Homework 5

Julissa Duenas

Due on November 29, 2020 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

Problem 1. Frequentist Coverage of The Bayesian Posterior Interval.

In quiz 1 we explored the importance and difficulty of well-calibrated prior distributions by examining the calibration of subjective intervals. Suppose that $y_1, .., y_n$ is an IID sample from a $Normal(\mu, 1)$. We wish to estimate μ.

1a. For Bayesian inference, we will assume the prior distribution $\mu \sim Normal(0, \frac{1}{\kappa_0})$ for all parts below. Remember, from lecture that we can interpret κ_0 as the pseudo-number of prior observations with sample mean $\mu_0 = 0$. State the posterior distribution of μ given $y_1, .., y_n$. Report the lower and upper bounds of the 95% quantile-based posterior credible interval for μ, using the fact that for a normal distribution with standard eviation σ, approximately 95% of the mass is between $\pm 1.96\sigma$.

Posterior Distribution: $p(\theta|y, \sigma^2) = L(\mu) * p(\mu)$

$Y_i and \mu$ are normal so posterior is also normally distributed $p(\mu|y, \sigma) \propto L(\mu) * p(\mu)$

$\propto exp(-\frac{(\bar{y}-\mu)^2}{\frac{2\sigma^2}{n}} * exp(-\frac{(\mu-\mu_0)^2}{2\tau^2})$

$\propto exp(-1/2(\frac{n}{\sigma^2} + \frac{1}{\tau^2})(\mu - \frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}})^2)$

$\tau^2 = \frac{1}{k_0}, \sigma^2 = 1, \mu_0 = 0$

$exp(-1/2(\frac{n}{1} + \frac{1}{\frac{1}{k_0}})(\mu - \frac{\frac{n\bar{y}}{1} + \frac{0}{\frac{1}{k_0}}}{\frac{n}{1} + \frac{1}{\frac{1}{k_0}}})^2)$

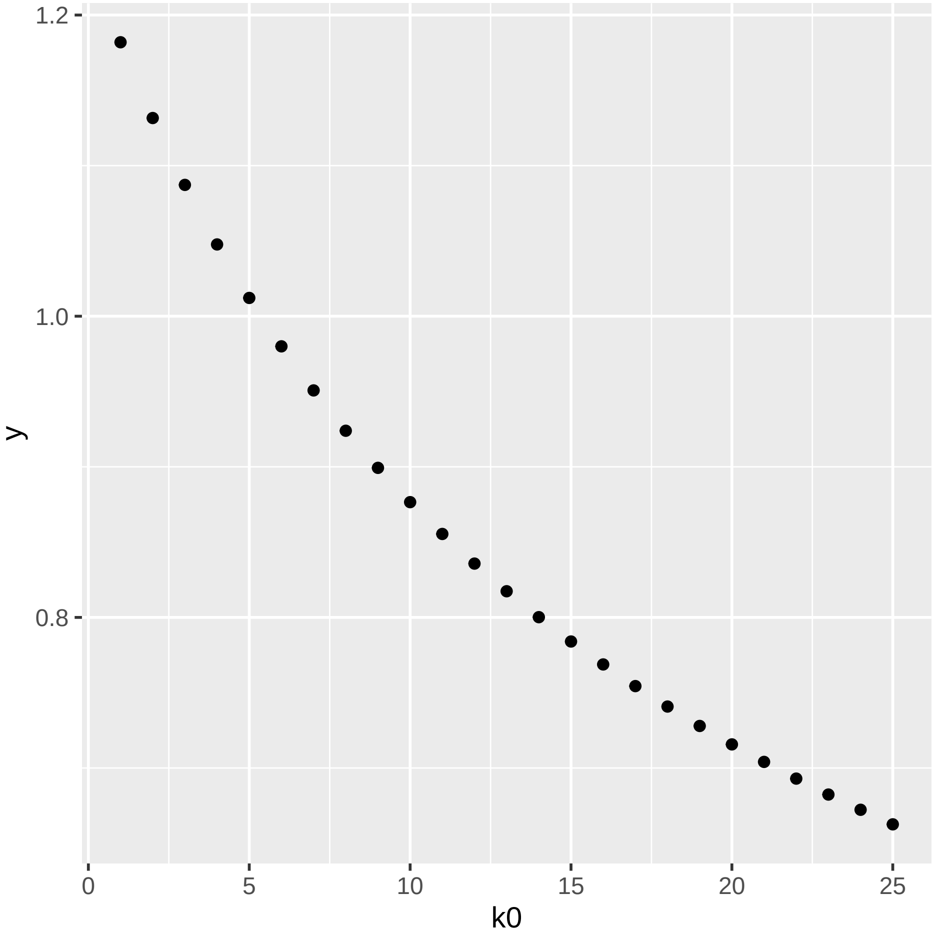
$= exp(-1/2(n + k_0)(\mu - \frac{n\bar{y}}{n+k_0}))$

$p(\mu|y_i, \sigma^2) = N(\frac{n\bar{y}}{n+k_0}, \frac{1}{n+k_0})$

Upper Bound: $1.96\sqrt{\frac{1}{n+k_0}}$ Lower Bound: $-1.96\sqrt{\frac{1}{n+k_0}}$

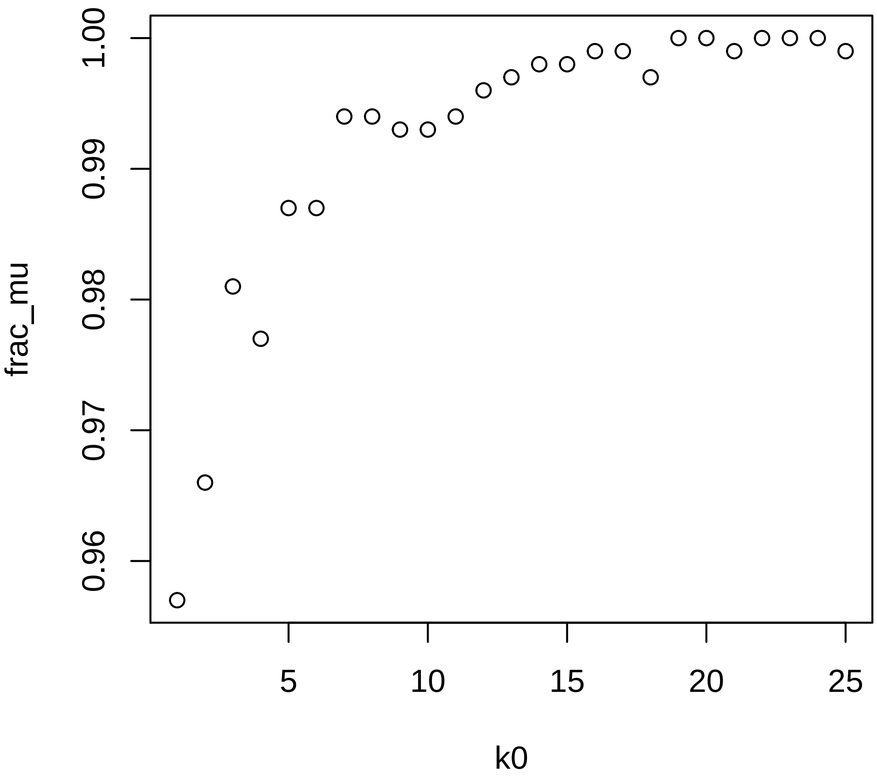
1b. Plot the length of the posterior credible interval as a function of κ_0, for $\kappa_0 = 1, 2, ..., 25$ assuming $n = 10$. Report how this prior parameter effects the length of the posterior interval and why this makes intuitive sense.

```
k0=seq(1,25)
n=10
values=c()
#Function
length_post_cred <- function(k0){
    2*1.96*sqrt(1/(n+k0))
}
#Plot
length_plot <- ggplot(data.frame(k0),aes(k0))+stat_function(fun=length_post_cred,geom='point',n=length(
length_plot
```



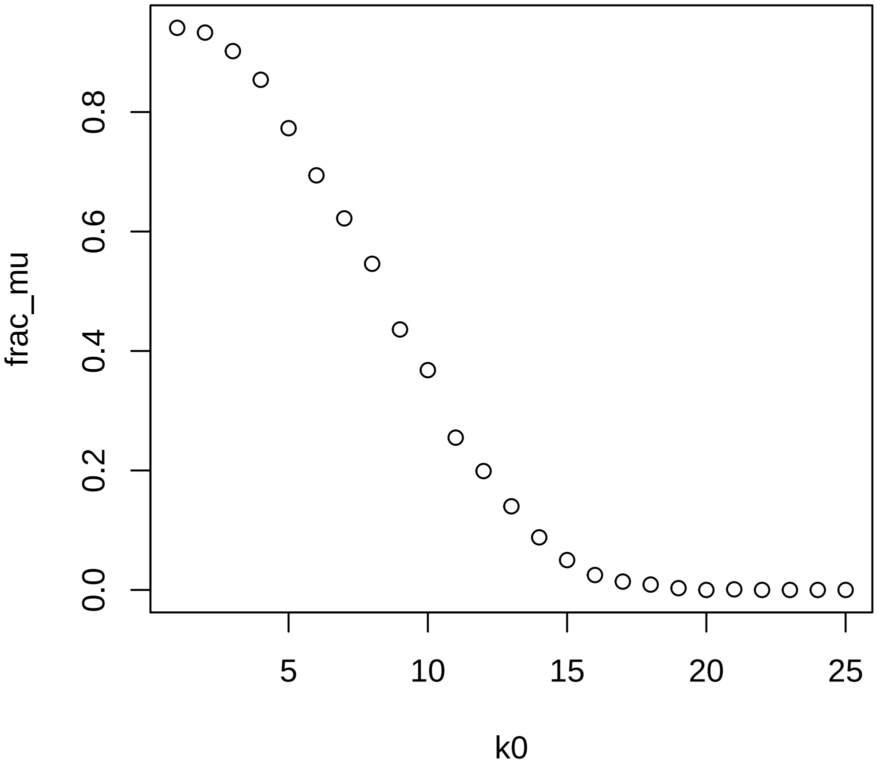
1c. Now we will evaluate the *frequentist coverage* of the posterior credible interval on simulated data. Generate 1000 data sets where the true value of $\mu = 0$ and $n = 10$. For each dataset, compute the posterior 95% interval endpoints (from the previous part) and see if it the interval covers the true value of $\mu = 0$. Compute the frequentist coverage as the fraction of these 1000 posterior 95% credible intervals that contain $\mu = 0$. Do this for each value of $\kappa_0 = 1, 2, ..., 25$. Plot the coverage as a function of κ_0.

```
true_mu <- 0
num_data <- 1000
dataset <- matrix(0,num_data,length(k0))
frac_mu <- rep(0,length(k0))
#function
for (k in k0) {
    for (data in seq(num_data)) {
        y <- rnorm(n,true_mu,1)
        post_mu <- (mean(y)*n/(k+n))
        cred_int <- qnorm(c(0.025,0.975),post_mu,sqrt(1/(k+n)))
        if(between(true_mu,cred_int[1],cred_int[2])==TRUE){
            dataset[data,k] <- 1
        }
    }
    frac_mu[k] <- sum(dataset[,k])/num_data
}
#plot
plot(k0,frac_mu)
```



1d. Repeat the 1c but now generate data assuming the true $\mu = 1$.

```
true_mu <- 1
num_data <- 1000
dataset <- matrix(0,num_data,length(k0))
frac_mu <- rep(0,length(k0))
#function
for (k in k0) {
    for (data in seq(num_data)) {
        y <- rnorm(n,true_mu,1)
        post_mu <- (mean(y)*n/(k+n))
        cred_int <- qnorm(c(0.025,0.975),post_mu,sqrt(1/(k+n)))
        if(between(true_mu,cred_int[1],cred_int[2])==TRUE){
            dataset[data,k] <- 1
        }
    }
    frac_mu[k] <- sum(dataset[,k])/num_data
}
#plot
plot(k0,frac_mu)
```



1e. Explain the differences between the coverage plots when the true $\mu = 0$ and the true $\mu = 1$. For what values of κ_0 do you see closer to nominal coverage (i.e. 95%)? For what values does your posterior interval tend to overcover (the interval covers the true value more than 95% of the time)? Undercover (the interval covers the true value less than 95% of the time)? Why does this make sense?

when $\mu = 0$, the posterior tends to over cover when k is greater than 2 and when it equals 1, it tends to undercover when k is greater than 2 aswell. as k increases towards 25, the intervals become smaller. Because of this, we can be much more certain of our prior belief that $\mu_0 = 0$

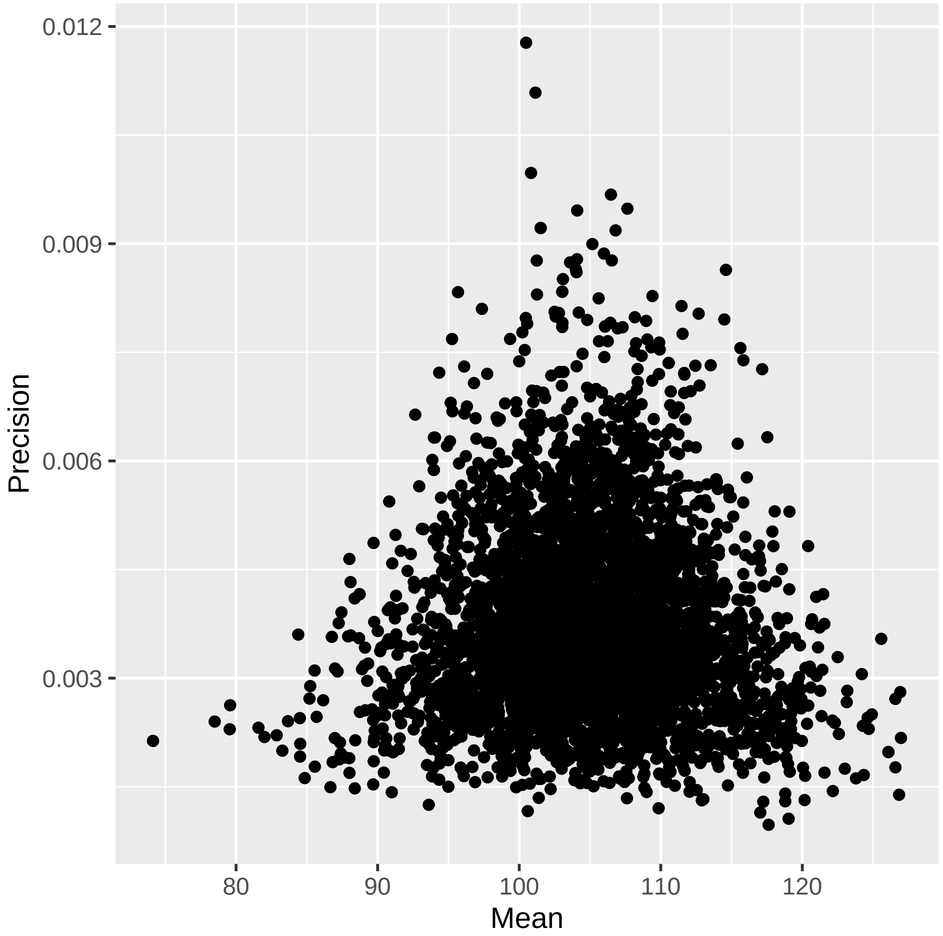
Bayesian inference for the normal distribution in Stan.

Create a new Stan file by selecting "Stan file" in the Rstudio menu. Save it as `IQ_model.stan`. We will make some basic modifications to the template example in the default Stan file for this problem. Consider the IQ example used from class. Scoring on IQ tests is designed to yield a N(100, 15) distribution for the general population. We observe IQ scores for a sample of n individuals from a particular town, $y_1, \ldots y_n \sim N(\mu, \sigma^2)$. Our goal is to estimate the population mean in the town. Assume the $p(\mu, \sigma) = p(\mu \mid \sigma)p(\sigma)$, where $p(\mu \mid \sigma)$ is $N(\mu_0, \sigma/\sqrt{\kappa_0})$ and $p(\sigma)$ is Gamma(a, b). Before you administer the IQ test you believe the town is no different than the rest of the population, so you assume a prior mean for μ of $\mu_0 = 100$, but you aren't to sure about this a priori and so you set $\kappa_0 = 1$ (the effective number of pseudo-observations). Similarly, a priori you assume σ has a mean of 15 (to match the intended standard deviation of the IQ test) and so you decide on setting $a = 15$ and $b = 1$ (remember, the mean of a Gamma is a/b). Assume the following IQ scores are observed:

```
y <- c(70, 85, 111, 111, 115, 120, 123)
n <- length(y)
```

2a. Make a scatter plot of the posterior distribution of the median, μ, and the precision, $1/\sigma^2$. Put μ on the x-axis and $1/\sigma^2$ on the y-axis. What is the posterior relationship between μ and $1/\sigma^2$? Why does this make sense? *Hint:* review the lecture notes.

```
#stan_model <- stan_model(file = "IQ_model.stan")
#a <- 15
#b <- 1
#k0 <- 1
#mu0 <- 100
#stan_fit <- rstan::sampling(stan_model,data=list(N=n,y=y,mu0=mu0,k0=k0,a=a,b=b),refresh=0)
#samples <- rstan::extract(stan_fit)
#save(samples,file='samples.Rdata')
load('samples.Rdata')
mu_samples <- samples$mu
sigma_samples <- samples$sigma
tibble(Mean=mu_samples,Precision=1/sigma_samples^2) %>%
    ggplot()+geom_point(aes(x=Mean,y=Precision))
```



when there is a high precision, there is a low posterior variability is μ. when there is a low precision, there is hgh uncertainty about μ. This makes sense because σ^2 is a measure of spread.

2b. You are interested in whether the mean IQ in the town is greater than the mean IQ in the overall population. Use Stan to find the posterior probability that μ is greater than 100.

```r
library(rstan)
y <- c(70, 85, 111, 111, 115, 120, 123)
n <- length(y)

post_prob_mu <- mean(mu_samples>100)
post_prob_mu
```

```
## [1] 0.78125
```

2c. You notice that two of the seven scores are significantly lower than the other five. You think that the normal distribution may not be the most appropriate model, in particular because you believe some people in this town are likely have extreme low and extreme high scores. One solution to this is to use a model that is more robust to these kinds of outliers. The Student's t distribution and the Laplace distribution are two so called "heavy-tailed distribution" which have higher probabilities of outliers (i.e. observations further from the mean). Heavy-tailed distributions are useful in modeling because they are more robust to outliers. Fit the model assuming now that the IQ scores in the town have a Laplace distribution, that is $y_1, \ldots, y_n \sim Laplace(\mu, \sigma)$. Create a copy of the previous stan file, and name it "IQ_laplace_model.stan". *Hint:* In the Stan file you can replace `normal` with `double_exponential` in the model section, another name for the Laplce distribution. Like the normal distribution it has two arguments, μ and σ. Keep the same prior distribution, $p(\mu, \sigma)$ as used in the normal model. Under the Laplace model, what is the posterior probability that the median IQ in the town is greater than 100? How does this compare to the probability under the normal model? Why does this make sense?

```
#laplace_stan_model <- stan_model("IQ_laplace_model.stan")
#laplace_stan_fit <- rstan::sampling(laplace_stan_model,data=list(N=n,y=y,mu0=mu0,k0=k0,a=a,b=b),refres
#laplace_samples <- rstan::extract(laplace_stan_fit)
#save(laplace_samples,file='laplace_samples.Rdata')
load('laplace_samples.Rdata')
laplace_mu_samples <- laplace_samples$mu
#probability
probability_laplace <- mean(laplace_mu_samples>100)
probability_laplace
```

```
## [1] 0.93
```

Logistic regression for pesticide toxicity data.

A environmental agency is testing the effects of a pesticide that can cause acute poisoning in bees, the world's most important pollinator of food crops. The environmental agency collects data on exposure to different levels of the pestidicide in parts per million (ppm). The agency also identifies collapsed beehives, which they expect could be due to acute pesticide poisoning. In the data they collect, each observation is pair (x_i, y_i), where x_i represents the dosage of the pollutant and y_i represents whether or not the hive survived. Take $y_i = 1$ means that the beehive has collapsed from poisoning and $y_i = 0$ means the beehive survived. The agency collects data at several different sites, each of which was exposed to a different dosages. The resulting data can be seen below:

```
x <- c(1.06, 1.41, 1.85, 1.5, 0.46, 1.21, 1.25, 1.09,
       1.76, 1.75, 1.47, 1.03, 1.1, 1.41, 1.83, 1.17,
       1.5, 1.64, 1.34, 1.31)

y <- c(0, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       1, 0, 0, 1, 1, 0, 0, 1, 1, 0)
```

Assume that beehiv collapse, y_i, given pollutant exposure level x_i, is $Y_i \sim \text{Bernoulli}(\theta(x_i))$, where $\theta(x_i)$ is the probability of death given dosage x_i. We will assume that $\text{logit}(\theta_i(x_i)) = \alpha + \beta x_i$ where $\text{logit}(\theta)$ is defined as $\log(\theta/(1-\theta))$. This model is known as *logistic regression* and is one of the most common methods for modeling probabilities of binary events.

3a. Solve for $\theta_i(x_i)$ as a function of α and β by inverting the logit function. If you haven't seen logistic regression before (it is covered in more detail in PSTAT 127 and PSTAT131), it is essentially a generalization of linear regression for binary outcomes. The inverse-logit function maps the linear part, $\alpha + \beta x_i$, which can be any real-valued number into the interval [0, 1] (since we are modeling probabilities of binary outcome, we need the mean outcome to be confined to this range).

$logit(\theta_i(x_i)) = \alpha + \beta x_i$

using $logit(\theta) = log(\frac{\theta}{1-\theta}$,

$log(\frac{\theta_i(x_i)}{1-\theta_i(x_i)}) = \alpha + \beta x_i$

$\frac{\theta_i(x_i)}{1-\theta_i(x_i)} = e^{\alpha+\beta x_i}$

$\theta_i(x_i) = e^{\alpha+\beta x_i} - \theta_i(x_i)e^{\alpha+\beta x_i}$

$\theta_i(x_i) + \theta_i(x_i)e^{\alpha+\beta x_i} = e^{\alpha+\beta x_i}$

$\theta_i(x_i)(1 + e^{\alpha+\beta x_i}) = e^{\alpha+\beta x_i}$

$\theta_i(x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$

3b The dose at which there is a 50% chance of beehvive collapse, $\theta(x_i) = 0.5$, is known as LD50 ("letha dose 50%"), and is often of interest in toxicology studies. Solve for LD50 as a function of α and β.

$\theta_i(x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = 0.5$

$e^{\alpha+\beta x_i} = 0.5(1 + e^{\alpha+\beta x_i}) = 0.5 + 0.5(e^{\alpha+\beta x_i})$

$0.5 = e^{\alpha+\beta x_i} - 0.5(e^{\alpha+\beta x_i})$

$0.5 = e^{\alpha+\beta x_i}(1 - 0.5)$

$1 = e^{\alpha+\beta x_i}$

$ln(1) = \alpha + \beta x_i$

$-\alpha = \beta x_i$

$\frac{-\alpha}{\beta} = x_i$

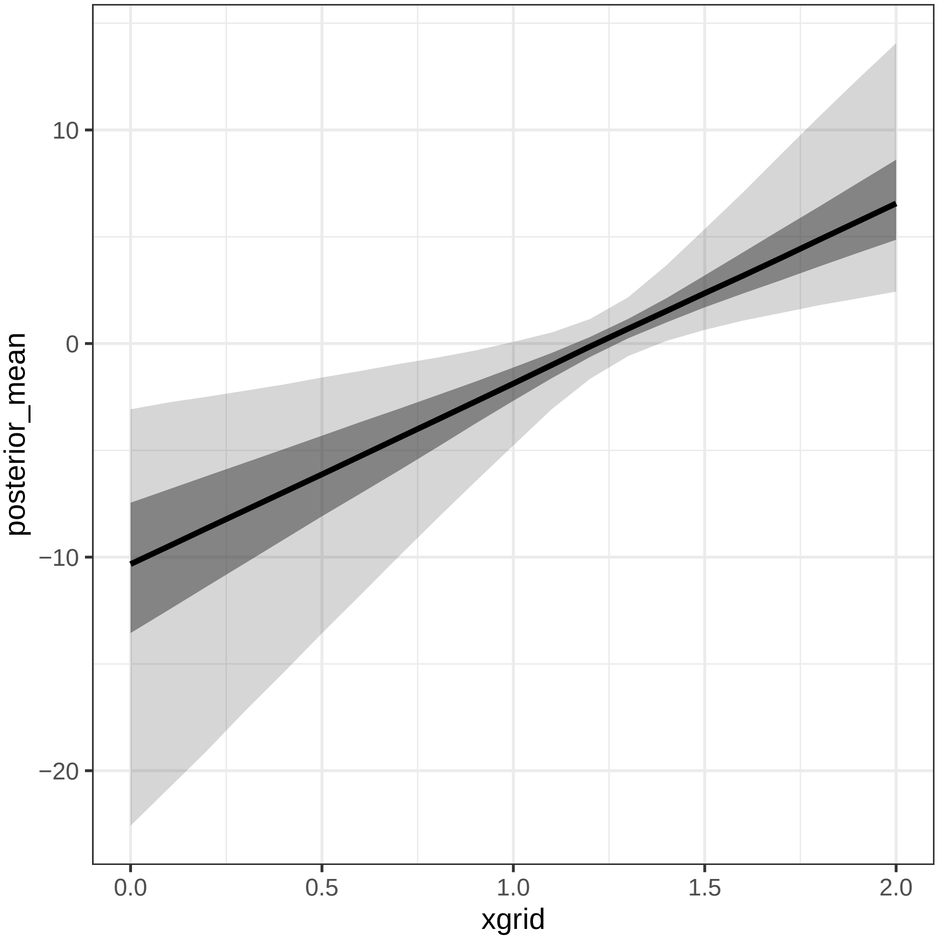
3c Implement the logistic regression model in stan by reproducing the stan model described here: https://mc-stan.org/docs/2_18/stan-users-guide/logistic-probit-regression-section.html. Run the stan model on the beehive data to get Monte Carlo samples. Compute Monte Carlo samples of the LD50 by applying the function derived in the previous part to your α and β samples. Report and estimate of the posterior mean of the LD50 by computing the sample average of all Monte Carlo samples of LD50.

```
n <- length(y)
#log_stan_model <- stan_model("log_model.stan")
#log_stan_fit <- rstan::sampling(log_stan_model,data=list(N=n,x=x,y=y),refresh=0)
#log_samples <- rstan::extract(log_stan_fit)
#save(log_samples, file='log_samples.Rdata')
load('log_samples.Rdata')
beta_samples <- log_samples$beta
alpha_samples <- log_samples$alpha

#Posterior mean
mean_LD50 <- mean(-alpha_samples/beta_samples)
mean_LD50
```

```
## [1] 1.211128
```

3d. Make a plot showing both 50% and 95% confidence band for the probability of a hive collapse as a function of pollutant exposure, $\Pr(y = 1 \mid \alpha, \beta, x)$. Plot your data on a grid of x-values from $x = 0$ to 2. *Hint:* see lab 7 for a similar example.

```
xgrid <- seq(0,2,0.1)
compute_curve <- function(sample){
    alpha <- sample[1]
    beta <- sample[2]
    y_values <- alpha+beta*xgrid
}
res <- apply(cbind(alpha_samples,beta_samples),1,compute_curve)
quantiles <- apply(res,1,function(x) quantile(x,c(0.025,0.25,0.75,0.975)))

posterior_mean <- rowMeans(res)
posterior_mean <- apply(res,1,median)
tibble(x=xgrid,
       q025=quantiles[1,],
       q25=quantiles[2,],
       q75=quantiles[3,],
       q975=quantiles[4,],
       mean=posterior_mean) %>%
    ggplot()+
    geom_ribbon(aes(x=xgrid,ymin=q025,ymax=q975),alpha=0.2)+
    geom_ribbon(aes(x=xgrid,ymin=q25,ymax=q75),alpha=0.5)+
    geom_line(aes(x=xgrid,y=posterior_mean),size=1)+
    theme_bw()
```


posterior_mean
xgrid