# Instructions

Please submit the markdown file and compiled pdf to canvas before Apr 16 at 11:59pm. For this assignment, you can discuss with classmates, but please at least attempt to go through it individually first so that you can see what you understand or don't understand. Ultimately, the final product you turn in should be your own work. So you can discuss questions with classmates, but your answers should be written in your own words.

# Intro

We will be examining data from "Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes" by Rafiei and Adeli (2018, Journal of Construction Engineering and Management) which can be accessed at:

https://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0001570

In particular, the authors aim to model the final cost of constructing a residual building. The covariates they use include a number of Economic Variables and Indices (EVI) as well as Physical and Financial (PF) variables. A description of each of the variables included is given in Table 1 of their paper. In their paper, they use pretty sophisticated prediction tools (neural networks) and they include EVI variables going back several time periods. In this assessment, we will be using linear regression and will only be considering the most recent EVI variables instead of the EVI variables from all time lags.

I've cleaned up the data a bit to make things easier, but you can access the raw data at: https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set

Take a few minutes to skim through the article to get an idea for the scientific problem of interest.

```
buildingCosts <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/Res



# Variables are grouped into PF and EVI
# Details on each variable can be found in Table 1 of the paper
names(buildingCosts)
```

```
##  [1] "PF1"       "PF2"       "PF3"       "PF4"       "PF5"       "PF6"
##  [7] "PF7"       "PF8"       "EVI1"      "EVI2"      "EVI3"      "EVI4"
## [13] "EVI5"      "EVI6"      "EVI7"      "EVI8"      "EVI9"      "EVI10"
## [19] "EVI11"     "EVI12"     "EVI13"     "EVI14"     "EVI15"     "EVI16"
## [25] "EVI17"     "EVI18"     "EVI19"     "finalCost"
```

```
#
dim(buildingCosts)
```

```
## [1] 372  28
```

```
# PF1 is a zip code, so it should be a factor
buildingCosts$PF1 <- factor(buildingCosts$PF1)
```

## Question 1 (2 points)

Give an example of a setting in which you (or someone else) might be interested in doing model selection. Would the primary goal be prediction or scientific discovery?

**Answer to question 1**

## Question 2 (2 points)

We fit two models. The first includes: * Total floor area of the building (PF2) * Lot area (PF3) * Duration of construction (PF7) * Consumer price index in the base year (EVI15). The second model also includes: Population of the city (EVI17).

Compare the $R^2$ of the two models. Explain why we may not prefer the model with the higher $R^2$?

```
# Two models of interest
mod1 <- lm(finalCost ~ PF2 + PF3 + PF7 + EVI15, data = buildingCosts)
mod2 <- lm(finalCost ~ PF2 + PF3 + PF7 + EVI15 + EVI17, data = buildingCosts)

### compare the R^2 for each of the models
```

**Answer to question 2**

## Question 3 (4 pts)

In the models above, we've only included a few of the covariates that we've recorded. Suppose we are trying to predict the final cost of new buildings which are not currently in our data set. We could add in some of the other covariates we've recorded to make a new model. Explain why including **more** covariates might potentially improve prediction for new buildings which are not in our data? Explain why including **less** covariates might potentially improve prediction for new buildings which are not in our data?

**Answer to question 3**

## Question 4 (1 point)

Consider the all models which could potentially be formed by including any of the covariates. Select a model using a forward selection procedure with BIC.

**Answer to question 4**

## Question 5 (1 point)

Consider the entire set of models which could potentially be formed by including any of the covariates. Select a model using a backward selection procedure with BIC.

**Answer to question 5**

## Question 6 (1 points)

In this case, the forward and backward search give different results. Which one do you think should be preferred? Why?

**Answer to question 6**

## Question 7 (2 point)

Using a branch and bound procedure, indicate which model would be selected by AIC and which model would be selected by BIC. Since the regsubsets procedure doesn't deal well with categorical variables, for now exclude the first variable which is zip code. You can remove the first column from the buildingCosts matrix using the following:

```r
## removes the first column
buildingCosts[, -1]

## removes the first and 28th column
buildingCosts[, -c(1, 28)]
```

**Answer to question 7**

## Question 8 (2 point)

There isn't one right answer, but for this specific problem, would you prefer to use AIC or BIC? Explain why?

**Answer to question 8**

## Question 9 (2 points)

In general, why might using a branch and bound procedure be preferred to using a forward or backward selection procedure? In what settings might you prefer using a forward or backward selection procedure?

**Answer to question 9**

## Question 10 (2 points)

Suppose your collaborator wants to do a t-test to see if the covariates which were selected using the branch and bound procedure are statistically significant. Explain to them why that would be a bad idea.

**Answer to question 10**

## Question 11 (2 points)

Suppose your collaborator didn't use a branch and bound procedure, but made plots of all the covariates and picked a few that looked promising. Would the hypothesis tests on those variables be valid? Explain why or why not?

**Answer to question 11**

## Question 12 (2 points)

Suppose your collaborator didn't look at your data at all, but did a literature review to help pick out the covariates they thought would be most relevant. Would the hypothesis tests on those variables be valid? Explain why or why not?

**Answer to question 12**