Third homework assignment

Professional Master in Economics | Bayesian Learning
Hedibert Freitas Lopes | Due date: 7:30pm, June 22nd, 2021

Prepare one (and only one) PDF file with your solutions

Send the file to our T.A. Igor Martins (`igorfbm@al.insper.edu.br`)

Assignments will be delivered in pairs (the pairs will be randomly assigned in class)

Bayesian linear regression: Comparing in-sample and out-of-sample fit

Let us revisit the `wage` data from our worked problem `http://hedibert.org/wp-content/uploads/2021/03/wage.txt`. The data is about monthly earnings, education, demographic variables, and IQ scores for $n = 935$ men in 1980[1] We used this dataset in `http://hedibert.org/wp-content/uploads/2021/03/wage-R.txt` to perform Bayesian Variable selection in multiple linear regression by comparing a few alternatives strategies: 1) BIC; 2) Mallows' Cp; 3) Horseshoe prior; and 4) Normal-Gamma prior; and using Bobby Gramacy's R Package `monomvn - https://bobby.gramacy.com/r_packages/monomvn/`. 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
x1 = data[,5]  # years of education
x2 = data[,7]  # years with current employer
x3 = data[,8]  # age in years
x4 = data[,9]  # =1 if married
x5 = data[,10] # =1 if black
x6 = data[,12] # =1 if live in SMSA

# Exploratory data analysis
par(mfrow=c(2,3))
plot(x1,y,xlab="Years of education",ylab="Standardized log wage")
plot(x2,y,xlab="Years with current employer",ylab="Standardized log wage")
plot(x3,y,xlab="Age in years",ylab="Standardized log wage")
boxplot(y~x4,outline=FALSE,names=c("Single","Married"),xlab="",ylab="Standardized log wage")
boxplot(y~x5,outline=FALSE,names=c("Not black","Black"),xlab="",ylab="Standardized log wage")
boxplot(y~x6,outline=FALSE,names=c("Not SMSA","SMSA"),xlab="",ylab="Standardized log wage")

# Ordinary least squares fit
summary(lm(y~x1+x2+x3+x4+x5+x6))
```

[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).

a) There are $p = 6$ covariates and, therefore, $2^6 - 1 = 63$ possible models (excluding the model with only the intercept!). 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) There are up to $q = p + 1$ regression coefficients. Let us use a conjugate prior for (β, σ^2) 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 $b_0 = 0_q$, $B_0 = 2I_q$, $c_0 = 2$ and $d_0 = 1$. Here, 0_q is a q-dimensional vector of zeros and I_q is the identity matrix of order q. For any one of the 62 sub-models, 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.

c) Now, let us pick the top 5 models based on the BIC and prior predictive and verify their out-of-sample root mean square error (RMSE) and mean absolute error (MAE) performances. 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 $r = 1, \ldots, R = 100$ and models $m = 1, \ldots, 5$,

$$RMSE_{ols}^{rm} = \sqrt{\frac{1}{467}\sum_{i=1}^{467}(y_{ir}^{test} - \hat{y}_{irm,ols}^{test})^2} \quad \text{and} \quad MAE_{ols}^{rm} = \frac{1}{467}\sum_{i=1}^{467}|y_{ir}^{test} - \hat{y}_{irm,ols}^{test}|,$$

similarly for $RMSE_{bayes}^{rm}$ and MAE_{bayes}^{rm}. Here y_{ir}^{test} is the i^{th} response/dependent variable in the r^{th} *testing set*, while $\hat{y}_{irm,ols}^{test}$ and $\hat{y}_{irm,ols}^{test}$ are the out-of-sample prediction based on the i^{th} *training set*. These out-of-sample (based on the *testing set*) 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.