Homework 11

Ethan Chang - ehc586

Please submit as a knitted HTML file on Canvas before the due date

For all questions, include the R commands/functions that you used to find your answer. Answers without supporting code will not receive credit.

> **Review of how to submit this assignment** All homework assignments will be completed using R Markdown. These `.Rmd` files consist of text/syntax (formatted using Markdown) alongside embedded R code. When you have completed the assignment (by adding R code inside codeblocks and supporting text outside of the codeblocks), create your document as follows:

> - Click the arrow next to the “Knit” button (above)
> - Choose “Knit to HTML” and wait; fix any errors if applicable
> - Go to Files pane and put checkmark next to the correct HTML file
> - Click on the blue gear icon (“More”) and click Export
> - Download the file and then upload to Canvas

For this homework assignment, we will look at data from the JOBS II study, which was concerned with the effect of attending a job-search skills seminar on an individual’s ability to find a job. Rather than focusing on this, however, we will focus on the relationship between several demographic variables - age, sex, race, education, and degree of economic hardship - of the study participants prior to the study. First, load the data:

```
## install.packages("mediation") RUN THIS LINE IF TO INSTALL THE PACKAGE
library(tidyverse)
library(caret)
library(rpart.plot)
jobs <- mediation::jobs
```

and familiarize yourself with the data by looking at the (admittedly sparse) documentation:

```
?mediation::jobs
```

Question 1 (1 pt)

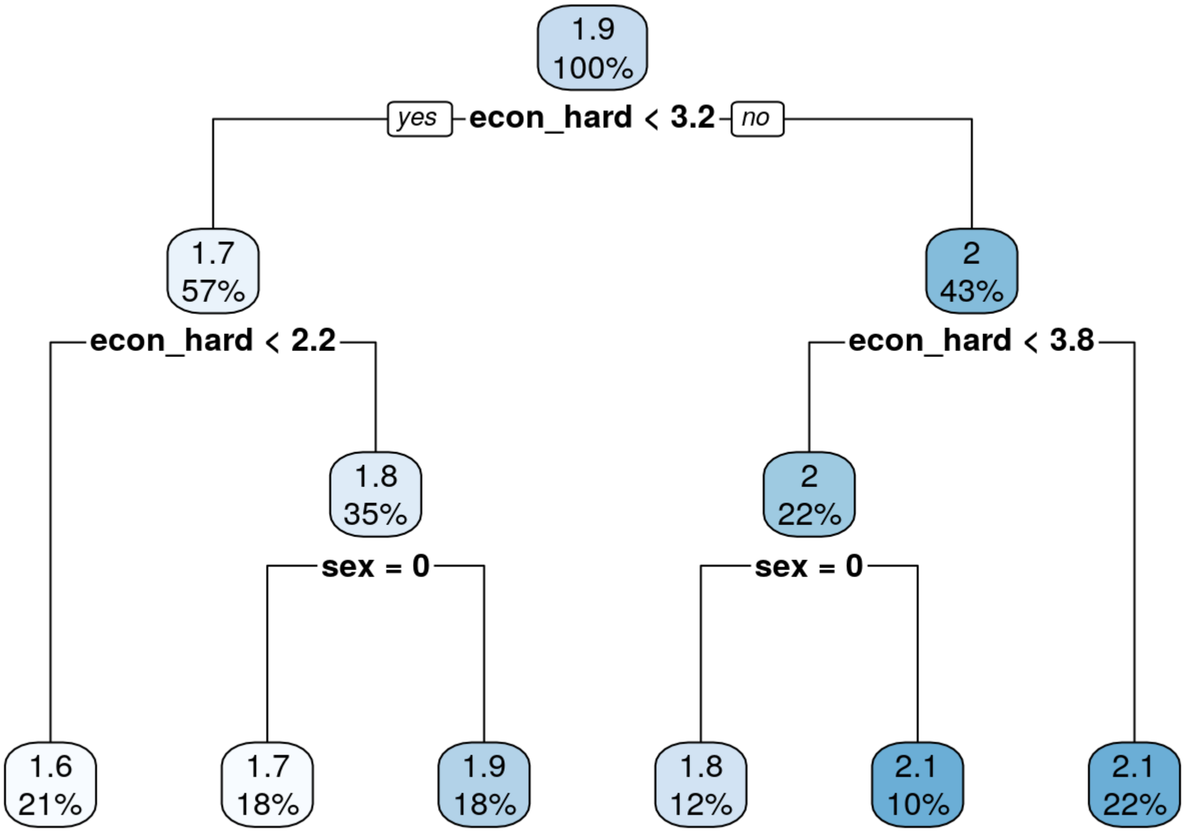
We are only interested in a small number of variables in this dataset: age, sex, race, education, degree of economic hardship, and depression level (`depress1`). Use `dplyr` to reduce the dataset to just these variables.

```
## Modify the code below and set eval = TRUE so that it runs when you knit
## your file
jobs <- jobs %>% select(age, sex, nonwhite, educ, econ_hard, depress1)
```

Question 2 (2 pt)

The depression score, `depress1` , is a measure of depression taking values between 1 and 5. Fit a classification and regression tree (CART) to this dataset using the `rpart` function and visualize the result using `rpart.plot` . Speaking qualitatively, which variables appear to be the most important for predicting depression?

```
## Your Code Here
rpart(depress1~., data = jobs) %>% rpart.plot
```



```
rpart(depress1~., data = jobs) %>% varImp()
```

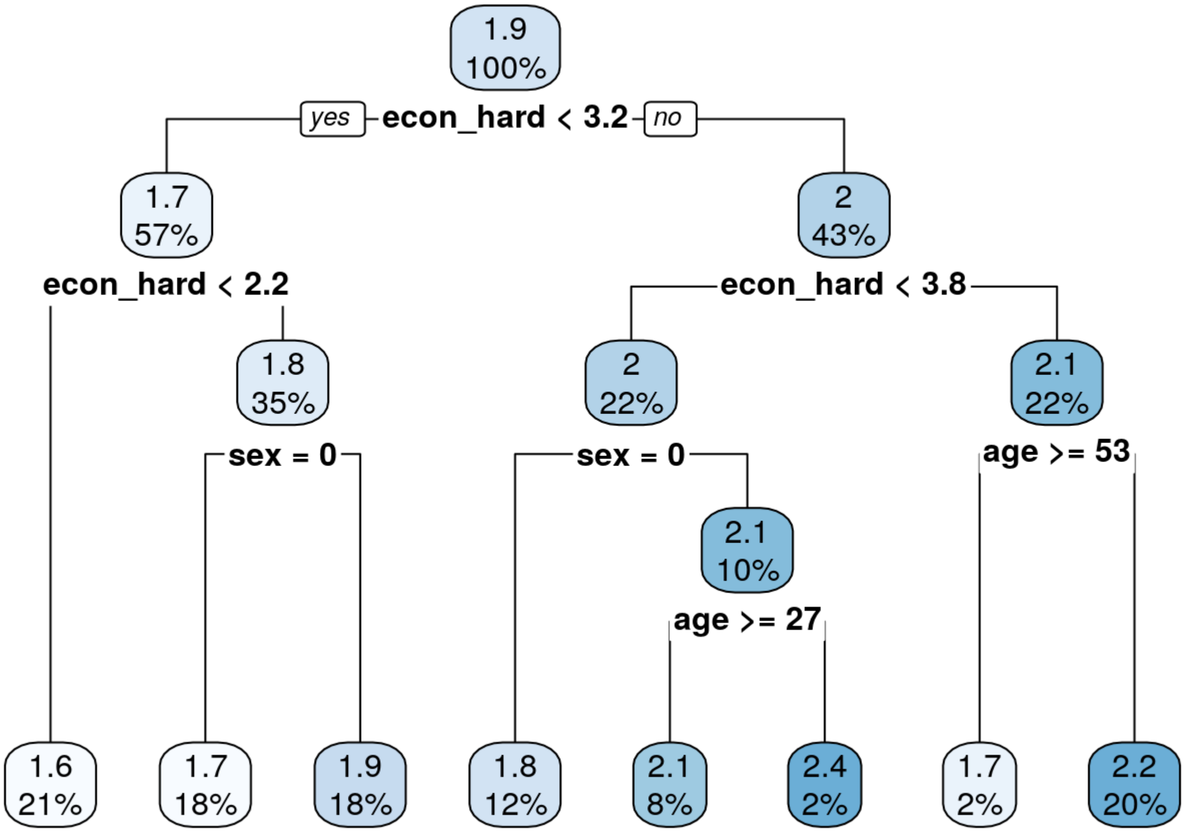
```
##              Overall
## age       0.05289128
## econ_hard 0.14553278
## educ      0.03146264
## nonwhite  0.04045579
## sex       0.14665377
```

Answer: Based on the decision tree above, the degree of economic hardship and sex variables appear to be the most important for predicting depression.

Question 3 (2pt)

Remember that how deep a decision tree is grown depends on the *complexity parameter* `cp` and the minimum number of observations associated to each node `minsplit`. Refit the CART model with `cp = 0.006` and `minsplit = 5` and plot the result. How do the results compare to the previous question?

```
## Your code here
rpart(depress1~., data = jobs, cp = 0.006, minsplit = 5) %>% rpart.plot
```



Answer: The resulting tree starts off exactly the same, with the left branch staying exactly the same. The difference though, is that the right branch branches off more and contains more options and results than the previous tree. It includes the age variable (either being >= 27 or 53) on some of the right branches.

Question 4 (1pt)

Referring to the figure in the previous question, what group of people does the decision tree determine to have the *lowest* level of depression on average? What about the *highest*?

Answer: The decision tree determines that people with a lower degree of economic hardship tend to have the lowest levels of depression on average, while those with higher degrees of economic hardship and are of lower ages tend to have the highest level of depression on average. Women also seem to have slightly higher levels of depression compared to men based on the decision tree. More specifically, the lowest level tend to belong to those with economic hardships below 2.2, whereas the highest tend to belong to those with economic hardships between 3.2 and 3.8, female, and ages lower than 27.

Question 5 (1pt)

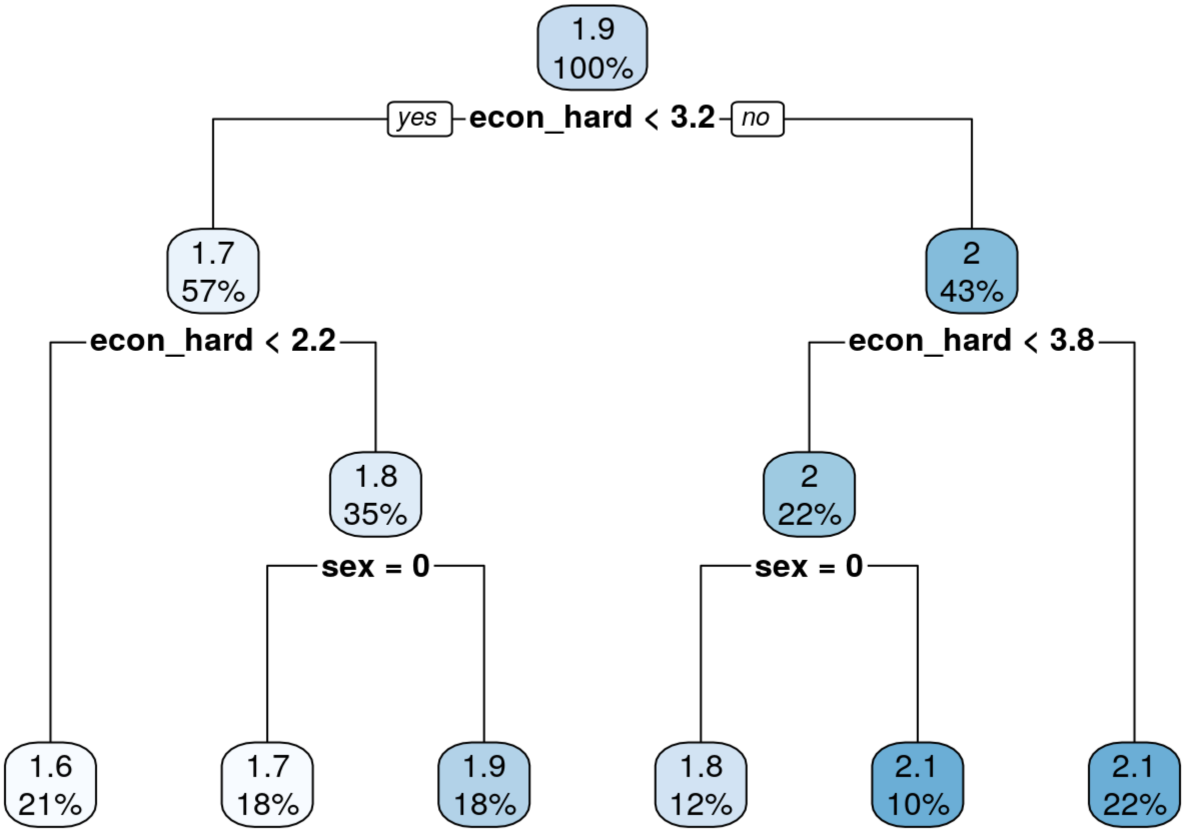
In order to prevent overfitting, it is ideal to choose `cp` using cross-validation. The following code creates a collection of `cp` 's to evaluate:

```
possible_cps <- data.frame(cp = seq(from = 0.001, to = 0.05, length = 20))
```

Use `10` replications of `20` -fold cross-validation to find the best value of `cp` with `minsplit = 2` . Then use `rpart.plot` to see how this compares with the default tree from Question 2. **How does this tree differ from the tree from Question 2?**

Hint: the `rpart` fit after using the `train` function will be in `$finalModel` .

```
## Setting up the cross-validation
set.seed(23849)
control <- rpart.control(minsplit = 2)
## Your R-code here
train_control <- trainControl(method = "repeatedcv", number = 20, repeats = 10)
tuned_rpart <- train(depress1~., method = "rpart", data = jobs, trControl = train_control, contr
ol = control, tuneGrid = expand.grid(cp = possible_cps))
rpart.plot(tuned_rpart$finalModel)
```

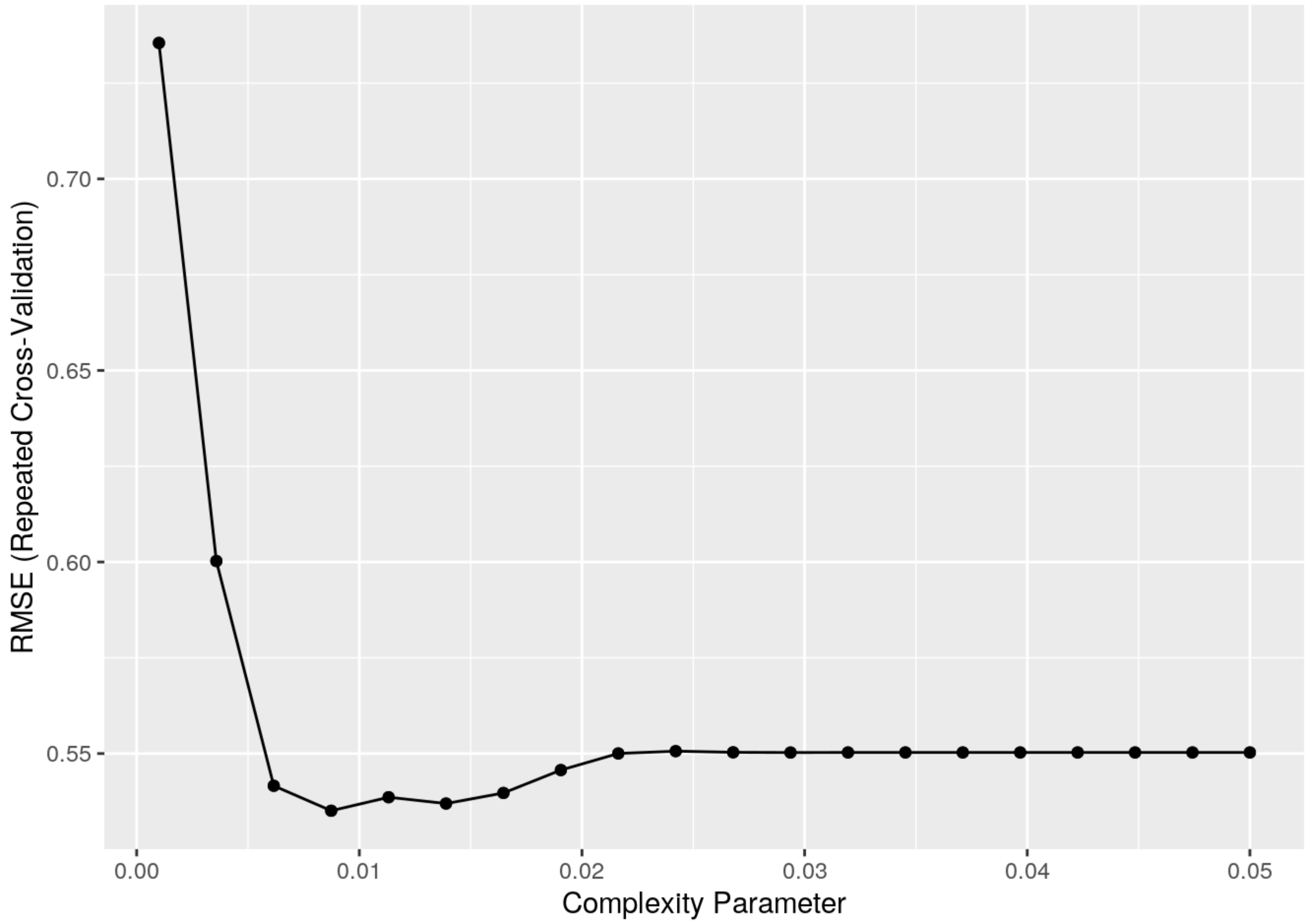


Answer: This tree does not differ at all from the tree from question 2.

Question 6 (1 pts)

Use `ggplot` on the model fit using the `train` function. What is the trend in the performance (in terms of RMSE) as we increase `cp` from `0` to `0.05` ?

```
## Your code here
ggplot(tuned_rpart, aes(x = cp, y = RMSE)) + geom_point()
```



Answer: As we increase the `cp` from 0 to 0.05, it can be seen that the performance (in terms of `RMSE`) improves overall as the `RMSE` drops drastically, and then levels off a bit after slightly increasing. Essentially, it improves immensely and then gets slightly worse, but stays consistently in that range.

Question 7 (1pt)

For comparison, let's see how this compares to the linear regression model. Fit a linear regression of `depress1` on everything and then run `summary()` on the fit.

```
## Your code here
lin_reg <- lm(depress1~., data = jobs)
summary(lin_reg)
```

```
##
## Call:
## lm(formula = depress1 ~ ., data = jobs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34804 -0.42736 -0.02045  0.40120  1.29141
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.430409   0.112907  12.669  < 2e-16 ***
## age                -0.002406   0.001725  -1.394 0.163527
## sex                 0.132871   0.035751   3.717 0.000215 ***
## nonwhitenon.white1 -0.163669   0.048291  -3.389 0.000732 ***
## educhighsc         -0.134336   0.081870  -1.641 0.101181
## educsomcol         -0.103611   0.080919  -1.280 0.200728
## educbach           -0.108621   0.087043  -1.248 0.212395
## educgradwk         -0.076728   0.090892  -0.844 0.398803
## econ_hard           0.195442   0.018306  10.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5303 on 890 degrees of freedom
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.122
## F-statistic: 16.59 on 8 and 890 DF,  p-value: < 2.2e-16
```

Based on the fit, given two individuals, one of whom is male and the other of whom is female, but who are otherwise identical in terms of sex, age, etc., how much higher/lower of a depression score is the female expected to have relative to the male?

Answer: Based on this fit, the female is expected to have a depression score that is higher than the male’s by 0.132871.

Question 8 (1pt)

Let’s now see which of the two approaches (linear regression or a decision tree) performs best in terms of predictive performance. Do the same cross-validation experiment on the linear regression model that you did on the CART. Which of the two methods performs best in terms of RMSE?

```
## Setting up the cross-validation
set.seed(23849)
## Your code here; we are using the same train_control from earlier
cv_lm <- train(depress1~., method = "lm", data = jobs, trControl = train_control)
cv_lm$results$RMSE
```

```
## [1] 0.5319177
```

```
min(tuned_rpart$results$RMSE)
```

```
## [1] 0.5350754
```

Answer: In terms of RMSE, it can be seen that the linear regression model performed better than CART as it had a lower RMSE (0.5319177 < 0.5350754)