Olin Yoder

Final Exam - Take Home

BSAN 450 (Spring 2023)

This Final Exam is due on May 12, 2023 at 6 PM Central. The total points possible are 100 and there are four (4) questions, each carrying equal points. You can discuss the exam with your friends or group members but you will submit your own solutions to Canvas, either in word or pdf. You are expected to write your own codes and produce your own report. Copying other student's code or report is not allowed and will constitute cheating.

Considerations in Grading:

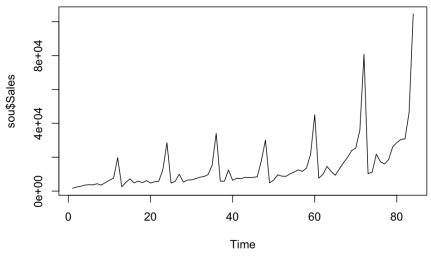
I am not just interested in the final output and R code, but also in the process you go through to arrive at your model. A significant amount of the grade for the exam will be determined by the process you use to come up with your final model. If you have a poor process or do not clearly describe that process your grade will be reduced significantly.

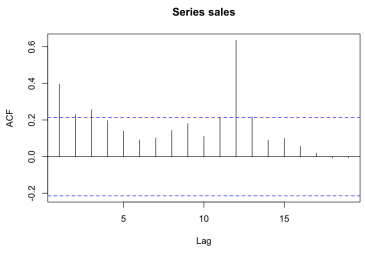
---

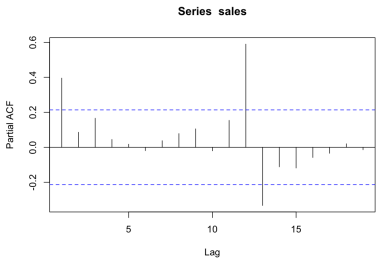
1. The time series of the monthly sales for a souvenir shop on the wharf at a beach resort town in Queensland, Australia is in the file named souvenir.csv and the name of the time series is Sales. Find a time series model that you believe is appropriate for this data. Document the steps you used to find this model and justify your choice of model.

   In developing this model you need to do the following: *For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model*.

   First, plot the data and the acf/pacf plots:

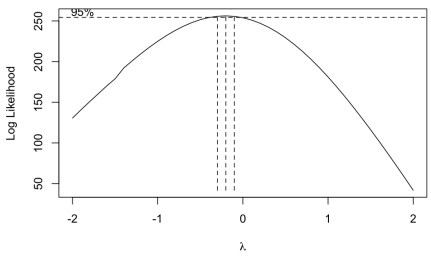




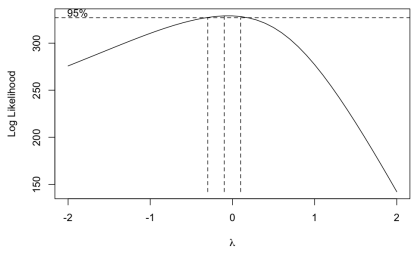


   - There is an obvious seasonal pattern, increasing mean, and non-constant variance.

   Yule-Walker



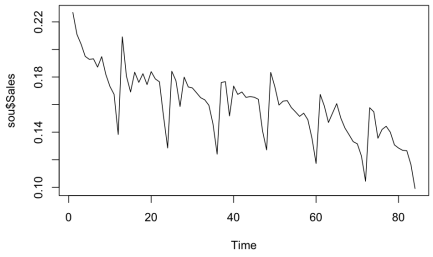
   OLS



First, we apply a transformation to fix the non-constant variance. Using the Yule-Walker method, the optimal lambda is -.2 with a CI of (-.3, -.1). Using OLS, we can use a log transformation.

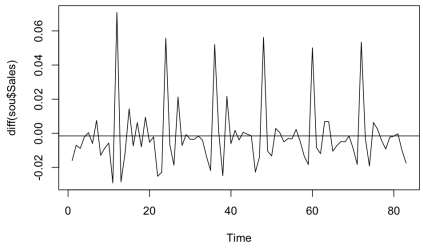
I chose to use the Yule-Walker method and take a transformation of ^-.2

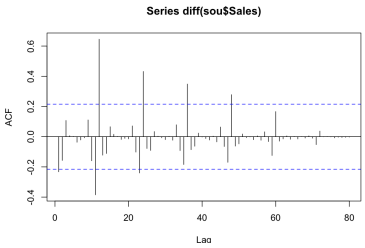
Transform and replot the data:



There is no longer the problem with non-constant variance, but the mean is still fluctuating and there is a seasonal component.

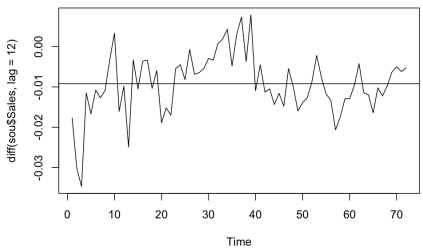
Taking just the first difference:

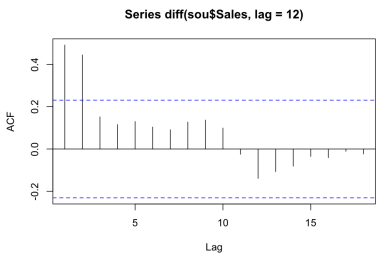




- The mean is roughly zero, but there is still an obvious seasonal component, so the series is nonstationary.

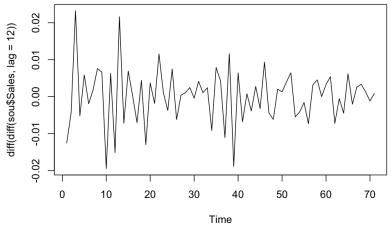
Taking just the 12th difference:

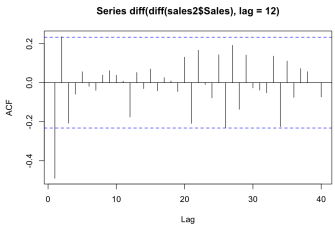




- The acf plot looks stationary, but the time series plot is questionably stationary. If I were to fit a model with only a 12th difference, it would involve a deterministic trend since the mean is not zero.

Taking both a first and twelfth difference:

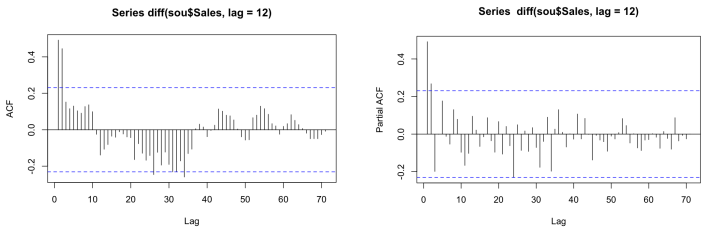




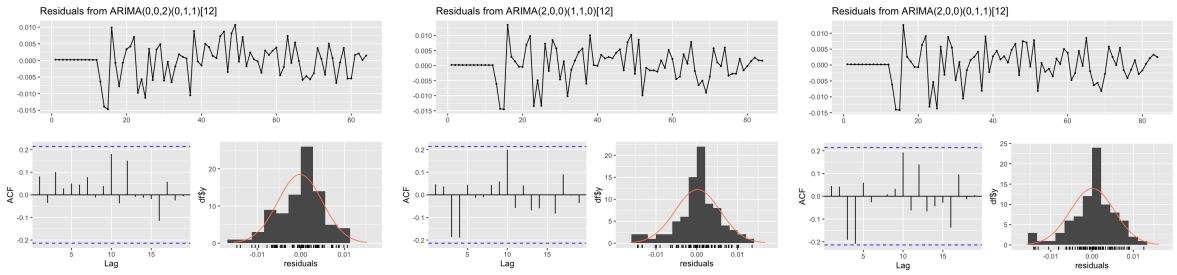
The mean is still zero and the series is roughly stationary now, although the variance may be slightly decreasing.

Now, based on the different differences, I am going to try two different approaches: 1) just using a 12th difference and 2) taking both a first and twelfth difference.

Fit a model with just a 12th difference & deterministic trend:

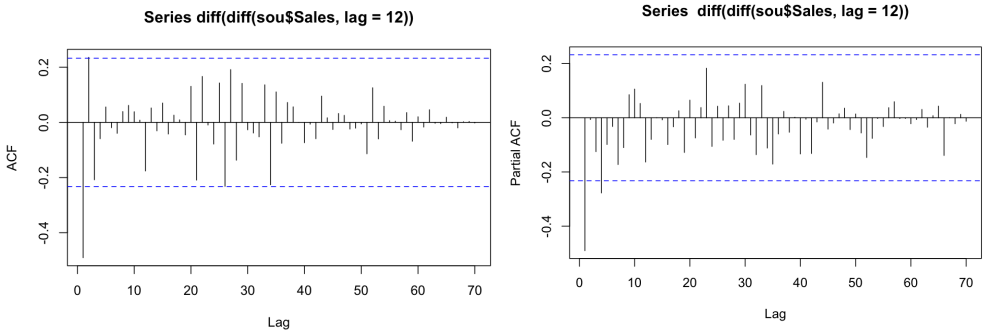


- Based on the ACF and PACF plots, which were a bit ambiguous, I fit an MA(2) * SMA(1), an AR(2) * SAR(1) and an AR(2) * SMA(1)
- While there are no significant seasonal lags, every model I fit without a seasonal component had a significant, or borderline significant, lingering autocorrelation around lag 12.



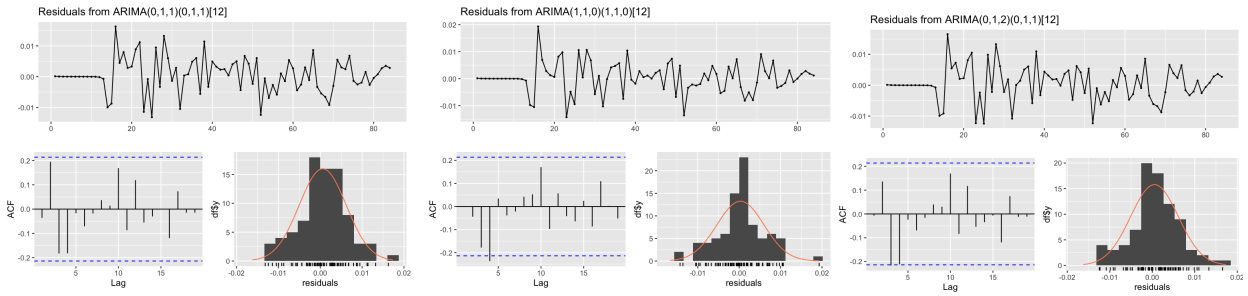
- There are no major issues with the models. All parameters were significant and they all returned a Box-Ljung p-value > .05.
- In the end, the MA(2) * SMA(1) model had the highest log likelihood and lowest AIC and sigma^2. As such, the best model using this approach is the MA(2) * SMA(1).

Fit a model taking the first & twelfth difference:



Like the first part, despite there being no significant season lags, all models without a seasonal component had significant lingering autocorrelations around lag 12. In the end, I fit an AR(1) * SAR(1), a MA(1) * SMA(1), and a MA(2) * SMA(1).

- Note: MA(2) in a MA(2) * SMA(1) was not significant



- Both Box-Ljung p-values were greater than .05. However, in the AR(1) * SAR(1) model there is significant lingering autocorrelations. The MA(1) * SMA(1) performed better in terms of sigma^2, aic, and log likelihood, too.
- However, when comparing the sigma^2, aic, and log likelihood to the best model fit with just a twelfth difference, the models using the first and twelfth difference performed worse. Therefore, moving forward, I would use the MA(2) * SMA(1) model with a twelfth difference and deterministic trend.

Check for outliers:

detectIO(sales2Model1)

[1] "No IO detected"

detectAO(sales2Model1)

[1] "No AO detected"

There are no outliers. My final model is sales^-.2 =

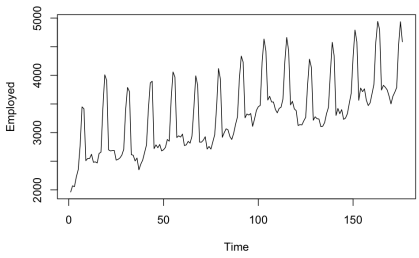
	Estimate	Std. Error	z value	Pr(>|z|)
ma1	**0.29949385**	0.08889933	3.3689	0.0007547 ***
ma2	**0.66955134**	0.09429663	7.1005	1.243e-12 ***
sma1	**-0.54865549**	0.17502700	-3.1347	0.0017204 **
xreg	**-0.00072669**	0.00013602	-5.3425	9.167e-08 ***

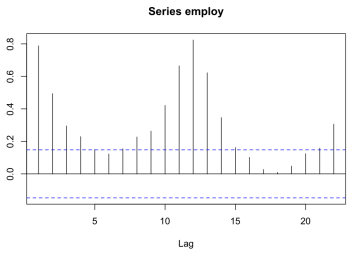
2. The time series of the number of males that are employed in non-agricultural industries is in the file named emales.csv and the time series name is Employed. Find a time series model that you believe is appropriate for this data. Document the steps you used to find this model and justify your choice of model.

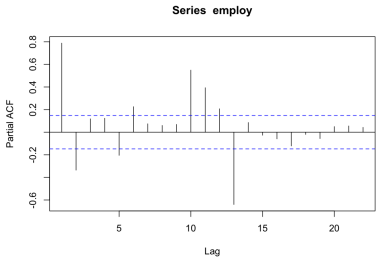
   In developing this model you need to do the following: *For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not*

describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model.

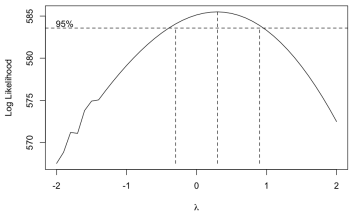
Plot the time series



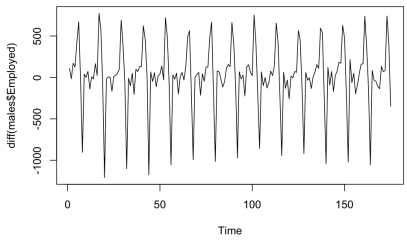


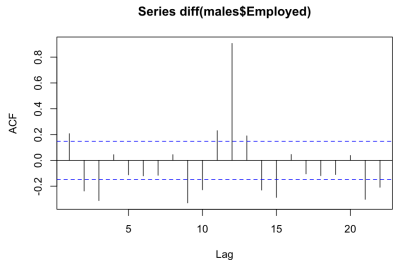


There is an obvious seasonal pattern and increasing mean. Although the Box-Cox CI does not include 1, the variance appears relatively constant so there is no need for a transformation.



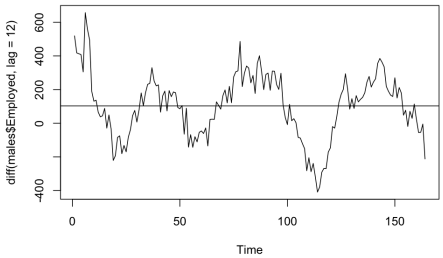
Take the first difference:

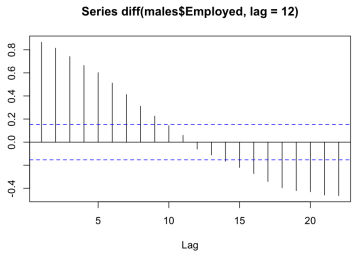




- There is still an obvious seasonal pattern, so the series is not stationary.

Take just the 12th difference:

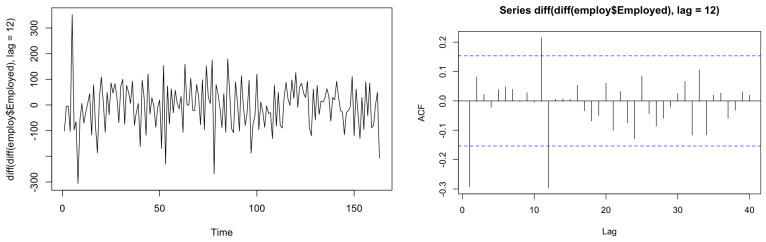




- The ACF plot does not decay slowly which indicates the series is non-stationary. Also notably, the mean is non zero.

Take the first and 12th difference:

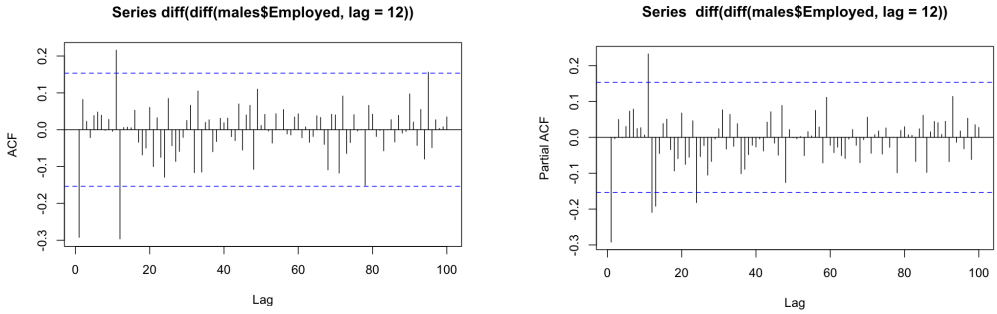
Olin Yoder



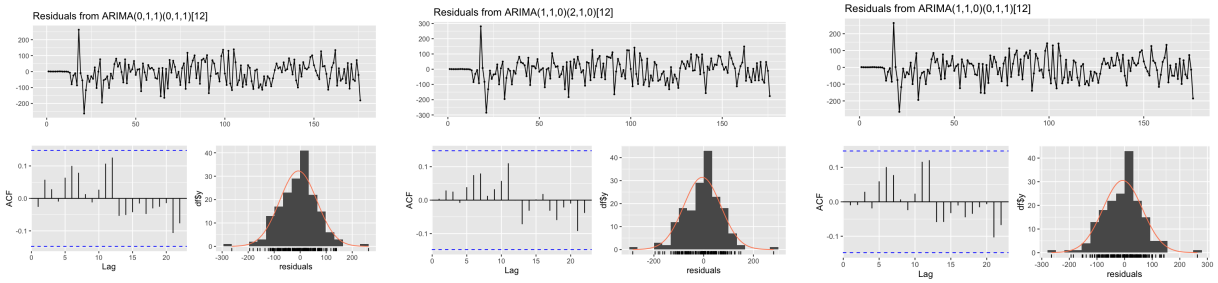
- The series appears to be stationary, although there may be an outlier at around t = 15.

Since neither the 1st difference or 12th difference alone doesn't appear to make the series stationary, I am instead going to fit a model using both a first and twelfth difference.

Both 1st and 12th difference:



Although there is no obvious solution, based on the significant lags, I fit an MA(1) * SMA(1), an AR(1) * SAR(2), and an AR(1) * SMA(1).



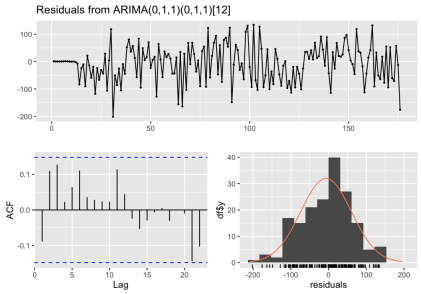
- No models had any major issues and all returned a Box-Ljung p-value > .05.
- The AR(1) * SAR(2) had the highest aic and sigma^2 and the lowest log-likelihood. The MA(1) * SMA(1) and AR(1) * SMA(1) had nearly identical sigma^2, aic, and log-likelihood values. Since they performed nearly identical, for the sake of simplicity, I choose to go with the MA(1) * SMA(1) model.

Check for outliers:

- At t = 18 and t = 21, there is potentially an innovative outlier.

Estimating the effects of the outliers:

- After estimating the effects of the outliers, sigma^2 and aic dropped and log likelihood increased which indicates that accounting for the effects of an innovative outlier at t = 18 and t = 21 improve the model. Additionally, both innovative outliers were significant.



The p-value for the Box-Ljung test was > .05 and there does not appear to be any major issues in the diagnostic plots. The only thing to note would be the nearly significant lingering autocorrelation at lag 21.

After rechecking for outliers, there were no further outliers.

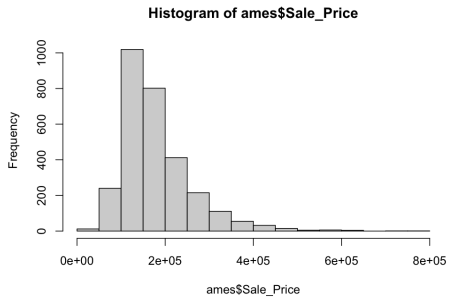
The final model:

	Estimate	Std. Error	z value	Pr(>|z|)
ma1	**-0.246629**	0.063087	-3.9093	9.255e-05 ***
sma1	**-0.610039**	0.078222	-7.7988	6.250e-15 ***
IO.18	**300.399537**	81.830023	3.6710	0.0002416 ***
IO.21	**-301.093422**	81.064685	-3.7142	0.0002038 ***

3. The data for this example is the Ames housing data available in ames.csv. The goal is to predict the Sale_Price using all independent variables. The data description is available at https://cran.r-project.org/web/packages/AmesHousing/AmesHousing. pdf. It is a good practice to make sure that your data has no missing values.

   #sum(is.na(ames))

   no n/a values



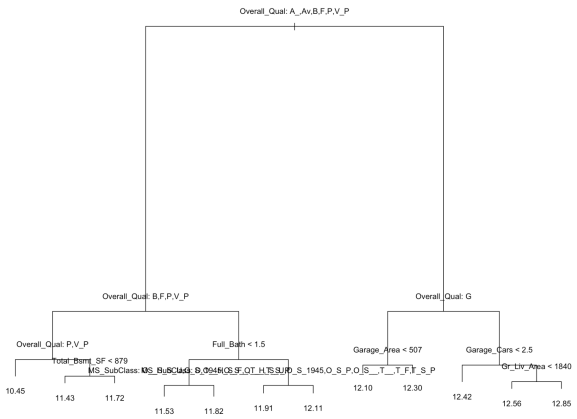
   #take a log transformation

a. Split this data into two equal parts: one for training and another for testing.

```
#Convert all characters classes to factors
ames[sapply(ames, is.character)] <- lapply(ames[sapply(ames, is.character)],
                                           as.factor)
set.seed(3456)
trainIndex <- createDataPartition(ames$Sale_Price, p = .5,
                                  list = FALSE,
                                  times = 1)
amesTrain <- ames[trainIndex,]
amesTest  <- ames[-trainIndex,]
```

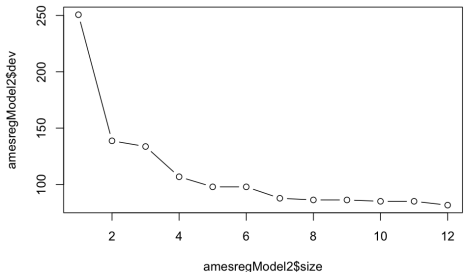
b. Fit a regression tree to the training data for predicting Sale_Price using all available independent variables. Plot the tree, discuss the size and interpret the tree that you constructed.

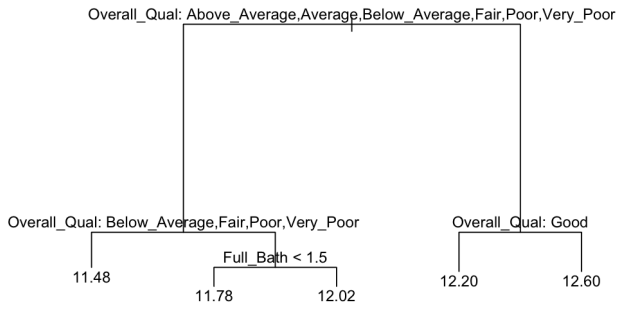
Using a log transformation, the tree has 12 terminal nodes and a depth of 4 (4 of 80 variables used to construct the tree). The tree itself is difficult to interpret.



c. Prune this tree using cross validation as discussed in class. Plot the pruned tree, discuss the size and interpret the pruned tree that you constructed. Now compare the MSE of the pruned tree and the unpruned tree on your test data. Which one has a lower MSE? Which tree would you recommend?

Set size = 5 since there is not a large improvement between 5 and 12 nodes.



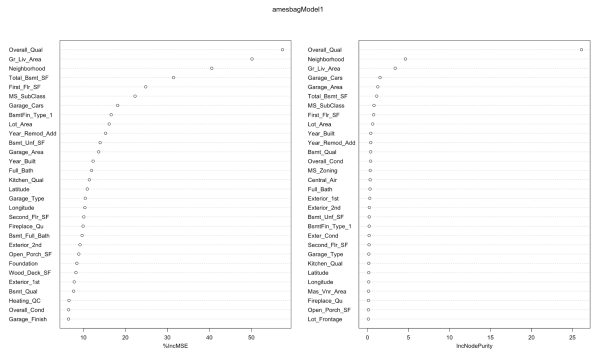


There are five terminal nodes. While the MSE for the pruned tree is slightly higher at 0.05655230 compared to the unpruned tree MSE (0.04556731), the pruned tree is more interpretable.

d. In this sub-part, you will construct a Bagged Regression tree on the training data and compute the MSE of your Bagged regression tree on the test data. Remember to carefully specify the mtry parameter.

mtry = 80 (every predictor is considered at each split)

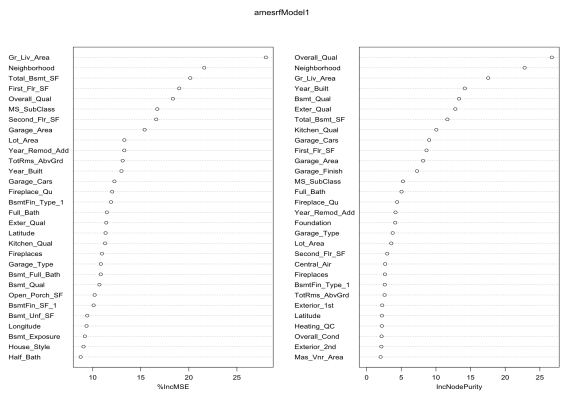
The MSE is 0.01763096.



e. In this sub-part, you will construct a Random Forest Regression tree on the training data and compute the MSE on the test data. Remember to carefully specify the mtry parameter.

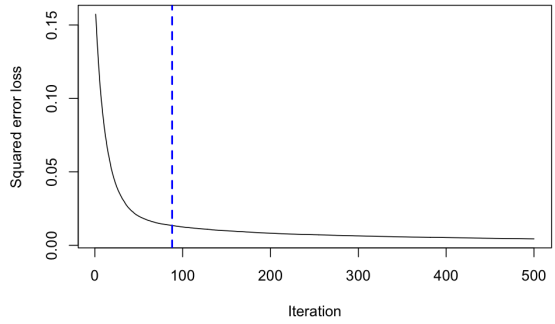
mtry = 9

The MSE is 0.01776218



The bagged tree and random forest have nearly identical mse's.

f. Finally, here you will develop a Boosted Regression tree on the training data and compute the MSE on the test data. You may try tweaking the n.tree and shrinkage parameters to see how it affects your test set MSE.



I found that when I had n.tree = 500 (more than 90 not necessary) and shrinkage = .06, it produced the lowest MSE (0.01292859).

g. Which tree amongst the four different regression trees that you constructed has the lowest MSE on the test set?

	Method	MSE
1	DecisionTree	0.04556731
2	PrunedDecisionTree	0.05655230
3	Bagged	0.01757415
4	RF	0.01786218
5	Boosted	0.01292859

The boosted model has the lowest MSE, and therefore is likely the best model for this data.

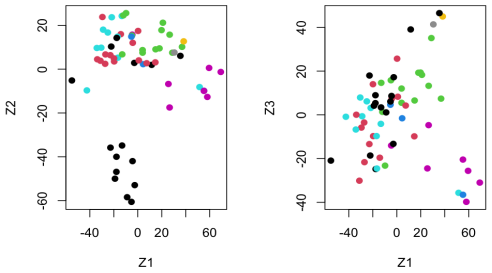
4. The data nci.data has 64 rows and 6,830 columns. Each row is a cell line for which we have gene expressions recorded for 6,830 genes. The nci.labs data has the cancer type for each of these 64 cell lines. Your analysis will rely primarily on nci.data and not on nci.labs as the latter are just labels.

```
library(ISLR) nci.labs=NCI60$labs
nci.data=NCI60$data
```

a. Perform PCA on nci.data with scale=TRUE in the prcomp function.

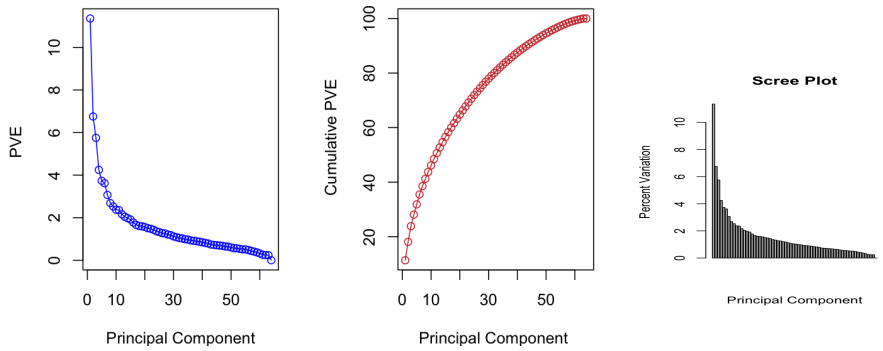
pr.out1 =prcomp(nci.data, scale=TRUE)

b. Plot the first two principal components. You will notice from your plot that observations belonging to a single cancer type tend to lie near each other in this low dimensional space.



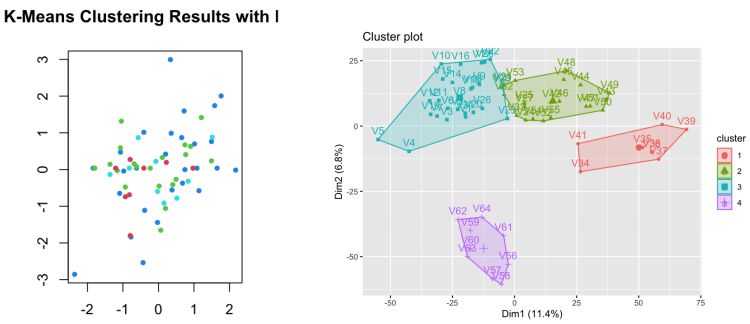
From the plots, we can see that cells of like cancers have similar genes.

c. As demonstrated in the lecture, plot the PVE of each principal component (i.e. a scree plot) and the cumulative PVE of each principal component. What is the % of variance explained by the first seven principal components? Based on the scree plot, how many principal components should we choose?



About 40% of the variance is explained by the first seven principal components. Based on the scree plot, we would recommend using the first four principal components, although the number of pc's depends on how much variability that needs to be explained.

d. Scale the nci.data such that the features have zero mean and 1 standard deviation. You will have to use the scale() function in R for this purpose. Then, on the scaled data, perform K-means clustering with K = 4. What are the cluster sizes that you obtain? Calculate the Between Cluster sum of squares as a % of total sum of squares.



Olin Yoder

Cluster sizes: 20, 9, 27, 8
Between cluster ss: 19.92218%