Jared Yu
Module 9 Assignment

1. Use any math/stat software of your choice to find a random number generator (e.g., www.numbergenerator.org/randomnumbergenerator) to randomly select only 14 rows of Table B.4 and randomly select 5 regressors from that table. Do Problem 10.6 (a) page 367 of Textbook, using your generated data.
    a. Use the all-possible-regressions method to find the "best" set of regressors.

Ans: (Reference: [1], [2])
The chosen rows are: 1, 2, 4, 5, 6, 7, 10, 11, 14, 16, 18, 19, 22, and 23. The chosen columns are: 1, 2, 4, 7, and 9.

The textbook mentions the leaps() package within the R programming language, therefore this was chosen for the assignment. The package has the regsubsets() function which performs the calculations necessary to find the best predictors for a regression model of a certain size. Since 5 predictors are randomly chosen, then the model sizes range from 1-5 predictors (not including the intercept).

According to [2], it makes the choice of best model based on the residual sums of squares calculation. The results of using regsubsets() show that the following set of predictors are the best for each number of predictors allowed in a model can be seen below in Table 1. It indicates that for a model with $1, 2, \cdots, 5$ predictors, which set from the randomly chosen x_1, x_2, x_4, x_7, x_9 will be best. Obviously when there are 5 predictors, they all will be considered.

Table 1 The below table shows the best regressors to include for a model give a set number of predictors.

Number of Predictors	1	2	3	4	5
Best Predictors to Use	x_1	x_1, x_2	x_1, x_2, x_9	x_1, x_2, x_4, x_7	x_1, x_2, x_4, x_7, x_9

The leaps() package will also calculate the R^2_{Adj}, C_p, and BIC for each of the models. The best model for each of these metrics are shown below in Table 2. Additionally, the AIC was manually calculated and is shown also. It indicates that for example with C_p, BIC, and AIC that the best model is the one with two predictors, which is x_1, x_2 and that according to C_p the best model is the one with four predictors, or x_1, x_2, x_4, x_7.

Table 2 The below table shows the best model for each of the metrics.

Metric	R^2_{Adj}	C_p	BIC	AIC
Number of Predictors in Chosen model	4	2	2	2

The result is that there is no one model selected based on the above metrics, however, the one with 2 predictors, x_1, x_2, seems to have the clear majority over the single time that an alternative does better.

2. Use any math/stat software of your choice to find a random number generator (e.g., www.numbergenerator.org/randomnumbergenerator) to randomly select only 14 rows of Table B.11. Do Problem 10.14 (a), page 368 of Textbook, using your generated data.

a. Build an appropriate regression model for quality y using the all-possible-regression approach. Use C_p as the model selection criterion, and incorporate the region information by using indicator variables.

Ans:

The chosen rows are: 1, 6, 8, 9, 11, 12, 15, 17, 18, 21, 29, 32, 34, and 36.

The same steps done in the previous problem are repeated here. In Table 3 below, it indicates the chosen model for each of the possible number of predictors in a model (where the max is 6). For convenience, let $x_1 = \text{Clarity}$, $x_2 = \text{Aroma}$, $x_3 = \text{Body}$, $x_4 = \text{Flavor}$, $x_5 = \text{Oakiness}$, $x_{6,1} = \text{Region 1}$, $x_{6,2} = \text{Region 2}$, and $x_{6,3} = \text{Region 3}$.

An issue however is that leaps() will force the maximum number of predictors to 7 in the full model. The reason seems to be according to linear dependencies, which can arise when categorical variables are substituted with dummy variables. Since none of the levels are dropped, the leaps() package seems to avoid linear dependencies in the inverse matrix calculation by not calculating the "actual" full model including all 8 predictors (the 5 numeric attributes plus 3 categorical indicator variables).

Table 3 The below table shows the best regressors to include for a model give a set number of predictors.

Number of Predictors	Best Predictors to Use
1	$x_{6,3}$
2	$x_4, x_{6,3}$
3	x_2, x_4, x_5
4	x_2, x_3, x_4, x_5
5	$x_2, x_3, x_4, x_5, x_{6,2}$
6	$x_2, x_3, x_4, x_5, x_{6,1}, x_{6,2}$
7	$x_1, x_2, x_3, x_4, x_5, x_{6,1}, x_{6,2}$

Like before, the R^2_{Adj}, C_p, BIC, and AIC for each of the models are shown below in Table 4. However, this problem says to only use the C_p metric to choose. Therefore, the chosen model is the one with three predictors, x_2, x_4, x_5, or Aroma, Flavor, and Oakiness. The resulting metric for C_p in this model is 0.5396424. It is interesting to see also that it is less decisive this time in comparison to before, since two metrics chose this model while two metrics chose another model.

Table 4 The below table shows the best model for each of the metrics.

Metric	R^2_{Adj}	C_p	BIC	AIC
Number of Predictors in Chosen model	4	3	3	4

Reference:

[1] http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/

[2] https://rpubs.com/davoodastaraky/subset

[3] https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/

Code Appendix:

```r
library(leaps)
### Problem 1
df <- MPV::table.b4
n <- 14; k <- 5
set.seed(1); chosen_rows <- sort(sample(seq(1, nrow(df)), n))
set.seed(1); chosen_cols <- sort(sample(seq(1, ncol(df)), k))
df1 <- df[chosen_rows, c(1, chosen_cols + 1)]
# Reference: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-reg
ression-essentials-in-r/
best1 <- regsubsets(x = y~., data = df1, nvmax = 5)
res.sum <- summary(best1)
p.m <- 2:6
aic <- n * log(res.sum$rss / n) + 2 * p.m
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$cp),
  BIC = which.min(res.sum$bic),
  AIC = which.min(aic)
)

### Problem 2
df <- MPV::table.b11
n <- 14
set.seed(2); chosen_rows <- sort(sample(seq(1,nrow(df)), n))
df2 <- df[chosen_rows,]

# Reference: https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/
best2 <- regsubsets(x = Quality ~ Clarity + Aroma + Body + Flavor + Oakiness +
  I(Region ==1) + I(Region ==2) + I(Region ==3), data = df2, nvmax = 8)
res.sum <- summary(best2)
p.m <- 2:8
aic <- n * log(res.sum$rss / n) + 2 * p.m
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$cp),
  BIC = which.min(res.sum$bic),
  AIC = which.min(aic)
)
```