UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSB 6816

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 2 — 24 May 2019

MEMORANDUM

TYD/TIME: 160 Minutes PUNTE/MARKS: 45

INSTRUCTIONS:

- Answer all questions in a single Word document. Please convert to .pdf at the end.
- Label questions clearly, as it is done on this question paper.
- All results accurate to 2 decimal places.
- Show all derivations, formulas, code, sources and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, devices or websites may be accessed prior to submission.

Question 1

This entire test is based on a supplied data set given on your LMS as 'Bayes2019Test2regData.csv'.

The target variable of interest is the Forced Exhalation Volume (FEV) of a sample of children of various ages. It is a proxy for lung capacity. In this question we will analyse FEV on the natural log scale with a Normal distribution.

Explanatory variables included in this data set include the age of the child, their height, their gender, and whether or not they smoke.

(a) Read in the data. Create a new variable lnFEV and show that it has a mean close to 1. [3]

```
md <- read.csv('Bayes2019Test2regData.csv')
md$lnFEV <- log(md$FEV)
mean(md$lnFEV)
```

✔✔✔

(b) Do an ordinary linear regression of lnFEV on age, height, gender, and smoking status. Mention which explanatory variables are deemed significant. [3]

```
names(md)
(summary1 <- summary(lm1 <- lm(lnFEV~age+ht+sex+smoke, data=md)))
```

✔

```
            Estimate Std.  Error t value Pr(>|t|)
(Intercept) -2.049564 0.202500 -10.121 <2e-16 ***
age 0.018095 0.007038 2.571 0.0113 *
ht 0.044456 0.003947 11.264 <2e-16 ***
sex 0.059533 0.029219 2.037 0.0437 *
smoke 0.009507 0.037327 0.255 0.7994
Residual standard error:  0.1634 on 125 degrees of freedom
Multiple R-squared:  0.7759, Adjusted R-squared:  0.7687
F-statistic:  108.2 on 4 and 125 DF, p-value:  < 2.2e-16
```

✔

All explanatory variables other than smoking status are deemed significant ✔as they have low p-values.

(c) Create a function that generates random initial values for each model parameter according to the estimates and standard errors reported by the regression above. [2]

```
library(MASS)
n <- nrow(md)
k <- 5
inits1 <- function() {list(B=mvrnorm(1,coef(lm1),vcov(lm1)), tau=(rchisq(1,(n-k))/sum
    (resid(lm1)^2)))}
```

✔ ✔

(d) Repeat the regression above by simulating at least 10000 samples from the joint posterior of the parameters, assuming independent, vague, but proper priors for all parameters. Give the posterior mean estimates of all the parameters. [7]

```
library(R2OpenBUGS)
reg1 <- function() {
  for (i in 1:n) {
    y[i]~dnorm(mu[i],tau)
    mu[i]<-inprod(X[i,1:k],B[1:k])
  }
  for (j in 1:k) {
    B[j]~dnorm(0,0.0001)
  }
  tau~dgamma(0.001,0.001)
}
write.model(reg1,'reg1.txt')
Xmat <- model.matrix(~age+ht+sex+smoke,data=md)
reg1data <- list(y=md$lnFEV, X=Xmat, n=n, k=k)
reg1out <- bugs(reg1data,inits1,c('B','tau'),5000,'reg1.txt',3,1500,1,debug=TRUE)
reg1out$mean
```

✔ ✔ ✔ ✔ ✔ ✔ ✔

$\boldsymbol{\beta} = [-2.04,\ 0.02,\ 0.04,\ 0.06,\ 0.01],\ \tau = 37.34$

(e) Use your simulations to calculate a Bayesian equivalent of p-values for each coefficient (to attempt to see how they differ from 0) and discuss any differences with the ordinary model. [4]

```
pvalfunc <- function(sims,target=0) { 2*min(mean(sims<target),mean(sims>target)) }
(pvalsBetaReg1 <- apply(reg1out$sims.list$B,2,pvalfunc))
```

✔ ✔

0.000000000 0.003238095 0.000000000 0.045523810 0.849714286 ✔

No differences in outcomes. ✔

(f) Simulate a sample of size at least 10000 from the posterior predictive distribution of every observation. Reverse the log transform and calculate 95% HPD intervals for every observed FEV value. [5]

```
source('hpd.r')
mumat <- reg1out$sims.list$B%*%t(Xmat)
predmat <- mumat + matrix(rnorm(length(mumat),0,1/sqrt(reg1out$sims.list$tau)),nrow(
    mumat),ncol(mumat))
intervals <- exp(apply(predmat,2,hpd.interval))
```

✔ ✔ ✔ ✔ ✔

(g) Calculate the empirical coverage of your fitted model. Is it in line with what you expected? [2]

```
(cover1 <- mean((intervals[1,]<md$FEV)&(intervals[2,]>md$FEV)))
```

✔

Something close to 0.95 ✔

(h) Report the DIC of your model. Now modify your model to allow for a variance that changes linearly with the age of the child. Give the new DIC and state whether it is an improvement. [7]

```
reg1out$DIC
reg2 <- function() {
  for (i in 1:n) {
    y[i]~dnorm(mu[i],tau[i])
    mu[i]<-inprod(X[i,1:k],B[1:k])
    tau[i]<-exp(v0+v1*age[i])
  }
  for (j in 1:k) {
    B[j]~dnorm(0,0.0001)
  }
  v0~dnorm(0,0.0001)
  v1~dnorm(0,0.0001)
}
write.model(reg2,'reg2.txt')
inits2 <- function() {list(B=mvrnorm(1,coef(lm1),vcov(lm1)), v0=log(rchisq(1,(n-k))/
    sum(resid(lm1)^2)), v1=0)}
reg2data <- list(y=md$lnFEV, X=Xmat, n=n, k=k, age=md$age)
reg2out <- bugs(reg2data,inits2,c('B','v0','v1'),5000,'reg2.txt',3,1500,2,debug=TRUE)
reg2out$DIC
```

✔ ✔ ✔ ✔ ✔

Model 1: -93.97 ✔

Model 2: -93.75, which is worse, but not much. ✔

Total for Question 1: 33

Question 2

In this question we want to model the FEV as given with a Gamma distribution. We would like the expected value to change with age and gender only, and we want the variance to be assumed constant throughout.

This model requires a parameter transformation and priors as follows:

$$\begin{aligned} FEV_i &\sim Gamma(\alpha, \lambda) \\ \alpha &= \mu^2\tau \\ \lambda &= \mu\tau \\ \mu &= exp\left(\beta_0 + \beta_1 age + \beta_2 gender\right) \\ \beta_i &\sim N(0, 10^2)\ \forall\ i \\ \tau &\sim Exp(0.1) \end{aligned}$$

(a) Simulate a sample from the joint posterior distribution (at least 10000 draws) and give Bayesian a equivalent of p-values for β_0, β_1, and β_2 (to attempt to see how they differ from 0). [7]

```
reg3 <- function() {
  for (i in 1:n) {
    y[i]~dgamma(alpha[i],lambda[i])
    alpha[i] <- pow(mu[i],2)*tau
    lambda[i] <- mu[i]*tau
    mu[i]<-exp(b0+b1*age[i]+b2*gender[i])
  }
  b0~dnorm(0,0.01)
  b1~dnorm(0,0.01)
  b2~dnorm(0,0.01)
  tau~dexp(0.1)
}
write.model(reg3,'reg3.txt')

lm2 <- lm(lnFEV~age+sex, data=md)
B2<-mvrnorm(1,coef(lm2),vcov(lm2))
yhat2 <- exp(model.matrix(~age+sex, data=md)%*%B2)
inits3 <- function() {list(b0=B2[1],b1=B2[2],b2=B2[3], tau=rchisq(1,(n-3))/sum((yhat2
    -md$FEV)^2))}
reg3data <- list(y=md$FEV, n=n, age=md$age, gender=md$sex)
reg3out <- bugs(reg3data,inits3,c('b0','b1','b2','tau'),8000,'reg3.txt',2,3000,2,
    debug=TRUE)

pvalfunc(reg3out$sims.list$b0)
pvalfunc(reg3out$sims.list$b1)
pvalfunc(reg3out$sims.list$b2)
```

✔ ✔ ✔ ✔ ✔

The p-values are all close to zero. ✔ ✔

(b) Simulate a sample of size at least 10000 from the posterior predictive distribution of every observation. Give 95% HPD intervals for every observed FEV value. Comment on the coverage. [5]

```
attach.bugs(reg3out)
nsims <- reg3out$n.sims
mumat <- exp(matrix(b0,nsims,n)+matrix(b1)%*%t(md$age)+matrix(b2)%*%t(md$sex))
taumat <- matrix(tau,nsims,n)
predmat <- matrix(rgamma(nsims*n,mumat^2*taumat,mumat*taumat),nsims,n)
intervals <- apply(predmat,2,hpd.interval)
(cover2 <- mean((intervals[1,]<md$FEV)&(intervals[2,]>md$FEV)))
```

✔ ✔ ✔

Coverage is about 96% which is a little high. ✔ ✔[This is an example answer.]

Total for Question 2: 12

Total half marks on memo = 90 vs. 90 = Double total margin points (=45).