AUSTRALIAN NATIONAL UNIVERSITY
RESEARCH SCHOOL OF FINANCE ACTUARIAL STUDIES, AND APPLIED STATISTICS

INTRODUCTION TO BAYESIAN DATA ANALYSIS (STAT3016/4116/7016)
SEMESTER 2 2017

ASSIGNMENT 4 - SOLUTIONS

see `assign4.R` for R code.

Problem 1

Let y_{ij} denote the number of hours spent on homework by student $i = 1, ..., n_j$ from school $j = 1, ..., 8$ (where n_j is the number of students sampled from school j). The hierarchical normal model is:

$$y_{ij}|\theta_j, \sigma^2 \sim \text{Normal}(\theta_j, \sigma^2) \text{ (within-group model)}$$

$$\theta_j|\mu, \tau^2 \sim \text{Normal}(\mu, \tau^2)$$

Using the standard semiconjugate normal and inverse-gamma prior distributions:

$$1/\sigma^2 \sim \text{gamma}(\nu_0/2 = 2/2, \nu_0\sigma_0^2/2 = 2 \times 15/2)$$

$$1/\tau^2 \sim \text{gamma}(\eta_0/2 = 2/2, \eta_0\tau_0^2/2 = 2 \times 10/2)$$

$$\mu \sim \text{normal}(\mu_0 = 7, \gamma_0^2 = 5)$$

Also m=8 (the number of schools). The full conditional distributions of the parameters are given below:

$$\mu\Big|\theta_1, ..., \theta_m, \tau^2 \sim \text{normal}\left(\frac{m \times \bar{\theta}\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, [m/\tau^2 + 1/\gamma_0^2]^{-1}\right)$$

$$1/\tau^2\Big|\theta_1, ..., \theta_m, \mu \sim \text{gamma}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum(\theta_j - \mu)^2}{2}\right)$$

$$\theta_j \Big| y_{1,j}, ..., y_{n_j,j}, \sigma^2 \sim \text{normal}\left(\frac{n_j \times \bar{y}_j \sigma^2 + 1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, [n_j/\sigma^2 + 1/\tau^2]^{-1}\right)$$

$$1/\sigma^2 \Big| \theta_1, ..., \theta_m, \mathbf{y}_1, ..., \mathbf{y}_m \sim \text{gamma}\left(\frac{1}{2}[\nu_0 + \sum_{j=1}^{m} n_j], \frac{1}{2}[\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2]\right)$$

(a) We run the Gibbs sampler for 5000 iterations. The diagnostic and metrics are below. All effective sizes are > 1000. The autocorrelation plots show no high autocorrelations to be concerned about and traceplots show good mixing of the respective chains for each parameter.

```
> apply(THETA,2, effectiveSize)
      theta1   theta2   theta3   theta4   theta5   theta6   theta7   theta8
[1] 4551.612 4968.519 5000.000 5465.115 4096.163 4625.084 5576.988 4776.939
> apply(SMT,2, effectiveSize)
      sigma2       mu     tau2
[1] 4739.021 4177.568 3618.343
```

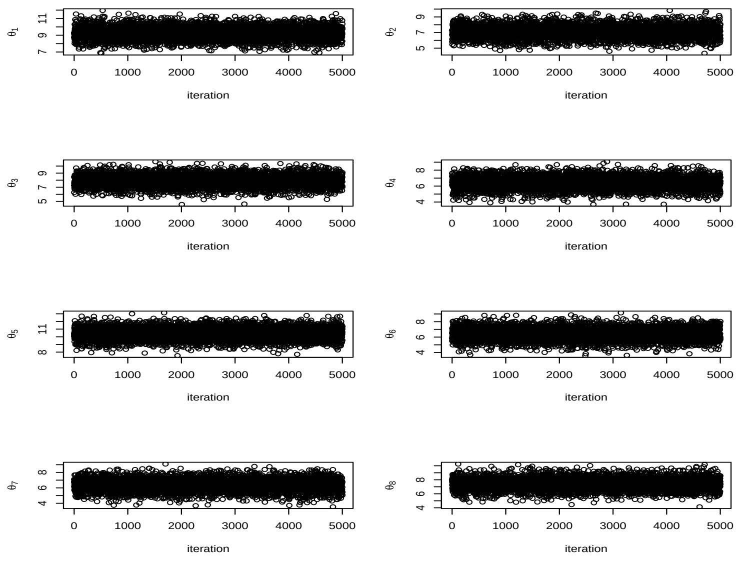

Figure 1: Traceplots - $\theta_1, ..., \theta_8$

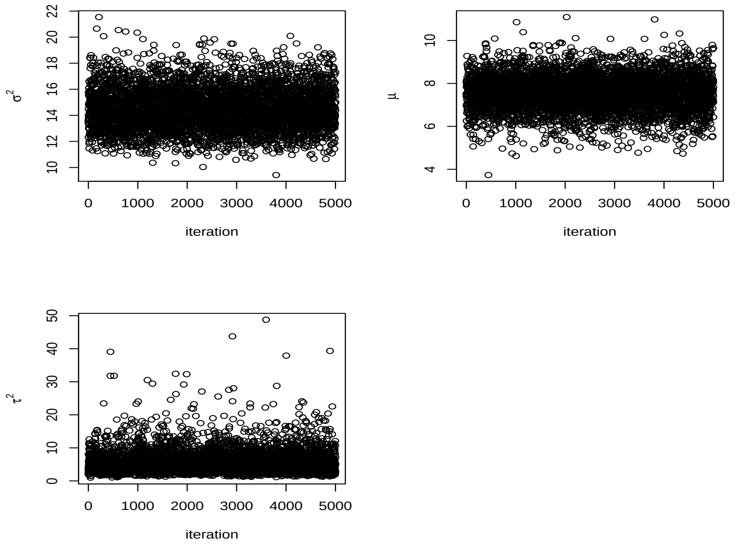

Figure 2: Traceplots - σ^2, μ, τ^2

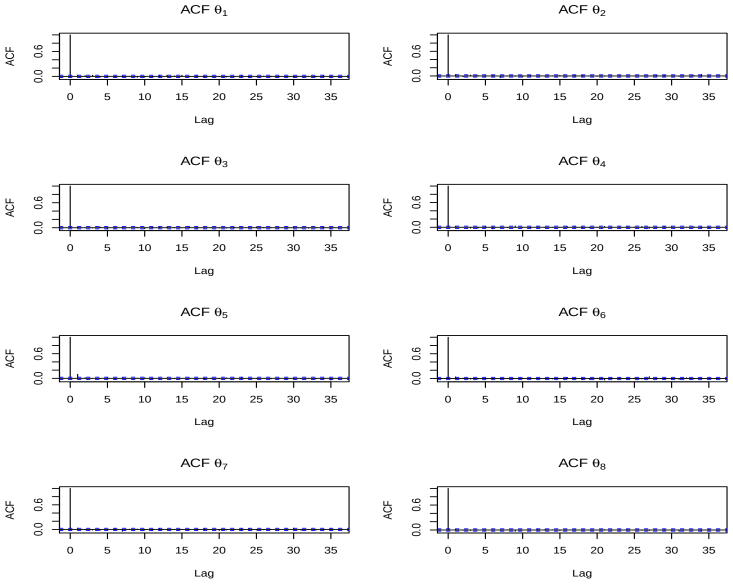

Figure 3: Autocorrelation plots - $\theta_1, ..., \theta_8$

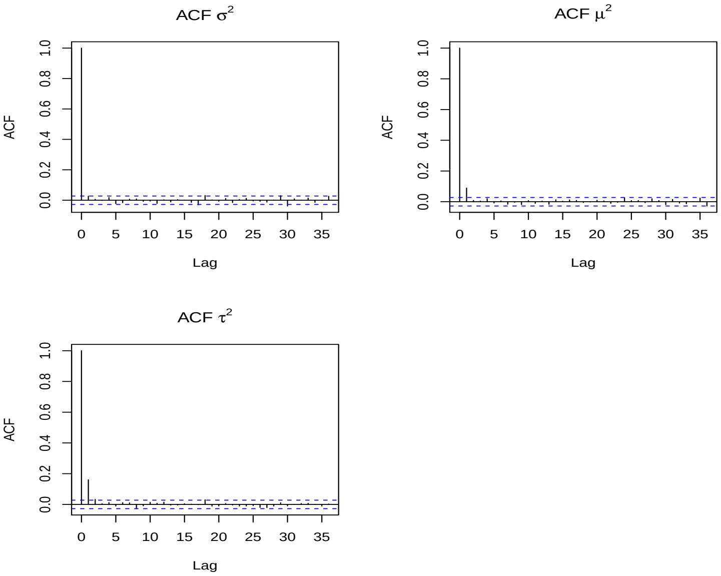

Figure 4: Autocorrelation plots - σ^2, μ, τ^2

(b)

```
>#  posterior means
> apply(SMT,2,mean)
     sigma2       mu     tau2
[1] 14.484101  7.548069  5.601912
> #posterior confidence regions
> apply(SMT,2,function(x) quantile(x,c(0.025,0.975)))
       sigma2       mu     tau2
2.5%  11.73272 5.914609  1.90546
97.5% 17.82171 9.127841 14.89749

 apply(prior,2,mean)
     sigma2       mu     tau2
[1]  6.041400  6.976526 86.502745
>
> #prior confidence regions
> apply(prior,2,function(x) quantile(x,c(0.025,0.975)))
       sigma2       mu     tau2
2.5%   1.855869  2.643866   2.715115
97.5% 18.044156 11.352968 391.676659
```

	prior mean	posterior mean	prior CI	posterior CI
σ^2	6.04	14.48	(1.86,10.04)	(11.73, 17.82)
μ	6.98	7.55	(2.64,11.35)	(5.91. 9.13)
τ^2	86.51	5.60	(2.72,391.68)	(1.91,14.90)

Table 1: Comparison of posterior means and confidence intervals - σ^2, μ, τ^2

The prior and posterior means for the variance parameters are very different, and the prior confidence interval estimates are a lot wider, particularly for τ^2. The prior and posterior mean estimates for μ are similar but again the prior confidence interval is wider.

(c) A comparison of R_{prior} and $R_{\text{posterior}}$ shows the data provide evidence of low to moderate between-school variation. The mode of the posterior density of R is around 0.2, indicating that approximately 20% of total variation is attributable to between-school variation. Contrast this to the relatively flat prior distribution on R, and we see that the data has provided evidence to reduce our uncertainty on the value of R.

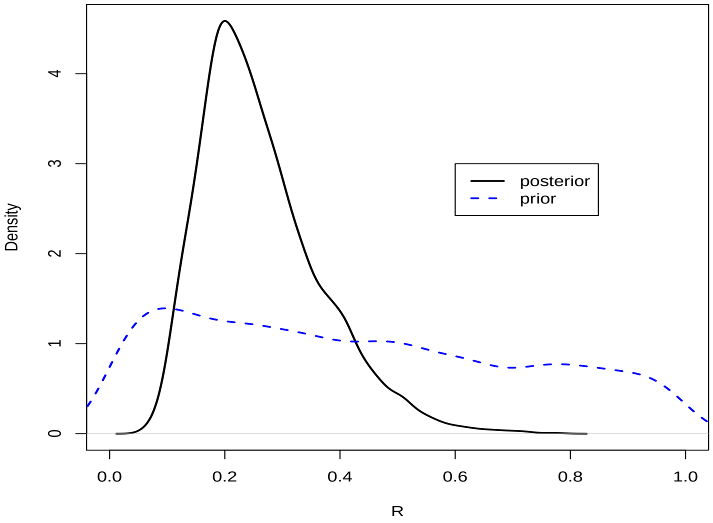

Figure 5: Comparison of R

(d)

```
> mean(THETA[,7]<THETA[,6])
[1] 0.5104
> min.theta<-apply(THETA[,-7],1,min)
> mean(THETA[,7]<min.theta)
[1] 0.3068
```

$$Pr(\theta_7 < \theta_6|\mathbf{y}_1, ..., \mathbf{y}_8) \approx 0.51$$

$$Pr(\theta_7 < \min(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_8)|\mathbf{y}_1, ..., \mathbf{y}_8) \approx 0.31$$

(e) The sample averages $\bar{y}_j$'s are very close to the posterior expectations $E[\theta_j|\mathbf{y}_1, ..., \mathbf{y}_8]$ (note that on the graph, the points have been jittered on the x-axis so that overlapping points are more clearly displayed).

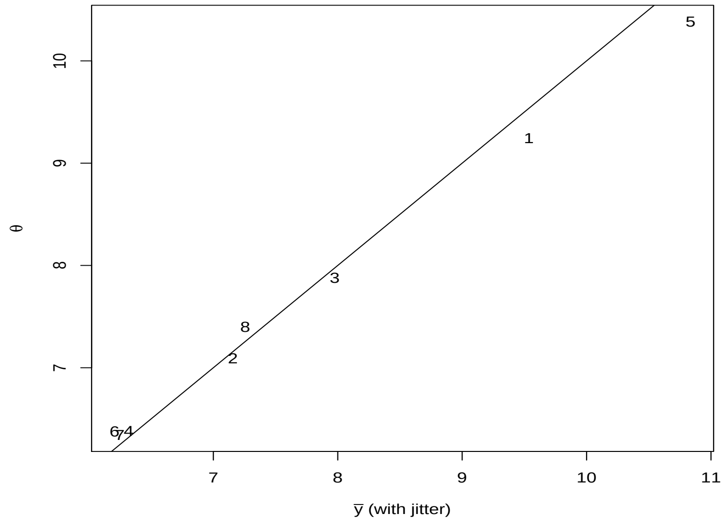

Figure 6: Comparison of $\bar{y}_j$'s and $E[\theta_j|\mathbf{y}_1, ..., \mathbf{y}_8]$

```
> mean(ybar)
[1] 7.64589
```

The sample mean is 7.65 and the posterior mean of μ is 7.55 (see part (b)). The two values are very close, indicating the strong influence of the likelihood on posterior inference.

The shrinkage effect for each school is given by $\frac{1/\tau^2}{n_j/\sigma^2+1/\tau^2}$. Below is a plot of the posterior mean of the shrinkage effects for each school versus the sample size. The shrinkage effects vary within a tight range of (0.11, 0.145) because the sample sizes only vary between (20, 25), however we do observe the downward trend in shrinkage effect (towards a common mean) as sample size increases.

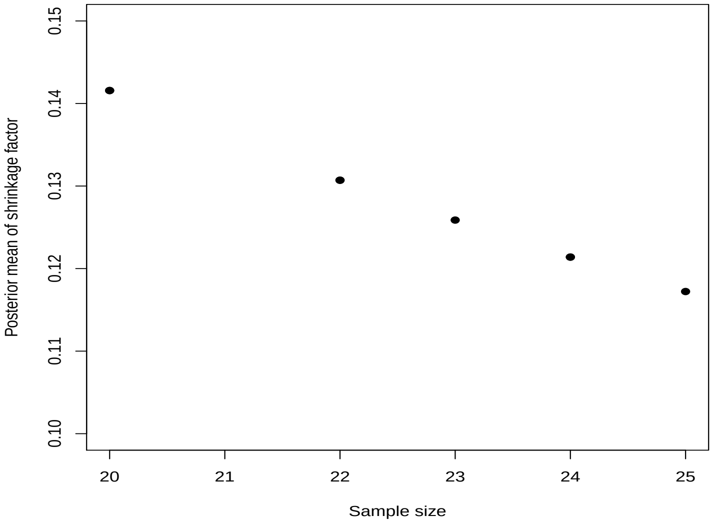

Figure 7: Posterior mean of shrinkage effect by school

Problem 2

(a)

$$\prod_{i=1}^{n} p(y_i|\alpha,\beta,x_i) \propto \prod_{i=1}^{n} \left(\frac{\exp^{\alpha+\beta x_i}}{1+\exp^{\alpha+\beta x_i}}\right)^{y_i} \left(\frac{1}{1+\exp^{\alpha+\beta x_i}}\right)^{1-y_i}$$

(b) Let's assume a weak informative prior, say $\alpha \sim N(0,10^2)$ and $\beta \sim N(0,10^2)$ and corr(α,β)=0. We could also assume a prior distribution on α and β that is independent and locally uniform in the two parameters, that is, $p(\alpha,\beta) \propto 1$. This represents the situation where we have no prior knowledge on the value of α and β.

Use a multivariate normal proposal distribution

$$\begin{pmatrix}\alpha^{(s+1)}\\ \beta^{(s+1)}\end{pmatrix} \sim N\left(\begin{pmatrix}\alpha^{(s)}\\ \beta^{(s)}\end{pmatrix}, \delta^2 s_y^2 (X^T X)^{-1}\right)$$

where $\delta = 6$ and $s_y^2(X^T X)^{-1}$ is an approximation of the variance-covariance matrix from the glm model.

```
> effectiveSize(BETA)
  alpha   beta
  1380    1361
> ac/S
[1] 0.42
```

After running the chain for 10000 iterations, the acceptance rate is 42% and the effective sizes are above 1000 for both parameters. Both MCMC diagnostics are acceptable for a two-parameter model.

(d) The priors are extremely flat (diffuse), while the posteriors are peaked. $Pr(\beta > 0|\mathbf{y}) = 0.99$ suggesting that as wing span increases, so does the probability of nesting success.

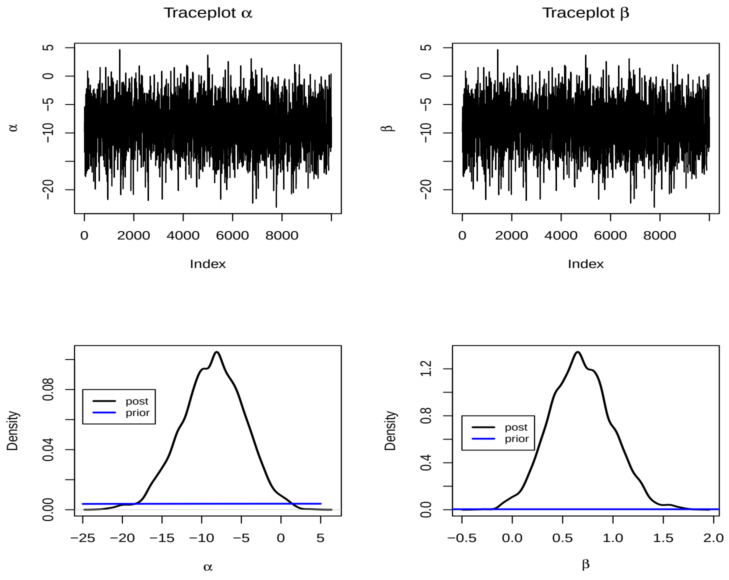

Figure 8: Traceplots and Prior/Posterior comparison

(e) Let's get a 95% confidence interval for $f_{\alpha,\beta}(x)$ when $x = \bar{x} = 12.96$.

```
x<-seq(10,15,0.02)
n.x<-length(x)
results<-NULL
for (i in 1: n.x){
results<-rbind(results,quantile(exp(BETA[,1]+x[i]*BETA[,2])/
          (1+exp(BETA[,1]+x[i]*BETA[,2])),prob=c(0.025,0.5,0.975)))
}
pdf("Fig15.pdf")
plot(x,results[,3],type="l",lwd=2,lty=2,
     ylim=c(min(results[,1]),max(results[,3])),ylab="",xlab="wingspan")
lines(x,results[,1],type="l",lwd=2,lty=2)
lines(x,results[,2],type="l",lwd=2,lty=1)
dev.off()
```

Observe that the confidence band gets larger for at the limits of the range of wingspan values.

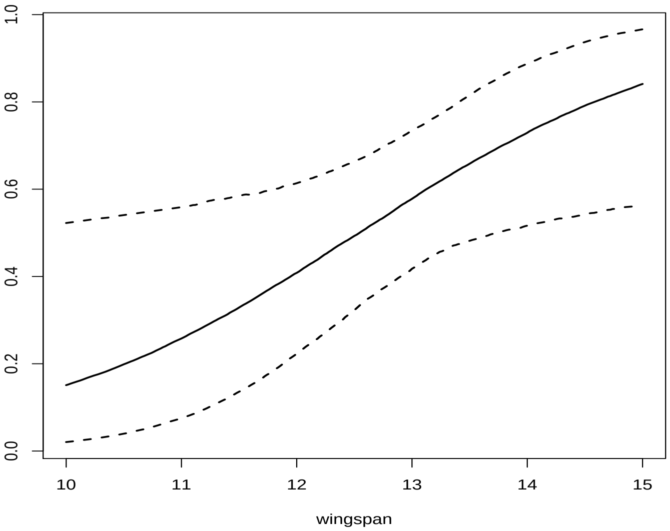

Figure 9: Predicted probability of mating by wingspan (95% posterior interval)

Problem 3

(a) Fitting the OLS model:

```
> p4data<-read.table("tplant.dat")
> colnames(p4data)<-c("height","time","pH")
> attach(p4data)
> m1<-lm(height~time+pH)
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2087     0.5891   12.24 7.45e-10 ***
time          3.9910     0.3280   12.17 8.14e-10 ***
pH            0.5778     0.1204    4.80 0.000167 ***

Residual standard error: 0.7334 on 17 degrees of freedom
Multiple R-squared: 0.9096,Adjusted R-squared: 0.899
F-statistic: 85.55 on 2 and 17 DF,  p-value: 1.339e-09
```

Both `time` and `pH` are significant predictors of tomato plant height, and both are positively associated with plant height. That is, more acidic soils are associated with taller tomato plants and obviously we expect tomato plants to grow over time. The model has high explanatory power ($R^2 = 0.9096$) and the estimate of the residual standard error is $\hat{\sigma} = 0.7334$.

(b) The residuals vs fitted values plot and the quantile plot of residuals show no patterns or deviation from normality to be concerned about. The '0' and '1' points on the residuals vs fitted values plot distinguish between data points at time==0 and time==1, and the separation of points by plotting symbol just indicates that data points where time==1 have higher fitted values.

However, on the autocorrelation plot of residuals, indicates non-trivial correlation at lags 2 and 4 between residuals which needs to be addressed in our model fit.

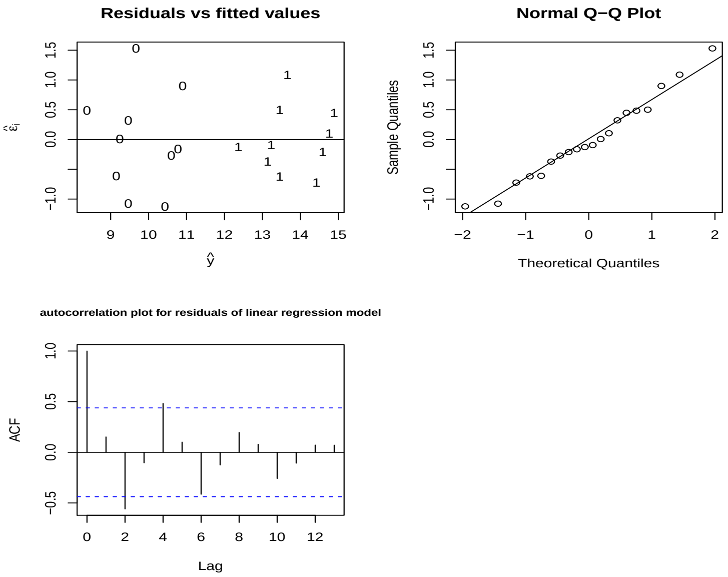

Figure 10: Residual diagnostic plots for tomato plant OLS model

(c) To allow for correlation between observations within a plant, we specify the variance -covariance matrix to be a block diagonal matrix as follows:

$$\Sigma = \sigma^2 \times \begin{pmatrix} 1 & \rho & \dots & \dots & \dots & \dots \\ \rho & 1 & \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & \rho & \dots & \dots \\ 0 & 0 & \rho & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

where ρ is a correlation parameter to be estimated, and the sampling model is $\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \Sigma)$.

To estimate the parameters in this model, we will run a Metropolis algorithm with diffuse prior distributions for the parameters $\boldsymbol{\beta}_0 = \mathbf{0}$, $\Sigma_0 = diag(1000)$, $\nu_0 = 1$, $\sigma_0^2 = 1$, and the prior for ρ will be the uniform distribution on (0,1).

Our proposal distribution for ρ will be $\rho^* \sim \text{Uniform}(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$. We set $\delta = 0.35$ to achieve good MCMC diagnostics.

We run the algorithm for 25000 iterations, and thin the sequence by storing every 25^{th} value. MCMC diagnostics are below. Effective sample sizes are all 1000 or above. The traceplots show good speed of mixing of the chains, and the autocorrelation plots show no high auto-correlations to be concerned about.

```
> apply(OUT.1000,2,effectiveSize )
[1] 1000.000 1096.028 1000.000 1000.000 1000.000
```

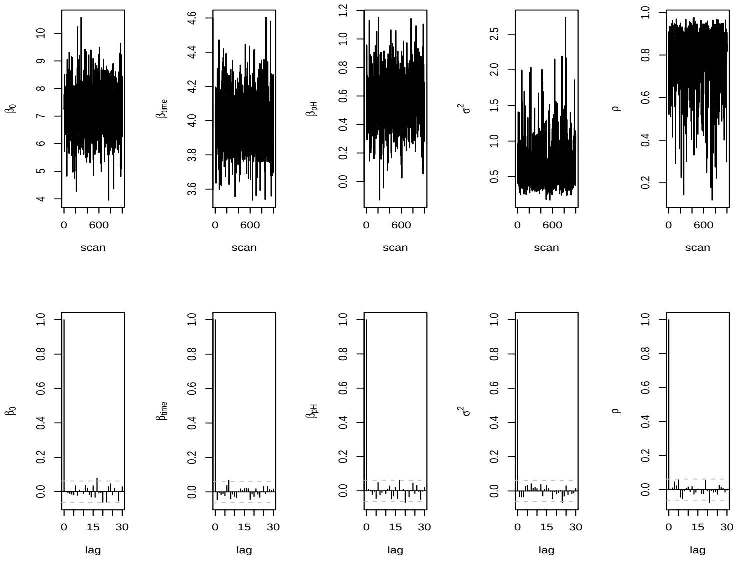

Figure 11: MCMC diagnostics

(d) The OLS estimates and posterior mean estimates from the Bayesian model are very similar. The standard error estimate associated with the coefficient of ‘pH’ is slightly bigger in the Bayesian model, but from both models we conclude a positive relationship between pH and tomato plant height. The posterior mean estimate of ρ is 0.79, indicating strong positive correlation between observations within a plant. On this point, we would favour the model with correlated responses to better represent the data.

```
> apply(OUT.1000,2,function(x) quantile(x,c(0.025,0.975)))
          [,1]     [,2]      [,3]      [,4]      [,5]
2.5%  5.612444 3.711481 0.2231237 0.2890406 0.4343778
97.5% 8.833429 4.300126 0.9090376 1.4962551 0.9466013

> apply(OUT.1000,2, mean)
[1] 7.1933650 3.9943800 0.5783851 0.6469995 0.7924479

> apply(OUT.1000,2,sd)
[1] 0.8341127 0.1513440 0.1753058 0.3154262 0.1336058
```

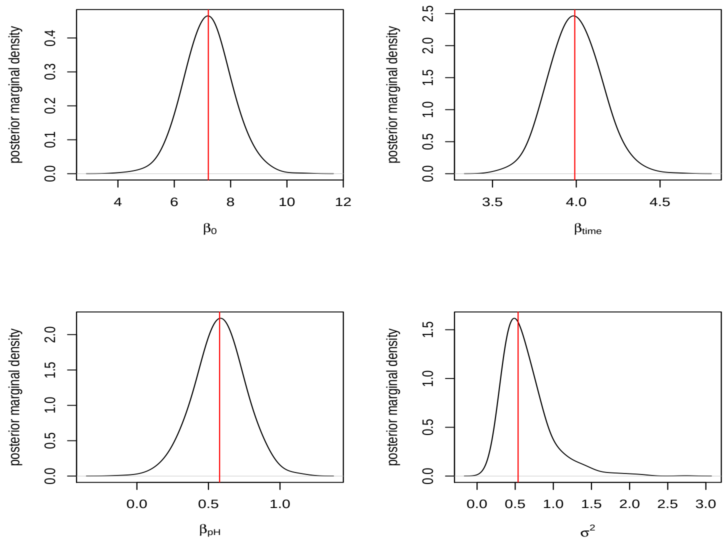

Figure 12: Marginal posterior densities

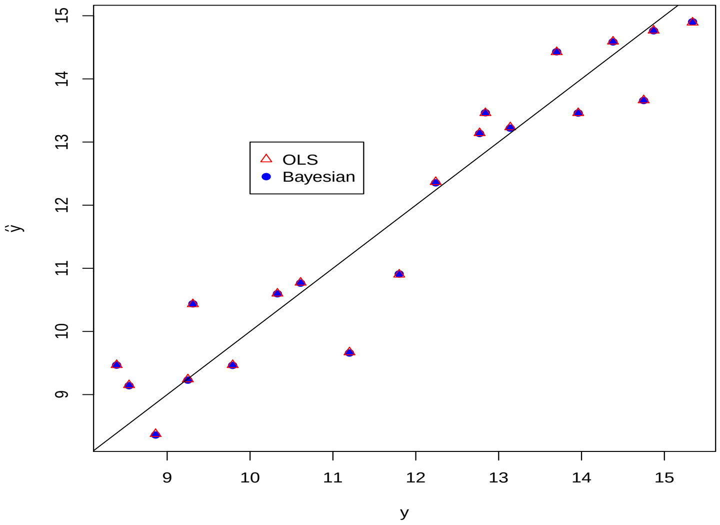

Figure 13: Comparison of fitted values

Problem 4

Let y_i be the observed number of deaths at hospital i, and let λ_i be the mortality rate per unit of exposure at hospital i. Let x_i denote the exposure at hospital i. The Bayesian hierarchical model assumptions are:

- Sampling model: $y_i \sim Pois(x_i\lambda_i)$
- Prior: $\lambda_i \sim Gamma(\alpha, \mu)$ where $E[\lambda_i] = \mu$ and $Var[\lambda_i] = \mu^2/\alpha$
- Hyperprior: $\mu \sim Norm(0, 1000)$ and $\alpha \sim Unif(0, 1000)$ (these are proper distributions with large uncertainties, to represent noninformative hyperprior distributions, but also to guarantee proper posterior distributions)

Then we can show that $\lambda_i|\alpha, \mu, y_i \sim Gamma(\alpha + y_i, \frac{\alpha}{\mu} + x_i)$

$$\mu|\alpha, \lambda_1, ..., \lambda_n, y \propto \mu^{-n\alpha} \exp(-\tfrac{1}{2}\mu^2 + \tfrac{\alpha}{\mu}\textstyle\sum_{i=1}^n \lambda_i)$$
$$\alpha|\mu, \lambda_1, ..., \lambda_n, y \propto \tfrac{\alpha^{\alpha n}}{\Gamma(\alpha)^n} \left(\textstyle\prod_{i=1}^n \lambda_i\right)^{\alpha-1} \exp(-\tfrac{\alpha}{\mu}\textstyle\sum_{i=1}^n \lambda_i)$$

To obtain posterior draws of $\mu, \alpha, \lambda_1, ..., \lambda_n$ we can use Gibbs sampling for $\lambda_1, ..., \lambda_n$ and a Metropolis algorithm for μ, α. We will use a normal jumping distribution for μ and a uniform jumping distribution for α. The respective jumping distributions will be centered at the current value $\mu^{(s)}$ or $\alpha^{(s)}$. We set the proposal variance for μ equal to the sample variance of the mortality rates across hospitals and the proposal variance for α equal to 1/6. These settings allowed us to achieve the desired acceptance rate (49% for μ and 50% for α).

MCMC diagnostics

The Metropolis-Gibbs algorithm was run for 10000 iterations. The full sequence of 10000 iterations showed high autocorrelation for most parameters, so we thinned the sequence, taking every 10th value to be used in posterior inference. Some autocorrelation plots and trace plots are shown below. Inspection of these plots shows that the sequence of draws for α still exhibits some relatively high autocorrelation and the variation in the trace plots could be reduced further, perhaps achievable by running the MCMC algorithm for more iterations.

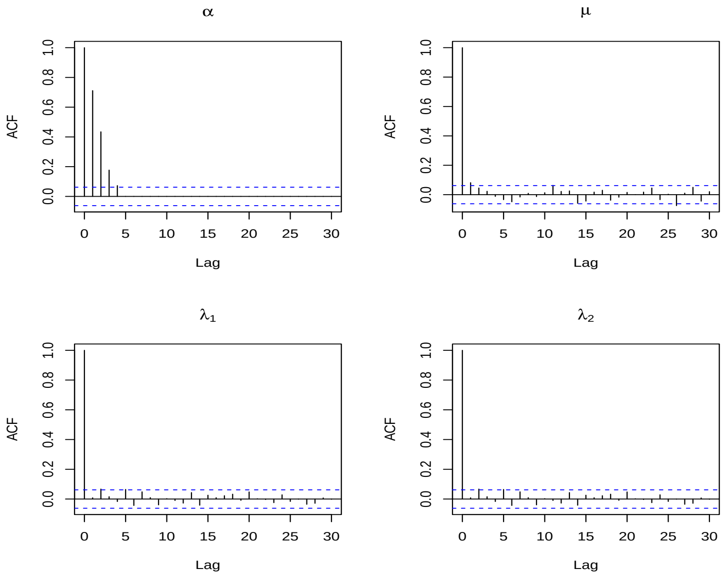

Figure 14: Autocorrelation plots

For each hospital $i = 1, ..., 94$ the posterior mean of λ_i is

$$E[\lambda_i|\alpha,\mu,y] = \frac{\alpha + y_i}{\alpha/\mu + x_i} = (1 - B_i)\frac{y_i}{x_i} + B_i\mu$$

where $B_i = \frac{\alpha}{\alpha+\mu x_i}$ is the shrinkage factor.

Below is a plot of the posterior mean shrinkages against the log of exposure. We can see that the shrinkage factor decreases as exposure increases. That is, for hospitals with small exposures, the Bayesian estimate shrinks the individual estimates more towards the combined estimate (although the mean shrinkage factor is at most 15%).

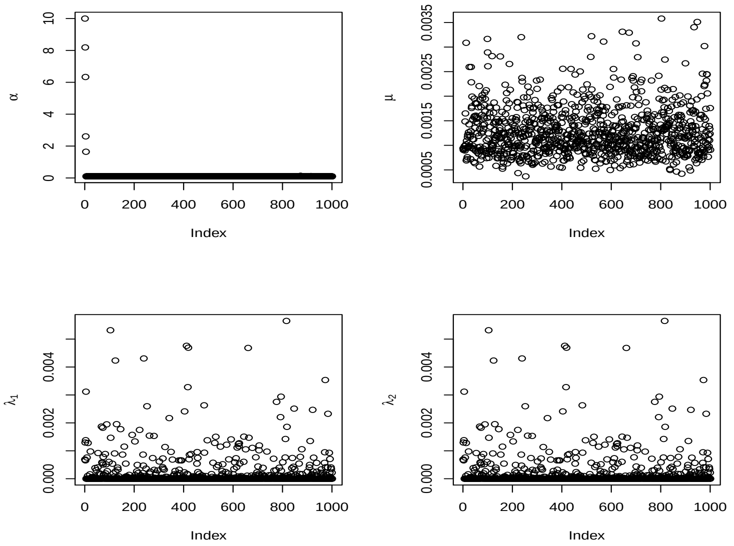

Figure 15: Traceplots

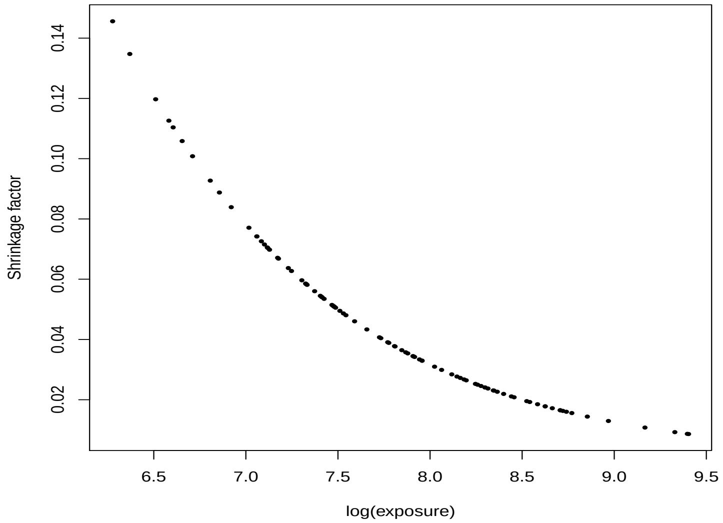

Figure 16: Mean shrinkage effects

To rank the hospitals by post transplant mortality, we can estimate the posterior mean of the mortality rate for each hospital. The number one hospital will have the lowest posterior mean. Below is a plot of the posterior means of λ_i by hospital. We can see that hospital 9 is the worst performing hospital with the highest posterior mean estimate of post transplant mortality rate. There are about 15 hospitals who equally have the best performance as measured by the lowest mean mortality rate estimates, including hospitals 27, 37, 63 and 85.

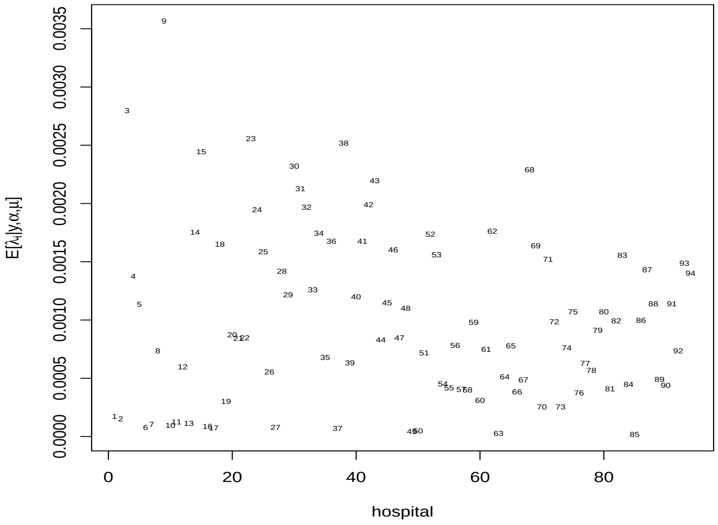

Figure 17: Posterior predictive p-values

For the posterior predictive checks, our test statistic is the observed number of deaths for each hospital. Below is a plot of the posterior predictive p-values for each hospital. Most of the p-values are within the range (0.2,0.4). There are a handful of hospitals (15 out of the 94), where the p-value is in the range (0.05, 0.10). For these hospitals, zero deaths were recorded in the observed data so the model does not perform as well (relatively speaking) in picking up the zero counts. As an alternative, a two-stage Bayesian model could be implemented where we first predict whether the count is zero or not, and then apply a truncated Poisson model to model the positive counts.

Problem 5 [STAT4116/STAT7016 ONLY]

(a)

$$\begin{aligned}
p(\boldsymbol{\theta}, \mu, \tau^2|\mathbf{y}) &\propto p(\boldsymbol{\theta}, \mu, \tau^2)p(\mathbf{y}|\boldsymbol{\theta}, \mu, \tau^2) \\
&= p(\mu, \tau^2)p(\boldsymbol{\theta}|\mu, \tau^2)p(\mathbf{y}|\boldsymbol{\theta}, \mu, \tau^2) \\
&\propto p(\mu, \tau^2)\prod_{j=1}^{J} \frac{1}{\theta_j(1-\theta_j)} \times \frac{1}{\tau}\exp\left(-\frac{1}{2\tau^2}(\text{logit}(\theta_j)-\mu)^2\right)\prod_{j=1}^{J}[\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}]
\end{aligned}$$

The factor $\frac{1}{\theta_j(1-\theta_j)}$ is $\frac{dx_j}{d\theta_j}$ (where $x_j = \log\frac{\theta_j}{1-\theta_j}$). We need this since we want $p(\theta_j|\mu, \tau^2)$ rather than p(logit$(\theta_j)|\mu, \tau^2)$

(b) Even though we can look at each of the J integrals individually - the integrand separates into independent factors - there is no obvious analytic technique which permits evaluation of these integrals. In particular, we cannot recognise a multiple of a familiar density function inside the integrals. One might try a substitution like $u_j = \text{logit}(\theta_j)$ or $v_j = \theta_j/(1-\theta_j)$, but neither substitution turns out to be helpful.

(c) In order for the expression $\frac{p(\boldsymbol{\theta},\mu,\tau^2|\mathbf{y})}{p(\boldsymbol{\theta}|\mu,\tau^2,\mathbf{y})}$ to be useful, we would need to know $p(\boldsymbol{\theta}|\mu, \tau^2, \mathbf{y})$ in exact form. Knowing it up to proportionality in $\boldsymbol{\theta}$ is insufficient because our goal is to use $p(\boldsymbol{\theta}|\mu, \tau^2, \mathbf{y})$ to find another density, $p(\mu, \tau^2|\mathbf{y})$, that depends on μ and τ^2, and the proportionality constant required in $p(\boldsymbol{\theta}|\mu, \tau^2, \mathbf{y})$ will depend on μ and τ^2.