Homework 3
AMS 207 Spring 2018
Wyara Vanesa Moura e Silva May 31

Question 16

Consider the data available as "birthweight" form the package "LearnBayes". Fit a linear regression that considers age and gender as explanatory variables for birth weight.

Some descriptive statistics about the data: The data is from a study where one is interested in predicting a baby's birthweight based on the gestational age and the baby's gender. Figure 1 shows the boxplot of the explanatory variable (gestational age) and the response variable (baby's birthweight) by the gender of the baby.

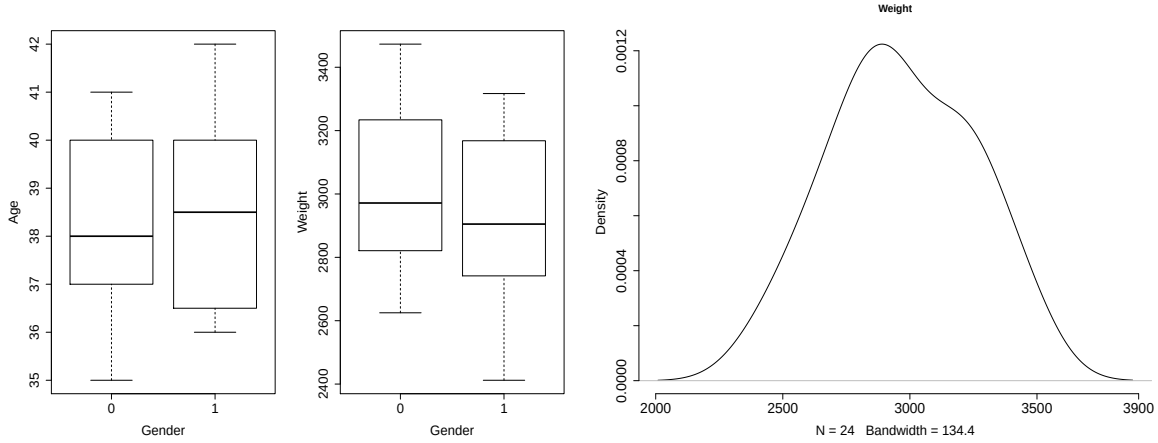

Figure 1: Boxplot of the age and weight by gender (left). Density of the response variable weight (right).

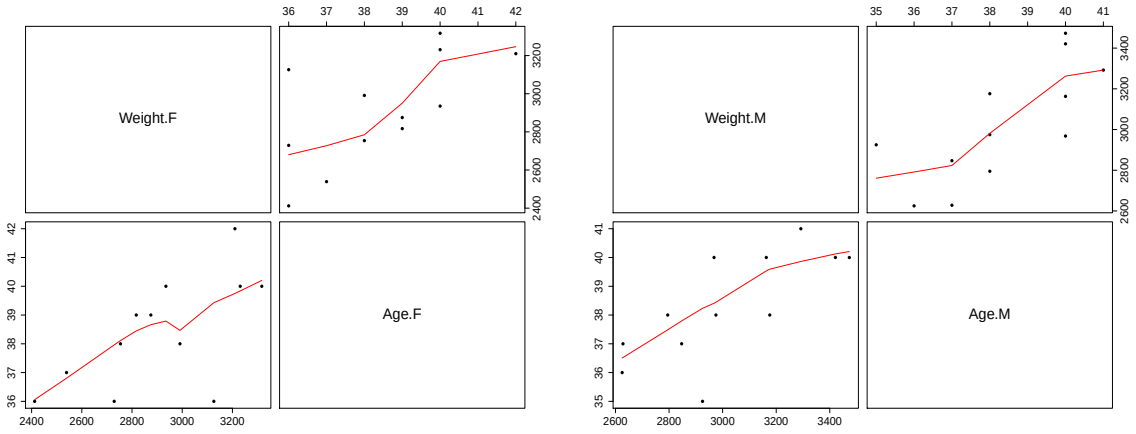

Figure 2: Pairs plots for the correlation between the gender (Female left side and Male right side) with weight and age.

It can be seen that both genders have about the same median when evaluating the age and

weight, however, there is a considerable variability in both variables. In particular, the weight of the male babies tend to be bigger than the female babies. The density of the response variable weight (Figure 1 right side) looks like a normal density.

In Figure 2 we can see the positive linear relationship between the pairs of gender with weight and age. In addition, if we focus in the scatter plot for the female there is an unusual point that has a small age but has great weight compared to other the other points, maybe this is a possible outlier. For the covariate male, we can see also a possible discrepant point, with the same behavior as in the female covariate.

16.1

Describe the posterior distribution of the regression parameters using a sample-based approach.
Solution:

In order to sample from the posterior distribution of the regression parameters we need some features, we use as reference book for this analysis Gelman et al. (2014).

First, in this case we are going to fit a normal linear regression, in which, we have a response variable $\mathbf{y}$ (weight) and 2 predictors variable x_1 and x_2, (age and gender). In which, the distribution of the y_i give $\mathbf{X}$ is describe by:

$$E(y_i|\boldsymbol{\beta}, \mathbf{X}) \quad = \quad \beta_0 + \beta_1 AGE_i + \beta_2 GENDER_i$$

In this model, the covariate gender are categorical, in which is represented by binary indicators; 0 (1) for male (female). Then, describing this as matrix, the likelihood for the $\mathbf{y}$ would be:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) \quad \sim \quad \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

The prior used in this normal regression model was based in a noninformative prior distribution given by:

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}) \quad \propto \quad \frac{1}{\sigma^2}$$

Then, in order to find the posterior distribution, for this model we are going to factor the joint posterior as:

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \quad \propto \quad p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X})$$

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \quad \propto \quad \pi(\boldsymbol{\beta}|\sigma^2, \mathbf{y})\pi(\sigma^2|\mathbf{y})$$

In which, we have as distribution for this two factors of the posterior distribution for the normal linear model:

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) \quad &\sim \quad \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta\sigma^2) \\ \pi(\sigma^2|\mathbf{y}) \quad &\sim \quad \mathcal{IG}((n-p)/2, (n-p) * \hat{\sigma}^2/2) \end{aligned} \tag{1}$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}; \quad V_\beta = (\mathbf{X}^T\mathbf{X})^{-1}; \quad \hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{2}$$

These posterior distributions are not difficult to find, we chose not to show the steps, but in Christensen (2011) a few steps can be found.

Now, we can draw samples from the posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ using the following algorithm (see the `R` code in Section 1):

1. It was computed the QR factorization, $\mathbf{X} = \mathbf{QR}$, using the function `lm` in `R`, where $\mathbf{Q}$ is an $n \times p$ matrix of orthonormal columns and $\mathbf{R}$ is a $p \times p$ upper triangular matrix.

2. Then, it was found the $\mathbf{R}^{-1}$ using the function `backsolve` in `R`, solving the system $\mathbf{R}\mathbf{R}^{-1} = \mathbf{I}$.

3. Now, we calculated the $V_\beta = \mathbf{R}^{-1}(\mathbf{R}^{-1})^T$.

4. Then we obtained $\hat{\boldsymbol{\beta}}$ solving the system $\mathbf{Q}^T\mathbf{y} = \mathbf{R}\hat{\boldsymbol{\beta}}$, using the function `backsolve` in `R`.

5. Then, it was computed the $\hat{\sigma}^2$ (see in equation (2)).

6. Finally, we sampled σ^2 (see distribution in (1)) first and then $\boldsymbol{\beta}$ (see distribution in (1)).

The QR factorization $\mathbf{X}$ turns out to be computationally efficient than using directly $\mathbf{X}$.

The sampling for the posterior distribution of σ^2 and $\boldsymbol{\beta}$ was computed by directly sampling. First, it was sampled $M = 10000$ of σ^2 and then $\boldsymbol{\beta}$ given the value of the σ^2 sampled, as shown in the distribution of the factorized posterior distributions (see the `R` code in Section 1).

Figure 3 displays the posterior distributions for the unknown parameters in the normal regression model $(\beta_0, \beta_1, \beta_2, \sigma)$.

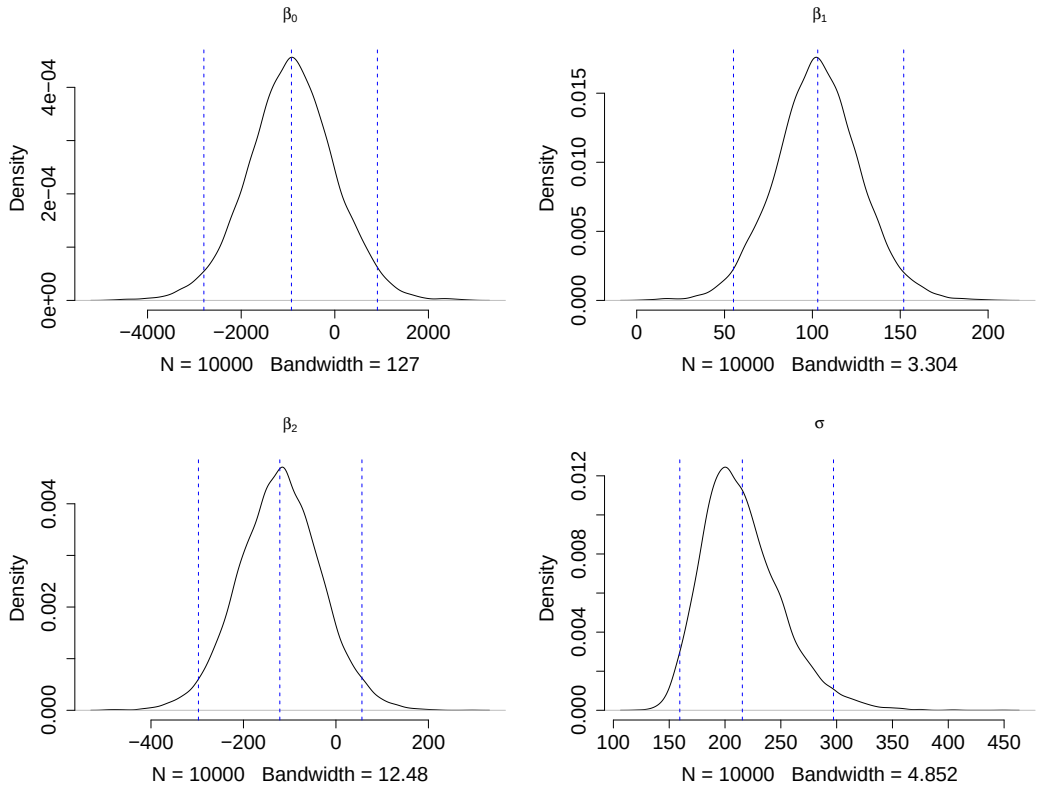

Figure 3: Posterior densities for the unknown parameters in the normal regression model $(\beta_0, \beta_1, \beta_2, \sigma)$. Blue dashed line correspond to the credibility interval (2.5%;97.5%) and mean

Table 1 shows some points inference for the posterior distribution of the parameters. We can notice that the mean posterior has close values to the one estimated by the `lm` function in `R`, which are, $\beta_0 = -924.94$, $\beta_1 = 103.02$, $\beta_2 = -121.25$. This happen because we used a noninformative prior. Also from the analysis using `lm` with all observations, it returns that only

the covariable age is significant, but without the 19th observation, all three betas have become significant.

Table 1: Summary of inference for the posterior mean and quantiles, for all parameters estimated.

	mean	standard deviation	2.5%	50%	97.5%
β_0	-925.4417	942.5644	-2778.5098	-933.8336	938.4001
β_1	103.0014	24.5126	54.8193	103.2046	151.3637
β_2	-120.3584	88.9555	-294.7337	-120.5255	55.2610
σ	215.2363	34.6620	159.9266	211.1457	294.7694

Also from the analysis using `lm` with all observations, it returns that only the covariate age is significant, but without the 19th observation, all three betas have become significant.

We tried another method by:

1. Compute the QR factorization, $\mathbf{X} = \mathbf{QR}$, using the function `lm` in `R`, where $\mathbf{Q}$ is an $n \times p$ matrix of orthonormal columns and $\mathbf{R}$ is a $p \times p$ upper triangular matrix.
2. Then, we can obtain $\hat{\boldsymbol{\beta}}$ solving the system $\mathbf{Q}^T\mathbf{y} = \mathbf{R}\hat{\boldsymbol{\beta}}$, using the function `backsolve` in `R`.
3. It was computed the $\hat{\sigma}^2$ (see in equation (2)).
4. Now, compute $z \sim \mathcal{N}(0, \mathbf{I})$. Then, we solve the system $\mathbf{R}\boldsymbol{\beta} = z$, using the function `backsolve` in `R`.
5. Sample σ^2 (see distribution in (1)). Then, do $\sigma\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}$.

The results from this method are similar to the previous method (see code in Section 1).

For a brief residuals analysis it was done two types of diagnostic. First, it was plotted the probability of a point to be an outlier and the standardized residuals versus the fitted values computed using the posterior distribution of the parameters.

The posterior probability that each observation to be an outlier was calculated by:

$$p_i = P(|\epsilon_i| > k\sigma|y) = \int (1 - \Phi(z_1) + \Phi(z_2))g(\sigma^2|y)d\sigma^2.$$

where,

$$z_1 = (k - \hat{\epsilon}_i/\sigma)/\sqrt{h_{ii}}, \quad z_1 = (-k - \hat{\epsilon}_i/\sigma)/\sqrt{h_{ii}}, \quad \text{and} \quad \hat{\epsilon}_i = y_i - x_i\hat{\beta}.$$

The h_{ii} correspond to the ith element of the hat matrix, and σ is the posterior distribution of this parameter. For the comparison, we used $k = 3$. See code in Section 1 for more details and also as reference book Albert (2009). The standardized residuals was calculated by: for the data point i, it was computed $(y_i - X_i\hat{\boldsymbol{\beta}})/\hat{\sigma} * \sqrt{1 - h_{ii}}$, see the respective equations for the parameters in (2) (Albert, 2009)

Figure 4 displays the two graphs of diagnostic, and we can confirm by both that the point 19 is indeed an outlier. As we can see the value for this observation in each diagnosis are far apart in relation to the other observations. So this observation has the response different from the predicted value. The 19th observation happens to be the observations point out in the begging of the analysis corresponding to a female baby who has a small age but has great

weight compared to other female babies. For more details about the calculation see the code in the Section 1.

We made a brief comparison for a model with and without the 19th observation, and the values in the posterior means for each parameter estimated obtained approximately an increase of 69.19%, 16.20% and 38.93%, for β_0, β_1 e β_2, respectively. But for σ we had a decrease of 12.47%.

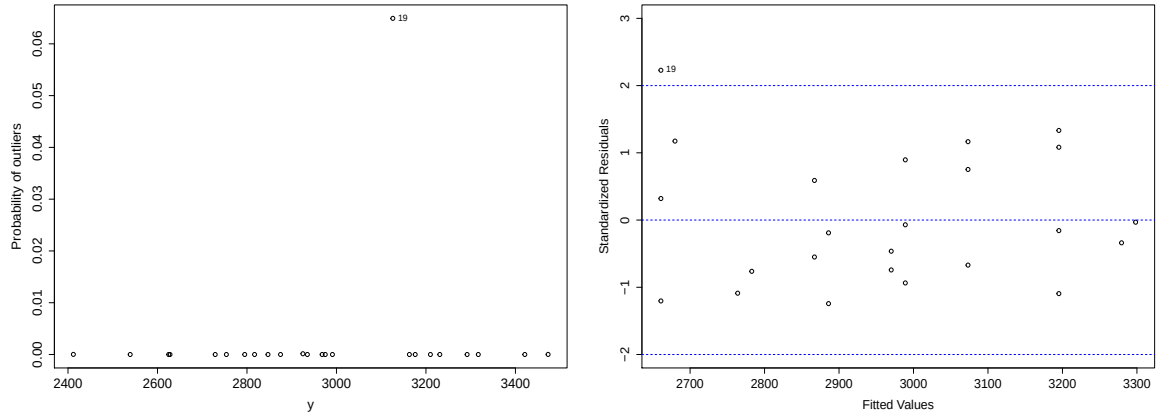

Figure 4: Plot of posterior probabilities of outliers for all observations (left side). Plot of standardized residuals versus fitted values.

16.2

Explore the predictive posterior distribution for the birth weight of children in the following four cases: (a) 36 week female/male; (b) 40 week female/male. Compare.
Solution:

First, we simulated replicas from the posterior predictive distribution conditional on the observed set of covariates $\mathbf{X}$, in order to evaluate the consistency of the model (Gelman et al., 2014).

For this part, it was done:

1. Computing σ^2 and $\boldsymbol{\beta}$ from the joint posterior distribution.
2. Computing $\mathbf{y}^{rep}$ from the density $\mathbf{y}^{rep} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, as noticed it was used the respective samples for $\boldsymbol{\beta}$ and σ^2 from the first step. The number of replicas was 10000.

Figure 5 shows on the right side the summary for each posterior predictive distribution sampled by the 5th and 95th quantiles. We can see how the observed data are consistent with corresponding predictive posterior distributions. However, we can notice that the observation 19th are outside of the 90% interval and this point can be seen as a possible outlier. In addition, on the left side of Figure 5 we have a plot of the observed values versus predicted values.

We can check the consistency of this particular point by plotting the replicate of the 19th marginal and calculate the p-value to check if the observed value is in the tail of this distribution, if so, then we will have another indication that this is an outlier. Figure 8 confirms (now using the posterior predictive distribution) that indeed this point has a great potential to be an outlier, as the p-value is 0.0266.

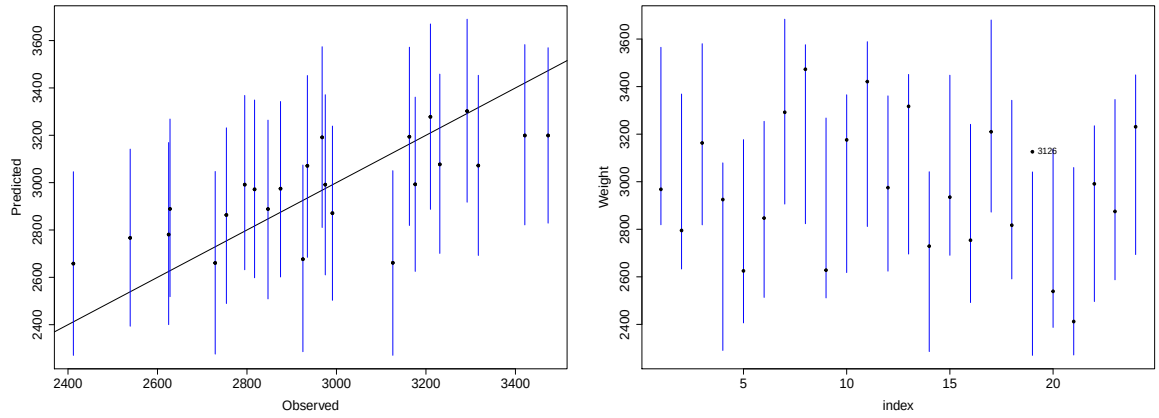

Figure 5: Observed values versus predicted values, with the 95th predictive percentile interval, and the predictive posterior mean indicated by the solid points (left side). Posterior predictive distributions of y_i^{rep} with the actual weight y_i indicated by the solid points. The weight that exceed the 95th percentile of the predictive distribution are labeled (right side).

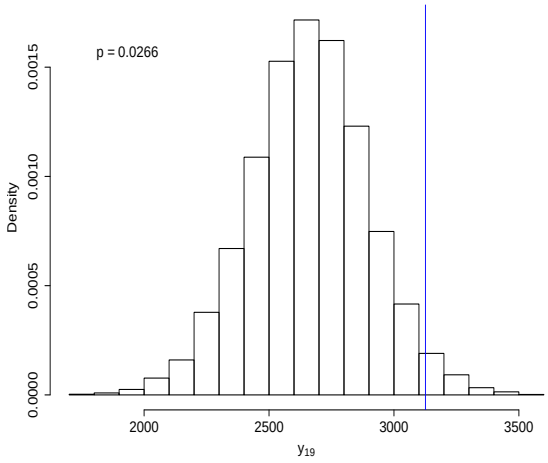

Figure 6: Histogram of samples from the posterior predictive distributions of y_{19}^{rep}. Blue solid lines corresponds to the observed values y_{19}. Also the p-values from the comparison.

Now we are going to explore the predictive posterior distribution for some specific cases of covariates. Therefore, we have four sets of covariates, which are given by:

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 36 & 0 \\ 1 & 36 & 1 \\ 1 & 40 & 0 \\ 1 & 40 & 1 \end{pmatrix}$$

These covariates are equivalent 36 and 40 weeks of gestational age evaluating when the gender of the baby is female or male (reminder: 0 (male) and 1 (female)).

In order to sample $\tilde{\mathbf{y}}$, i.e. the posterior predictive for the matrix $\mathbf{X}_1$ of covariates, we need first to sample $(\boldsymbol{\beta}, \sigma)$ from the joint posterior distribution, which we have from the earlier analysis. Then, we can draw $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{X}_1\boldsymbol{\beta}, \sigma^2\mathbf{I})$ (Gelman et al., 2014).

For the expected response for each set of covariate values, it was done by sampling from $\mathbf{y}^* = \mathbf{X}_1\boldsymbol{\beta}$, which $\boldsymbol{\beta}$ is from the posterior distribution (Albert, 2009).

We can see in Figure 8 that the predictive distribution for future observations is wider than the mean response distributions given in Figure 7, in all four sets of covariates.

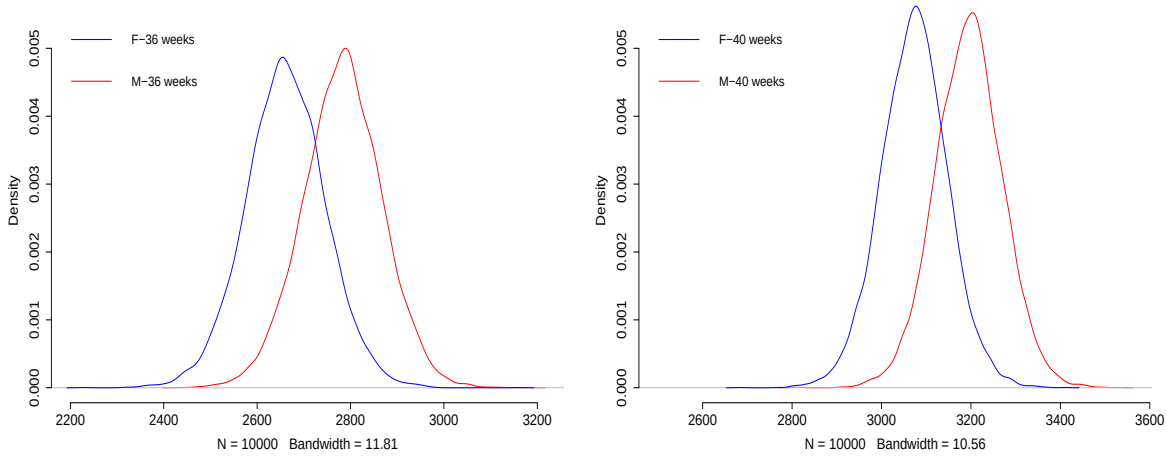

Figure 7: Densities from the simulated draws of the posterior of the mean weights ($\mathbf{y}^* = \mathbf{X}_1\boldsymbol{\beta}$) of the four sets of covariate values.

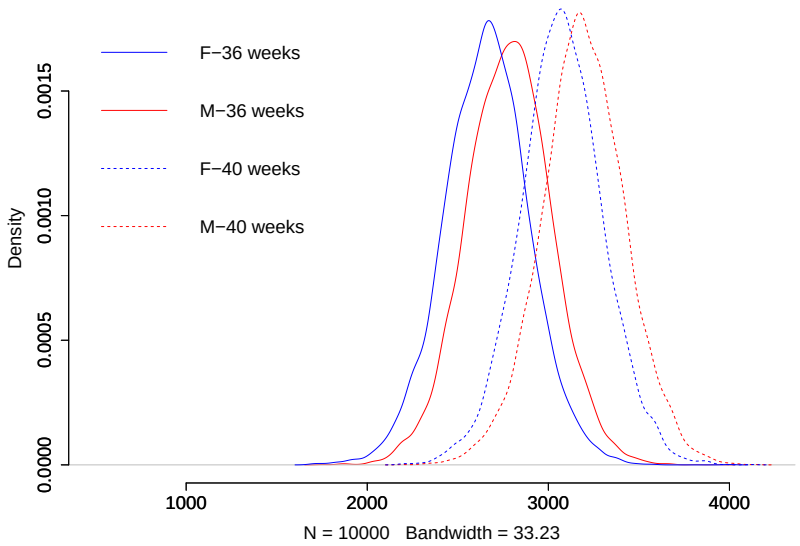

Figure 8: Densities from the simulated draws of the predictive distribution of a future weight $\tilde{\mathbf{y}}$ for the four sets of covariate values.

Table 2 shows prediction intervals for the birth weight of the expected response for each set of the covariate, as well as the prediction intervals for a future response for each covariate. We can notice the same behavior observed in the density graphs.

Table 2: Prediction intervals for the birth weight based on the mean weights and the posterior predictive distribution for future weight simulated for the four sets of covariate.

	Expected			Prediction		
	2.5%	50%	97.5%	2.5%	50%	97.5%
36 weeks - Female	2492.012	2660.363	2831.135	2203.435	2663.092	3130.062
36 weeks - Male	2614.856	2783.519	2948.560	2325.581	2783.886	3250.574
40 weeks - Female	2924.937	3074.022	3221.273	2619.504	3074.272	3527.866
40 weeks - Male	3046.684	3195.803	3343.292	2744.469	3198.762	3656.216

We can conclude that when the time of gestational weeks increases, the birth weight of the baby tends to increase too. In addition, the weight of the male babies tends to be bigger than the female babies in both times of gestational weeks. In which this behavior was observed in the first descriptive analysis of the boxplots.

References

Albert, J. (2009). Bayesian computation with R. Springer Science & Business Media.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 3). Boca Raton, FL: CRC press.

1 R Code

See below all the R Code used to perform the analysis.

```
# HW 3 - AMS 207

rm(list=ls(all=TRUE))
set.seed(7)

library(LearnBayes)
library(mvtnorm)

data(birthweight)
attach(birthweight)
gender = factor(gender)

#birthweight = birthweight[-19,]
n = nrow(birthweight)

par(mfrow=c(1,2))
boxplot(age ~ gender, ylab="Age", xlab="Gender",  cex.axis=1.5, cex.lab=1.5)
boxplot(weight ~ gender, ylab="Weight",xlab="Gender", cex.axis=1.5, cex.lab=1.5)
dev.off()

plot(density(weight), axes=FALSE, main="Weight", cex.axis=1.5, cex.lab=1.5)
axis(1, at=c(2000, 2500, 3000, 3500, 3900), cex.axis=1.6)
axis(2, cex.axis=1.6)

# scatter plots
ageM = split(age, gender)$"0"
weightM = split(weight, gender)$"0"
ageF = split(age, gender)$"1"
weightF = split(weight, gender)$"1"
varM = cbind(weightM, ageM)
varF = cbind(weightF, ageF)
pairs(varM, pch=16, panel=panel.smooth, labels=c("Weight.M","Age.M"),
cex.axis=1.5, cex.lab=1.5)
pairs(varF, pch=16, panel=panel.smooth, labels=c("Weight.F","Age.F"),
cex.axis=1.5, cex.lab=1.5)

# fitting the model using lm
fit = lm(weight ~ age + gender, qr=TRUE, x=TRUE, y=TRUE)
  ls(fit)

nlm.coef = function(fit){
  X = model.matrix(fit)
  y = fit$y

  #rank
```

```
  rank = qr(fit$x)$rank

  # extracting the decomposition Q and R
  Q = qr.Q(qr(X))
  R = qr.R(qr(X))

  # computing beta.hat best way
  beta.qr = backsolve(R,(t(Q)%*%y))

  # residual sums of squares
  rss = t(y-X%*%beta.qr)%*%(y-X%*%beta.qr)
  # estimate of residuals
  sigma = rss/(nrow(X) - qr(X)$rank)

  # inverse of R
  Rinv = backsolve(r = R, x = diag(ncol(R)))

  # identity matrix to confirm
  Rind = round(R%*%Rinv, 3)

  # full variance-covariance solved system
  vcov = Rinv%*%t(Rinv)*as.vector(sigma)^2

  list(X=X, y=y, beta.hat=beta.qr, sigma=sigma, Rinv=Rinv, Rind=Rind,
  vcov=vcov, rank=rank)
}

# parameters used in the posterior
beta.hat = nlm.coef(fit)
X = beta.hat$X
y = beta.hat$y
p = beta.hat$rank
beta.mean = beta.hat$beta.hat
n.beta = nrow(beta.mean)
R = beta.hat$Rinv
sigma.hat = beta.hat$sigma

nmc = 10000

post.sigma = rep(NA, nmc)
post.beta = array(NA, dim=c(nmc, n.beta))

post.sigma = sqrt(1/rgamma(nmc, ((n-p)/2), (n-p)/2*sigma.hat))

# slow
for(i in 1:nmc)
  post.beta[i,] = rmvnorm(1, mean = t(beta.mean),
  sigma = tcrossprod(R)*(post.sigma[i])^2)
```

```
# much faster
post.beta = rmnorm(nmc, mean = rep(0,p), varcov = tcrossprod(R))
post.beta = array(1, c(nmc, 1))%*%t(beta.mean) +
array(post.sigma, c(nmc, p))*post.beta

# another way
R = qr.R(qr(X))
z = rmnorm(nmc,rep(0,n.beta),diag(ncol(R)))

beta = backsolve(R,t(z))

post.beta = array(NA, dim=c(nmc, n.beta))
post.beta = post.sigma*t(beta) + array(1, c(nmc, 1))%*%t(beta.mean)

apply(post.beta,2,mean)
apply(post.beta,2,sd)
apply(post.beta,2,quantile, probs=c(0.025,0.5, 0.975))
mean(post.sigma)
quantile(post.sigma, probs=c(0.025, 0.5,0.975))

write.table(post.beta,file="beta.txt")
write.table(post.sigma,file="sigma.txt")

# densities
par(mfrow=c(2,2))
for(i in 1:3){
  j = i-1
  plot(density(post.beta[,i]), axes=FALSE, cex.axis=1.5, cex.lab=1.5,
  main=substitute(paste(beta[j]),list(j=j)))
  axis(1, cex.axis=1.5)
  axis(2, cex.axis=1.5)
  abline(v=quantile(post.beta[,i], probs=c(0.025, 0.975)), lty=2, col=4)
  abline(v=mean(post.beta[,i]), lty=2, col=4)
}
plot(density(post.sigma), axes=FALSE,  cex.axis=1.5, cex.lab=1.5,
main=substitute(paste(sigma)))
axis(1, cex.axis=1.5)
axis(2, cex.axis=1.5)
abline(v=quantile(post.sigma, probs=c(0.025, 0.975)), lty=2, col=4)
abline(v=mean(post.sigma), lty=2, col=4)
dev.off()

# histograms
par(mfrow=c(2,2))
for(i in 1:3){
  j = i-1
  hist(post.beta[,i], axes=FALSE, cex.axis=1.5, cex.lab=1.5,
  xlab=substitute(paste(beta[j]),list(j=j)), main="", freq=FALSE)
  axis(1, cex.axis=1.5)
```

```
  axis(2, cex.axis=1.5)
  abline(v=quantile(post.beta[,i], probs=c(0.025, 0.975)), lty=2, col=4)
  abline(v=mean(post.beta[,i]), lty=2, col=4)
}
hist(post.sigma, axes=FALSE, cex.axis=1.5, cex.lab=1.5,xlab=substitute(paste(sigma)),
main="",freq=FALSE)
axis(1, cex.axis=1.5)
axis(2, cex.axis=1.5)
abline(v=quantile(post.sigma, probs=c(0.025, 0.975)), lty=2, col=4)
abline(v=mean(post.sigma), lty=2, col=4)
dev.off()

# Posterior predictions
pred.y = matrix(0, n, nmc)
for (i in 1:nmc)
  pred.y[,i] = rnorm(n, X%*% post.beta[i,], 1*post.sigma[i])

# summary of observed versus predicted
pred.mean = apply(pred.y, 1, mean)
pred.ci=apply(pred.y,1,quantile,c(.05,.95))
plot(y, pred.mean, ylim = range(pred.ci), cex.axis=1.5, cex.lab=1.5,pch = 20,
ylab="Predicted", xlab="Observed")
abline(0, 1)
segments(x0 = y, y0 = pred.ci[1,], x1 = y, y1 = pred.ci[2,], col = 'blue')
points(y, pred.mean, pch = 20, col = 'black')

# summary of each predictive distribution by the 5th and 95th quantiles
pred.sum=apply(pred.y,1,quantile,c(.05,.95))
par(mfrow=c(1,1))
ind=1:length(y)
matplot(rbind(ind,ind),pred.sum,type="l",cex.axis=1.5, cex.lab=1.5,lty=1,col=4,xlab="index"
points(ind,y,pch=20)
out=(y>pred.sum[2,])
text(ind[out], y[out], label=y[out], pos = 4)

# possible outlier
hist(pred.y[19,], freq=FALSE, main="",cex.axis=1.5, cex.lab=1.5,
xlab=expression(paste('y'[19])))
abline(v=y[19], col=4)
pv.y19 = length(which((pred.y[19,]>(y[19])) == TRUE))
pv.y19 = round(pv.y19/nmc,4)
legend("topleft",pch=NA, cex=1.5, box.lty = 0, legend=bquote("p =" ~ .(pv.y19)),
bty='n',col=c(2))

# 36 and 40 weeks combined with Female and Male
cov1 = c(1,36,0)
cov2 = c(1,36,1)
cov3 = c(1,40,0)
cov4 = c(1,40,1)
```

```
X1 = rbind(cov1, cov2, cov3, cov4)

# expected response for the covariates
pred.cov = array(0, c(nmc, nrow(X1)))
for (i in 1:nrow(X1))
  pred.cov[,i] = t(X1[i,] %*% t(post.beta))

plot(density(pred.cov[,1]),xlim=c(2200,max(density(pred.cov[,1])$x)),
     ylim=c(min(density(pred.cov[,1])$y),max(density(pred.cov[,3])$y)),axes=FALSE,
     cex.lab=1.5,cex.main = 1.5, col="red",main="")
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)
lines(density(pred.cov[,2]),cex.lab=1.5,cex.main = 1.5, col="blue",main="Female")
legend("topleft",lwd=1, lty=c(1,1), cex=1.3, box.lty = 0,
       legend=c(expression('F-36 weeks'), expression('M-36 weeks')),bty='n',
       col=c(4,2))

plot(density(pred.cov[,3]),xlim=c(2500,max(density(pred.cov[,3])$x)),axes=FALSE,
cex.lab=1.5,cex.main = 1.5, col="red", lty=1,main="")
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)
lines(density(pred.cov[,4]),cex.lab=1.5,cex.main = 1.5, col="blue",lty=1,main="Female")
legend("topleft",lwd=1, lty=c(1,1), cex=1.3, box.lty = 0,
       legend=c(expression('F-40 weeks'), expression('M-40 weeks')),bty='n',col=c(4,2))

apply(pred.cov,2,quantile, probs=c(0.025,0.5, 0.975))

# Posterior predictions
pred.cov = array(0, c(nmc, nrow(X1)))
for (i in 1:nrow(X1))
  pred.cov[,i] = t(X1[i,] %*% t(post.beta)) + rnorm(nmc)*post.sigma

plot(density(pred.cov[,1]),xlim=c(500,max(density(pred.cov[,4])$x)),
     ylim=c(min(density(pred.cov[,1])$y),max(density(pred.cov[,3])$y)),axes=FALSE,
     cex.lab=1.5,cex.axis=1.5,cex.main = 1.5, col="red",main="")
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)
lines(density(pred.cov[,2]),cex.lab=1.5,cex.main = 1.5, col="blue",main="Female")

lines(density(pred.cov[,3]),cex.lab=1.5,cex.main = 1.5, col="red", lty=2,
main="40 weeks")
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)
lines(density(pred.cov[,4]),cex.lab=1.5,cex.main = 1.5, col="blue",lty=2,
main="Female")
legend("topleft",lwd=1, lty=c(1,1,2,2), cex=1.5, box.lty = 0,
legend=c(expression('F-36 weeks'), expression('M-36 weeks'),expression('F-40 weeks'),
expression('M-40 weeks')),bty='n',col=c(4,2,4,2))
```

```
apply(pred.cov,2,quantile, probs=c(0.025,0.5, 0.975))

# histograms
par(mfrow=c(2,2))
hist(pred.cov[,1],cex.lab=1.5,cex.axis=1.5,cex.main = 1.5, freq=FALSE,
col="#92C5DE", main="36 weeks - Female", xlab = "Weight")
hist(pred.cov[,2],cex.lab=1.5,cex.axis=1.5,cex.main = 1.5, freq=FALSE,
col="#92C5DE", main="36 weeks - Male", xlab = "Weight")
hist(pred.cov[,3],cex.lab=1.5,cex.axis=1.5,cex.main = 1.5, freq=FALSE,
col="#92C5DE", main="40 weeks - Female", xlab = "Weight")
hist(pred.cov[,4],cex.lab=1.5,cex.axis=1.5,cex.main = 1.5, freq=FALSE,
col="#92C5DE", main="40 weeks - Male", xlab = "Weight")
dev.off()

# residuals
# hat matrix
hat.matrix = function(fit.qr) {
  # Q factor
  Q = qr.qy(fit.qr, diag(1, nrow = nrow(fit.qr$qr), ncol = fit.qr$rank))
  # QQ'
  tcrossprod(Q)
}

hat.m = hat.matrix(fit$qr)

outlier.prob = function (fit,beta.hat, hat.m, post.sigma, k)
{
  e.hat = as.vector(beta.hat$y-beta.hat$X%*%beta.hat$beta.hat)
  h = diag(hat.m)
  prob = 0 * e.hat
  for (i in 1:length(prob)) {
    z1 = (k - e.hat[i]/post.sigma)/sqrt(h[i])
    z2 = (-k - e.hat[i]/post.sigma)/sqrt(h[i])
    prob[i] = mean(1 - pnorm(z1) + pnorm(z2))
  }

  list(prob=prob, e.hat=e.hat, h=h)
}

res.sum = outlier.prob(fit, beta.hat, hat.m, post.sigma, 3)
prob.outlier = res.sum$prob

par(mfrow=c(1,1))
plot(y,prob.outlier, ylab="Probability of outliers", cex.axis=1.5, cex.lab=1.5,
cex.main = 1.5)
out = (prob.outlier > 2*pnorm(-3))
text(y[out], prob.outlier[out], label=which(out == TRUE), pos = 4)

# fitted values
```

```
pred.fitted = array(0, c(nmc, nrow(X)))
for (i in 1:nrow(X))
  pred.fitted[,i] = t(X[i,] %*% t(post.beta))

# residuals against the fitted values
pred.mean.fitted = apply(pred.fitted, 2, mean)
plot(pred.mean.fitted,res.sum$e.hat, ylim=c(-300,510), cex.axis=1.5, cex.lab=1.5,
cex.main = 1.5,ylab="Residuals", xlab="Fitted Values")
abline(h=0, lty=2, col=4)
identify(pred.mean.fitted,res.sum$e.hat)

#r = res.sum$e.hat/(as.numeric(sqrt(beta.hat$sigma))*sqrt(1-res.sum$h))
r = res.sum$e.hat/(as.numeric(sqrt(beta.hat$sigma)))
plot(pred.mean.fitted,r, ylim=c(-2,3), cex.axis=1.3, cex.lab=1.4,cex.main = 1.5,
ylab="Standardized Residuals", xlab="Fitted Values")
abline(h=c(-2,0,2), lty=2, col=4)
identify(pred.mean.fitted,r)

e.bayes = array(0,c(nmc,n))
for(i in 1:nmc){
  e.bayes[i,] = (beta.hat$y-beta.hat$X%*%post.beta[i,])
}

cor.pred = array(0,c(nmc))
for(j in 1:nmc){
  cor.pred[j] = cor(e.bayes[j,],pred.fitted[j,])
}

plot(density(cor.pred), cex.axis=1.5, cex.lab=1.5,cex.main = 1.4,
main = "Correlation")
abline(v=0, lty=2, col=4)
```