BIOSTAT/STAT 570: Coursework 4 Key

To be submitted to the course canvas site by 11:59pm Monday 1st November, 2021.

1. This question is intended to illustrate how OLS estimation is affected by misspecification of the variance-covariance of the error terms.

   Consider the simple linear regression model

   $$Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1(t_i - \bar{t}) + \epsilon_i, \tag{1}$$

   where t_i represents time and the error terms ϵ_i are normal and are such that E$[\epsilon_i] = 0$, $i = 1, ..., n$. In the following assume that $n = 5, 10, 20, 30, 40, 50$, t_i equally spaced in $(-2, 2)$, $\beta_0 = 3$, $\beta_1 = 1.5$ and $\sigma^2 = 1$.

   Consider the following three forms for the variance-covariance:

   I. var$(\epsilon_i) = \mu_i\sigma^2$, and cov$(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.
   II. var$(\epsilon_i) = \mu_i^2\sigma^2$, and cov$(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.
   III. var$(\epsilon_i) = \sigma^2$, and cov$(\epsilon_j, \epsilon_k) = \sigma^2\rho^{|t_j - t_k|}$ with $-1 < \rho < 1$.

   (a) Simulate data from the above models and estimate β_0 and β_1 using OLS (use the values $\rho = 0.1, 0.5, 0.9$). Examine the 95% confidence interval coverage for β_0 and β_1.
   Answer: Figures 1 and 2 show the coverage for the three models. These were the result of 10,000 simulations, which leads to negligibly small MC error.

   (b) Summarize your conclusions.
   [Hint: For model III, note that the marginal distribution of $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\intercal$ is an n-dimensional zero mean normal with

   $$\text{var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \cdots & \cdots & \rho^{n-1}\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \cdots & \cdots & \rho^{n-2}\sigma^2 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ \rho^{n-1}\sigma^2 & \cdots & \cdots & \cdots & \cdots & \sigma^2 \end{bmatrix}$$

   which will help with data simulation.]
   Answer: Examining the top row of Figure 1, we see that for Model I, in spite of the mean-variance relationship, we still obtain approximately correct coverage for β_0 and β_1 with increasing sample sizes. Sandwich intervals are included for comparison. In Model II (second row, Figure 1) we see that the coverage for the intercept is

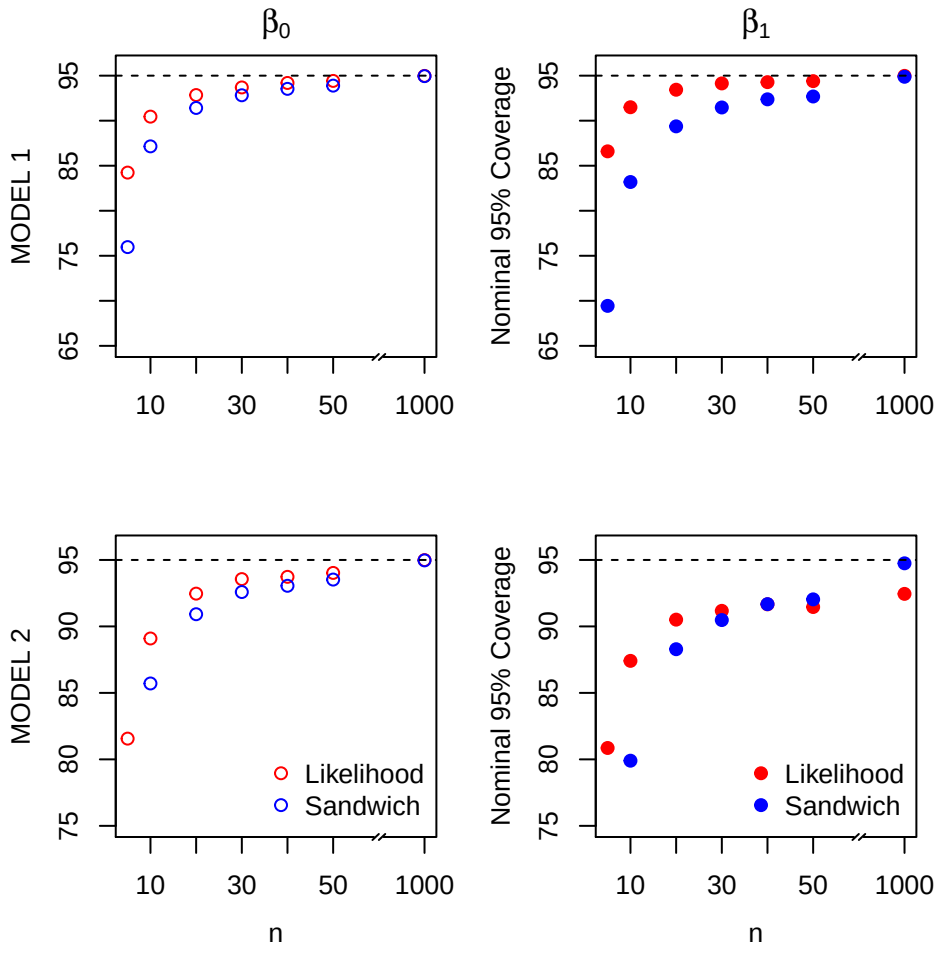

Figure 1: Coverage for β_0 and β_1 for Models I and II

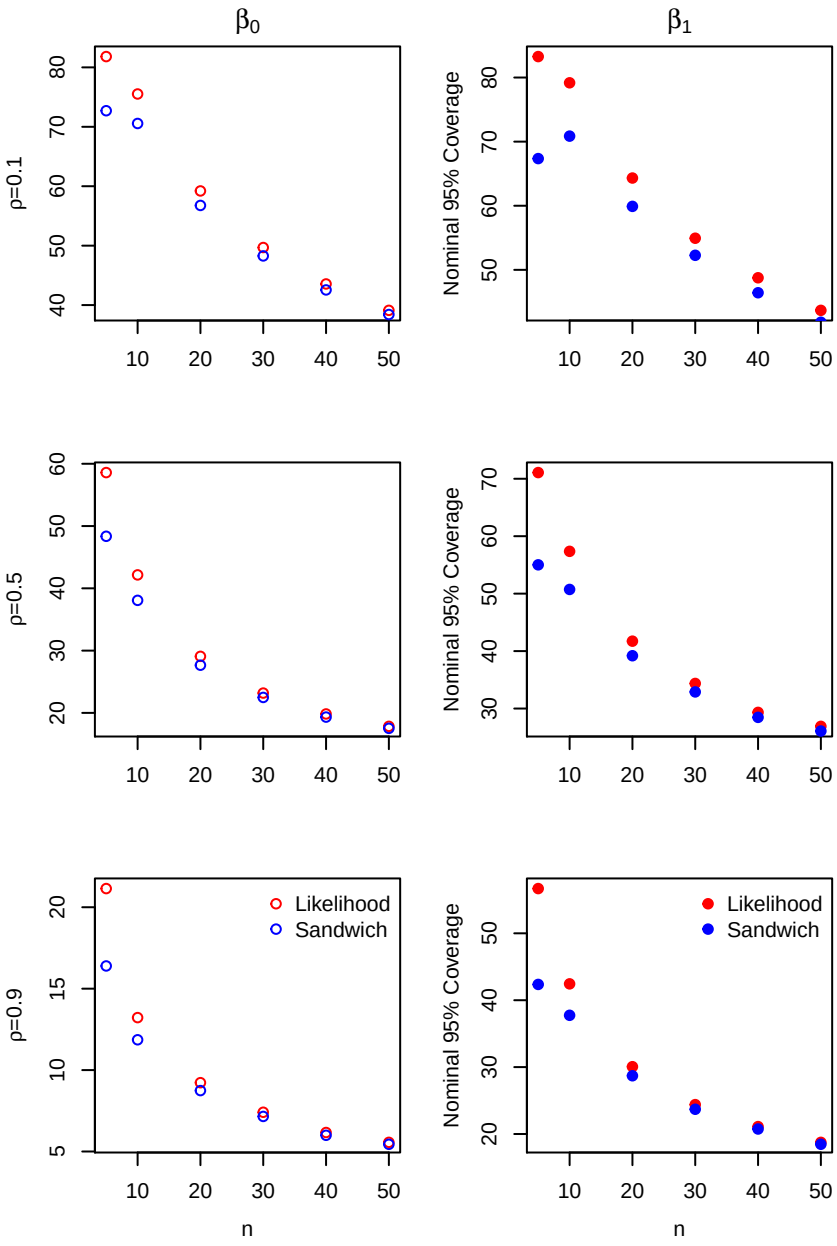

Figure 2: Coverage for β_0 and β_1 for Model III

approaching 95% for β_0, but at $n = 50$ we have under coverage for β_1 that is further aggravated at $n = 1000$.

It is important to note that this is not the case in general. The OLS is not typically robust to heteroscedasticity and few student's gave consideration to why it might be the case here. If you re-run the simulation with $t = i$ or `t <- seq(-1,2,length=n)`, you will see that the likelihood-based standard errors provide under coverage. Similarly, if we increase β_1, we see more undercoverage (especially for Model II). Why is this the case? In brief, I think (feel free to email if you disagree) what is happening is that we end up with approximately the same number of observations that underestimate and overestimate the true variance, in some sense canceling each other out and giving an "almost consistent" estimate of the variance. The terms which overestimate the variance end up being more severe in Model II, which is why we see the slight undercoverage.

For Model III (results in Figure 2), we see that the coverage is terrible for both parameters and gets worse with both increasing ρ and n. As ρ increases, each additional observation is telling us essentially the same thing as the previous, which we are not accounting for in our estimation procedures leading to terrible calibration. As n increases, so too does the degree of the autocorrelation (by design) which leads again to terrible coverage. You'll learn (much) more on how to deal with correlated observations in 571.

2. In this question we will consider inference when the sampling model is multivariate hypergeometric. Suppose a population contains objects of K different types, with $X_1, \ldots, X_K$ being the number of each type, $\sum_{k=1}^{K} X_k = N$. A simple random sample of size n is taken and the number of each type, $Y_1, \ldots, Y_K$, is recorded (so that $\sum_{k=1}^{K} y_k = n$).

   An obvious model for $Y_1, \ldots, Y_k$, is the multivariate hypergeometric distribution:

$$\Pr(Y_1 = y_1, \ldots, Y_K = y_K | x_1, \ldots, x_K) = \frac{\prod_{k=1}^{K} \binom{x_k}{y_k}}{\binom{N}{n}},$$

   with means and variances:

$$\mathsf{E}[Y_k|x_k] = n\frac{x_k}{N} \tag{2}$$

$$\mathsf{var}(Y_k|x_k) = n\frac{x_k}{N}\left(1 - \frac{x_k}{N}\right)\frac{N-n}{N-1} \tag{3}$$

   Suppose we take a sample from a population of K distinct objects, and record $y_1, \ldots, y_K$, but the numbers $X_1, \ldots, X_K$ are unknown (but N is known).

   (a) Using (2), write down an estimator for X_k, $k = 1 \ldots, K$. We will refer to this as a method of moments estimator. Using (3) give a from for the variance of this estimator, along with an estimator of this variance.

Answer: Since we have $\mathsf{E}[Y_k|x_k] = n\frac{x_k}{N}$, then

$$\hat{x}_k = \frac{N}{n} Y_k$$

is a methods of moments estimator of X_k. The variance of the MME is

$$\begin{aligned}\mathsf{var}(\hat{x}_k) &= \frac{N^2}{n^2}\mathsf{var}(Y_k|x_k) = \frac{N^2}{n^2} n \frac{x_k}{N}\left(1 - \frac{x_k}{N}\right)\frac{N-n}{N-1} \\ &= \frac{x_k}{N}\left(1 - \frac{x_k}{N}\right)\frac{N^2(N-n)}{n(N-1)}\end{aligned}$$

Then the estimator of variance is

$$\begin{aligned}\hat{\mathsf{var}}(\hat{x}_k) &= \frac{\hat{x}_k}{N}\left(1 - \frac{\hat{x}_k}{N}\right)\frac{N^2(N-n)}{n(N-1)} \\ &= \frac{Y_k}{n}\left(1 - \frac{Y_k}{n}\right)\frac{N^2(N-n)}{n(N-1)}\end{aligned}$$

(b) We now consider a Bayesian approach to inference. Consider a multinomial distribution for counts $X_1, \ldots, X_K$,

$$\Pr(X_1 = x_1, \ldots, X_K = x_K | p_1, \ldots, p_K) = \frac{N!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k}, \tag{4}$$

with $p_k > 0$ and $\sum_{k=1}^K p_k = 1$. Show that the Dirichlet:

$$\pi(p_1, \ldots, p_K) = \frac{\Gamma(\alpha_+)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}, \tag{5}$$

where $\alpha_k > 0$, $k = 1, \ldots, K$, and $\alpha_+ = \sum_{k=1}^K \alpha_k$, is the conjugate distribution to the multinomial sampling model.

[One interpretation of this set up is that $p_1, \ldots, p_K$ are the proportions in each category in a hypothetical infinite population of objects.]

Answer: Here we use vectors $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{x} = (x_1, \ldots, x_K)$. By Bayesian fomular,

$$\begin{aligned}p(\boldsymbol{p}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{p})\pi(\boldsymbol{p})}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\boldsymbol{p})\pi(\boldsymbol{p}) \\ &= \frac{N!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k} \frac{\Gamma(\alpha_+)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1} \\ &\propto \prod_{k=1}^K p_k^{\alpha_k + x_k - 1}\end{aligned}$$

which is kernel of Dirichlet distribution with parameters $\{\alpha_k + x_k\}$, thus Dirichlet is the conjugate distribution to the multinumial samplling model.

(c) The compound multinomial distribution, CMult$(N, \boldsymbol{\alpha})$, is defined as

$$\Pr(X_1 = x_1, \ldots, X_K = x_K) = \frac{N!\Gamma(\alpha_+)}{\Gamma(N + \alpha_+)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{x_k!\Gamma(\alpha_k)},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. Show that the prior predictive distribution, obtained as the marginal distribution when the likelihood is (4) and the prior is (5), is of compound multinomial form with parameters that you should identify.

Answer: The prior predictive distribution is the marginal distribution $p(\boldsymbol{x})$. Using Bayesian fomular,

$$\begin{aligned}
p(\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{p})\pi(\boldsymbol{p})}{p(\boldsymbol{p}|\boldsymbol{x})} \\
&= \frac{\frac{N!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k} \frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}}{\frac{\Gamma(\alpha_+ + \sum x_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k + x_k)} \prod_{k=1}^{K} p_k^{\alpha_k + x_k - 1}} \\
&= \frac{N!\Gamma(\alpha_+)}{\Gamma(\alpha_+ + \sum_1^K x_k)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + x_k)}{x_k!\Gamma(\alpha_k)} \\
&= \frac{N!\Gamma(\alpha_+)}{\Gamma(\alpha_+ + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + x_k)}{x_k!\Gamma(\alpha_k)}
\end{aligned}$$

Thus the marginal distribution is of compound multinomial distribution with parameters $(N = \sum_{k=1}^{K} x_k, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ comes from the parameters of Dirichlet distribution.

(d) Find the mean E$[X_k]$ and variance var(X_k), $k = 1, \ldots, K$, of a compound multinomial distribution.

[Hint: you may quote without proof the means and variances of the multinomial and Dirichlet distributions.]

Answer: The mean and variance of multinomial distribution are

$$\mathsf{E}[X_k|\boldsymbol{p}] = Np_k \qquad\qquad \mathsf{var}(X_k|\boldsymbol{p}) = Np_k(1 - p_k)$$

The mean and variance of Dirichlet distribution are,

$$\mathsf{E}[p_k] = \frac{\alpha_k}{\alpha_+} \qquad\qquad \mathsf{var}(p_k) = \frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)}$$

The mean is

$$\begin{aligned}\mathsf{E}[X_k] &= \mathsf{E}_p[\mathsf{E}[X_k|\boldsymbol{p}]] \\ &= \mathsf{E}[Np_k] = N\mathsf{E}[p_k] \\ &= N\frac{\alpha_k}{\alpha_+}\end{aligned}$$

The variance is

$$\begin{aligned}\mathsf{var}(X_k) &= \mathsf{var}(\mathsf{E}[X_k|p]) + \mathsf{E}[\mathsf{var}(X_k|p)] \\ &= \mathsf{var}(Np_k) + \mathsf{E}[Np_k(1-p_k)] \\ &= N^2\mathsf{var}(p_k) + N\mathsf{E}[p_k] - N\mathsf{E}[p_k^2] \\ &= N^2\mathsf{var}(p_k) + N\mathsf{E}[p_k] - N\mathsf{var}(p_k) - N(\mathsf{E}[p_k]^2) \\ &= N(N-1)\frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)} + N\frac{\alpha_k}{\alpha_+}(1 - \frac{\alpha_k}{\alpha_+}) \\ &= \frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2}\frac{N(\alpha_+ + N)}{\alpha_+ + 1}\end{aligned}$$

(e) Let $W_k = X_k - y_k$ represent the unobserved counts, $k = 1, \ldots, K$. Show that the posterior distribution $\Pr(W_1, \ldots, W_K | y_1, \ldots, y_K)$ is compound multinomial CMult$(N - n, \boldsymbol{\alpha} + \boldsymbol{y})$, where $\boldsymbol{y} = (y_1, \ldots, y_K)$.

Answer: First we consider the posterior distribution of $(\boldsymbol{x}|\boldsymbol{y})$, and we have

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})}{p(\boldsymbol{y})}$$

Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{y}$, then $x_k = w_k + y_k$ for each k. Then the Jacobian matrix $\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{w}} = \boldsymbol{I}_K$. Hence

$$\begin{aligned}p(\boldsymbol{w}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{y})\pi(\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{y})}{p(\boldsymbol{y})}\boldsymbol{I}_K \propto p(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{y})\pi(\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{y}) \\ &\propto \frac{\prod_{k=1}^K \binom{w_k + y_k}{y_k}}{\binom{N}{n}} \frac{N!\Gamma(\alpha_+)}{\Gamma(\alpha_+ + N)} \prod_{k=1}^K \frac{\Gamma(\alpha_k + w_k + y_k)}{(w_k + y_k)!\Gamma(\alpha_k)} \\ &= \prod_{k=1}^K \frac{(w_k + y_k)!}{y_k! w_k!} \frac{\Gamma(w_k + \alpha_k + y_k)}{(w_k + y_k)!} \frac{N!(N-n)!n!}{N!} \\ &= \prod_{k=1}^K \frac{\Gamma(w_k + \alpha_k + y_k)}{w_k!}\end{aligned}$$

$\prod_{k=1}^{K} \frac{\Gamma(w_k+\alpha_k+y_k)}{w_k!}$ is the kernel of compound multinomial. Since $\sum_{k=1}^{K} w_k = \sum_{k=1}^{K}(X_k - y_k) = N - n$, $p(\boldsymbol{w}|\boldsymbol{y} \sim CMult(N-n, \boldsymbol{\alpha}+\boldsymbol{y})$.

(f) Write down the posterior mean and posterior variance of X_k, $k = 1, \ldots, K$. Comment on the case when $\alpha_k = 0$, $k = 1, \ldots, K$.

Answer: Since we know that $(\boldsymbol{W}|\boldsymbol{y}) \sim CMult((N-n), \boldsymbol{\alpha}+\boldsymbol{y})$, and $\boldsymbol{X} = \boldsymbol{W} + \boldsymbol{y}$, we can calculated the posterior expectation and variance of $\boldsymbol{X}$ from part (d).

$$\begin{aligned}
\mathsf{E}[W_k|\boldsymbol{y}] &= (N-n)\frac{\alpha_k + y_k}{\alpha_+ + \sum_1^K y_i} = (N-n)\frac{\alpha_k + y_k}{\alpha_+ + n} \\
\mathsf{var}(W_k|\boldsymbol{y}) &= (N-n)\frac{(\alpha_k + y_k)(\alpha_+ + n - \alpha_k - y_k)}{(\alpha_+ + n)^2}\frac{\alpha_+ + N}{\alpha_+ + n + 1}
\end{aligned}$$

The posterior mean is

$$\mathsf{E}[X_k|\boldsymbol{y}] = \mathsf{E}[W_k|\boldsymbol{y}] + y_k = (N-n)\frac{\alpha_k + y_k}{\alpha_+ + n} + y_k$$

The posterior variance is

$$\mathsf{var}(X_k|\boldsymbol{y}) = \mathsf{var}(W_k|\boldsymbol{y}) = (N-n)\frac{(\alpha_k + y_k)(\alpha_+ + n - \alpha_k - y_k)}{(\alpha_+ + n)^2}\frac{\alpha_+ + N}{\alpha_+ + n + 1}$$

When $\boldsymbol{\alpha} = 0$,

$$\begin{aligned}
\mathsf{E}[X_k|\boldsymbol{y}] &= (N-n)\frac{y_k}{n} + y_k = \frac{N y_k}{n} \\
\mathsf{var}(W_k|\boldsymbol{y}) &= (N-n)\frac{y_k(n-y_k)}{n^2}\frac{N}{n+1} = \frac{y_k}{n}(1 - \frac{y_k}{n})\frac{N(N-n)}{n+1}
\end{aligned}$$

A certain infectious disease can be caused by one of three different pathogens, A, B, or C. Over a 1 year period population surveillance is carried out, and 750 individuals are observed to be infected. A random sample of 62 cases is selected for lab testing, i.e., to determine the pathogen responsible. Of these 62 selected cases, the numbers who were infected by pathogens A, B, C, were 43, 19, 0, respectively.

We wish to estimate the numbers of the total population of cases that were infected by each of the pathogens.

(g) Calculate the method of moments estimators and the associated standard errors.

Answer: In this question, we need to specify each parameter of the model: $N = 750$, $n = 62$, $K = 3$, $(y_1, y_2, y_3) = (43, 19, 0)$. Using the results from part (a), we have

	A	B	C
MME	520.16	229.84	0.00
SE	42.09	42.09	0.00

Table 1: Results of part(g)

(h) Calculate the Bayesian posterior mean and posterior standard deviation, with prior specification, $\alpha_k = 1$, $k = 1, \ldots, K$. Which estimates are the most reasonable.

Answer: Using the results from part (f) and letting $\alpha_k = 1$ for $k = 1, 2, 3$, we have

	A	B	C
Posterior mean	508.72	230.69	10.58
Posterior sd	41.43	40.89	10.90

Table 2: Results of part (h)

Bayes estimates give more reasonable point and uncertainty estimates for pathogen C. Both methods have similar estimates for pathogen A and B.

(i) Devise a sampling-based method for sampling from the posterior, and present histogram representations of the posterior distributions of $X_k|\boldsymbol{y}$, $k = 1, \ldots, K$.

Answer: Figure 3 shows the histogram representations of the posterior distributions of $X_k|\boldsymbol{y}$, $k = 1, \ldots, 3$.

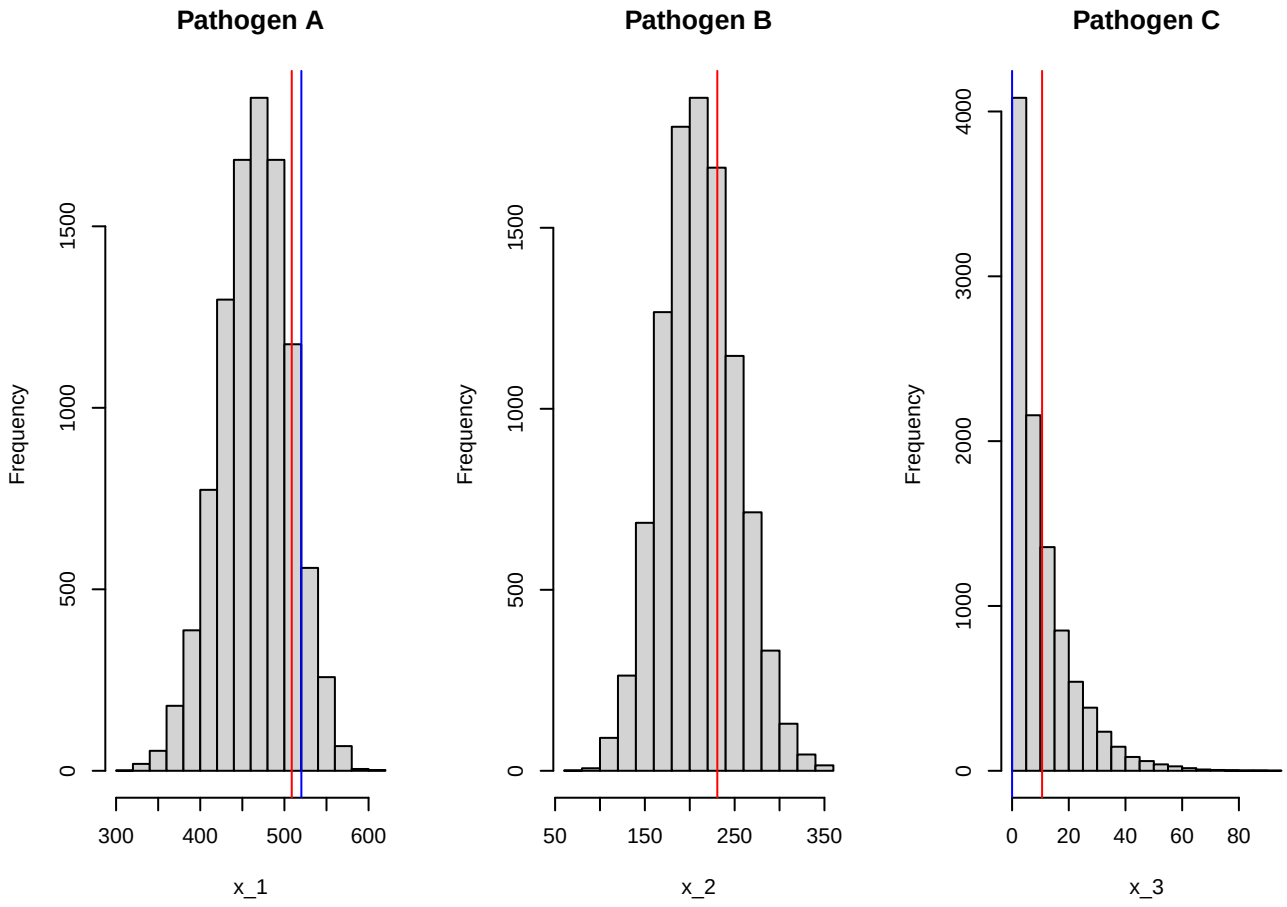

Figure 3: Histogram representations of posterior distributions. Vertical lines represent posterior mean (red) and the method of moments estimator of the mean (blue). For Pathogen B the red and blue lines are very close and separation is not visualized.

R Code

```
### QUESTION 1 ###

library(sandwich)
library(MASS)
library(mvtnorm)

do.one <- function(n,model,b0=3,b1=1.5,sig=1,rho){
  t <- seq(-2,2,length=n)
  mu <- b0 + b1 * (t-mean(t))

  # set up the variances of each model
  if(model==1){
    var.eps <- mu*sig
    eps <- rnorm(n,mean=0,sd=sqrt(var.eps))
  }else if(model==2){
    var.eps <- mu^2*sig
    eps <- rnorm(n,mean=0,sd=sqrt(var.eps))
  }else if(model==3){
    mat <- diag(n)
    for(i in 1:n){
      for(j in 1:n){
        mat[i,j] <- rho^abs(t[i]-t[j])
      }
    }
    eps <- rmvnorm(1,sigma=mat)
  }

  #gernerate the data
  Y <- as.vector(mu + eps)

  fm <- lm(Y~t)

  b.hat <- fm$coef
  var.mod <- diag(vcov(fm)) #likelihood based variance est.
  est.mod <- b.hat + sqrt(var.mod)%o%c(-1.96,0,1.96) #95% CI
  cov.mod <- c(b0,b1)>est.mod[,1] & c(b0,b1)<est.mod[,3] #coverage for b0 and b1
```

```
  var.sand <- diag(vcovHC(fm,type="HC0")) #sandwich based variance est.
  est.sand <- b.hat + sqrt(var.sand)%o%c(-1.96,0,1.96)
  cov.sand <- c(b0,b1)>est.sand[,1] & c(b0,b1)<est.sand[,3]


  return(c(est.mod[1,],cov.mod[1],est.mod[2,],cov.mod[2],est.sand[1,]
           ,cov.sand[1],est.sand[2,],cov.sand[2]))
}


#helper function to summarize results given by replicate()
rslt.summarize <- function(rslt){
  bias.b0 <- 3-mean(rslt[2,])
  bias.b1 <- 1.5-mean(rslt[6,])
  width.mod.b0 <- mean(rslt[3,]-rslt[1,])
  width.mod.b1 <- mean(rslt[7,]-rslt[5,])
  width.sand.b0 <- mean(rslt[11,]-rslt[9,])
  width.sand.b1 <- mean(rslt[15,]-rslt[13,])
  cov.mod.b0 <- mean(rslt[4,])
  cov.mod.b1 <- mean(rslt[8,])
  cov.sand.b0 <- mean(rslt[12,])
  cov.sand.b1 <- mean(rslt[16,])

  return(c(mod,n,rho,bias.b0,bias.b1,width.mod.b0,
           width.mod.b1, width.sand.b0, width.sand.b1, cov.mod.b0, cov.mod.b1,
           cov.sand.b0, cov.sand.b1))
}

#now run the models n.reps times under different sample sizes, etc...
n.reps <- 50000
rslt.table <- NULL
for(mod in 1:3){
  print(mod)
  for(n in c(5,10,20,30,40,50)){
    print(n)
    if(mod==3){
      for(rho in c(0.1,0.5,0.9)){
        rslt <- replicate(n.reps,do.one(n=n,model=mod,rho=rho))
        rslt.table <- rbind(rslt.table,rslt.summarize(rslt))
        print(rho)
```

```
      }
    }else{
      rho <- 0
      rslt <- replicate(n.reps,do.one(n=n,model=mod,rho=0))
      rslt.table <- rbind(rslt.table,rslt.summarize(rslt))
      print(rho)
    }
  }
}

write.table(rslt.table,"HW2_Q1_sim.txt",row.names=F,sep="\t")

# the code below generates the tables given in the key
q1 <- read.table("HW2_Q1_sim_FINAL.txt",sep="\t",header=T)
colnames(q1) <- c("model","n","rho","bias.b0","bias.b1","width.mod0"
                  ,"width.mod1","width.sand0","width.sand1","cov.mod0","cov.mod1","cov.sand0
                  "cov.sand1")

pdf("q1_1.pdf",height=5,width=5)
layout(matrix(c(1,3,2,4),2,2))
par(mgp=c(2.5,1,0), mar = c(3.6,3.5,2,1.5))

plot(q1$n[1:6],100*q1$cov.mod0[1:6],pch=1,col="red",lwd=1,xlab="",
     ylab="MODEL 1",ylim = c(65,96),main=expression(paste(beta[0])))
points(q1$n[1:6],100*q1$cov.sand0[1:6],pch=1,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=1,
#  col=c("red","blue"))

plot(q1$n[1:6],100*q1$cov.mod1[1:6],pch=19,col="red",lwd=1,xlab="",
     ylab="Nominal 95% Coverage",ylim = c(65,96),main=expression(paste(beta[1])))
points(q1$n[1:6],100*q1$cov.sand1[1:6],pch=19,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=19,
#  col=c("red","blue"))

# model 2 plots
plot(q1$n[7:12],100*q1$cov.mod0[7:12],pch=1,col="red",lwd=1,
     xlab="n",
     ylab="MODEL 2",ylim = c(75,96))
```

```
points(q1$n[7:12],100*q1$cov.sand0[7:12],pch=1,col="blue",lwd=1)
abline(h=95,lty=2)
legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=1,
       col=c("red","blue"))

plot(q1$n[7:12],100*q1$cov.mod1[7:12],pch=19,col="red",lwd=1,xlab="n",
     ylab="Nominal 95% Coverage",ylim = c(75,96))
points(q1$n[7:12],100*q1$cov.sand1[7:12],pch=19,col="blue",lwd=1)
abline(h=95,lty=2)
legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=19,
       col=c("red","blue"))

dev.off()

# model 3 plot
rho1 <- seq(13,30,3)
rho5 <- seq(14,30,3)
rho9 <- seq(15,30,3)

pdf("q1_2.pdf",width=4.5,height=6.5)
layout(matrix(c(1,3,5,2,4,6),3,2))
par(mgp=c(2.5,1,0), mar = c(3.6,3.5,2,1.5))

plot(q1$n[rho1],100*q1$cov.mod0[rho1],pch=1,col="red",lwd=1,
     main=expression(paste(beta[0])), xlab="",ylab=expression(paste(rho,"=0.1")) )#,ylim = c
points(q1$n[rho1],100*q1$cov.sand0[rho1],pch=1,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=1,
#  col=c("red","blue"))

plot(q1$n[rho1],100*q1$cov.mod1[rho1],pch=19,col="red",lwd=1,
     main=expression(paste(beta[1])), xlab="", ylab="Nominal 95% Coverage" )#,ylim = c(65,96

points(q1$n[rho1],100*q1$cov.sand1[rho1],pch=19,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=19,
#  col=c("red","blue"))

#rho = 0.5
plot(q1$n[rho5],100*q1$cov.mod0[rho5],pch=1,col="red",lwd=1,
```

```
     xlab="", ylab=expression(paste(rho,"=0.5")) )#, ylim=c(40,80))
points(q1$n[rho5],100*q1$cov.sand0[rho5],pch=1,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=1,
#  col=c("red","blue"))

plot(q1$n[rho5],100*q1$cov.mod1[rho5],pch=19,col="red",lwd=1,
     xlab="",  ylab="Nominal 95% Coverage" )#, ylim=c(40,80))
points(q1$n[rho5],100*q1$cov.sand1[rho5],pch=19,col="blue",lwd=1)
abline(h=95,lty=2)
#legend(c("Likelihood","Sandwich"),x="bottomright",bty="n",pch=19,
#  col=c("red","blue"))

#rho = 0.9
plot(q1$n[rho9],100*q1$cov.mod0[rho9],pch=1,col="red",lwd=1,
     xlab="n", ylab=expression(paste(rho,"=0.9")) )#,ylim=c(10,30))
points(q1$n[rho9],100*q1$cov.sand0[rho9],pch=1,col="blue",lwd=1)
abline(h=95,lty=2)
legend(c("Likelihood","Sandwich"),x="topright",bty="n",pch=1,
       col=c("red","blue"))

plot(q1$n[rho9],100*q1$cov.mod1[rho9],pch=19,col="red",lwd=1,
     xlab="n",  ylab="Nominal 95% Coverage" )#, ylim=c(30,60))
points(q1$n[rho9],100*q1$cov.sand1[rho9],pch=19,col="blue",lwd=1)
abline(h=95,lty=2)
legend(c("Likelihood","Sandwich"),x="topright",bty="n",pch=19,
       col=c("red","blue"))

dev.off()


### QUESTION 2 ###

### Data ###
y <- c(43,19,0)
N <- 750
ns <- sum(y)

### Part g ###
# Method of moms
```

```
q <- y/ns
N*q
# 576 224   0

# se
sqrt(N^2*q*(1-q)*(N-ns)/(ns*(N-1)))
# 49.21613 49.21613  0.00000

### Part h ###
# Bayes Means
alpha <- c(1,1,1)
(postmean <- (N-ns)*(y+alpha)/(sum(y)+sum(alpha)) + y)
# 559.58491 226.26415  14.15094

# Post vars
postvar <- (N-ns)*(alpha+y)*(sum(alpha)+ns-alpha-y)*(sum(alpha)+N)/((sum(alpha)+ns)^2*(sum(a
sqrt(postvar)
# 48.48155 47.57218 14.36870

### Extra: Prior sensitivity ###
alpha2 <- c(2,2,2)
(postmean2 <- (N-ns)*(y+alpha2)/(sum(y)+sum(alpha2)) + y)
# 544.92857 228.28571  26.78571

# Post vars
postvar2 <- (N-ns)*(alpha2+y)*(sum(alpha2)+ns-alpha2-y)*(sum(alpha2)+N)/((sum(alpha2)+ns)^2*
sqrt(postvar2)
# 48.09509 46.52246 19.11105

# Samples from the posterior
library(DCluster)
nsim <- 10000
library(VGAM) # To access the rdiric function
p <- rdiric(nsim,alpha+y)
w <- matrix(0,nrow=nsim,ncol=3)
for (i in 1:nsim){
  w[i,] <- rmultin(N-ns,p[i,])
}
par(mfrow=c(1,3))
hist(w[,1],main="Pathogen A",xlab="x_1") # Samples from posterior
```

```
abline(v=postmean[1],col="red") # Posterior Mean
abline(v=N*q[1],col="blue") # MoM
hist(w[,2],main="Pathogen B",xlab="x_2") # Samples from posterior
abline(v=postmean[2],col="red") # Posterior Mean
abline(v=N*q[2],col="blue")) # MoM
hist(w[,3],main="Pathogen C",xlab="x_2") # Samples from posterior
abline(v=postmean[3],col="red") # Posterior Mean
abline(v=N*q[3],col="blue") # MoM
par(mfrow=c(1,1))
```