```
  Project 3: Simulation & Bootstrapping
     A Case Study: Transplant Center


Team 31: Fan Feng, Maxwell Weiss, Jordan Lopez
```

Read in the data

Read data

```r
r11xplant <- read.table("C:/Users/Think/Documents/Transplant/R11xplant.csv",
sep = ",", header = T)

r11donor<-read.table("C:/Users/Think/Documents/Transplant/R11donor.csv", sep
= ",", header = T)

uva <- read.table("C:/Users/Think/Documents/Transplant/UVAxplant.csv", sep =
",", header = T)

duke <- read.table("C:/Users/Think/Documents/Transplant/Dukexplant.csv", sep
= ",", header = T)

mcv <- read.table("C:/Users/Think/Documents/Transplant/MCVxplant.csv", sep =
",", header = T)

unc <- read.table("C:/Users/Think/Documents/Transplant/UNCxplant.csv", sep =
",", header = T)
```

Source the bootstrapping functions

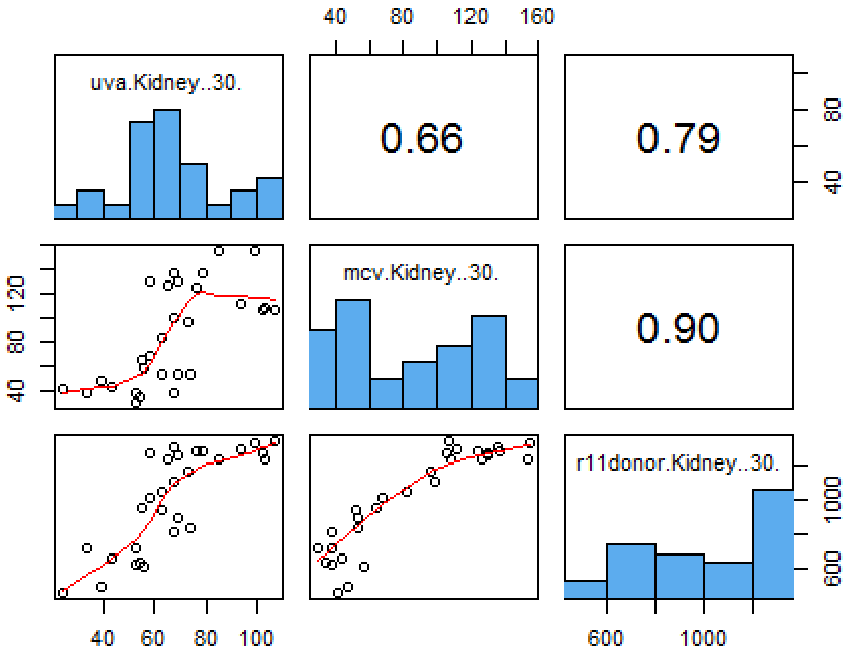
```r
library(boot)

source("C:/Users/Think/Downloads/TSbootfunctions.R")

source("C:/Users/Think/Downloads/SPM_Panel.R")

source("C:/Users/Think/Downloads/Transplant.plots.R")
```

Part 1: Basic Statistics

Step 1.1 Compare the performance of UVa with MCV kidney transplants

Get the distribution of uva$Kidney, r11donor$Kidney. What do you observe?

```
kidney<-data.frame(uva$Kidney[-30],mcv$Kidney[-30],r11donor$Kidney[-30])
uva.pairs(as.matrix(kidney))
```



We observed that the distribution for the uva center is much more symmetric than the other two. The distribution for the mcv center is very asymmetric, with higher density on the extremes. The mcv center has the highest correlation with region 11 kidney donors.

On average, how many kidney transplants are performed at UVa per year? MCV?

```
mean(uva$Kidney[-30])

## [1] 67.31034

mean(mcv$Kidney[-30])

## [1] 84.68966
```

Excluding year 2017 data, 67.31 kidney transplants are performed at UVA per year and 84.69 kidney transplants are performed at MCV per year, on average.

Perform a paired t-test between uva and mcv kidney transplants What is the result?

```
t.test(uva$Kidney[-30],mcv$Kidney[-30],paired=T)
```

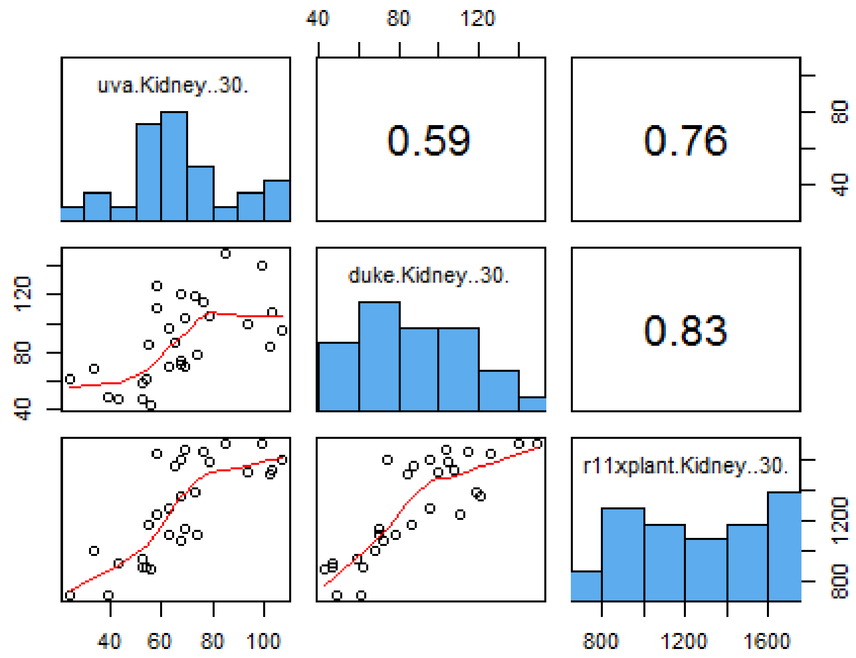
```
##
##  Paired t-test
##
## data:  uva$Kidney[-30] and mcv$Kidney[-30]
## t = -2.9737, df = 28, p-value = 0.005994
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -29.350877  -5.407743
## sample estimates:
## mean of the differences
##               -17.37931
```

The paired t-test resulted in a p-value of about .006. We reject the null hypothesis, indicating that there is a significant difference in the means of UVA and MCV kidney transplants.

Step 1.2 Compare the performance of UVa with Duke kidney transplants

Get the distribution of uva$Kidney, r11xplant$Kidney. What do you observe?

```
kidney1<-data.frame(uva$Kidney[-30],duke$Kidney[-30],r11xplant$Kidney[-30])
uva.pairs(as.matrix(kidney1))
```



We observed that the dist. for the UVA center is much more symmetric than the other two. The dist. for the Duke center is roughly uniform, with a peak on the lower end of the range. The Duke center has the higher correlation with region 11 kidney explants.

On average, how many kidney transplants are performed at UVa per year? Duke?

```
mean(uva$Kidney[-30])

## [1] 67.31034

mean(duke$Kidney[-30])

## [1] 87.7931
```

Excluding year 2017 data, 67.31 kidney transplants are peformed at UVA per year and 87.79 kidney transplants are performed at Duke per year.

Perform a paired t-test between uva and duke kidney transplants. What is the result?

```
t.test(uva$Kidney[-30],duke$Kidney[-30],paired=T)

## 
##  Paired t-test
## 
## data:  uva$Kidney[-30] and duke$Kidney[-30]
## t = -4.6855, df = 28, p-value = 6.551e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -29.43734 -11.52817
## sample estimates:
## mean of the differences 
##               -20.48276
```

The paired t-test resulted in a p-value of about .000065 we reject the null hypothesis, indicating that there is a significant difference in the means of UVA and Duke kidney transplants.

Step 1.3 Use bootstrapping to test the hypothesis: there is not a significant difference between UVa and MCV kidney transplants.

What are the standard errors of the mean? What're the 95% confidence intervals? Do you accept or reject the null hypothesis?

```
bs.mean<-function(x,i)
{
  return(mean(x[i]))
}
uvamcv.diff<-ts(uva$Kidney-mcv$Kidney,1988,2016)
bs.uvamcv.diff<-boot(uvamcv.diff,bs.mean,R=10000)
bs.uvamcv.diff

## 
## ORDINARY NONPARAMETRIC BOOTSTRAP
## 
## 
## Call:
```

```
## boot(data = uvamcv.diff, statistic = bs.mean, R = 10000)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* -17.37931 -0.05074483    5.654799
```

```r
boot.ci(bs.uvamcv.diff,0.95,type=c('bca','perc'))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bs.uvamcv.diff, conf = 0.95, type = c("bca",
##     "perc"))
##
## Intervals :
## Level     Percentile            BCa
## 95%   (-28.62,  -6.59 )   (-29.03,  -6.83 )
## Calculations and Intervals on Original Scale
```

The standard errors of the mean is about 5.76. The 95% confidence intervals are (-28.93, -6.41) [Percentile]; (-29.21, -6.72) [BCa]. We reject the null hypothesis because the confidence intervals don't include zero.

Step 1.4 Use bootstrapping to test the hypothesis: There is not a significant difference between UVa and Duke kidney transplants.

What are the standard errors of the mean? What're the 95% confidence intervals? Do you accept or reject the null hypothesis?

```r
uvaduke.diff<-ts(uva$Kidney-duke$Kidney,1988,2016)
bs.uvaduke.diff<-boot(uvaduke.diff,bs.mean,R=10000)
bs.uvaduke.diff
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = uvaduke.diff, statistic = bs.mean, R = 10000)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* -20.48276 -0.01981724    4.364546
```

```r
boot.ci(bs.uvaduke.diff,0.95,type=c('bca','perc'))
```

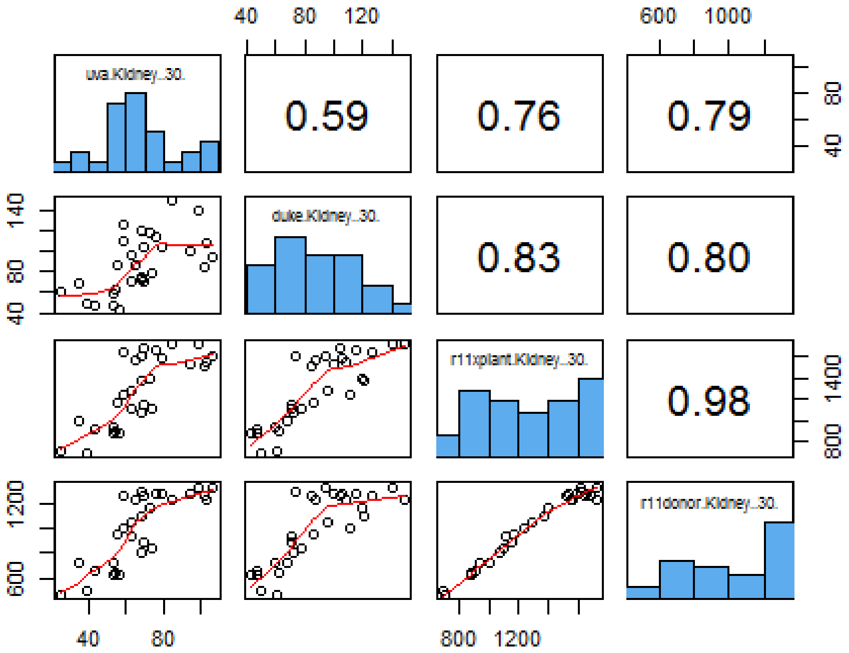
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
```

```
##
## CALL :
## boot.ci(boot.out = bs.uvaduke.diff, conf = 0.95, type = c("bca",
##     "perc"))
##
## Intervals :
## Level     Percentile            BCa
## 95%   (-29.21, -12.10 )   (-29.45, -12.34 )
## Calculations and Intervals on Original Scale
```

The standard error of the mean is about 4.29. The 95% confidence intervals are (-29.07,-12.21) [Percentile]; (-29.38,-12.59) [BCa]. We reject the null hypothesis because the confidence intervals don't include zero.

Step 1.5 Get the scatter plot matrix with the above 4 variables (UVA kidney, Duke kidney, R11 trnasplants, R11 donors). Describe what you observe. You can use either uva.pairs() {source("SPM_Panel.R")} or pairs().

```
uva.pairs(as.matrix(data.frame(uva$Kidney[-30],duke$Kidney[-30],r11xplant$Kid
ney[-30],r11donor$Kidney[-30])))
```



We observe that the Duke center is more correlated with both Region 11 kidney transplants, as well as Region 11 kidney donors, than UVA. The dist. for kidney transplants at UVA is somewhat symmetric with extreme modes in the center. The dist. for kidney transplants at Duke is much less symmetric, closer to being uniform The UVA and Duke centers are almost equally correlated with Region 11 Donors. The R11 Transplants and R11 donors for kidney transplants are extremely correlated (almost 1).

Part 2: Linear Regression Models

Test the hypothesis: There is no difference between the forecasted numbers of kidney transplants that will performed at UVA and at MCV in 2017.

Step 2.1 Build a linear model: uva$Kidney = b0+b1*r11donor$Kidney+e. Call it uva.kidney.lm. Analyze the result: R^2, model utility test, t-tests, etc.

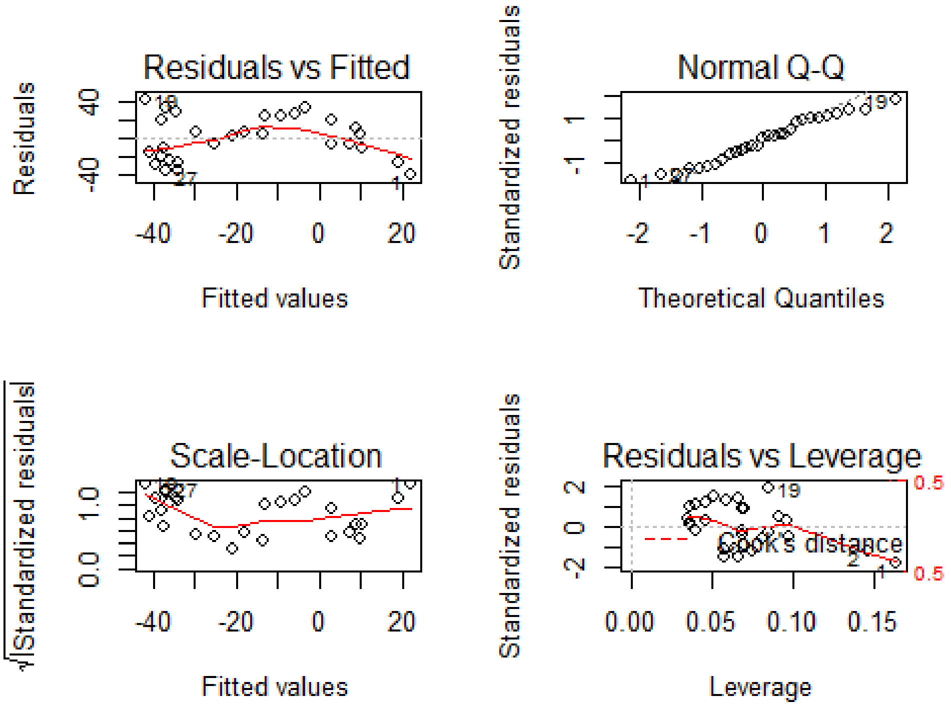
```
uva.mcv.diff<-(uva$Kidney[-30]-mcv$Kidney[-30])
uva.kidney.lm<-lm(uva.mcv.diff~r11donor$Kidney[-30])
summary(uva.kidney.lm)
```

```
##
## Call:
## lm(formula = uva.mcv.diff ~ r11donor$Kidney[-30])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.960 -20.072   2.055  21.085  41.838
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          54.98127   16.54883   3.322 0.002571 **
## r11donor$Kidney[-30] -0.07242    0.01594  -4.542 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.13 on 27 degrees of freedom
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4122
## F-statistic: 20.63 on 1 and 27 DF,  p-value: 0.0001042
```

We obtained an Adjusted R^2 value of .4122 and a Multiple R^2 value of .4332, indicating somewhat good fit The model utility test resulted in a p-value of .0001042, indicating that the model is highly significant, even at the .01 level. The t-test for the r11donor variable results in a p-value of .000104, indicating that the paramater is different from zero. These results indicate that the r11 donor variable is very useful in predicting the difference between the mean of uva and mcv kidney transplants. For every increase in r11donor$kidney, uva kidney transplants decreases relative to mcv by -.072.

Step 2.2 Generate the diagnostic plots. Interpret the results of the diagnostic plots. Do you see any problem?

```
par(mfrow=c(2,2))
plot(uva.kidney.lm)
```



```
par(mfrow=c(1,1))
```

The Residuals vs Fitted plot shows a lack of fit and heteroscendasticity. The QQ plot shows that the tails of the distribution are not Gaussian. The Residuals vs Leverage plot shows there are multiple outliers, none of which are influential.

Step 2.3 Estimate the model with bootstrapping (by residuals). Is b1 significant?

```
# Get the fitted values from the regression model
uvamcv.fit<-fitted(uva.kidney.lm)
#    Get the residuals from the regression model
uvamcv.res<-residuals(uva.kidney.lm)
#    Get the regression model
uvamcv.mod<-model.matrix(uva.kidney.lm)
#   Bootstrapping LM
uvamcv.boot<-RTSB(uva.mcv.diff,r11donor$Kidney[-30],uvamcv.fit,uvamcv.res,uva
mcv.mod,5000)
uvamcv.boot

## 
## ORDINARY NONPARAMETRIC BOOTSTRAP
## 
## 
## Call:
## boot(data = cbind(resp, pred), statistic = coef.fun, R = num, 
##     fit = fit, resid = resid, X2 = X)
## 
```

```
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1* 54.98126654 -5.061152e-02 16.00718162
## t2* -0.07241551  8.658413e-05  0.01546475
```

```
#    95% CI of r11donor
boot.ci(uvamcv.boot, .95, index=2)
```

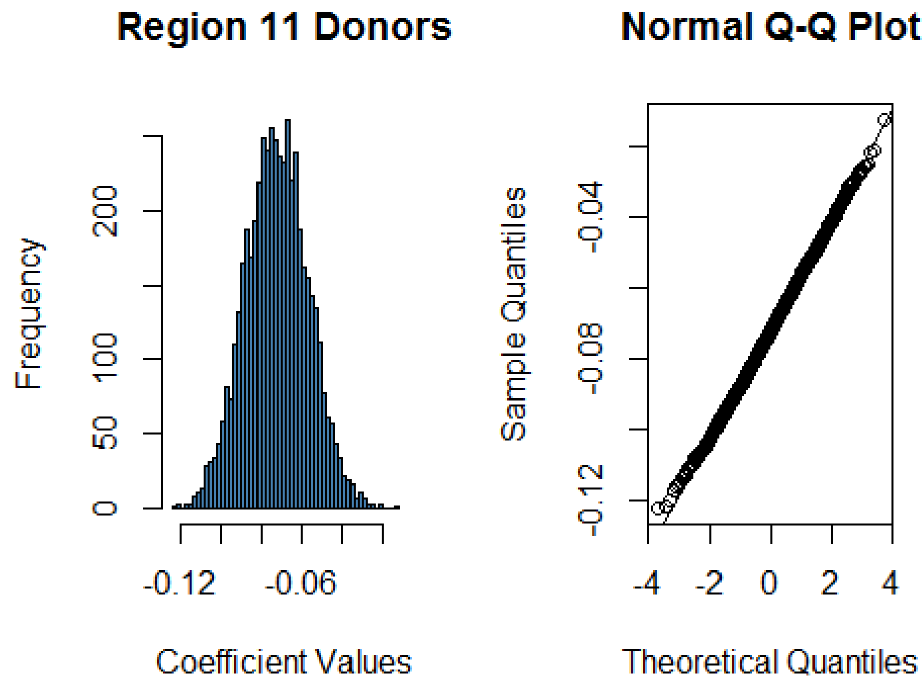
```
## Warning in boot.ci(uvamcv.boot, 0.95, index = 2): bootstrap variances
## needed for studentized intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = uvamcv.boot, conf = 0.95, index = 2)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.1028, -0.0422 )   (-0.1028, -0.0422 )
##
## Level     Percentile            BCa
## 95%   (-0.1027, -0.0420 )   (-0.1044, -0.0434 )
## Calculations and Intervals on Original Scale
```

None of the confidence intervals include zero as a value.

Distribution of b1

```
par(mfrow = c(1,2))
hist(uvamcv.boot$t[,2], main = "Region 11 Donors",xlab ="Coefficient Values",
   col = "steelblue", breaks = 50)
qqnorm(uvamcv.boot $t[,2])
qqline(uvamcv.boot $t[,2])
```



```
par(mfrow = c(1,1))
```

Is b1 significant? b1 is significant because all four of the 95% confidence intervals for b1 estimate do not include zero, implying that b1 is not equal to zero the histogram for region 11 donor coefficient values forms almost a normal distribution, centered around -.07, implying that b1 is significant.

Step 2.4* (bonus) What about Duke? Repeat the above steps and compare the results.

Test the hypothesis: There is no difference between the forecasted numbers of kidney transplants that will be performed at UVA and at Duke in 2017.

Build a linear model and analyze the results

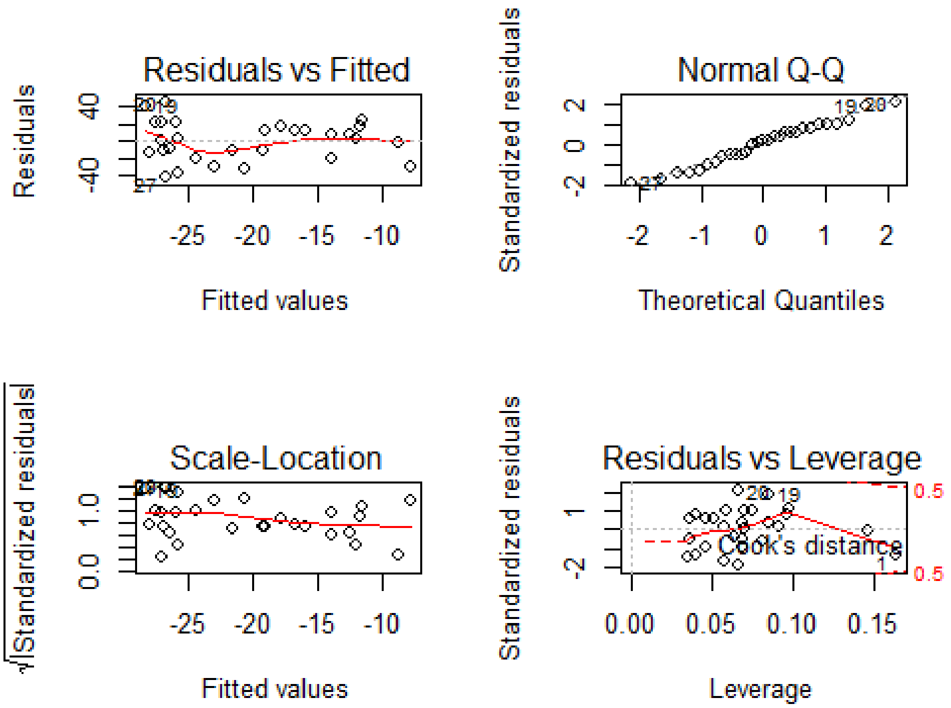
```
uva.duke.diff<-uva$Kidney[-30]-duke$Kidney[-30]
uva.duke.lm<-lm(uva.duke.diff~r11donor$Kidney[-30])
summary(uva.duke.lm)

##
## Call:
## lm(formula = uva.duke.diff ~ r11donor$Kidney[-30])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.231 -12.978   3.817  16.923  44.722
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.71586   15.77331   0.172    0.865
## r11donor$Kidney[-30] -0.02322    0.01520  -1.528    0.138
##
## Residual standard error: 23 on 27 degrees of freedom
## Multiple R-squared:  0.07957,    Adjusted R-squared:  0.04548
## F-statistic: 2.334 on 1 and 27 DF,  p-value: 0.1382
```

We obtained an Adjusted R^2 value of .04548 and a Multiple R^2 value of .079, indicating a worse fit than the previous model. The model utility test resulted in a p-value of .1382, indicating that the model is not significant, contrasted to the previous model comparing UVA and MCV. The t-test for the r11donor variable results in a p-value of .138, indicating that the paramater is not different from zero. These results indicate that the r11 donor variable is not useful in predicting the difference between the mean of uva and duke kidney transplants. The r11donor variable is much worse at predicting the difference between UVA and Duke than it is at predicting UVA and MCV kidney transplants.

Generate the diagnostic plots. Interpret the results of the diagnostic plots. Do you see any problem?

```
par(mfrow=c(2,2))
plot(uva.duke.lm)
```



```
par(mfrow=c(1,1))
```

The Residuals vs Fitted plot also shows a lack of fit and heteroscendasticity. The QQ plot shows that the distrbution is Gausian, even at the tails, different from the previous model.

The Residuals vs Leverage plot again shows there are multiple outliers, none of which are influential.

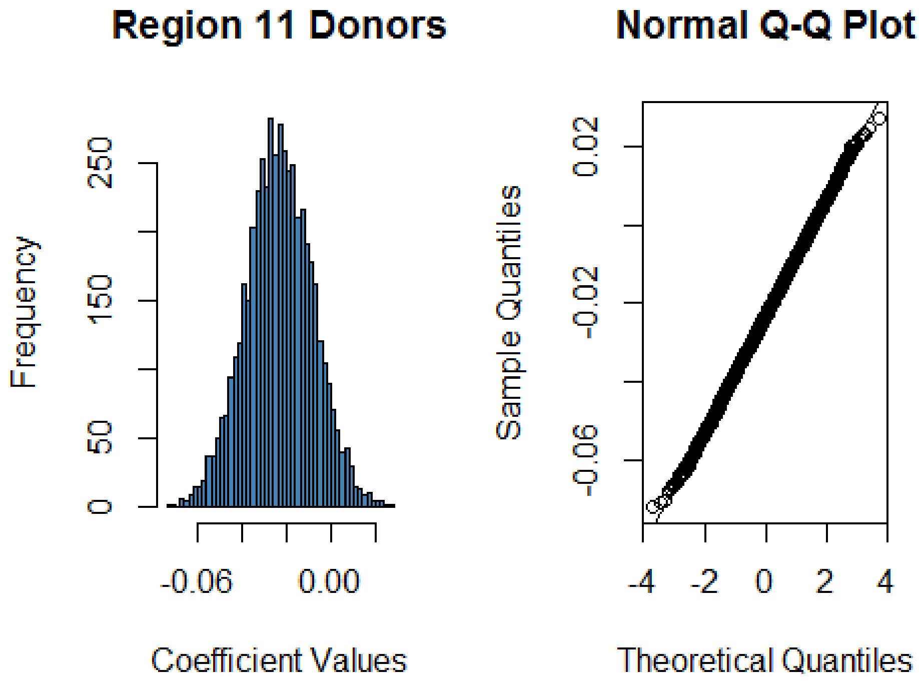
Estimate the model with bootstrapping (by residuals). Is b1 significant?

```
# Get the fitted values from the regression model
uvaduke.fit<-fitted(uva.duke.lm)
#Get the residuals from the regression model
uvaduke.resid<-residuals(uva.duke.lm)
#Get the regression model
uvaduke.mod<-model.matrix(uva.duke.lm)
#Bootstrapping LM [UVA and Duke]
uvaduke.boot<-RTSB(uva.duke.diff,r11donor$Kidney[-30],uvaduke.fit,uvaduke.res
id,uvaduke.mod,5000)
uvaduke.boot

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = cbind(resp, pred), statistic = coef.fun, R = num,
##     fit = fit, resid = resid, X2 = X)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1*  2.71585614  1.682581e-02  15.4069112
## t2* -0.02321623 -4.496154e-05   0.0149151

#95% CI of r11donor
boot.ci(uvaduke.boot, .95, index=2)

## Warning in boot.ci(uvaduke.boot, 0.95, index = 2): bootstrap variances
## needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = uvaduke.boot, conf = 0.95, index = 2)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.0524,  0.0061 )   (-0.0526,  0.0064 )
##
## Level     Percentile            BCa
## 95%   (-0.0528,  0.0061 )   (-0.0520,  0.0070 )
## Calculations and Intervals on Original Scale
```

All 4 confidence intervals include zero as a value, in contrast to the confidence interavals for the previous model.

Distribution of b1

```
par(mfrow = c(1,2))
hist(uvaduke.boot$t[,2], main = "Region 11 Donors",xlab ="Coefficient Values",
   col = "steelblue", breaks = 50)
qqnorm(uvaduke.boot $t[,2])
qqline(uvaduke.boot $t[,2])
```



```
par(mfrow = c(1,1))
```

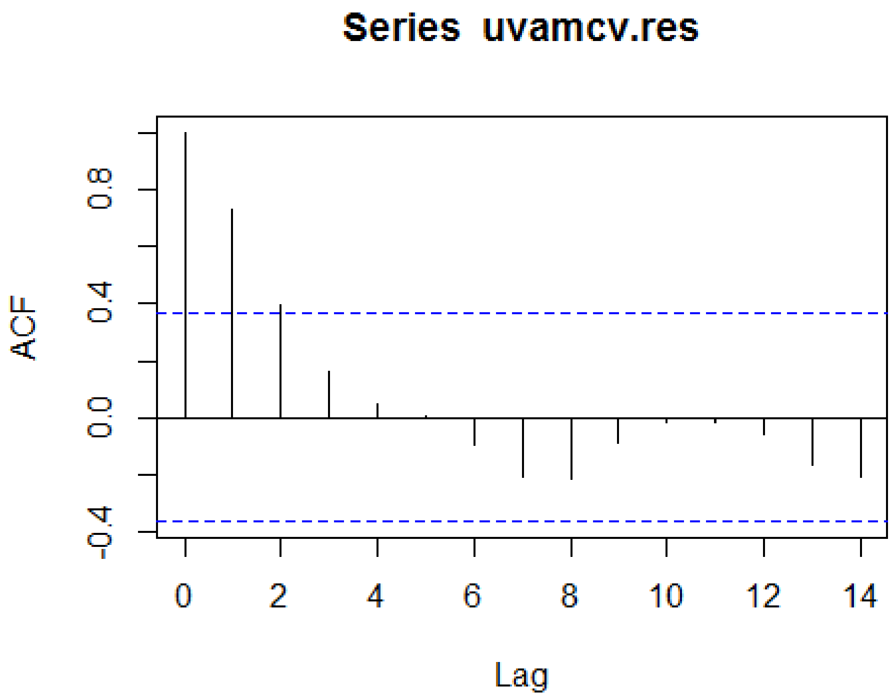
Is b1 significant?

B1 is not significant for the model predicting the differences between UVA and Duke, in contrast to the previous model predicting the differences between UVA and MCV (which was significant). This is because all four of the confidence intervals for b1 included zero as a value. In addition, the histogram for b1 has a much greater frequency for a coefficeint value of zero than for the previous model. And the normal QQ plot has a sample quantile of zero that is much closer to a zero theoretical quantile than the previous model.

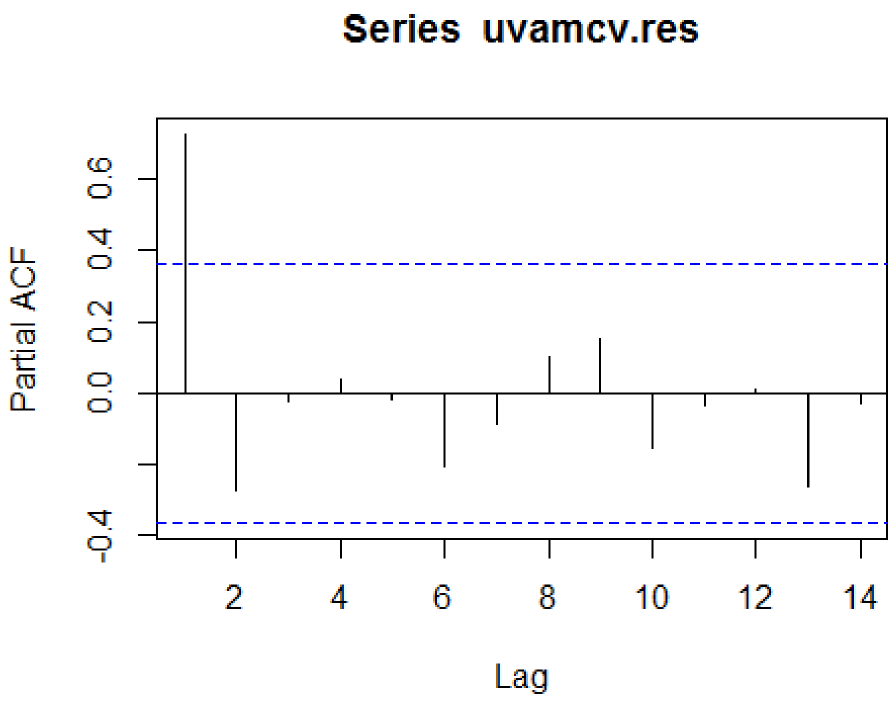
Part 3: Time Series Models

Step 3.1 Generate the ACF and PACF plots of the residuals from your part 2 linear model for UVA and MCV kidney transplants. Interpret the results of the ACF and PACF plots. Do you recommend modeling the residuals? If so, what kind of model should you try based on these plots?

```
acf(uvamcv.res)
```



```
pacf(uvamcv.res)
```



The acf and pacf plots indicate that the the residuals are stationary - exponential decay in the ACF plot. However, the plots show that there are correlated residuals, indicating that there is still structure to be modeled. We would recommend modeling the residuals with an AR(1) model because the PACF plot cuts off after lag 1, but the ACF shows sinusoidal decay.

Step 3.2 Based on the above ACF and PACF plots, fit an ar model to the residuals

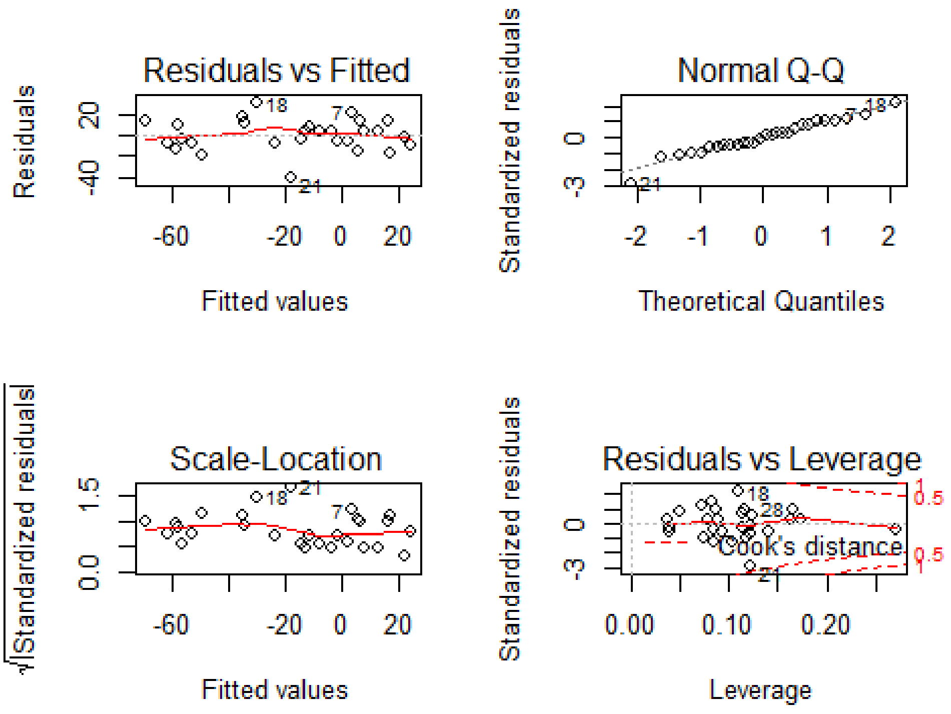
Add the AR model of the residuals to regression linear model based on the acf/pacf plots. Call this model uvamcv.kidney.lm2. Analyze the regression results.

```
uva.mcv.diff2<-(uva$Kidney[2:29]-mcv$Kidney[2:29])
uvamcv.kidney.lm2<- lm(uva.mcv.diff2~r11donor$Kidney[2:29]+uva.kidney.lm$resi
duals[1:28])
summary(uvamcv.kidney.lm2)

##
## Call:
## lm(formula = uva.mcv.diff2 ~ r11donor$Kidney[2:29] + uva.kidney.lm$residua
ls[1:28])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.559  -7.609   0.911  10.577  29.942
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    68.51887   11.04990   6.201 1.74e-06 ***
## r11donor$Kidney[2:29]          -0.08473    0.01050  -8.073 1.99e-08 ***
## uva.kidney.lm$residuals[1:28]  0.73848    0.11884   6.214 1.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.79 on 25 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.7872
## F-statistic: 50.93 on 2 and 25 DF,  p-value: 1.523e-09
```

The new model has an Adjusted R^2 of .787, indicating a great fit. A model utility test resulted in a extremely significant p-value (very close to zero), indicating that the model is very significant. Lastly, the p-values for both the r11donor variable and the residuals variable are significant, implying that the added residuals model is useful in predicting the difference in UVA and MCV kidney transplants.

Generate diagnostic plots for uvamcv.kidney.lm2. What are your observations?

```
par(mfrow=c(2,2))
plot(uvamcv.kidney.lm2)
```



```
par(mfrow=c(1,1))
```

The residuals vs fitted plot indicates that while there is some heteroscendasticity, there is much less than in the previous model. The normal QQ plot shows that the distribution is relatively normal. There are also a few outliers, with one having a Leverage of close to .5. Overall, this model is a better fit than the previous model without the added residuals.

Step 3.3 Bootstrap the above time series model. Are the coefficients significant?

```
#    Get the fitted values from the regression model
uvamcv.lm2.fit<-fitted(uvamcv.kidney.lm2)
#    Get the residuals from the regression model
uvamcv.lm2.resid<-residuals(uvamcv.kidney.lm2)
#    Get the regression model
uvamcv.lm2.mod<-model.matrix(uvamcv.kidney.lm2)
#     Use the RTSB function to obtain the bootstrap
uvamcv.lm2.boot<-RTSB(uva.mcv.diff2,r11donor$Kidney[2:29],uvamcv.lm2.fit,uvam
cv.lm2.resid,uvamcv.lm2.mod,5000)
#     The estimates
uvamcv.lm2.boot
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = cbind(resp, pred), statistic = coef.fun, R = num,
##     fit = fit, resid = resid, X2 = X)
##
##
## Bootstrap Statistics :
##        original        bias     std. error
## t1* 68.51886681 -2.289802e-01 10.306782771
## t2* -0.08473044  2.540858e-04  0.009808855
## t3*  0.73847881 -1.832219e-05  0.113437794
```

```
summary(uvamcv.kidney.lm2)
```

```
##
## Call:
## lm(formula = uva.mcv.diff2 ~ r11donor$Kidney[2:29] + uva.kidney.lm$residua
ls[1:28])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.559  -7.609   0.911  10.577  29.942
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   68.51887   11.04990   6.201 1.74e-06 ***
## r11donor$Kidney[2:29]         -0.08473    0.01050  -8.073 1.99e-08 ***
## uva.kidney.lm$residuals[1:28]  0.73848    0.11884   6.214 1.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.79 on 25 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.7872
## F-statistic: 50.93 on 2 and 25 DF,  p-value: 1.523e-09
```

```r
uvamcv.lm2.boot$t
```

```
##               [,1]        [,2]      [,3]
##    [1,]  68.58376 -0.08569992 0.7666128
##    [2,]  66.45976 -0.08462506 0.5693139
##    [3,]  69.15767 -0.08100353 0.6974397
##    [4,]  64.30341 -0.08135681 0.6029577
##    [5,]  66.59142 -0.08617817 0.6049983
##    [6,]  72.51649 -0.08534610 0.7034461
##    [7,]  73.59860 -0.09112669 0.6507626
##    [8,]  63.46417 -0.08046237 0.7433778
##    [9,]  77.88243 -0.09860112 0.6141399
##   [10,]  62.49586 -0.08341954 0.6688954
#############[Omitted the bootstrap results in between]######################
## [4998,]  62.14758 -0.07648758 0.7608643
## [4999,]  69.00107 -0.08361433 0.7204979
## [5000,]  68.23365 -0.08224317 0.7751768
```

```r
sqrt(abs(var(uvamcv.lm2.boot$t)))
```

```
##             [,1]        [,2]        [,3]
## [1,] 10.3067828 0.312652419 0.166540247
## [2,]  0.3126524 0.009808855 0.005302774
## [3,]  0.1665402 0.005302774 0.113437794
```

```r
boot.ci(uvamcv.lm2.boot, .95, index=2) #for r11donor variable
```

```
## Warning in boot.ci(uvamcv.lm2.boot, 0.95, index = 2): bootstrap variances
## needed for studentized intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = uvamcv.lm2.boot, conf = 0.95, index = 2)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.1042, -0.0658 )   (-0.1043, -0.0659 )
##
## Level     Percentile            BCa
## 95%   (-0.1036, -0.0652 )   (-0.1036, -0.0652 )
## Calculations and Intervals on Original Scale
```

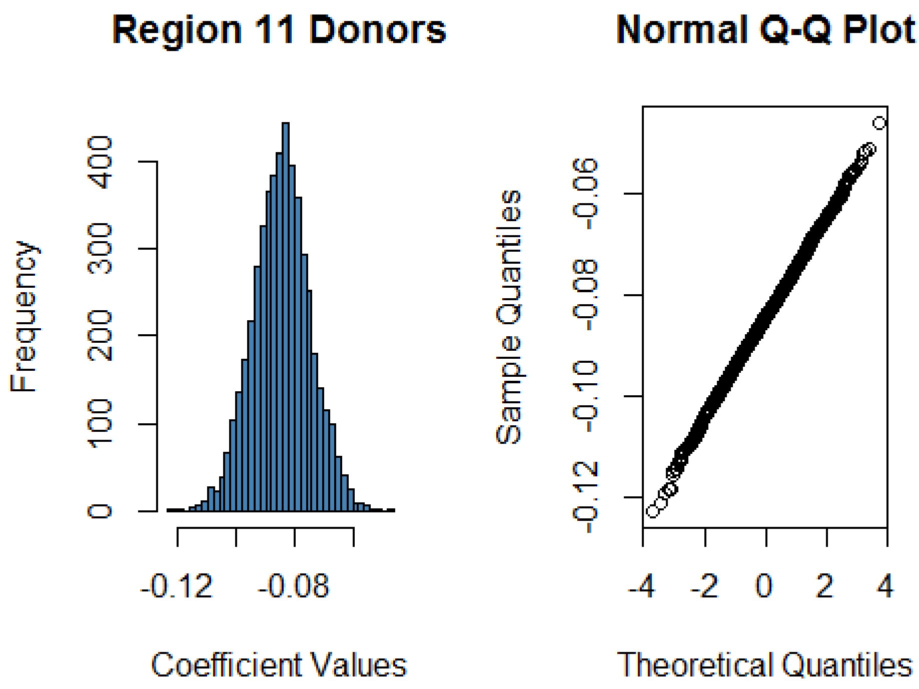
```r
boot.ci(uvamcv.lm2.boot, .95, index=3) #for residual variable
```

```
## Warning in boot.ci(uvamcv.lm2.boot, 0.95, index = 3): bootstrap variances
## needed for studentized intervals
```

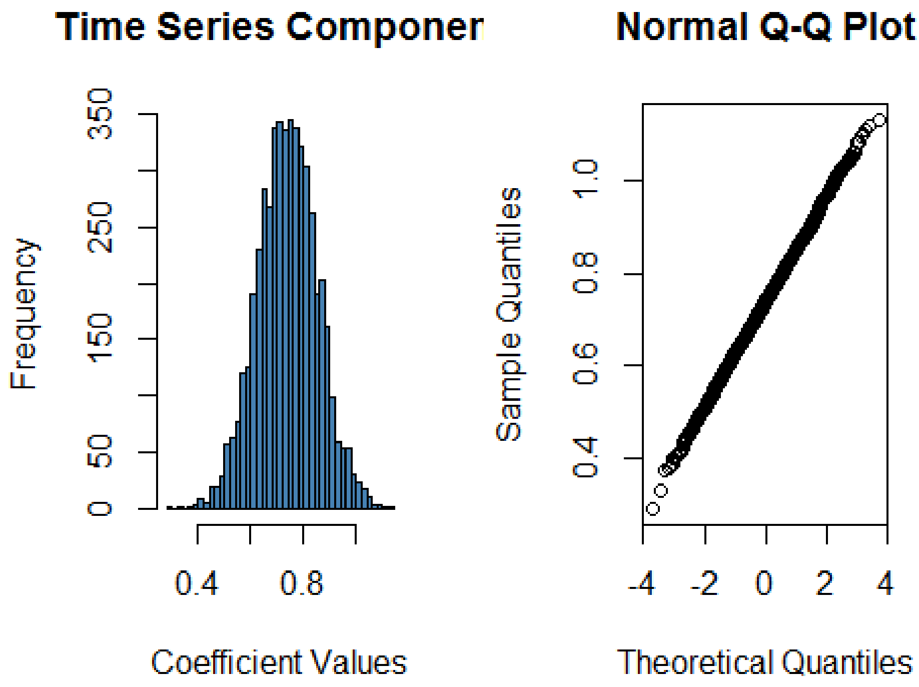
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
```

```
## CALL :
## boot.ci(boot.out = uvamcv.lm2.boot, conf = 0.95, index = 3)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.5162,  0.9608 )   ( 0.5109,  0.9654 )
##
## Level     Percentile            BCa
## 95%   ( 0.5116,  0.9661 )   ( 0.5201,  0.9730 )
## Calculations and Intervals on Original Scale
```

```
#    Plot the results for the coeffeciecient for region 11 donors
par(mfrow = c(1,2))
hist(uvamcv.lm2.boot$t[,2], main = "Region 11 Donors",xlab ="Coefficient Valu
es",   col = "steelblue", breaks = 50)
qqnorm(uvamcv.lm2.boot$t[,2])
```



```
par(mfrow = c(1,1))
#    Plot the results for the coeffeciecient for time series components
par(mfrow = c(1,2))
hist(uvamcv.lm2.boot$t[,3], main = "Time Series Component",xlab ="Coefficient
 Values",   col = "steelblue", breaks = 50)
qqnorm(uvamcv.lm2.boot$t[,3])
```



```
par(mfrow = c(1,1))
```

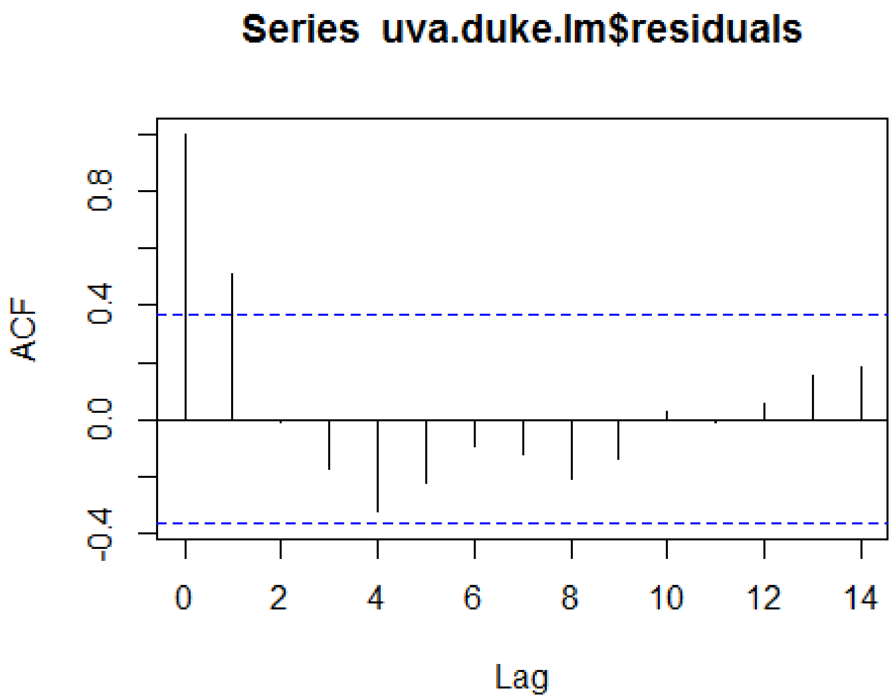
Are the coefficients significant?

The coefficient for region 11 donors is very significant, due to the histogram of its coefficient values; there is a frequency of about zero for a coefficient value of 0 The confidence intervals for the region 11 donors coefficient also do not contain zero in them. The coefficient for the time series components is also very significant due to the histogram being centered around a coefficient value of .8. The confidence intervals for the time sereies component coefficeint also do not contain zero in them.

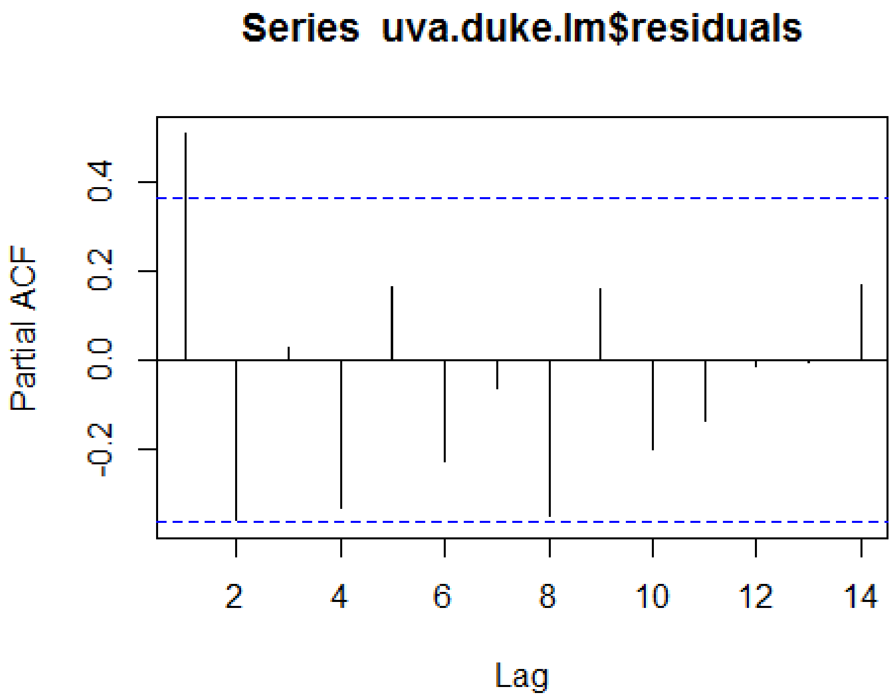
Step 3.5* (bonus) What about Duke? Repeat the above steps and compare the results.

Generate the ACF and PACF plots of the residuals from your part 2 linear model for UVA and Duke kidney transplants. Interpret the results of the ACF and PACF plots. Do you recommend modeling the residuals? If so, what kind of model should you try based on these plots?

```
acf(uva.duke.lm$residuals)
```



```
pacf(uva.duke.lm$residuals)
```



The plots show that there are correlated residuals, indicating that there is still structure to be modeled. We would recommend modeling the residuals with an AR(2) model because the PACF plot cuts off after lag 2. This is different with UVA vs MCV in which we modeled the residuals with an AR(1) model.

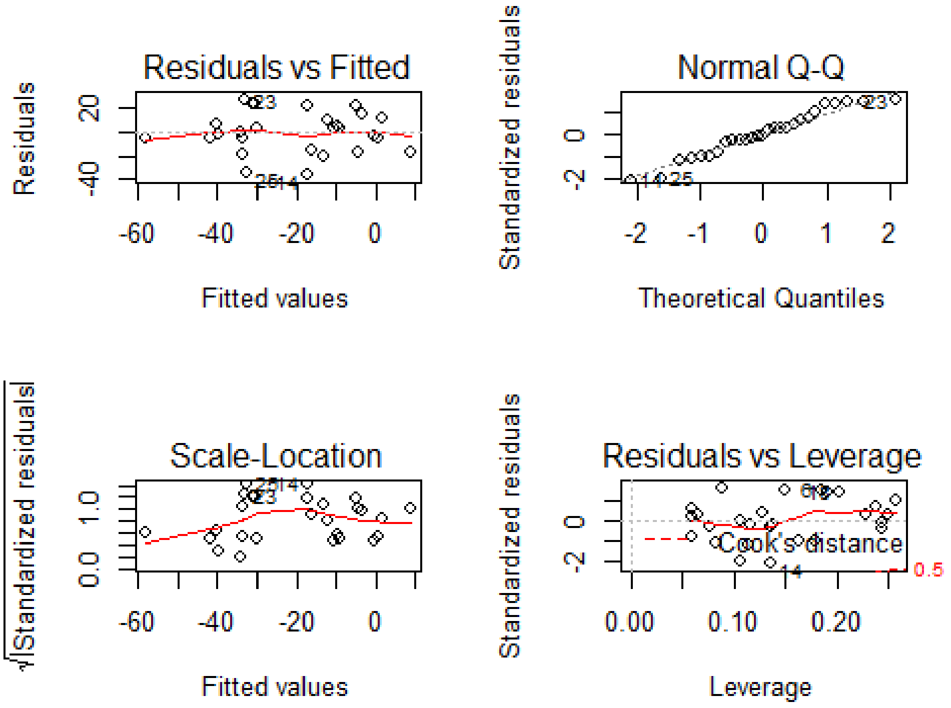
```
uva.duke.diff2<-(uva$Kidney[3:29]-duke$Kidney[3:29])
uvaduke.kidney.lm2<- lm(uva.duke.diff2~r11donor$Kidney[3:29]+uva.duke.lm$resi
duals[1:27]+uva.duke.lm$residuals[2:28])
summary(uvaduke.kidney.lm2)
```

```
##
## Call:
## lm(formula = uva.duke.diff2 ~ r11donor$Kidney[3:29] + uva.duke.lm$residual
s[1:27] +
##     uva.duke.lm$residuals[2:28])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.570 -10.290  -0.737  11.394  27.027
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.31900   15.39060   0.411 0.685184
## r11donor$Kidney[3:29]       -0.02597    0.01439  -1.805 0.084162 .
## uva.duke.lm$residuals[1:27] -0.41587    0.19343  -2.150 0.042306 *
## uva.duke.lm$residuals[2:28]  0.74838    0.19046   3.929 0.000671 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.54 on 23 degrees of freedom
## Multiple R-squared:  0.4778, Adjusted R-squared:  0.4097
## F-statistic: 7.014 on 3 and 23 DF,  p-value: 0.00162
```

The new model has an Adjusted R^2 of .41, indicating a great fit. A model utility test resulted in a extremely significant p-value (very close to zero), indicating that the model is very significant. Lastly, the p-values for both the r11donor variable and the residuals variable are significant, implying that the added residuals model is useful in predicting the difference in UVA and MCV kidney transplants.

Generate diagnostic plots for uvamcv.kidney.lm2. What are your observations?

```
par(mfrow=c(2,2))
plot(uvaduke.kidney.lm2)
```



```
par(mfrow=c(1,1))
```

The residuals vs fitted plot indicates that while there is some heteroscendasticity, there is much less than in the previous model, the normal QQ plot shows that the distribution is relatively normal there are also a few outliers, with one having a Leverage of close to .5. Overall, this model is a better fit than the previous model without the added residuals.

Bootstrapping

```
#    Get the fitted values from the regression model
uvaduke.lm2.fit<-fitted(uvaduke.kidney.lm2)
#    Get the residuals from the regression model
uvaduke.lm2.resid<-residuals(uvaduke.kidney.lm2)
#    Get the regression model
uvaduke.lm2.mod<-model.matrix(uvaduke.kidney.lm2)
#     Use the RTSB function to obtain the bootstrap
uvaduke.lm2.boot<-RTSB(uva.duke.diff2,r11donor$Kidney[3:29],uvaduke.lm2.fit,u
vaduke.lm2.resid,uvaduke.lm2.mod,5000)
#     The estimates
uvaduke.lm2.boot

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = cbind(resp, pred), statistic = coef.fun, R = num,
##     fit = fit, resid = resid, X2 = X)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1*  6.31899624 -0.1308063461 14.14608671
## t2* -0.02596803  0.0001252857  0.01318283
## t3* -0.41587097 -0.0002160284  0.17832983
## t4*  0.74837591  0.0013779390  0.17689176
```

```
summary(uvaduke.kidney.lm2)
```

```
##
## Call:
## lm(formula = uva.duke.diff2 ~ r11donor$Kidney[3:29] + uva.duke.lm$residual
s[1:27] +
##     uva.duke.lm$residuals[2:28])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.570 -10.290  -0.737  11.394  27.027
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.31900   15.39060   0.411 0.685184
## r11donor$Kidney[3:29]       -0.02597    0.01439  -1.805 0.084162 .
## uva.duke.lm$residuals[1:27] -0.41587    0.19343  -2.150 0.042306 *
## uva.duke.lm$residuals[2:28]  0.74838    0.19046   3.929 0.000671 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.54 on 23 degrees of freedom
## Multiple R-squared:  0.4778, Adjusted R-squared:  0.4097
## F-statistic: 7.014 on 3 and 23 DF,  p-value: 0.00162
```

```
boot.ci(uvaduke.lm2.boot, .95, index=2)
```

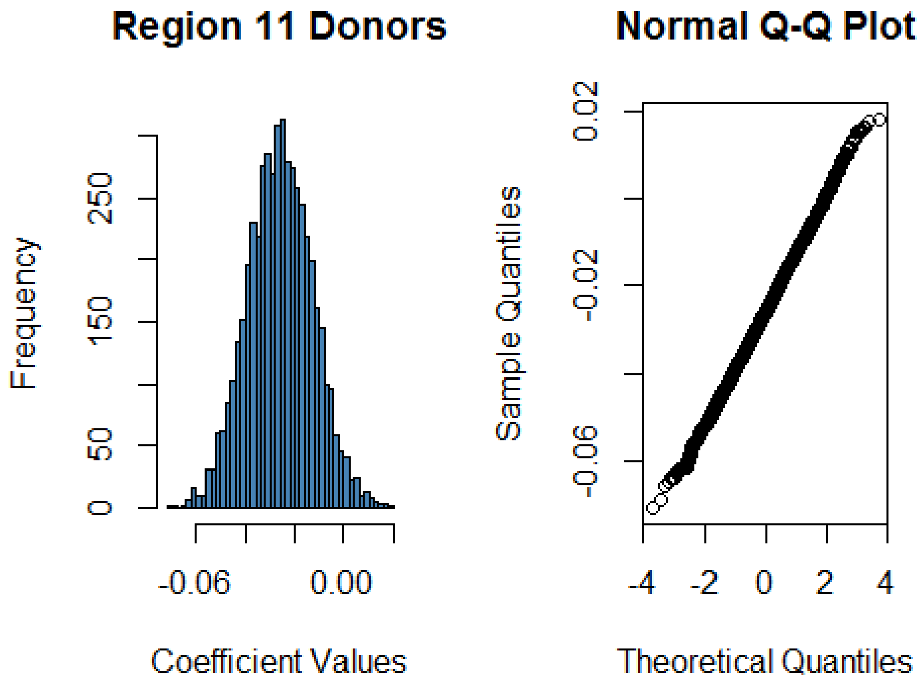
```
## Warning in boot.ci(uvaduke.lm2.boot, 0.95, index = 2): bootstrap variances
## needed for studentized intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = uvaduke.lm2.boot, conf = 0.95, index = 2)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.0519, -0.0003 )   (-0.0523, -0.0004 )
##
## Level     Percentile            BCa
## 95%   (-0.0515,  0.0003 )   (-0.0516,  0.0002 )
## Calculations and Intervals on Original Scale
```

According to the confidence intervals, the r11donor variable is significant because Percentile and BCa don't bracket zero.

Plot the results for the coeffiecient for region 11 donors

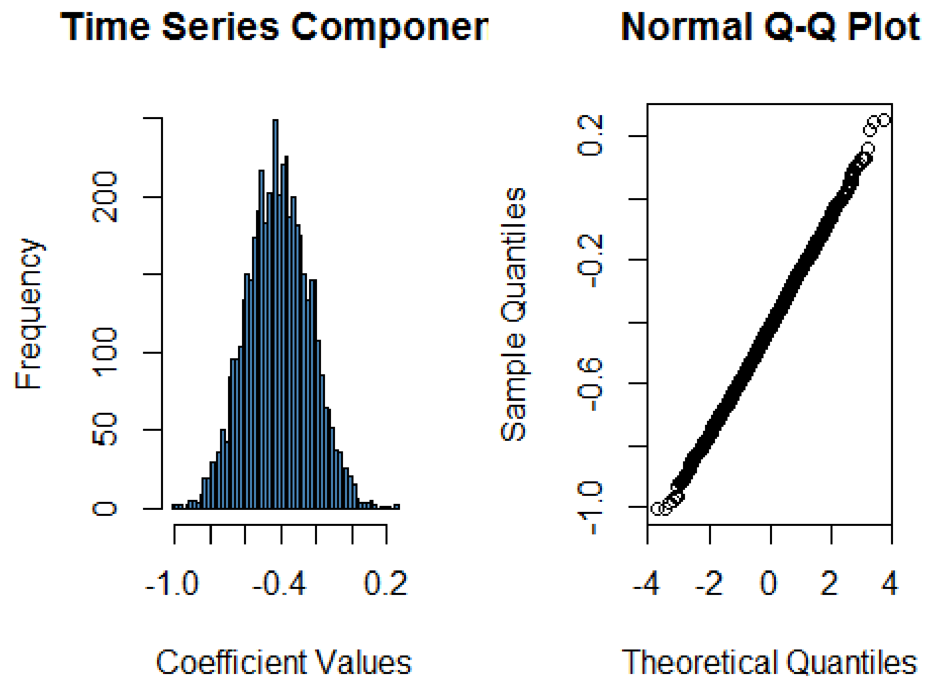
```
par(mfrow = c(1,2))
hist(uvaduke.lm2.boot$t[,2], main = "Region 11 Donors",xlab ="Coefficient Val
ues",   col = "steelblue", breaks = 50)
qqnorm(uvaduke.lm2.boot$t[,2])
```



```
par(mfrow = c(1,1))
```

Plot the results for the coeffiecient for time series components

```
par(mfrow = c(1,2))
hist(uvaduke.lm2.boot$t[,3], main = "Time Series Component",xlab ="Coefficien
t Values",   col = "steelblue", breaks = 50)
qqnorm(uvaduke.lm2.boot$t[,3])
```



```r
par(mfrow = c(1,1))
```

Are the coefficients significant?

The coefficient for region 11 donors is significant, due to the histogram of its coefficient values; there is a frequency of about zero for a coefficient value of 0. The confidence intervals for the region 11 donors coefficient also do not contain zero in them. The coefficient for the time series components is also significant due to the histogram being centered around a coefficient value of .8. The confidence intervals for the time sereies component coefficeint also do not contain zero in them.

Part 4: Predicting Differences in Kidney Transplants Part 1

Step 4.1 Build an AR model to predict the difference in 2017

```r
uvamcv.diff<-ts(uva$Kidney-mcv$Kidney,1988,2017)
uvamcv.ar<-ar(uvamcv.diff,method="yule-walker")
```

Step 4.2 Use the predict function with ar model to forecast 2017 differences between UVA and MCV

```r
uvamcv.pred<-predict(uvamcv.ar,newdata=uvamcv.diff)
uvamcv.pred<-predict(uvamcv.ar,se.fit=T,interval="predict")
uvamcv.pred
```

```
## $pred
## Time Series:
## Start = 2018
## End = 2018
## Frequency = 1
## [1] -11.95251
##
## $se
## Time Series:
## Start = 2018
## End = 2018
## Frequency = 1
## [1] 16.81561
```

Calculate the CI and plot the time series and prediction and CIs on a graph

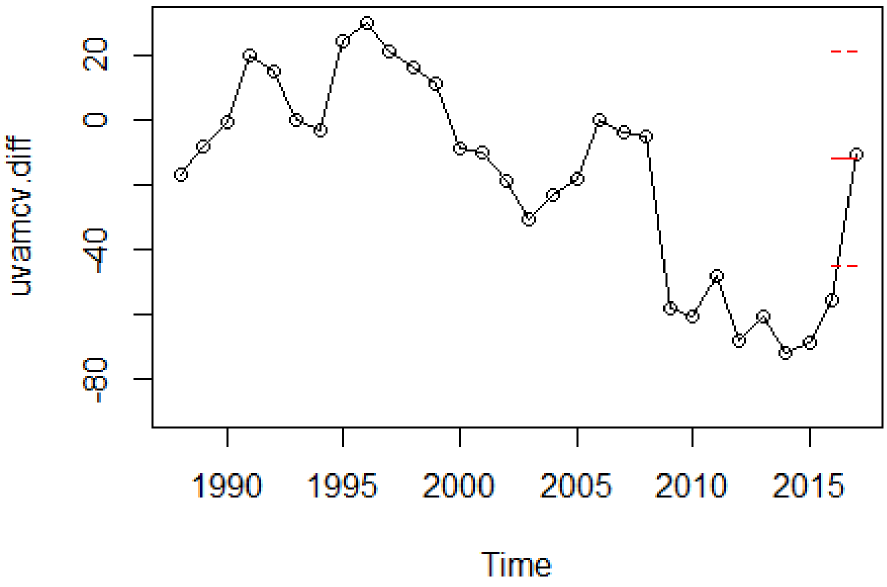
Plot the historical time series, the new prediction, and CIs on a graph

```
uvamcv.pred$pred+1.96*uvamcv.pred$se #upper CI

## Time Series:
## Start = 2018
## End = 2018
## Frequency = 1
## uvamcv.pred$pred
##         21.00609

uvamcv.pred$pred-1.96*uvamcv.pred$se #lower CI

## Time Series:
## Start = 2018
## End = 2018
## Frequency = 1
## uvamcv.pred$pred
##        -44.91111
```

The CI is (-45, 21).

What do you observe?

```
plot(uvamcv.diff,ylim=c(-90,30),type='o') #historical time series
segments(2016,uvamcv.pred$pred,col = "red",2017) # prediction
segments(2016,uvamcv.pred$pred+1.96*uvamcv.pred$se,col = "red",2017,lty = "da
shed") # upper CI
segments(2016,uvamcv.pred$pred-1.96*uvamcv.pred$se,col = "red",2017,lty = "da
shed") # lower CI
```



We observed that the predicting for 2017 is very close to the actual 2017 value, which is much greater than for 2016. The confidence interval is also very wide, with the lower bound being much around -45 and the upper bound being around 21.

Step 4.3 Bootstrapping the difference of UVa and MCV in 2017

To obtain a bootstrap estimate of the prediction for 2017 use the TSB function in the source file.

```
uvamcv.boot<-TSB(uvamcv.diff,uvamcv.pred$pred,5000)
uvamcv.boot

## 
## MODEL BASED BOOTSTRAP FOR TIME SERIES
## 
## 
## Call:
## tsboot(tseries = ts.res, statistic = tspred.fun, R = boot.number, 
##     sim = "model", n.sim = 3 * length(tsint), orig.t = FALSE, 
##     ran.gen = ts.sim, ran.args = list(ts = tsint, model = ts.model), 
##     tsnew = oth.arg)
## 
## 
## Bootstrap Statistics :
##          mean std. error
## t1* -13.01386   2.385828
```

```
tsboot.ci(uvamcv.boot)

##          5%        95%
## -17.082627  -9.334297
```

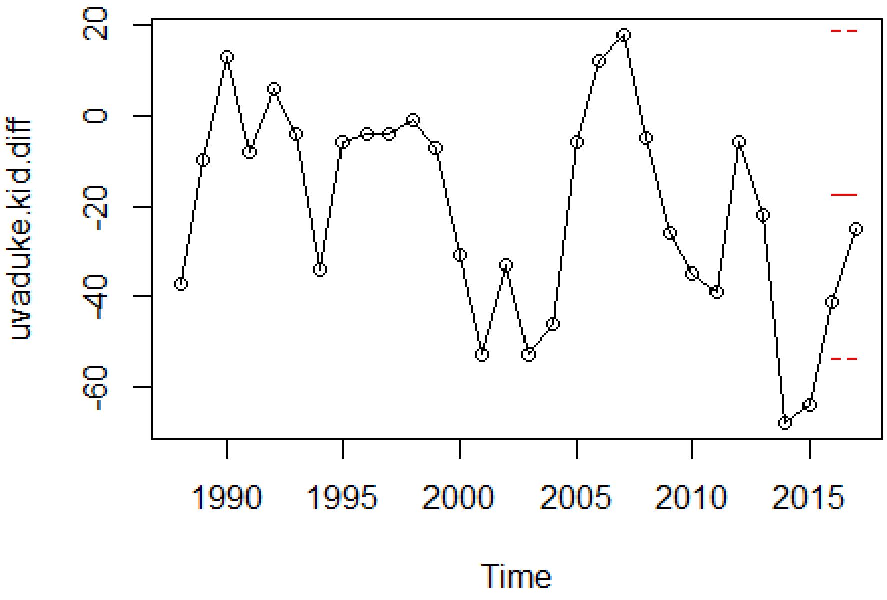
Interpret the results

Is there a significant difference between UVA and MCV in 2017?

There is a significant difference between UVA and MCV in 2017 the mean difference was estimated as -13, and the 95% CI was (-17.14,-9.3) which doesn't include zero. The actual 2017 difference falls with in the bootstrapping interval.

Step 4.4* (bonus) What about Duke? Repeat the above steps and compare for Duke.

```
uvaduke.kid.diff<-ts(uva$Kidney-duke$Kidney,1988,2017)
uvaduke.ar.kid<-ar(uvaduke.kid.diff, method = "yule-walker")

uvaduke.kidney.pred<-predict(uvaduke.ar.kid,newdata=uvaduke.kid.diff)

plot(uvaduke.kid.diff,type='o')
segments(2016,uvaduke.kidney.pred$pred,col = "red",2017) # Prediction
segments(2016,uvaduke.kidney.pred$pred+1.96*uvaduke.kidney.pred$se,col = "red
",2017,lty = "dashed") # upper CI
segments(2016,uvaduke.kidney.pred$pred-1.96*uvaduke.kidney.pred$se,col = "red
",2017,lty = "dashed") # lower CI
```



The prediction is very close to the actual 2017 difference. The 95% CI is very wide and contains the actual 2017 difference.

```
diff.boot<-TSB(uvaduke.kid.diff,uvaduke.kidney.pred$pred,5000)
diff.boot

##
## MODEL BASED BOOTSTRAP FOR TIME SERIES
##
##
## Call:
## tsboot(tseries = ts.res, statistic = tspred.fun, R = boot.number,
##     sim = "model", n.sim = 3 * length(tsint), orig.t = FALSE,
##     ran.gen = ts.sim, ran.args = list(ts = tsint, model = ts.model),
##     tsnew = oth.arg)
##
##
## Bootstrap Statistics :
##          mean std. error
## t1* -19.28815   1.978364

tsboot.ci(diff.boot)

##        5%       95%
## -22.66402 -16.17790
```

There is a significant difference between UVA and Duke in 2017. The mean difference was estimated as -20.96, and the 95% CI was (-24.22,-17.65) which doesn't bracket zero

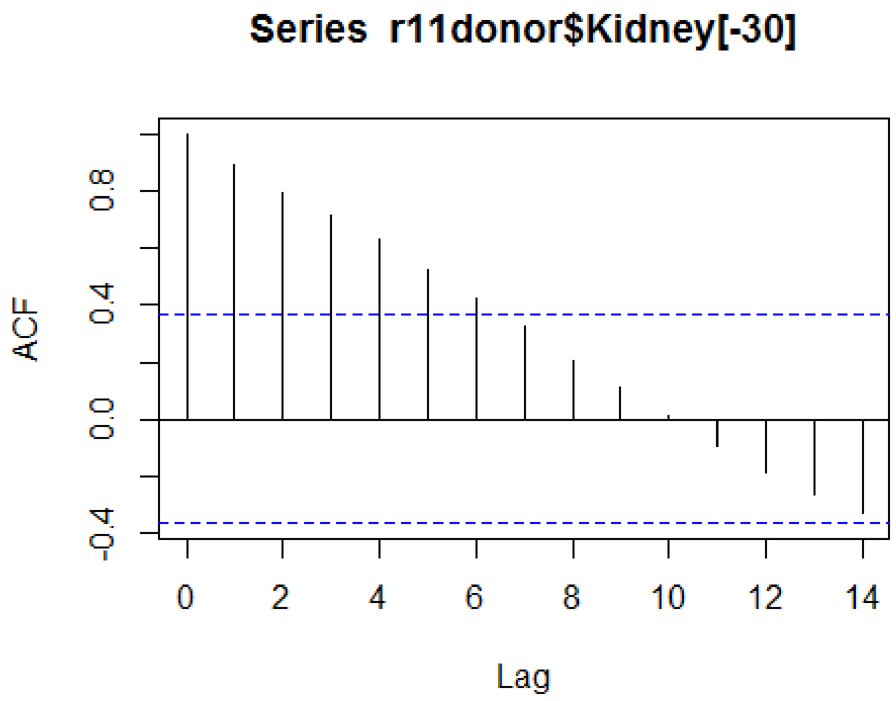
---

Part 5: Predicting Differences in Kidney Transplants Part 2

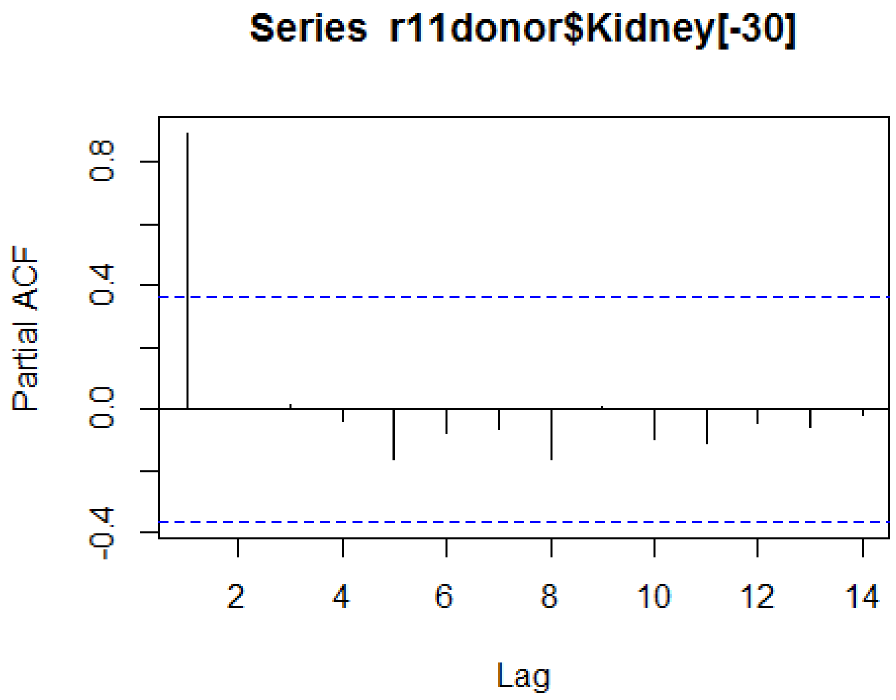
Step 5.1 Develop an AR model of region 11 kidney donors

Plot an acf/pacf to estimate the model order for region 11 kidney donors - what model order do you suggest?

```
acf(r11donor$Kidney[-30])
```



```
pacf(r11donor$Kidney[-30])
```



An AR(1) model is recommended because the pacf cuts off after lag 1

Use ar() to fit an ar model to region 11 kidney donors from 1988-2016

```
r11donor.ar<-ar(r11donor$Kidney[-30],method="yule-walker")
r11donor.ar

##
## Call:
## ar(x = r11donor$Kidney[-30], method = "yule-walker")
##
## Coefficients:
##      1
## 0.8923
##
## Order selected 1  sigma^2 estimated as  17285
```

Step 5.2 Forecast the R11 donors and standard errors for 2017 using your ar model from step 5.1. Use forecast from the library(forecast).

```
library(forecast)

## Warning: package 'forecast' was built under R version 3.4.2

r11donor.pred <- forecast(r11donor.ar,h=1)
r11donor.pred

##    Point Forecast    Lo 80    Hi 80   Lo 95    Hi 95
## 30       1289.039 1120.551 1457.526 1031.36 1546.718
```

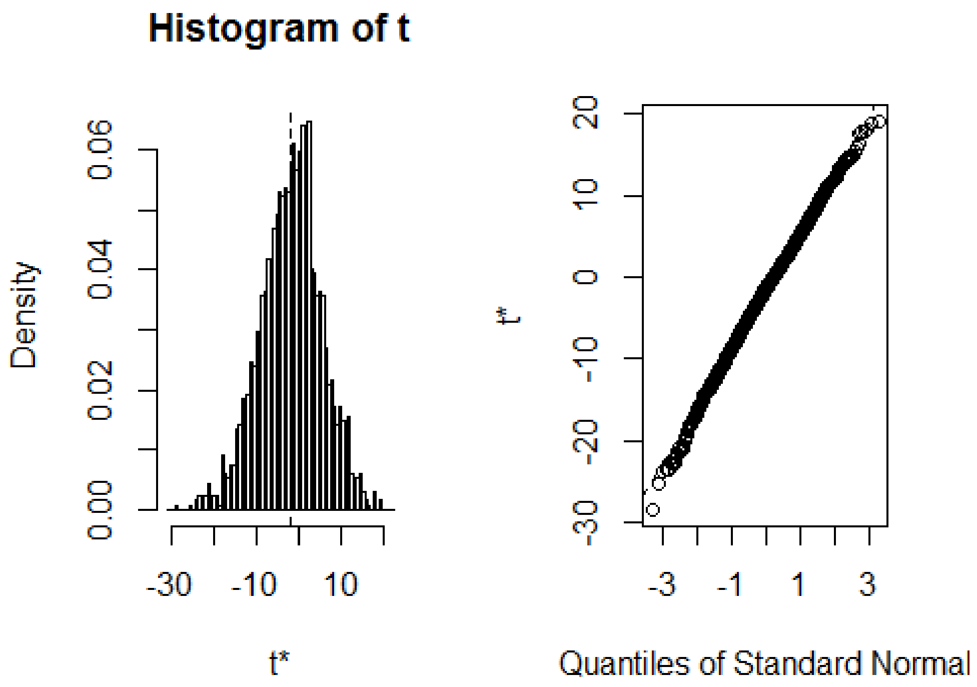
Step 5.3 Use the linear model from part 3.2 combined with the forecast of region 11 kidney donors to forecast the differences in number of kidney transplants between UVa and MCV for 2017. Use the predict() function

```
#   Creating the new data frame
uvamcv.nd<-data.frame(r11k=r11donor.pred$mean,uvamcv.lm.e1=uva.kidney.lm$resi
duals[29])
#   Predict the linear model with the time series
uvamcv.new<-predict(uvamcv.kidney.lm2,newdata=uvamcv.nd,num=5000)

## Warning: 'newdata' had 1 row but variables found have 28 rows
```

Step 5.4 Bootstrap the Forecast from the linear model combined with the forecast of region 11 kidney donors to forecast the differences in number of kidney transplants between UVa and MCV for 2017.

```
#   Bootstrap prediction
r11donor.kidney<-r11donor$Kidney[-30]
r11k <- r11donor.kidney[2:29]
uvamcv.dm <- data.frame(uvamcv.diff[2:29], r11k, uva.kidney.lm$residuals[1:2
8])
r11donor.boot <- RFB(uvamcv.dm, model = uvamcv.kidney.lm2, ndata = uvamcv.nd,
 num=2000)
```

```
## Warning: 'newdata' had 1 row but variables found have 28 rows

#   Bootstrap plot
plot(r11donor.boot, index = 1)
```



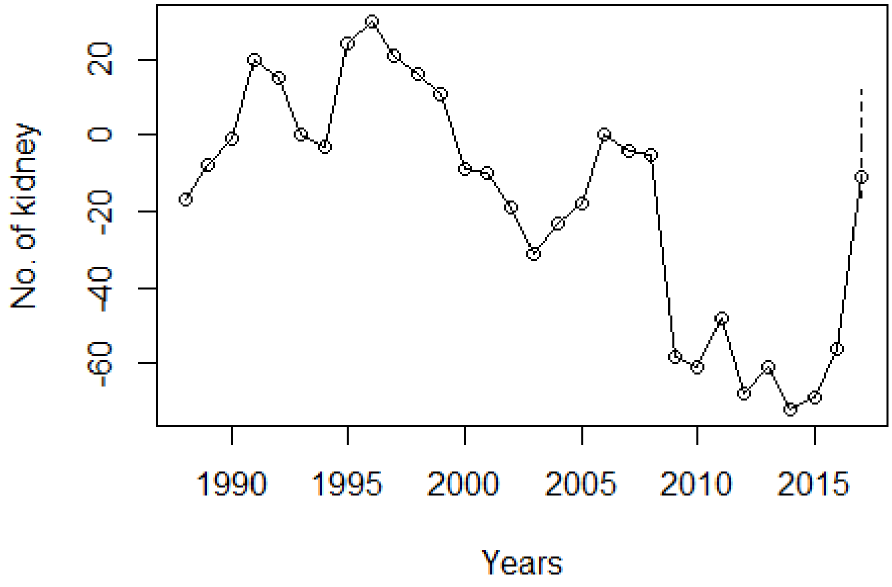
```
#   Bootstrap confidence intervals
r11donor.boot.ci<-boot.ci(r11donor.boot, index=1)

## Warning in boot.ci(r11donor.boot, index = 1): bootstrap variances needed
## for studentized intervals

r11donor.boot.ci

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = r11donor.boot, index = 1)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-16.154,  11.707 )   (-15.916,  12.211 )
##
## Level     Percentile            BCa
## 95%   (-16.260,  11.868 )   (-17.774,  10.795 )
## Calculations and Intervals on Original Scale
```

Interpret the results

The confidence interval for bootstrap is large, and the actual value falls into the confidence interval.

Step 5.5 Plot the current and predictions for each value along with the confidence intervals. Describe your observations.

```
uvamcv<-uva$Kidney-mcv$Kidney
plot(uvamcv~uva$Year,xlim=c(1988,2017), type="o",xlab="Years",ylab="No. of ki
dney")
segments(2017,r11donor.boot$t0[1],2017, r11donor.boot.ci$percent[4],lty="dash
ed")
segments(2017,r11donor.boot$t0[1],2017, r11donor.boot.ci$percent[5],lty="dash
ed")
```



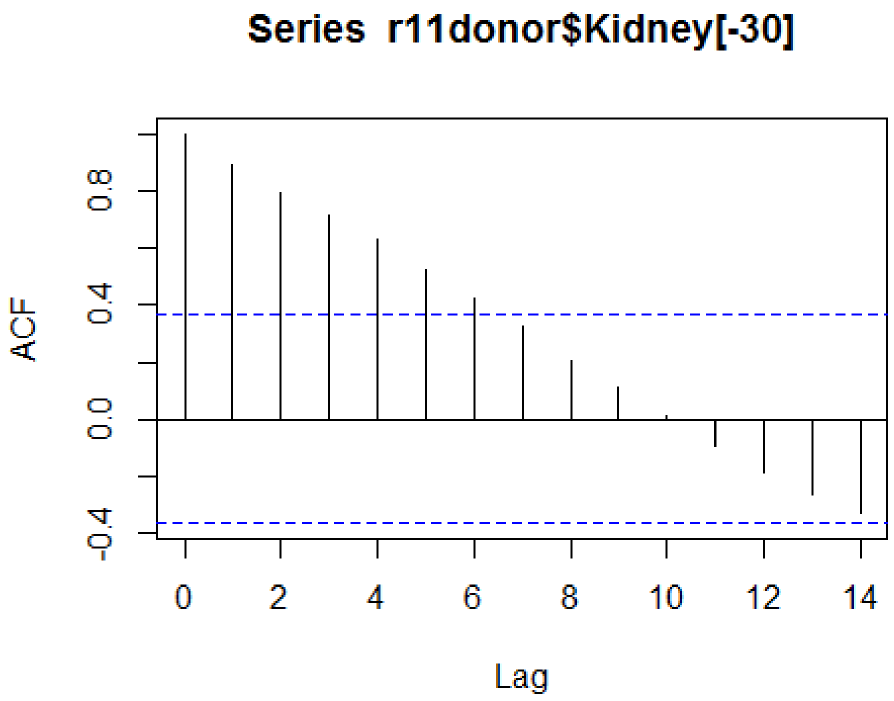
Observations?

The actual difference for 2017 falls in the bootstrapping confidence interval

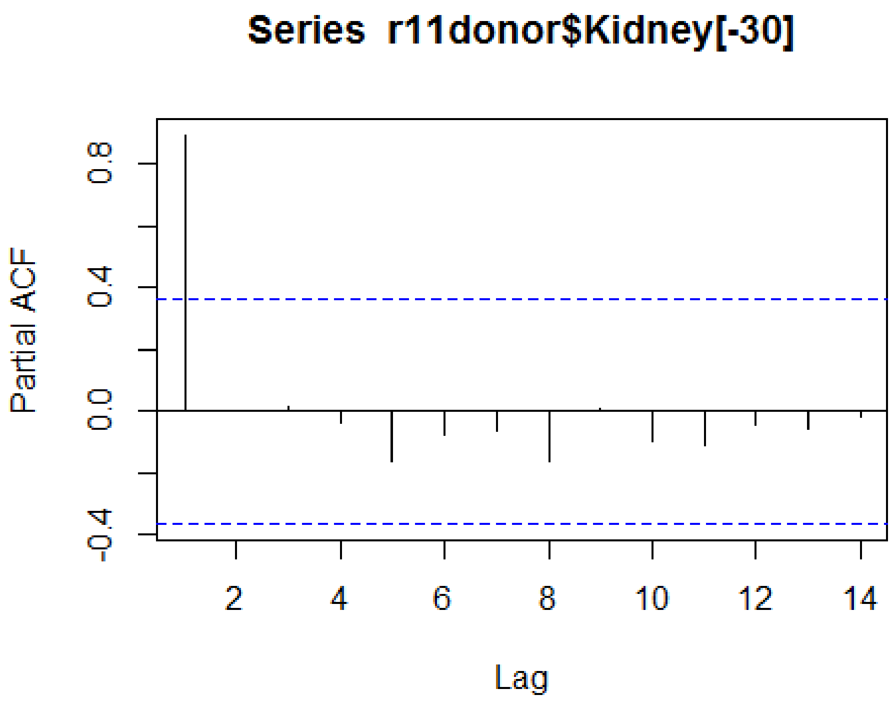
Step 5.6* (bonus) What about Duke? Repeat the above steps for Duke.

Plot an acf/pacf to estimate the model order for region 11 kidney donors - what model order do you suggest?

```
acf(r11donor$Kidney[-30])
```



```
pacf(r11donor$Kidney[-30])
```

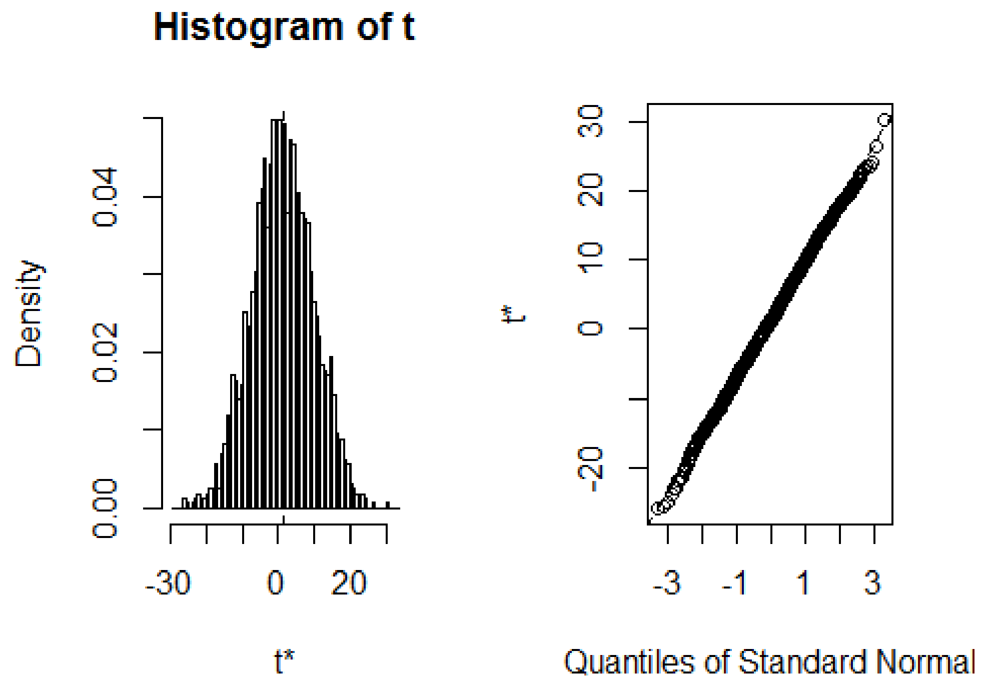


An AR(1) model is recommended because the pacf cuts off after lag 1.

```
#Use ar() to fit an ar model to region 11 kidney donors from 1988-2016
r11donor.ar<-ar(r11donor$Kidney[-30],method="yule-walker")
r11donor.ar

##
## Call:
## ar(x = r11donor$Kidney[-30], method = "yule-walker")
##
## Coefficients:
##      1
## 0.8923
##
## Order selected 1  sigma^2 estimated as  17285

r11donor.pred <- forecast(r11donor.ar,h=1)
r11donor.pred

##    Point Forecast    Lo 80    Hi 80   Lo 95    Hi 95
## 30       1289.039 1120.551 1457.526 1031.36 1546.718
```

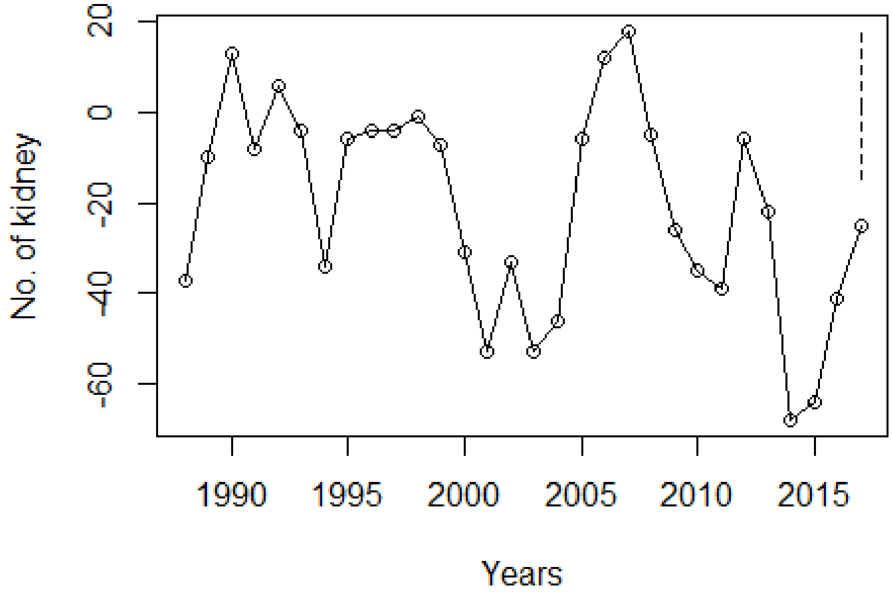
Use the predict() function

```
#   Creating the new data frame
uvaduke.nd<-data.frame(r11k=r11donor.pred$mean,uvamcv.lm.e1=uva.kidney.lm$residuals[29])

#   Predict the linear model with the time series
uvaduke.new<-predict(uvaduke.kidney.lm2,newdata=uvaduke.nd,num=5000)

## Warning: 'newdata' had 1 row but variables found have 27 rows

uvaduke.diff<-ts(uva$Kidney-duke$Kidney,1988,2017)
uvaduke.dm <- data.frame(uvaduke.diff[3:29], r11k[1:27], uva.duke.lm$residuals[1:27])
r11donor.duke.boot <- RFB(uvaduke.dm, model = uvaduke.kidney.lm2, ndata = uvaduke.nd, num=2000)

## Warning: 'newdata' had 1 row but variables found have 27 rows


#   Bootstrap plot
plot(r11donor.duke.boot, index = 1)
```



```
#   Bootstrap confidence intervals
r11donor.duke.boot.ci<-boot.ci(r11donor.duke.boot, index=1)

## Warning in boot.ci(r11donor.duke.boot, index = 1): bootstrap variances
## needed for studentized intervals

r11donor.duke.boot.ci

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = r11donor.duke.boot, index = 1)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-14.834,  18.076 )   (-14.671,  17.940 )
##
## Level     Percentile            BCa
## 95%   (-15.072,  17.539 )   (-15.035,  17.692 )
## Calculations and Intervals on Original Scale

uvaduke<-uva$Kidney-duke$Kidney
plot(uvaduke~uva$Year,xlim=c(1988,2017), type="o",xlab="Years",ylab="No. of k
idney")
segments(2017,r11donor.duke.boot$t0[1],2017, r11donor.duke.boot.ci$percent[4],
lty="dashed")
```

```
segments(2017,r11donor.duke.boot$t0[1],2017, r11donor.duke.boot.ci$percent[5],
lty="dashed")
```



Observations?

The actual difference for 2017 does not fall into the bootstrapping confidence interval.