MAST20005/MAST90058: Week 11 Lab

Goals: Bayesian inference for: (i) proportions; (ii) mean of a normal distribution; (iii) simple linear regression.

1 Bayesian estimation of proportions

Let Y be the number of successes in n independent Bernoulli trials. Then the data and prior distributions are, respectively,

$$Y \mid \theta \sim \text{Bi}(n, \theta), \quad \theta \sim \text{Beta}(\alpha, \beta).$$

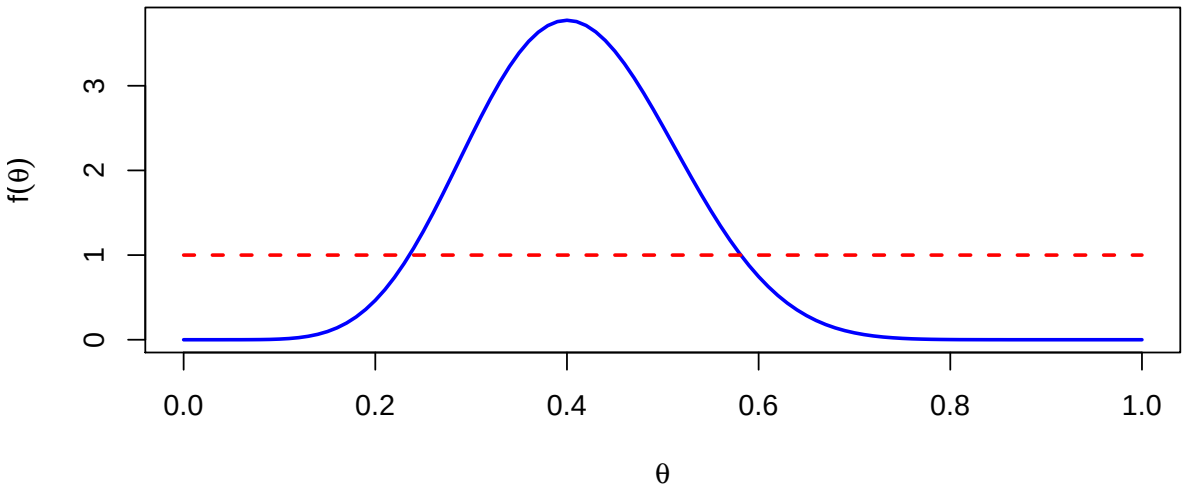
The posterior pdf can be obtained as,

$$f(\theta \mid y) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + y)\Gamma(\beta + n - y)} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \quad 0 < \theta < 1,$$

i.e. $\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$.

1. Suppose we observe $y = 8$ successes out of $n = 20$ trials. Plot the prior and posterior densities assuming $\alpha = \beta = 1$. What kind of information is given by this prior? Comment on the information gain on θ gained after observing the data.

```
a <- 1
b <- 1
y <- 8
n <- 20
theta <- seq(0, 1, 0.01)
prior     <- dbeta(theta, a, b)
posterior <- dbeta(theta, a + y, b + n - y)
par(mar = c(4, 4, 1, 1))  # tighter margins
plot(theta, posterior, type = "l", lwd = 2, col = "blue",
     xlab = expression(theta),
     ylab = expression(f(theta)))
points(theta, prior, type = "l", lty = 2, lwd = 2, col = "red")
```



Note that $\text{Beta}(1, 1) = \text{Unif}(0, 1)$, meaning that we consider all values of θ to be equally plausible before observing any data. After observing the data we obtain a posterior centered around 0.4 and the uncertainty on θ is much reduced.

2. Find the posterior mean.

```
(a + y) / (a + b + n)

## [1] 0.4090909
```

3. Find the central 95% credible interval.

```
qbeta(c(0.025, 0.975), a + y, b + n - y)

## [1] 0.2181969 0.6156456
```

4. Estimate the posterior odds $\mathbb{E}\left(\frac{\theta}{1-\theta} \mid y\right)$ by simulating from the posterior distribution.

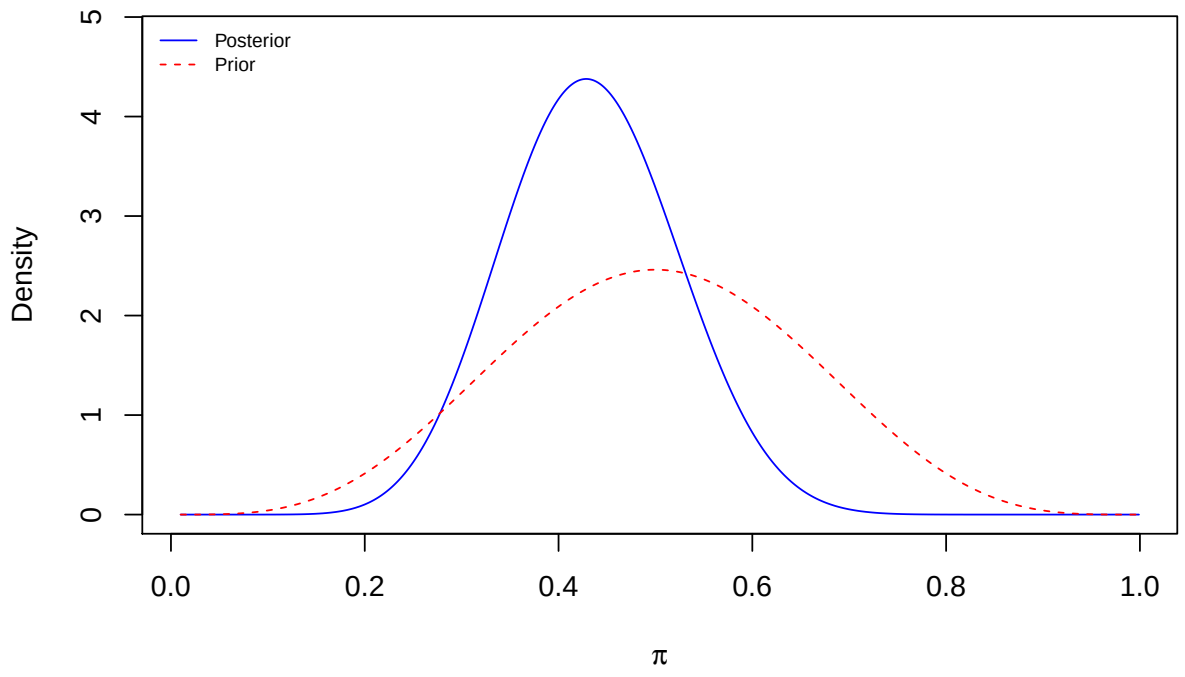
```
sample.p <- rbeta(1000, a + y, b + n - y)
mean(sample.p / (1 - sample.p))

## [1] 0.7544855
```

In the above code, we generated 1000 observations from the posterior, $\theta_1, \ldots, \theta_{1000} \sim \text{Beta}(\alpha + y, \beta + n - y)$ and then obtained,

$$E\left(\frac{\theta}{1-\theta} \,\middle|\, y\right) \approx \frac{1}{1000} \sum_{i=1}^{1000} \frac{\theta_i}{1-\theta_i}.$$

5. The package `Bolstad` contains some useful routines for simple Bayesian models (if it is not installed, run `install.packages("Bolstad")` to install it). For example the following carries out a Bayesian analysis for proportions:

```
library(Bolstad)
par(mar = c(4, 4, 1, 1))  # tighter margins
binobp(8, 20, 5, 5, plot = TRUE)
```



```
## Posterior Mean            :  0.4333333
## Posterior Variance        :  0.0079211
## Posterior Std. Deviation  :  0.0890008
##
## Prob. Quantile
## ---------------
## 0.005 0.2190745
## 0.010 0.2369340
## 0.025 0.2644553
## 0.050 0.2892715
## 0.500 0.4318325
## 0.950 0.5825361
## 0.975 0.6106372
## 0.990 0.6425740
## 0.995 0.6637734
```

Note that the third and fourth arguments in `binobp` are α and β. Try different values for the prior parameters and report how the information affects the posterior distribution. Particularly, recall that the posterior mean can be written as,

$$\mathbb{E}\left(\theta \mid y\right) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

i.e. a weighted average between the prior mean $\alpha/(\alpha+\beta)$ and the MLE y/n.

6. The estimator $\tilde{\theta} = (Y+2)/(n+4)$ is sometimes used instead of the sample proportion $\hat{\theta} = Y/n$. Note that $\tilde{\theta}$ is the posterior mean when using Beta$(2,2)$ as the prior. Using this prior, compare the central 95% credible interval with the confidence interval $\tilde{\theta} \pm 1.96 \times \sqrt{\tilde{\theta}(1-\tilde{\theta})/n}$, by simulation for different values of the true proportion θ. The following code computes estimates of the coverage probability for the two types of intervals.

```
theta <- 0.1 # true parameter
n <- 20      # sample size
a <- 2       # prior parameter (alpha)
b <- 2       # prior parameter (beta)
nsim <- 1000 # number of simulations

# Simulate data.
y <- rbinom(nsim, size = n, prob = theta)

# Calculate limits for credible interval.
l1 <- qbeta(0.025, a + y, b + n - y)
u1 <- qbeta(0.975, a + y, b + n - y)

# Calculate limits for confidence interval.
p.tilde <- (y + 2) / (n + 4)
se <- sqrt(p.tilde * (1 - p.tilde) / n)
l2 <- p.tilde - 1.96 * se
u2 <- p.tilde + 1.96 * se
```

```r
# Calculate coverage probabilities.
mean(l1 < theta & theta < u1)
```

```
## [1] 0.868
```

```r
mean(l2 < theta & theta < u2)
```

```
## [1] 0.986
```

Comment on the accuracy of the interval estimates obtained by the two methods for different choices of θ and n.

2 Normal model

To illustrate how one might use the posterior distribution for inference, consider a study analysing the effect of specific training programs before a standardised test (Gelman, 2006). Assume a random sample of 240 differences of scores. Let $y_i = z_{2i} - z_{1i}$, where z_{i1} is the score for individual i before being coached, and the score z_{2i} is recorded after being coached. Suppose that the sample mean of these differences is $\bar{y} = 7.87$ and the standard deviation for the differences, $\sigma = 64.5$, is known. Particularly, we assume the data model $\bar{y} \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n)$ and prior $\theta \sim \mathrm{N}(\theta_0, \sigma_0^2)$. Then the posterior distribution is a normal distribution with mean,

$$E(\theta \mid \bar{y}) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\, y + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\, \theta_0$$

and variance,

$$\mathrm{var}(\theta \mid \bar{y}) = \frac{(\sigma^2/n)\, \sigma_0^2}{\sigma^2/n + \sigma_0^2}.$$

1. What are the implications of a large value of σ_0^2?
2. Find a 95% credible interval for θ when $\sigma_0^2 \to \infty$ and compare such an interval with the confidence interval for θ. Note that the posterior in this case is,

$$\theta \mid \bar{y} \sim \mathrm{N}(\bar{y}, \sigma^2/n).$$

Thus a 95% credible interval is obtained as follows:

```r
post.mean <- 7.87   # posterior mean
post.mean
```

```
## [1] 7.87
```

```r
post.sd <- 64.5 / sqrt(240)  # posterior standard deviation
post.sd
```

```
## [1] 4.163457
```

```r
qnorm(c(0.025, 0.975), post.mean, post.sd)
```

```
## [1] -0.290226 16.030226
```

Note that this is the same as the confidence interval $\bar{y} \pm 1.95\,\sigma/\sqrt{n}$, which we usually calculate using:

```
7.87 + c(-1, 1) * 1.96 * 64.5 / sqrt(240)

## [1] -0.2903759 16.0303759
```

3. How does the Bayesian interval change if $\sigma_0 = 5.1$ and the prior mean is $\theta_0 = 15.2$?

4. This example can be extended to illustrate how one can use a Bayesian approach to consider differences in means of multiple populations as in ANOVA. Suppose coaching programs were implemented in 4 schools, with school averages given in the following table.

School	i	Difference ($\bar{y}_{i\cdot}$)	Squared standard error (s_i^2/n_i)
A	1	28	15
B	2	8	10
C	3	-3	16
D	4	7	11

Assume:

$$y_{ij} \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$$

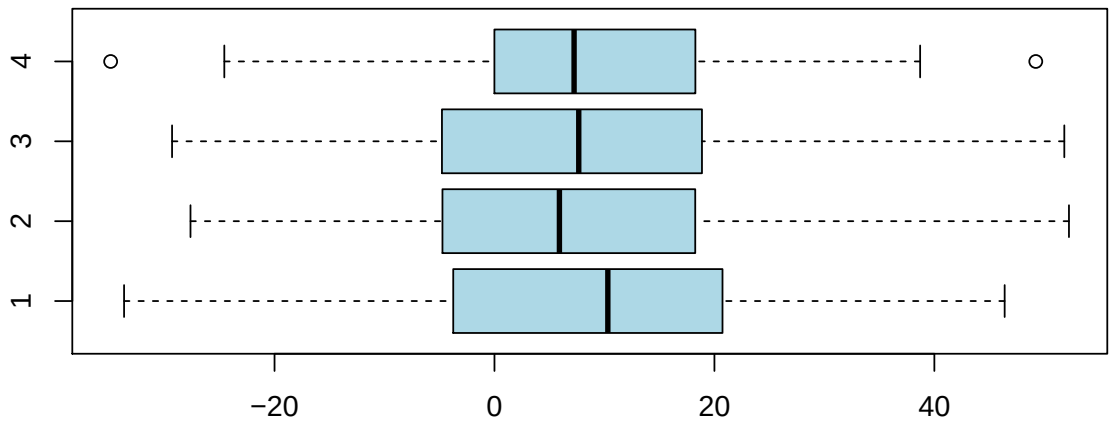
where y_{ij} is the jth observation from the ith school, θ_i is the mean effect of school i and σ_i^2 is the known variance of school i. Further, assume normal priors for each θ_i so that

$$\theta_i \sim \mathrm{N}(7.87, 17.35^2)$$

for all $i = 1, \ldots, 4$.

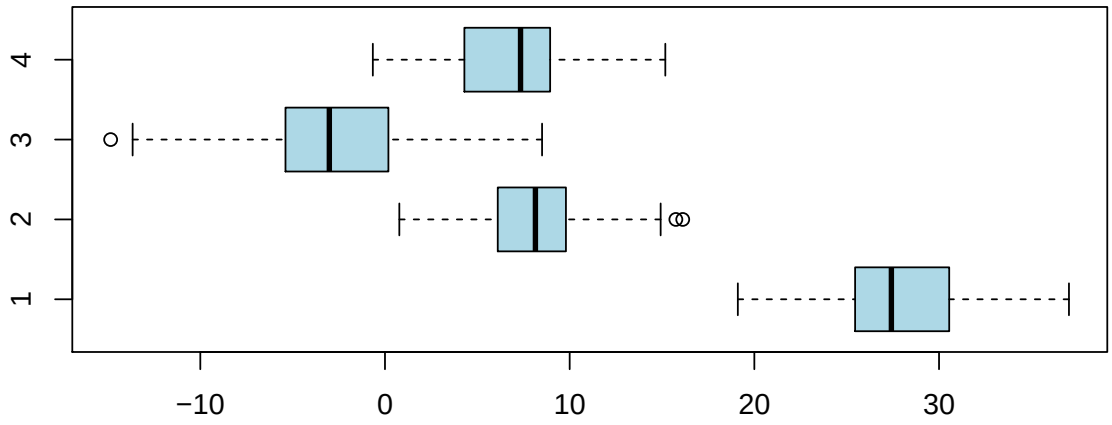
Give a graphical illustration of the prior and posterior distributions using boxplots for each school. A graph is simply obtained by sampling from the priors:

```
A <- rnorm(100, 7.87, 17.35)
B <- rnorm(100, 7.87, 17.35)
C <- rnorm(100, 7.87, 17.35)
D <- rnorm(100, 7.87, 17.35)
par(mar = c(4, 4, 1, 1))  # tighter margins
boxplot(A, B, C, D, horizontal = TRUE, col = "lightblue")
```



To represent the posteriors, we assume $\sigma_i^2/n_i = s_i^2/n_i$ and draw from the posteriors:

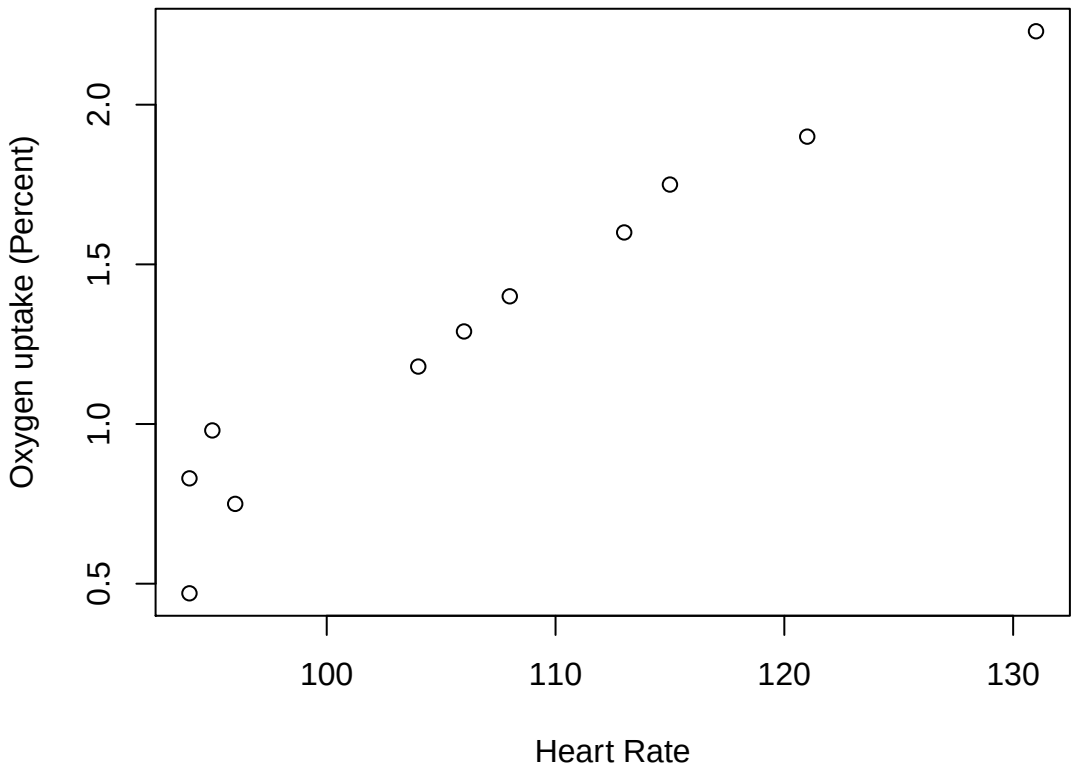
```
y.bar <- c(28, 8, -3, 7)
se2 <- c(15, 10, 16, 11)
theta.0 <- rep( 7.87, 4)
sigma.0 <- rep(17.35, 4)
post.means <- (sigma.0^2 / (sigma.0^2 + se2)) * y.bar +
              (      se2 / (sigma.0^2 + se2)) * theta.0
post.sd <- sqrt(sigma.0^2 * se2 / (sigma.0^2 + se2))
A <- rnorm(100, post.means[1], post.sd[1])
B <- rnorm(100, post.means[2], post.sd[2])
C <- rnorm(100, post.means[3], post.sd[3])
D <- rnorm(100, post.means[4], post.sd[4])
par(mar = c(4, 4, 1, 1))  # tighter margins
boxplot(A, B, C, D, horizontal = TRUE, col = "lightblue")
```



3 Bayesian regression

In this section we look at the familiar regression model $Y_i \sim \mathrm{N}(\alpha+\beta(x_i-\bar{x}), \sigma^2)$ from a Bayesian perspective. To this end, one can consider priors $\alpha \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$, and $\beta \sim \mathrm{N}(\mu_\beta, \sigma_\beta^2)$, while σ is regarded as known. The function `bayes.lin.reg()` computes posterior distributions for α and β and it enables us to specify either specific parameters for the prior or improper uniform ('flat') priors. If unspecified, the regression variance σ^2 is estimated as usual from the residuals. Consider 11 measurements on heart rate vs oxygen uptake.

```
O2 <- c(0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23)
HR <- c(94, 96, 94, 95, 104, 106, 108, 113, 115, 121, 131)
plot(HR, O2, xlab = "Heart Rate", ylab = "Oxygen uptake (Percent)")
```

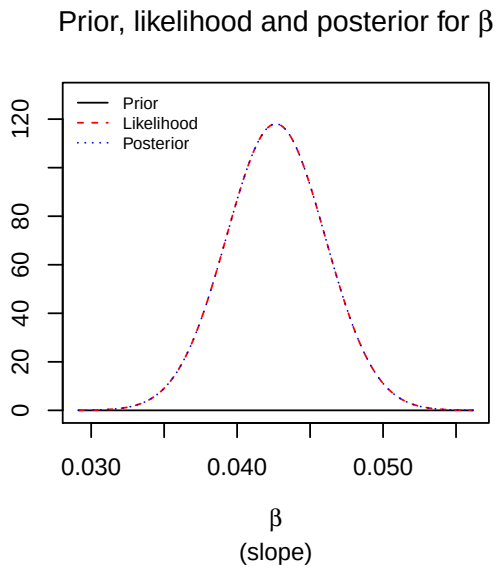


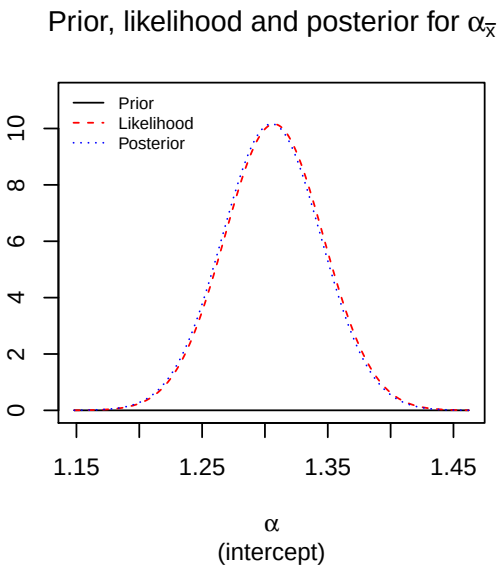

```
x <- HR - mean(HR)
y <- O2
coef(summary(lm(y ~ x)))  # usual regression fit

##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 1.30727273 0.039282918 33.27840 9.840739e-11
## x           0.04265141 0.003379797 12.61952 5.009294e-07
```

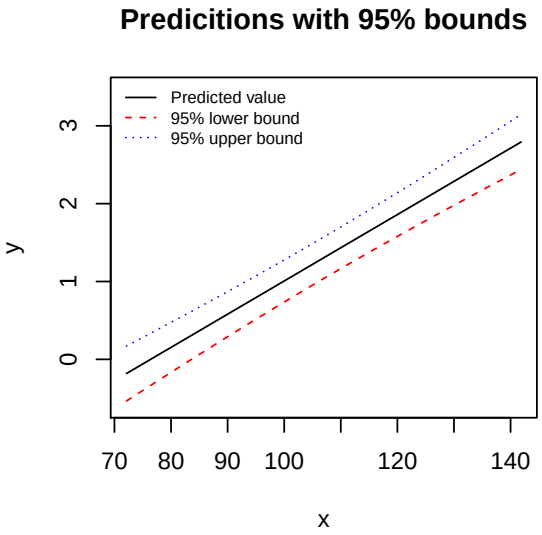
Next compute posterior distributions for the parameters under flat and informative priors. Explore the effect of different choices of prior on the final inference for the regression slope.

1. Fit the Bayesian regression model with flat priors.

```
bayes.lin.reg(y = O2, x = HR, slope.prior = "flat", intcpt.prior = "flat")

## Standard deviation of residuals:  0.13
##              Posterior Mean Posterior Std. Deviation
##              -------------- ------------------------
## Intercept:   1.305          0.039253
## Slope:       0.04265        0.0033798
```

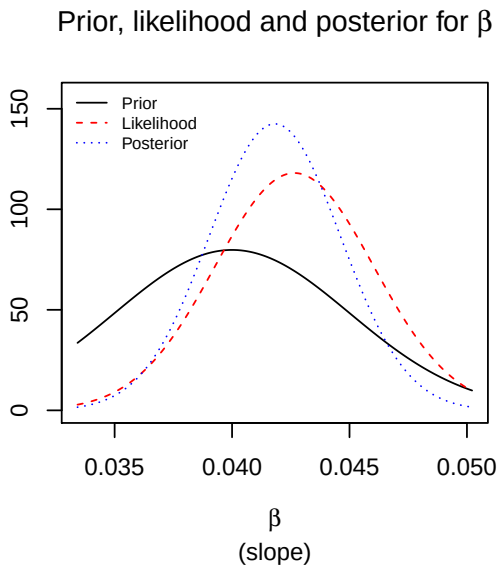


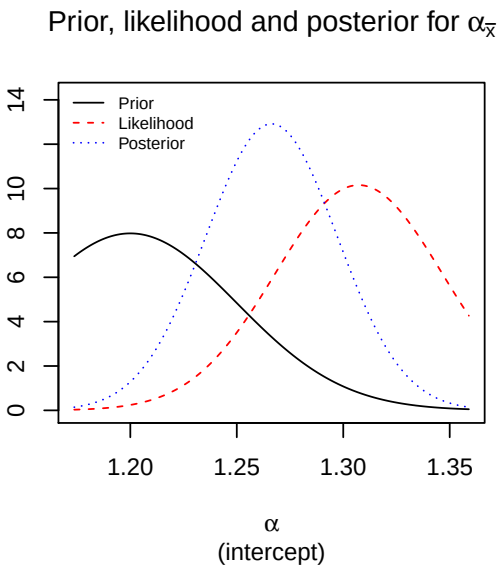


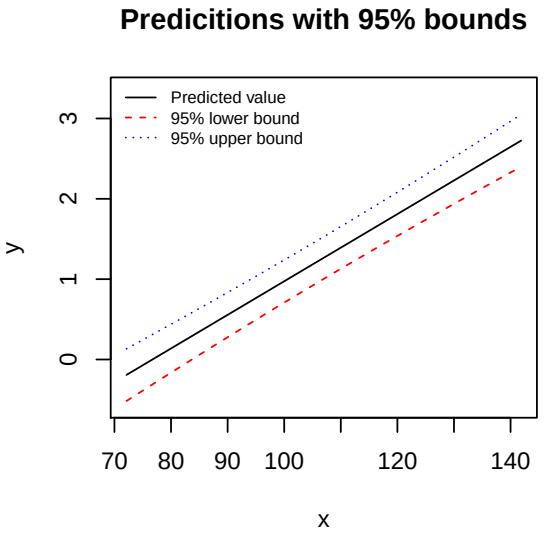


2. Re-do it with prior parameters $\mu_\alpha = 1.2$, $\sigma_\alpha = 0.05$, $\mu_\beta = 0.04$, $\sigma_\beta = 0.005$:

```
bayes.lin.reg(O2, HR, slope.prior = "n", intcpt.prior = "n",
              ma0 = 1.2, mb0 = 0.04, sa0 = 0.05, sb0 = 0.005)

## Standard deviation of residuals:  0.13
##             Posterior Mean Posterior Std. Deviation
##             -------------- ------------------------
## Intercept:  1.266          0.03089
## Slope:      0.04182        0.0028001
```







Exercises

1. Consider question 2 from the tutorial problems. For both choices of prior, do the following:

   (a) Draw a plot showing the prior and posterior distributions.

   (b) Calculate a 95% credible interval.

   (c) For the drug the be commercially successful, you need it to be effective in at least 3% of patients. What is your posterior probability that it will be successful?

2. Consider part 6 from Section 1. Estimate, via simulations, the coverage of the two interval estimates for $n = 20$ and $\theta \in \{0.05, 0.10, \ldots, 0.95\}$. Draw a plot comparing these results. What do you notice? What are the relative widths of these intervals?