# Module 5 Assessment (Key)

Include your name

3/29/2022

## Instructions

Please submit the markdown file and compiled pdf to canvas before Apr 12 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** This answer can be quite broad, but generally for prediction one is more interested in the specific $\hat{y}$ that is output from the model regardless of the covariates included. For scientific discovery, the specific covariates which are included are the interesting part.

1 pt for situation 1 pt for primary goal

## 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 $R^2$?

```
# Includes
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
summary(mod1)
```

```
##
## Call:
## lm(formula = finalCost ~ PF2 + PF3 + PF7 + EVI15, data = buildingCosts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -193.83  -53.41   -6.98   38.23  598.81
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.246e+02  1.662e+01  -7.495 5.01e-13 ***
## PF2          6.152e-02  7.973e-03   7.716 1.14e-13 ***
## PF3         -1.845e-01  2.908e-02  -6.344 6.59e-10 ***
## PF7          1.507e+01  2.216e+00   6.799 4.29e-11 ***
## EVI15        2.660e+00  9.928e-02  26.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.09 on 367 degrees of freedom
## Multiple R-squared:  0.7163, Adjusted R-squared:  0.7132
## F-statistic: 231.7 on 4 and 367 DF,  p-value: < 2.2e-16
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = finalCost ~ PF2 + PF3 + PF7 + EVI15 + EVI17, data = buildingCosts)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -202.32  -53.55   -7.00   41.06  593.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.269e+02  1.671e+01  -7.593 2.62e-13 ***
## PF2          6.251e-02  8.005e-03   7.808 6.16e-14 ***
## PF3         -1.887e-01  2.925e-02  -6.450 3.54e-10 ***
## PF7          1.538e+01  2.229e+00   6.902 2.26e-11 ***
## EVI15        2.845e+00  1.775e-01  16.025  < 2e-16 ***
## EVI17       -2.417e-03  1.926e-03  -1.255     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.02 on 366 degrees of freedom
## Multiple R-squared:  0.7175, Adjusted R-squared:  0.7137
## F-statistic: 185.9 on 5 and 366 DF,  p-value: < 2.2e-16
```

**Answer to question 2** Because model 1 is nested within model 2, the RSS for model 2 will never be smaller than model 1 even if the additional variable EVI17 is not actually in the true data generating procedure for final cost. Thus, we must balance the increased $R^2$ (which in this case is quite small) against the cost of the additional complexity. If we are trying to select the model for scientific reasons, we would have to think about whether the increased $R^2$ is simply due to chance. If we are building a model for prediction, we would want to guard against overfitting to the data in a way that might make our predictions on new data worse.

1 point for comparing RSS and seeing that model 2 is smaller 1 point for mentioning tradeoff between RSS and increased complexity

## 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** When using a linear model for prediction, it may be that additional covariates contain information about the dependent variable that is not captured by the covariates which are already included. Thus, adding those covariates could improve the predictions **if** we can estimate the coefficients for those covariates well.

However, in general, when we add covariates but keep the total number of observations the same, we generally are able to to estimate each individual coefficient with less precision. Thus, if the decrease in precision outweighs the additional information from the new covariate, the predictions for new buildings might actually be worse.

2 points for each answer

## Question 4 (1 point)

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

```
## smallest model to consider
intOnly <- lm(finalCost~ 1, data = buildingCosts)
## largest model to consider
mod <- lm(finalCost ~ ., data = buildingCosts)

out_forward_bic <- step(object = intOnly, direction = "forward",
                        scope = formula(mod), trace = T, k = log(nrow(buildingCosts)))

summary(out_forward_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.

```
out_backward_bic <- step(object = mod, direction = "backward",
                        scope = formula(mod), trace = T, k = log(nrow(buildingCosts)))
summary(out_backward_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** The forward and backward procedure output different final models. However, we can still compare the BIC for each of the models that are output. In this case, R uses negative AIC/BIC, so we would prefer the model with the smaller value. We see the output for forward has a value of 2538.81 and the output for backward has a value of 2535.82. Thus, we would prefer the model output by backward search.

## 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:

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

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

```
n <- nrow(buildingCosts)
p <- ncol(buildingCosts[, -c(1, 28)])

out_leaps <- leaps::regsubsets(x = buildingCosts[,-c(1, 28)],
                               y = buildingCosts$finalCost, nvmax = p,
```

```
                                names = colnames(buildingCosts[,-c(1, 28)]))

sout <- summary(out_leaps)
sout


aic <- -n/2 * log(sout$rss / n) - (1:p + 2)
bic <- -n/2 * log(sout$rss / n) - log(n) /2 * (1:p + 2)


# model selected by aic
which.max(aic)
# model selected by bic
which.max(bic)

# Model
names( buildingCosts[, -c(1, 28)] )[ sout$which[which.max(aic), -1] ]

names( buildingCosts[, -c(1, 28)] )[ sout$which[which.max(bic), -1] ]
```

**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** Potential answer: In this case, we see that the model selected by AIC is larger than the model selected by BIC (as is typical), but there is still quite a bit of overlap. It seems that the authors are more concerned with prediction rather than selecting the exactly correct model for what affects prices. Thus, because the sample size is moderate (I realize this is somewhat subjective), then AIC may have better predictive performance and BIC's guarantee on selecting the "right model" is less important. Thus AIC might be preferred.

Potential answer: Perhaps we are most interested in seeing what covariates are/are not driving the cost of construction. Then, we might be more interested in using BIC because it has a theoretical guarantee that it will select the "right model" (though the theoretical guarantee may not be as valuable if the assumptions are not be exactly satisfied). In addition, we see that the model selected by BIC is smaller, so it could be easier to interpret and explain.

## Question 8 (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 8** Branch and bound would be preferred whenever it is computationally feasible because it always returns the optimal model in terms of AIC/BIC. However, there are some cases where the number of covariates is too large for branch and bound, but forward and backward selection might still be possible. In this case, although forward and backward selection may only find a sub-optimal model, it may be the only choice (although Lasso or Ridge may also be useful here). Typically the results of forward/backward search are reasonable, even they may not be the best.

## Question 9 (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 9** The tests are not valid because the covariates of interest and t-tests are performed using the same data. The model selection selects covariates whose correlation with the dependent variable is unusually large (unusual if the null hypothesis of no association is true). Thus, the results of the t-test for these selected variables will confirm that the estimated coefficients are indeed unusually large. Naively applied, this will result in a Type I error rate which is larger than we specify. Essentially, the tests are not valid because the model selection procedure used the data which was used to caluclate the p-values.

## Question 10 (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 10** (1 point) No the tests are not valid (1 point) The tests are not valid because even though a automated model selection was not used, the collaborator's exploratory analysis used the data which was used to caluclate the p-values

## Question 11 (2 points)

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

**Answer to question 11** (1 point) Yes the tests are valid (1 point) The tests are valid because the model selection didn't use the data which was used to calculate the p-values