PROBLEM SET # 9

EC/ACM/CS 112: Bayesian Statistics

Due date	Tuesday, March 9, 10:30 am
Submission instructions	>> Create a pdf of the R Notebook with your solutions (details below) >> Submit in Canvas
Additional files included in the problem set package	++ solutions template ++ dataset *Wages1.csv*

QUESTION 1. MODEL COMPARISON USING WAIC AND CV

LEARNING GOALS

- Practice using WAIC and CV in model comparison using a familiar dataset.

DATASET. The dataset for this problem is "Wages1.csv", which we have been using in class to study linear regression.

It contains measures of the following variables for 3,294 individuals:

- X = individual identifier
- wage = hourly wage per-hour (in $)
- schooling = years of schooling
- experience = years of full time work experience
- sex = "male" or "female"

The dataset has no missing observations.

MODELS

We are going to compare two linear regression models of this dataset.

The first basic model involves a regression of logWage on a constant and the predictors schooling and experience.

The second model involves a regression of logWage on a constant, schooling, experience, an indicator for gender (female = 1, male = 0), an interaction of schooling and the gender indicator, and an interaction of experience and the gender indicator.

STEPS

Step 1. Compute the WAIC estimate of the deviance using both models.

Step 2. Use cross-validation to compute the out-of-sample deviance measure of predictive fit in both models. Use 9 equal sized folds.

Step 3. What do the results of Steps 1 and 2 suggest about which model has better predictive accuracy out of sample? Do they provide consistent answers?

QUESTION 2. THE FREEDMAN PARADOX

LEARNING GOALS

- Deepen your understanding of the problems associated with a common practice in model selection

NOTE

- This problem is based on the paper "A Note on Screening Regression Equations", D. Feedman, The American Statistician, 1983.

BACKGROUND

- Consider the following common scenario. A researcher has a dataset containing a large number of observations for a variable of interest y, and a large number of potential predictors, x_1 to x_K.
- The researcher is interested in understanding which variables have a significant effect in predicting y in a linear regression model.
- In order to address the question, the researcher uses the following 4-step procedure:
  (Step 1) Estimate a linear regression of y on a constant and all the available predictors
  (Step 2) Identify the subset of regressors with estimated coefficients that are significant at a very lax criterion, such as the 25% level.
  (Step 3) Estimate a new linear regression of y on a constant and the regressors identified in Step 2.

(Step 4) Use the results of the second regression to draw conclusions about which variables are significant predictors of y in a linear regression model.

- Friedman's Paradox refers to the fact that this procedure can introduce a substantial bias towards false discovery of significant predictors.
- The goal of this problem is to carry out simulations to explore the issues with this widely used practice.

STEPS

Step 1. Carry out 1000 simulations of the analysis process described above under the following assumptions:

- Simulate a dataset with 100 observations and fifty explanatory variables/regressors.
- The variable y and the regressors x_1 to x_{50} should be independently and identically distributed as a standard normal distribution (i.e., with mean = 0 and SD=1). Note that this implies that in the true data generating process the true regression coefficients are zero for all predictors!
- Estimate a linear regression of y on a constant and x_1 to x_{50} using frequentist methods (recall the **lm()** command that you have encountered in previous sets).
- Use the results of the estimated model to identify the subset of predictors that have significant coefficients at the 25% level.
- Estimate a new linear regression model using only a constant and the regressors that survived the previous step.
- Store the number of estimated regression coefficients that are significant at the widely used 5% level in both the full model and the smaller second model.

Visualize the results of the simulation by displaying the distribution of the number of significant coefficients in the regression models estimated before and after the variable selection step. To maximize the effectiveness of the plot, you should use a single plot for both densities and add a legend.

TIP:

- In R, the command **summary(model)$coefficients** returns the parameters of fitted regression models, including the results of significant tests for all of the estimated coefficients.

Step 2. The next step is to explore the role that the number of available regressors plays in Freedman's paradox. To do this, repeat the simulations from part 1 for the case of a dataset with 25, 50, or 75 regressors. All other parameters remain the same. You should perform 1000 simulations for each of the sample size levels.

For each simulation, compute a variable called *ensp* given by the number of significantly estimated coefficients after the variable selection MINUS the number of significantly estimated coefficients before the variable selection.

Visualize the results by displaying the distribution of *ensp* for the case of 25, 50, and 75 regressors. To maximize the effectiveness of the plot, you should use a single plot for all densities and add a legend.

Step 3. Repeat Part 2, but now keep the number of regressors at 50, and vary the standard deviation with which the regressors are sampled. In particular, compare the cases in which the each regressor x is sampled with a standard deviation of 1, 5 or 10. All other details are as above.

Visualize the results by displaying the distribution of *ensp* for the three different values of the standard deviation. To maximize the effectiveness of the plot, you should use a single plot for all densities and add a legend.

Why is this pattern so different from the one in Part 2?