Assignment 2 with Solutions

STAT314/STAT461

Set: Tue, Aug-10. Due: Fri Aug-20

NB. The exact numbers will depend on your seed for the random number generator as well as on your prior assumptions. However, given the amount of the data, the prior assumptions should not make too much of a difference (and neither should the seed).

Let's get the data:

You may need to install the package `palmerpenguins` either via the Packages menu or directly via the Console before you start this exercise.

```
install.packages("palmerpenguins")
```

```
library(palmerpenguins)
data(penguins)
```

Remember some basic commands such as

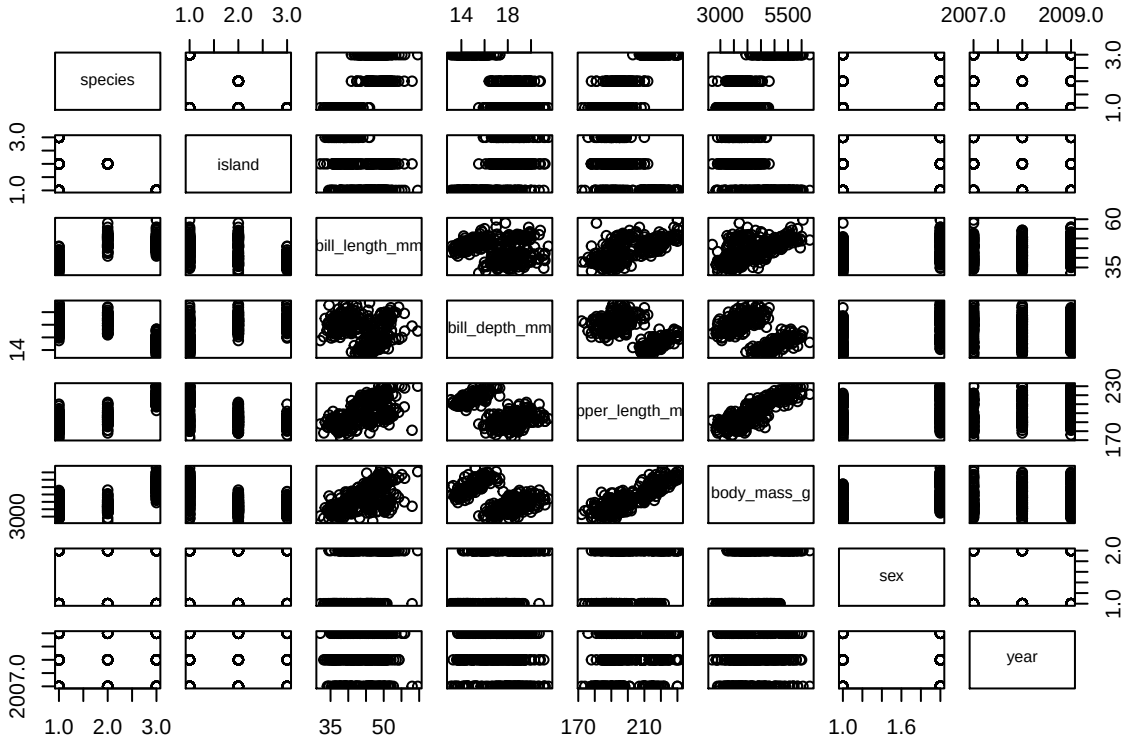
```
str(penguins)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    344 obs. of  8 variables:
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num  39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num  18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int  181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int  3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

```
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
plot(penguins)
```



for a quick look at the data. Also, keep in mind that there are always several different ways to do things in R, so if you prefer to use other commands/packages, it's fine as long as it works quickly enough and correctly.

Problem 1. Gaussian model,

For now, we are going to focus on **male Adelie** penguins only. So let's create a subset

```
dmA <- penguins[(penguins$species=='Adelie')&(penguins$sex=='male'),]
```

and let's drop any lines which contain missing body mass:

```
dmA <- dmA[!is.na(dmA$body_mass_g),]
```

Part (a):

Let x_i represent the body mass. Assuming the Bayesian conjugate model:

$$x_i|\mu \sim N(\mu, 350^{-2})$$

and

$$\mu \sim N(\mu_0, \tau_0).$$

Encyclopedia Britannica says that Adelie penguins weight 4-6 kg with males being slightly heavier. So, I am going to go for $\mu_0 = 5000$ and $\tau_0 = 1/500^2$. (Your assumptions might be different. Note, that the 500

is probably too wide a range, more appropriate for individual penguin weight rather than the population average. However, not knowing much about them, I prefer to err on the side of caution.)

The posterior distribution can then be derived as

$$\mu|x \sim N(\frac{n\bar{x}\tau + \mu_0\tau_0}{n\tau + \tau_0}, n\tau + \tau_0).$$

So, we need island-specific n and $\bar{x}$:

```
(n <- table(dmA$island))
```

```
##
##    Biscoe     Dream Torgersen
##        22        28        23
```

and

```
(x.bar <- tapply(dmA$body_mass_g,dmA$island,mean))
```

```
##    Biscoe     Dream Torgersen
##  4050.000  4045.536  4034.783
```

We can then obtain the parameters for the posterior distribution. The mean:

```
mu0 <- 5000; tau0 <- 1/500^2
tau <- 1/350^2
(post.mean <- (n*x.bar*tau+mu0*tau0)/(n*tau+tau0))
```

```
##
##    Biscoe     Dream Torgersen
##  4070.698  4061.952  4054.917
```

... and the precision, which can also be transformed into standard deviation:

```
(post.prec <- (n*tau+tau0))
```

```
##
##       Biscoe        Dream    Torgersen
## 0.0001835918 0.0002325714 0.0001917551
```

```
1/sqrt(post.prec)
```

```
##
##    Biscoe     Dream Torgersen
##  73.80288  65.57251  72.21485
```

So, my posterior distributions for the average population weights are:

$$\mu_{\text{Biscoe}}|x \sim N(4070.70, 73.8^{-2})$$

$$\mu_{\text{Dream}}|x \sim N(4061.95, 65.6^{-2})$$

and

$$\mu_{\text{Torgersen}}|x \sim N(4054.92, 72.2^{-2})$$

Part (b):

Based on the above, the Biscoe island has the heaviest penguins on average. But let's use simulations to figure out how certain we are about it. Let's simulate 10^4 observations from the island-specific posterior distributions, and see which island wins in each.

```
mu.Biscoe <- rnorm(10^4,post.mean[1],1/sqrt(post.prec[1]))
mu.Dream <- rnorm(10^4,post.mean[2],1/sqrt(post.prec[2]))
mu.Torgersen <- rnorm(10^4,post.mean[3],1/sqrt(post.prec[3]))

mu.array <- cbind(mu.Biscoe,mu.Dream,mu.Torgersen)

heaviest <- apply(mu.array,1,which.max)
table(heaviest)/10^4
```

```
## heaviest
##      1      2      3
## 0.3893 0.3239 0.2868
```

Seems like the first island, Biscoe, has the heaviest male Adelie penguins, on average, ($P = 0.3893$).

Problem 2. Population means vs. random individual penguins.

The Gentoo penguins were only found on Biscoe island. Let's look at them now.

(a) Again, assuming Gaussian distribution for the body mass, derive posterior island-specific distributions for the mean population weight of male and female Gentoo penguins respectively. Explain your prior assumptions. (**1 pt**)

Similar to the above:

```
d.Gentoo <- penguins[penguins$species=='Gentoo',]
d.Gentoo <- d.Gentoo[!is.na(d.Gentoo$body_mass_g),]
```

```
(n <- table(d.Gentoo$sex))
```

```
##
## female   male
##     58     61
```

and

```
(x.bar <- tapply(d.Gentoo$body_mass_g,d.Gentoo$sex,mean))
```

```
##   female     male
## 4679.741 5484.836
```

Wikipedia says that the average weight of a Gentoo penguin is about 6.5 kg. And, again, I am going to give it a fairly vague prior $\pm 2kg$.

We can then obtain the parameters for the posterior distribution. The mean:

```
mu0 <- 6500; tau0 <- 1/500^2
tau <- 1/500^2
(post.mean <- (n*x.bar*tau+mu0*tau0)/(n*tau+tau0))
```

```
##
##   female     male
## 4710.593 5501.210
```

```
(post.tau <- (n*tau+tau0))
```

```
##
```

```
##   female     male
## 0.000236 0.000248
```

```
1/sqrt(post.tau)
```

```
##
##   female     male
## 65.09446 63.50006
```

(b) What is the posterior probability that the males are **on average** heavier than females? (**1 pt**)

Let's simulate from posterior distribution for mu and see how often the random realisation for males is greater than that for females:

```
mu.F <- rnorm(10^4,post.mean[1],1/sqrt(post.tau[1]))
mu.M <- rnorm(10^4,post.mean[2],1/sqrt(post.tau[2]))

mean(mu.M > mu.F)
```

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

It's always true for our 10^4 simulations. So, in the report, I would put the following:

- The Gentoo male penguins are on average heavier than the females $P > .9999$.*

(c) What is the posterior probability that a random individual male is heavier than a random individual female? (Hint: use posterior predictive distribution.) (**1pt**)

Now, let's simulate from the posterior predictive distribution for $\tilde{x}$. That means, we first simulate μ (already done above), and then simualte $\tilde{x}$ from the likelihood conditional on μ.

```
x.tilde.F <- rnorm(10^4,mu.F,500)
x.tilde.M <- rnorm(10^4,mu.M,500)

mean(x.tilde.M > x.tilde.F)
```

```
## [1] 0.8644
```

Now, the probability that a random male penguin is heavier than a random female penguin is only 0.86. Because there is variability among individual penguins. (Try plotting posterior density for μ and posterior predictive density for $\tilde{x}$ to compare the overlap.)

Problem 3. Simple Linear Regression.

getting rid of incomplete observations

```
penguins.clean <- penguins[(!is.na(penguins$bill_length_mm))&
                           (!is.na(penguins$bill_depth_mm))&
                           (!is.na(penguins$species)),]
```

and fitting a simple linear model:

$$\text{Length}_i = a + b\text{Depth}_i + \epsilon_i$$

```
library(MCMCglmm)
m1 <- MCMCglmm(bill_length_mm ~ bill_depth_mm, data=penguins.clean,verbose=F)
summary(m1)$sol
```

```
##                post.mean   l-95% CI   u-95% CI eff.samp pMCMC
## (Intercept)   55.2303672 50.2469309 59.3339760     1000 0.001
## bill_depth_mm -0.6599362 -0.9140136 -0.3889183     1000 0.001
```

Each additional mm of bill depth is associated with an average 0.65 mm decrease in bill length. Note, that the 95% CI does not include 0, so we are pretty certain about this. Hmm... Shouldn't they be positively correlated!?

(b) Now, let's fit a model which takes species into account:

```
m2 <- MCMCglmm(bill_length_mm ~ species*bill_depth_mm,
               data=penguins.clean,verbose=F)
summary(m2)$sol
```

```
##                                 post.mean    l-95% CI  u-95% CI eff.samp pMCMC
## (Intercept)                    22.9827812  17.2527073 28.466232 1000.000 0.001
## speciesChinstrap               -9.4740306 -19.6080536  1.716248 1000.000 0.076
## speciesGentoo                  -5.7365071 -14.8871570  2.610391 1093.413 0.198
## bill_depth_mm                   0.8616577   0.5645156  1.180486 1000.000 0.001
## speciesChinstrap:bill_depth_mm  1.0553878   0.4820975  1.649581 1000.000 0.001
## speciesGentoo:bill_depth_mm     1.1586908   0.5849837  1.644253 1000.000 0.001
```

So, for Adelie penguins, each additional mm of bill depth is associated with an average 0.86 mm increase in bill length. For Chinstrap penguins, each additional mm of bill depth is associated with an average $1.06 + 0.86$ mm increase in bill length. And for Gentoo penguins, each additional mm of bill depth is associated with an average $1.16 + 0.86$ mm increase in bill length.

There is, in fact strong evidence for positive correlation between the variables.

(c) Let's plot the data and the two models. (There are various ways to do this.)

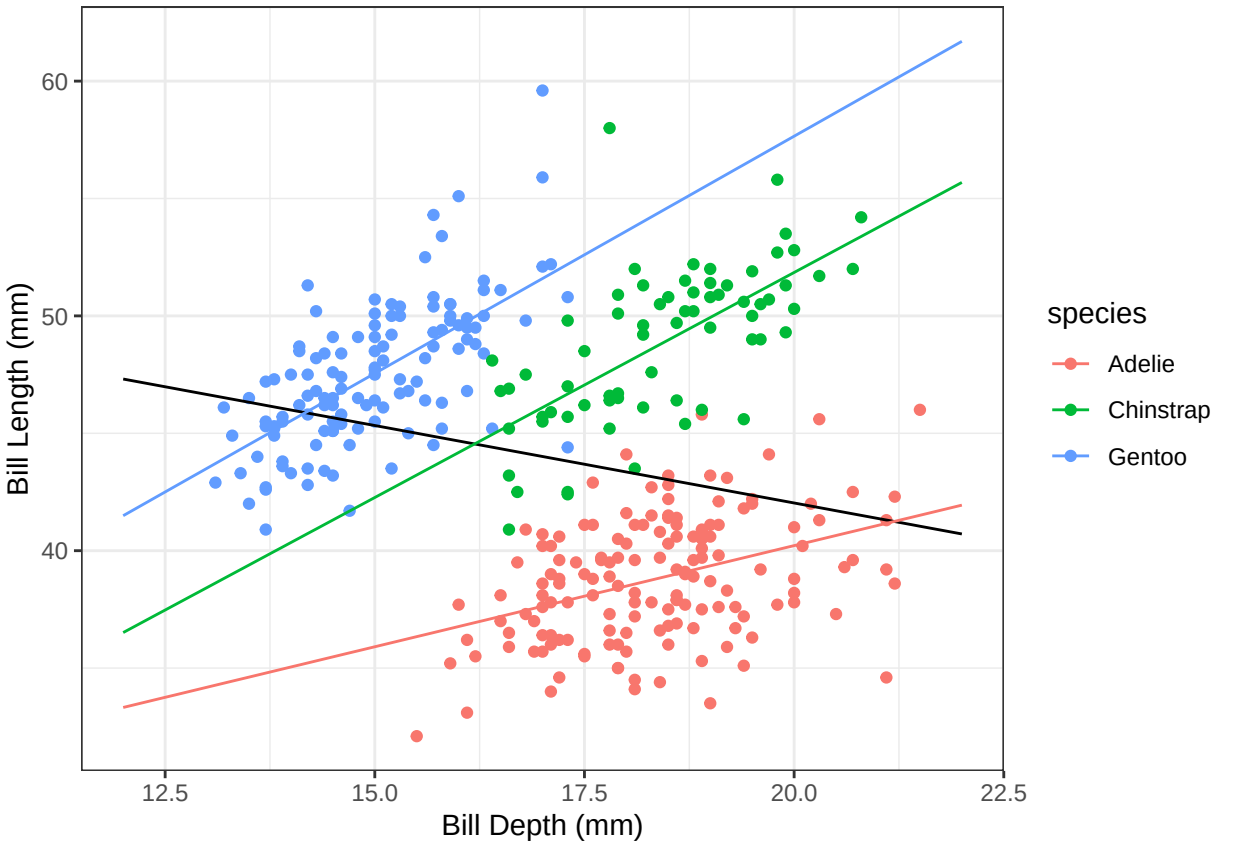
```
data.pred <- expand.grid(species=unique(penguins.clean$species),
                         bill_depth_mm=seq(12,22,.1))

# to get the posterior means for a and b from the first model
a1 <- summary(m1)$sol[1,1]
b1 <- summary(m1)$sol[2,1]

# and for the second model
a2 <- summary(m2)$sol[1:3,1];
  # adding the baseline
  a2[2:3] <- a2[2:3]+a2[1]
b2 <- summary(m2)$sol[4:6,1]
  b2[2:3] <- b2[2:3]+b2[1]


data.pred$m1 <- a1+b1*data.pred$bill_depth_mm
data.pred$m2 <- a2[as.numeric(data.pred$species)]+
                     b2[as.numeric(data.pred$species)]*data.pred$bill_depth_mm
```

```
library(ggplot2)
ggplot(data=penguins.clean,aes(x=bill_depth_mm,y=bill_length_mm))+
  geom_point(aes(group=species,col=species))+
  geom_line(data=data.pred,aes(y=m1))+
  geom_line(data=data.pred,aes(y=m2,group=species,col=species))+
  xlab('Bill Depth (mm)')+
  ylab('Bill Length (mm)')+
  theme_bw()
```



And therein lies the lesson: there is no point in fitting bi-variate simple regressions to a complex phenomenon. Start with the "complicated" model. Otherwise, you may get completely spurious correlations (and omit influential variables).