Homework 4

Instructor: Hedibert Freitas Lopes
Course: STP 598 Advanced Bayesian Statistical Learning (Class # 31199)
Semester: Spring 2022
Due date: 1:30pm, April 11th, 2022.

Bayesian linear regression: We will use the `wage` data, which includes monthly earnings, education, demographic variables, and IQ scores for $n = 935$ men[1]. Below is a short script for you to get started:

```
data = read.table("http://hedibert.org/wp-content/uploads/2021/03/wage.txt")

# Dependent variable - standardized log wage
y  = data[,1]
y  = (y-mean(y))/sd(y)
n  = length(y)

# Predictors
X = matrix(0,n,6)
X[,1] = data[,5]  # years of education
X[,2] = data[,7]  # years with current employer
X[,3] = data[,8]  # age in years
X[,4] = data[,9]  # =1 if married
X[,5] = data[,10] # =1 if black
X[,6] = data[,12] # =1 if live in SMSA

# Exploratory data analysis
par(mfrow=c(2,3))
plot(X[,1],y,xlab="Years of education",ylab="Standardized log wage")
plot(X[,2],y,xlab="Years with current employer",ylab="Standardized log wage")
plot(X[,3],y,xlab="Age in years",ylab="Standardized log wage")
boxplot(y~X[,4],names=c("Single","Married"),xlab="",ylab="Standardized log wage")
boxplot(y~X[,5],names=c("Not black","Black"),xlab="",ylab="Standardized log wage")
boxplot(y~X[,6],names=c("Not SMSA","SMSA"),xlab="",ylab="Standardized log wage")
```

[1]For further details, see Blackburn and Newmark (1992) Unobserved ability, efficiency wages and interindustry wage, Quarterly Journal of Economics, 107, 1421-36 and Wooldridge (2012) Introductory Econometrics: A Modern Approach (5th edition).

Ordinary least squares

```
X = cbind(x1,x2,x3,x4,x5,x6)
summary(lm(y~X)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9950 -0.5832 -0.1119  0.4564  5.2903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.846950   0.372983 -10.314  < 2e-16 ***
x1           0.135740   0.013531  10.032  < 2e-16 ***
x2           0.018402   0.005972   3.081 0.002122 **
x3           0.037512   0.009747   3.849 0.000127 ***
x4           0.435239   0.094900   4.586 5.13e-06 ***
x5          -0.455911   0.088981  -5.124 3.65e-07 ***
x6           0.438294   0.064932   6.750 2.60e-11 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.8874 on 928 degrees of freedom
Multiple R-squared:  0.2175,Adjusted R-squared:  0.2124
F-statistic: 42.99 on 6 and 928 DF,  p-value: < 2.2e-16
```

a) R^2 **and BIC selection:** There are $p = 6$ covariates and, therefore, $2^6 - 1 = 63$ possible models (excluding the model with only the intercept!). Your first task is to fit all 63 models to the whole data and compare them in terms of adjusted R^2 and BIC. List the top 5 models and comment your findings.

b) **Bayesian model selection:** Let us now use a conjugate prior for $(\beta, \sigma^2) \in (\Re^{1+p}, \Re^+)$ for the full model, i.e.

$$\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0) \quad \text{and} \quad \sigma^2 \sim IG(c_0, d_0)$$

where $b_0 = 0_{1+p}$, $B_0 = 2I_{1+p}$, $c_0 = 2$ and $d_0 = 1$. Here, 0_{1+p} is a $(1 + p)$-dimensional vector of zeros and I_{1+p} is the identity matrix of order $1 + p$. For any one of the 62 sub-models (excluding the model with only the intercept), consider subsets of b_0 and B_0 corresponding the the sub-model. Your job is to compute the prior predictive $p(y|X, \mathcal{M}_i)$ for all sub-models $i = 1, \ldots, 63$ and rank them all. Compare the top 5 models (with the largest prior predictive densities) with the above top 6 models ranked according to R^2 and BIC. Recall that, for the Bayesian analysis of the linear and Gaussian regression with conjugate prior, all the derivations are obtained in closed form, including the evaluation of the prior predictive. We should in class that $p(y|X, \mathcal{M}_i)$ is multivariate Student's t.

c) **Out-of-sample study:** Based on a) and b), pick the top $M = 3$ models based on two of the above three criteria: BIC and prior predictive. Your job here is to verify their out-of-sample performances based on root mean square error (RMSE) and mean absolute error (MAE) criteria. In order to do that, let us randomly split the data into a *training set* with $n_1 = 468$ observations and a *testing set* with $n_2 = 467$ observations. Repeat the split $R = 100$ times. More precisely, for $r = 1, \ldots, R$ and models $m = 1, \ldots, M$, compute

$$RMSE_{ols}^{rm} = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}(y_{ir} - \hat{y}_{irm,ols})^2} \quad \text{and} \quad MAE_{ols}^{rm} = \frac{1}{n_2}\sum_{i=1}^{n_2}|y_{ir} - \hat{y}_{irm,ols}|,$$

similarly for $RMSE_{bayes}^{rm}$ and MAE_{bayes}^{rm}. The observation y_{ir} is the actual i^{th} response/dependent variable in the r^{th} *testing set*, while $\hat{y}_{irm,ols}$ and $\hat{y}_{irm,bayes}$ are the out-of-sample prediction based on the r^{th} *training set* and OLS and Bayes estimation, respectively. These out-of-sample (based on *testing sets*) estimates are computed as

$$\hat{y}_{irm,ols} = x'_{irm}\hat{\beta}_{rm,ols} \quad \text{and} \quad \hat{y}_{irm,bayes} = x'_{irm}\tilde{\beta}_{rm,bayes},$$

where $\hat{\beta}_{rm,ols}$ and $\hat{\beta}_{rm,bayes}$ are, respectively, OLS estimate and posterior mean of β_{rm} based on the *training set* of split r and model m. Report and discuss you findings.