicted proportion yielded by Model 1 is way too high, so Model 2 is preferred here.

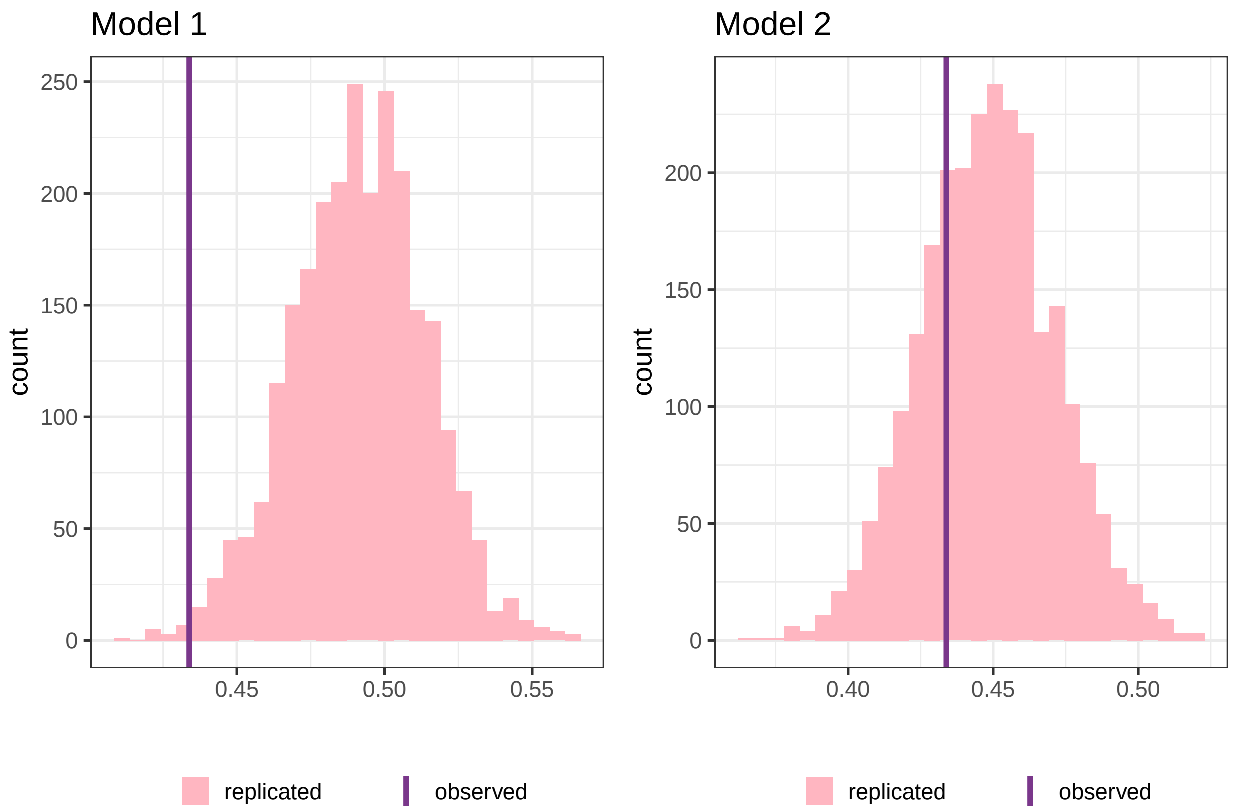
- Model 1: $P\left(t\left(\boldsymbol{y}^{rep}\right) < t(\boldsymbol{y})\right) = 0.0052$
- Model 2: $P\left(t\left(\boldsymbol{y}^{rep}\right) < t(\boldsymbol{y})\right) = 0.2612$

```r
stat_compare <- function(y, t){
  mean(y < t)
}
stat_compare(trep1, t_y)
```

```
## [1] 0.0052
```

```r
stat_compare(trep2, t_y)
```

```
## [1] 0.2612
```



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

The $elpd_{LOO}$ is higher for Model 2, so it is preferred.

```
##        elpd_diff se_diff
## model2   0.0       0.0  
## model1 -15.5       4.4  
```

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

In Figure 7, we see $ELPD_i$'s of Model 2 overlap those of Model 1 for most households, but Model 2 seems to yield higher predictive accuracy for the probabilities of switching of households that didn't switch to safer wells, as indicated by the red points scattering away from the line.

In Figure 8, as all $elpd_diff$ are less than 4, the difference between the two models is generally small. However, for households whose wells were contaminated with arsenic levels on the log scale between 0 and approximately 1.2, Model 1 appears to estimate the probability of switching better for those who didn't switch, while Model 2 does a better job for those that switched. For households whose wells were contaminated with logged arsenic levels less than 0 or approximately higher than 1.2, the reverse is true. The two plots are consistent in the sense that they both suggests the large difference in the posterior predictive accuracy between the two models for households that were exposed to high levels of arsenic and did not switch to safer wells, as indicated by some outliers households by the red dots scattering away from the y-intercept line.

Figure 7

ELDP

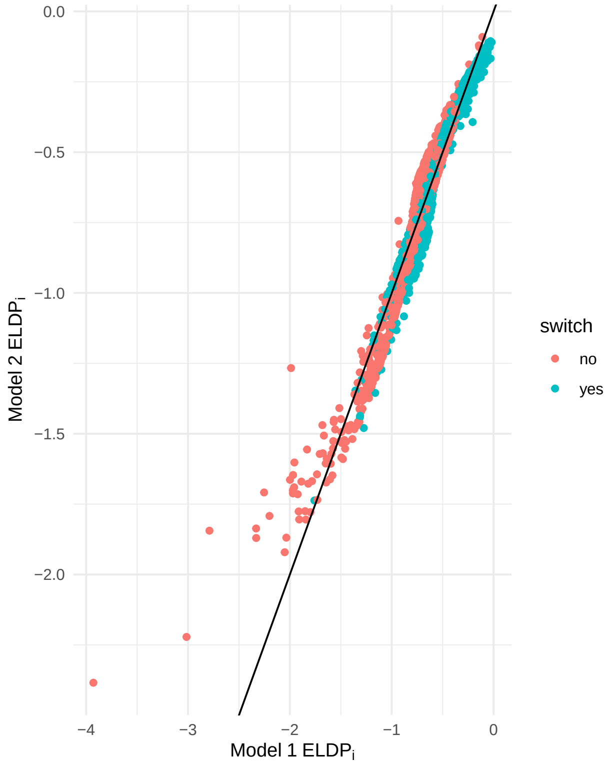
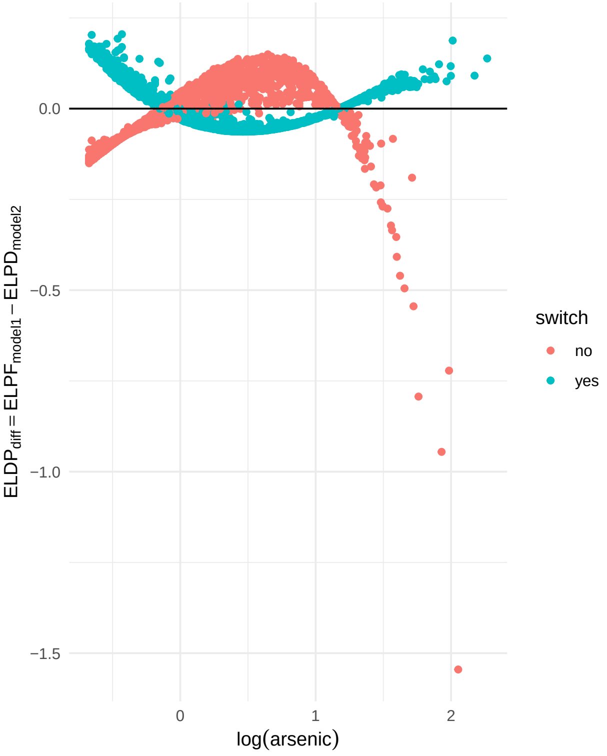


Figure 8

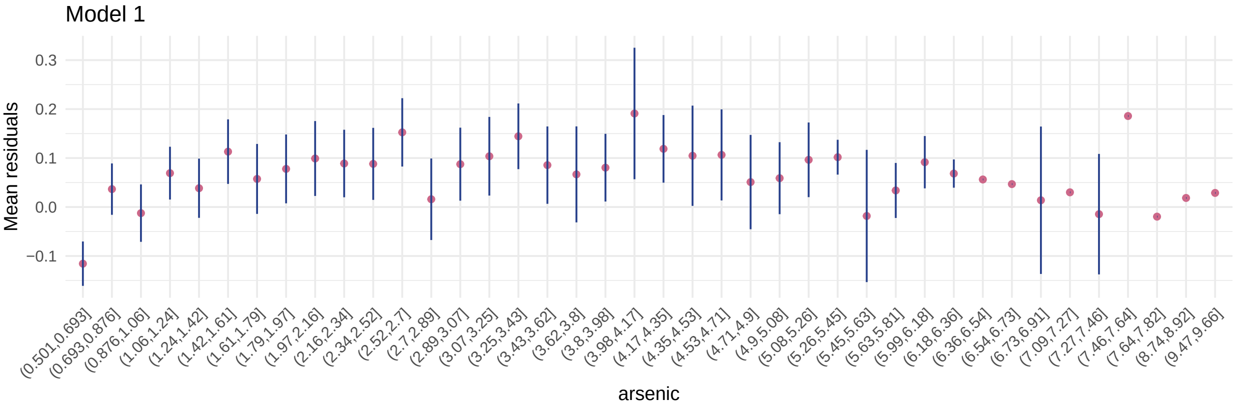
log(arsenic) and ELDP$_{diff}$

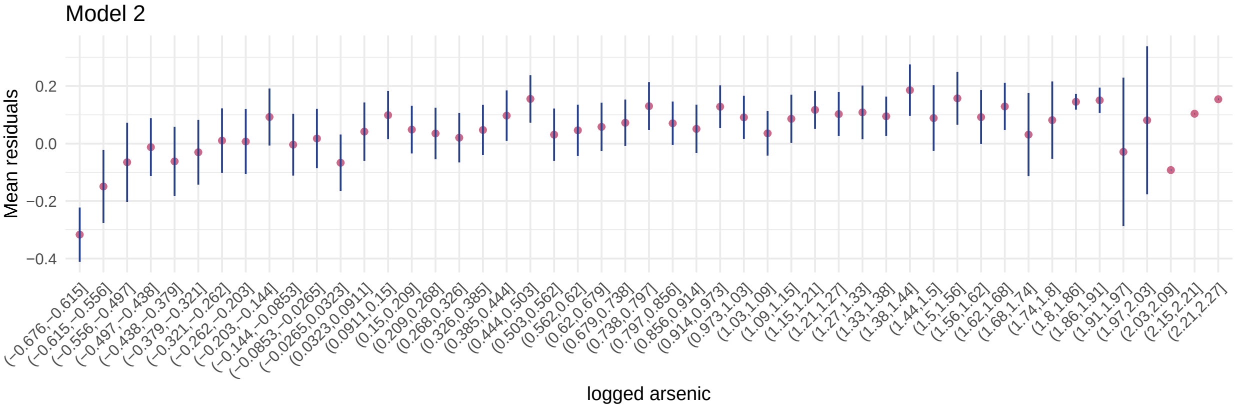


f) Given the outcome in this case is discrete, we can directly interpret the $ELPD_i$s. In particular, what is $\exp(ELPD_i)$?

Since $ELPD_i = log[p(y_i|\mathrm{y}_{-i})]$, we intepret $e^{ELPD_i} = p(y_i|\mathrm{y}_{-i})$ as the leave-one-out predictive probability of switching well of the i^{th} household (i.e. probability that the i^{th} household decided to switche wells based on the decisions either switch or not of all other households.)

g) For each model recode the $ELPD_i$'s to get $\hat{y}_i = E\left(Y_i|\boldsymbol{y}_{-i}\right)$. 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.


Model 1
Mean residuals
(0.501,0.693]
(0.693,0.876]
(0.876,1.06]
(1.06,1.24]
(1.24,1.42]
(1.42,1.61]
(1.61,1.79]
(1.79,1.97]
(1.97,2.16]
(2.16,2.34]
(2.34,2.52]
(2.52,2.7]
(2.7,2.89]
(2.89,3.07]
(3.07,3.25]
(3.25,3.43]
(3.43,3.62]
(3.62,3.8]
(3.8,3.98]
(3.98,4.17]
(4.17,4.35]
(4.35,4.53]
(4.53,4.71]
(4.71,4.9]
(4.9,5.08]
(5.08,5.26]
(5.26,5.45]
(5.45,5.63]
(5.63,5.81]
(5.99,6.18]
(6.18,6.36]
(6.36,6.54]
(6.54,6.73]
(6.73,6.91]
(7.09,7.27]
(7.27,7.46]
(7.46,7.64]
(7.64,7.82]
(8.74,8.92]
(9.47,9.66]
arsenic


Model 2
Mean residuals
(−0.676,−0.615]
(−0.615,−0.556]
(−0.556,−0.497]
(−0.497,−0.438]
(−0.438,−0.379]
(−0.379,−0.321]
(−0.321,−0.262]
(−0.262,−0.203]
(−0.203,−0.144]
(−0.144,−0.0853]
(−0.0853,−0.0265]
(−0.0265,0.0323]
(0.0323,0.0911]
(0.0911,0.15]
(0.15,0.209]
(0.209,0.268]
(0.268,0.326]
(0.326,0.385]
(0.385,0.444]
(0.444,0.503]
(0.503,0.562]
(0.562,0.62]
(0.62,0.679]
(0.679,0.738]
(0.738,0.797]
(0.797,0.856]
(0.856,0.914]
(0.914,0.973]
(0.973,1.03]
(1.03,1.09]
(1.09,1.15]
(1.15,1.21]
(1.21,1.27]
(1.27,1.33]
(1.33,1.38]
(1.38,1.44]
(1.44,1.5]
(1.5,1.56]
(1.56,1.62]
(1.62,1.68]
(1.68,1.74]
(1.74,1.8]
(1.8,1.86]
(1.86,1.91]
(1.91,1.97]
(1.97,2.03]
(2.03,2.09]
(2.15,2.21]
(2.21,2.27]
logged arsenic