Exam 2 – ST440/540 – Instructor Solution

(1) **Model description**: Let Y_i be the log CRP for subject i and X_{ij} be the j^{th} log exposure variable (heavy metal and PFAS variables). The response and all covariates are standardized to have mean zero and variance one. We consider several models that are special cases of the second-order model

$$Y_i = \alpha_0 + age_i\alpha_1 + income_i\alpha_2 + \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{j=1}^{p} {X_{ij}}^2\gamma_j + \sum_{j<k} X_{ij}X_{ik}\delta_{jk} + \varepsilon_i$$

where β_j, γ_j and δ_{jk} are the linear, quadratic and interaction effects, respectively, and $\varepsilon_i \sim N(0, \sigma^2)$, independent over i. We select uninformative priors $\alpha_k \sim N(0, 100^2)$ and $\beta_j, \gamma_j, \delta_{jk} \sim N(0, \tau^2)$ with $\tau^2, \sigma^2 \sim InvG(0.1, 0.1)$. The three models are: (1) Linear with $\gamma_j = \delta_{jk} = 0$, (2) Quadratic with $\delta_{jk} = 0$ and (3) the full model. Each model is fit using Gibbs sampling in JAGS with 10,000 MCMC iterations after a burn-in of 1,000 iterations.

(2) **Model comparisons**: Models are compared using DIC. The DIC (pD) for the linear, quadratic and full models are 7461 (9.9), 7443 (17.3) and 7450 (40.5), so we select the quadratic model for the analysis.

(3) **Goodness of fit**: We use posterior predictive checks to study the residual normality assumption. For quantile levels 0.00, 0.05, …, 1.00 we use the sample quantile as the summary statistic. The Bayesian p-values in Figure 1 (left) show some lack of fit for lower quantiles with p-values near zero, but the fit is overall reasonable.

4. **Variable importance**: The only two terms in Table 1 with posterior 95% intervals that exclude zero are the linear term for cadmium and the quadratic term for mercury. Figure 1b shows that cadmium has an increasing effect, with possible plateau for large values. Mercury has a strong quadratic effect with low mean CSP for both large and small values of mercury. None of the PFAS exposures are statistically significant, but the quadratic term for PFDeA and the linear term for nPFOA are close.

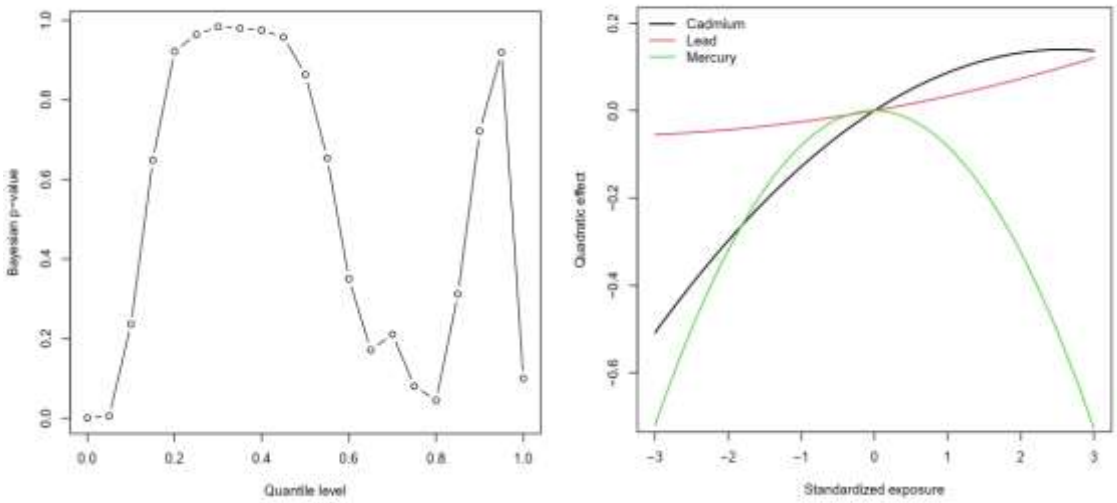

Figure 1: Bayesian p-values by quantile level (left) and estimated exposure response curve for the heavy metals from the quadratic model (right).

	Linear				Quadratic			
	Median	2.5%	97.5%	Prob>0	Median	2.5%	97.5%	Prob>0
cadmium	**0.108**	**0.063**	**0.151**	**1.000**	-0.021	-0.050	0.008	0.077
lead	0.029	-0.014	0.074	0.909	0.004	-0.023	0.031	0.600
mercury	-0.001	-0.047	0.047	0.481	**-0.081**	**-0.112**	**-0.047**	**0.000**
PFDeA	-0.018	-0.075	0.038	0.263	-0.023	-0.052	0.007	0.067
PFHxS	0.018	-0.034	0.071	0.749	-0.006	-0.036	0.023	0.340
PFNA	0.013	-0.041	0.070	0.676	0.014	-0.016	0.045	0.816
nPFOA	-0.050	-0.110	0.008	0.047	0.015	-0.012	0.041	0.876
sbPFOA	0.007	-0.081	0.105	0.552	-0.004	-0.022	0.013	0.324

Table 1: Posterior median, 95% credible interval, and probability of a positive effect for the linear (β_j) and quadratic (γ_j) terms; 95% intervals that exclude zero are in bold.

R code

```
# Load and scale the data

 dat <- read.csv("E2.csv")
 Y   <- scale(log(dat$CRP))
 Y   <- as.vector(Y)
 n   <- length(Y)
 X   <- scale(log(dat[,3:10]))
 Z   <- dat[,11:12]
 p   <- ncol(X)
 lab <- colnames(X)

# Covariates for the linear model

 X1 <- X

# Covariates for the quadratic model

 X2           <- cbind(X,X^2) # Quadratic
 colnames(X2) <- c(lab,paste0(lab,"^2"))

# Covariates for the full second-order model

 X3   <- NULL
 labs <- NULL
 for(j in 1:p){for(k in 1:p){if(j<k){
     X3   <- cbind(X3,X[,j]*X[,k])
     labs <- c(labs,paste0(lab[j],"x",lab[k]))
 }}}
 X3   <- cbind(X2,X3)
 colnames(X3) <- c(colnames(X2),labs)

# Function to the a multiple linear regression model in JAGS

 Bayes_MRL <- function(X,Y,m,burn=1000,n.iter=50000){

  library(rjags)

  model_string <- textConnection("model{

    # Likelihood
     for(i in 1:n){
       Y[i]   ~ dnorm(mu[i],taue)
       mu[i] <- alpha+inprod(X[i,],beta[])
     }
    # Priors
     for(j in 1:p){
       beta[j] ~ dnorm(0,taub)
     }
     alpha ~ dnorm(0,0.001)
     taue  ~ dgamma(0.1, 0.1)
     taub  ~ dgamma(0.1, 0.1)

   # Posterior predictive checks
    for(i in 1:n){
     Yp[i]    ~ dnorm(mu[i],taue)
    }
    ord  <- order(Yp) # PPD quantiles
    for(k in 1:M){
      D[k] <- Yp[ord[m[k]]]
    }
  }")

  n     <- nrow(X)
```

```
  p     <- ncol(X)
  M     <- length(m)
  data  <- list(Y=Y,X=X,n=n,p=p,M=M,m=m)
  model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
  update(model, burn, progress.bar="none")
  samps <- coda.samples(model, variable.names=c("beta","D"),
                        n.iter=n.iter, progress.bar="none")[[1]]
  DIC   <- dic.samples(model,n.iter=n.iter,progress.bar="none")
  beta  <- samps[,1:p+M]
  colnames(beta) <- colnames(X)
  out   <- list(beta=beta,D=samps[,1:M],DIC=DIC)
 return(out)}

# Fit the three models

 m    <- round(seq(1,length(Y),length=21))
 fit1 <- Bayes_MRL(X1,Y,m)
 fit2 <- Bayes_MRL(X2,Y,m)
 fit3 <- Bayes_MRL(X3,Y,m)

# Compare DIC

 fit1$DIC
 fit2$DIC
 fit3$DIC

# Posterior checks for model 2
 D0   <- Y[order(Y)[m]]
 pval <- colMeans(sweep(fit2$D,2,D0,"-")>0)
 plot(m/n,pval,type="b",xlab="Quantile level",ylab="Bayesian p-value")

# Summarize effects
 q   <- apply(fit2$beta,2,quantile,c(0.50,0.025,0.975))
 p   <- colMeans(fit2$beta>0)
 out <- round(cbind(t(q),p),3)
 out

 x    <- seq(-3,3,.1)
 cad  <- q[1,c(1,9)]
 lead <- q[1,c(1,9)+1]
 merc <- q[1,c(1,9)+2]
 plot(x,cad[1]*x + cad[2]*x^2,type="l",lwd=2,ylim=c(-0.7,0.2),
      xlab="Standardized exposure",ylab="Quadratic effect")
 lines(x,lead[1]*x+lead[2]*x^2,col=2,lwd=2)
 lines(x,merc[1]*x+merc[2]*x^2,col=3,lwd=2)
 legend("topleft",c("Cadmium","Lead","Mercury"),lwd=2,col=1:3,bty="n")
```