Assignment2 – Regression

Brian Ritz

Friday, April 24, 2015

Dataset: Hollywood Movies

Objective:

1 Plot the independent variables X2, X3 X4 together in a single plot - what do you conclude in terms of relationship between them? 2 Scatter plot among the variables 3 ADF of the independent variables 4 Regression output - R2 and any other metric you want to mention 5 Regression coefficients 6 p-value of the coefficients 7 What can you comment about multicollinearity? 8 plot the ACF of the residuals

1 Plot the independent variables X2, X3, and X4 together in a single plot - what do you conclude in terms of relationship between them?

```
library(tidyr) #data munging
```

```
## Warning: package 'tidyr' was built under R version 3.1.3
```

```
library(dplyr) #data munging
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
```

```
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2) #plotting
library(tseries) #adf.test
```

```
## Warning: package 'tseries' was built under R version 3.1.3
```

```
library(usdm) # vif
```

```
## Warning: package 'usdm' was built under R version 3.1.3
```

```
## Loading required package: raster
```

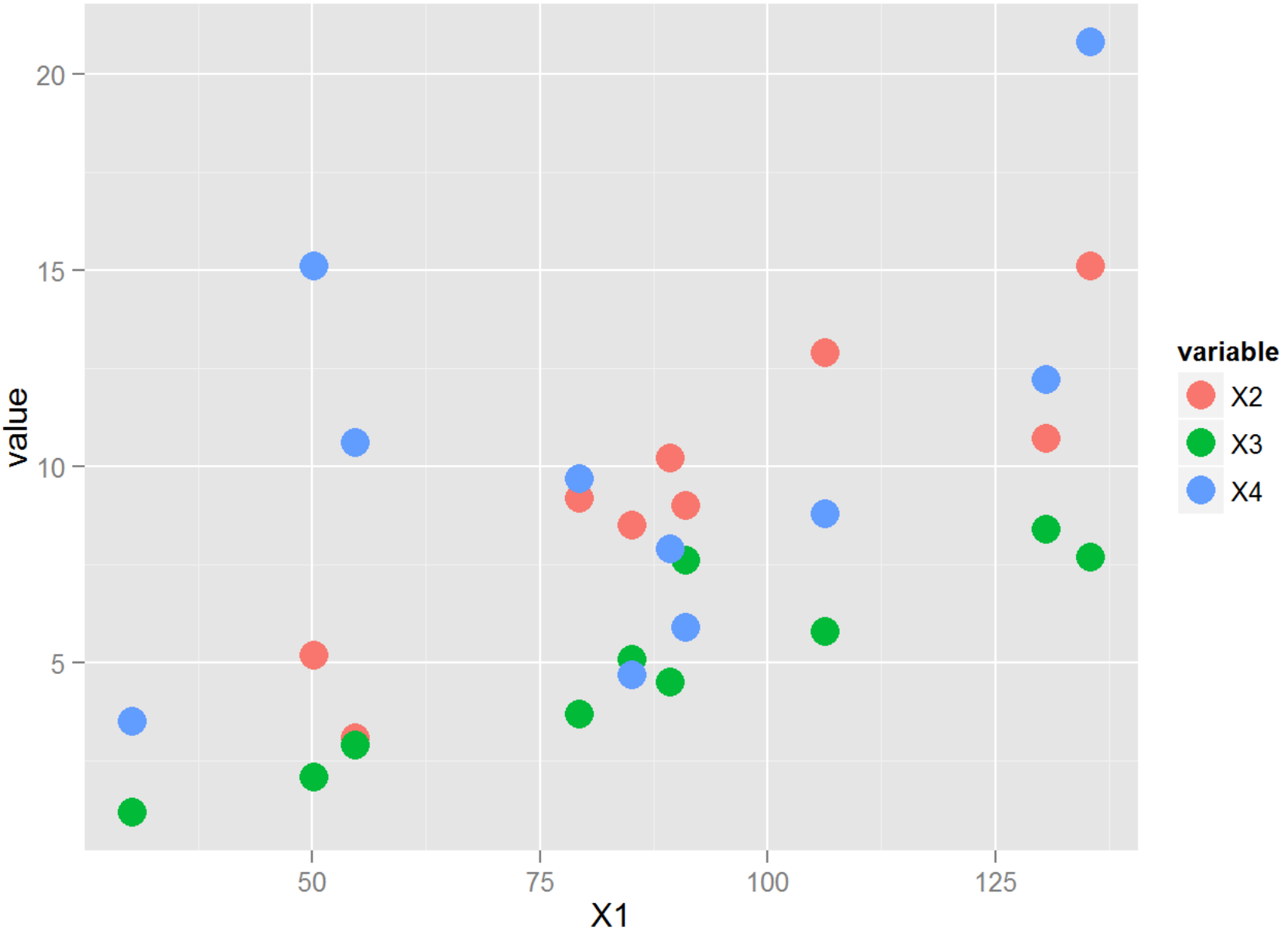
```
## Warning: package 'raster' was built under R version 3.1.3
```

```
## Loading required package: sp
```

```
## Warning: package 'sp' was built under R version 3.1.3
```

```
##
## Attaching package: 'raster'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

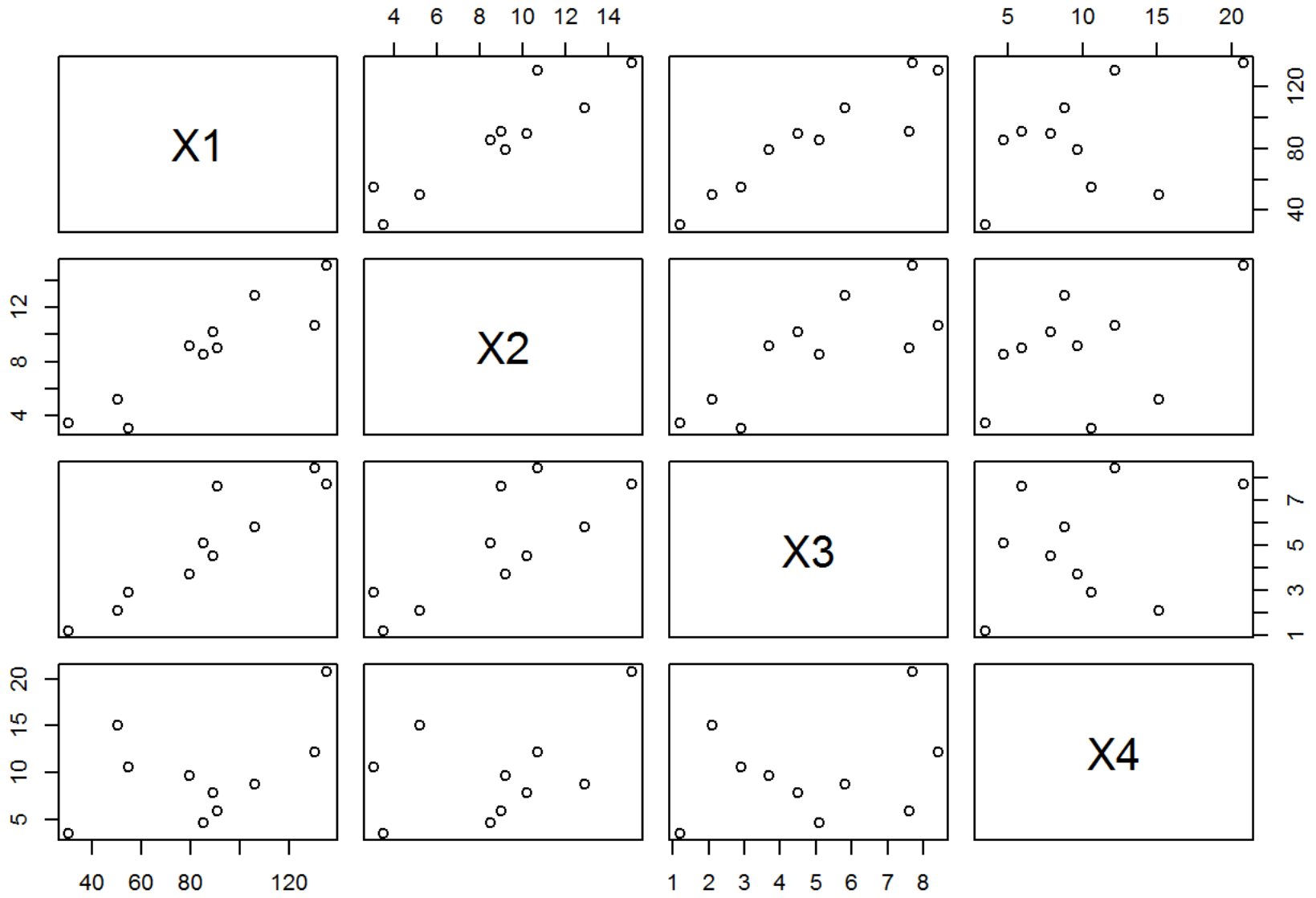
```
movie.dataset <- read.csv("mlr04.csv", header=T)
gather(movie.dataset, "variable","value", 2:4) %>%
  ggplot(aes(x=X1, y=value, col=variable)) + geom_point(size=5)
```



WHAT DO I SEE? All variables have a positive relation ship with X1. X4 has the most variance, while X2 and X3 have less variance but similar variance. X2 and X3 are more highly correlated with eachother than X4. All variables show a positive relationship to X1.

2 Scatter plot among the variables

```
plot(movie.dataset)
```



WHAT DO I SEE? X1, X2, and X3 all have positive relationships with each other. But X4 does not have as apparent positive relationship with either X1, X2, or X3.

3 ADF of the independent variables

```
# X2
adf.test(movie.dataset$X2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  movie.dataset$X2
## Dickey-Fuller = -0.8326, Lag order = 2, p-value = 0.9452
## alternative hypothesis: stationary
```

```
#X3
adf.test(movie.dataset$X3)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  movie.dataset$X3
## Dickey-Fuller = -1.454, Lag order = 2, p-value = 0.7804
## alternative hypothesis: stationary
```

```
#X4
adf.test(movie.dataset$X4)
```

```
## Warning in adf.test(movie.dataset$X4): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  movie.dataset$X4
## Dickey-Fuller = -10.7343, Lag order = 2, p-value = 0.01
## alternative hypothesis: stationary
```

WHAT DO I SEE? The null hypothesis for the Augmented Dickey Fuller test is that the series is not stationary (that there is a unit root of the time series). We reject the null hypothesis for X2 and X3, but fail to reject it at X4. This means that we have reason to believe X4 is stationary, while X2 and X3 are not stationary.

4 Regression output - R2 and any other metric you want to mention

```
summary(model <- lm(X1~X2+X3+X4, data=movie.dataset, y=T, x=T))
```

```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4, data = movie.dataset, x = T,
##     y = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4384  -3.1695   0.8499   3.5134   9.6207
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6760     6.7602   1.135   0.2995
## X2            3.6616     1.1178   3.276   0.0169 *
## X3            7.6211     1.6573   4.598   0.0037 **
## X4            0.8285     0.5394   1.536   0.1754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.541 on 6 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9502
## F-statistic: 58.22 on 3 and 6 DF,  p-value: 7.913e-05
```

The R-squared is .96, which is relatively high. You can interpret this as the model explains 96% of the variation in X1 through variation in the independent variables. The F-statistic has a small p-value, indicating that the model fits the data well.

5 Regression coefficients

```
model$coefficients
```

```
## (Intercept)          X2          X3          X4
##   7.6760285   3.6616040   7.6210513   0.8284681
```

The coefficients can be interpreted in the following way: for a one unit increase in the independent variable, the dependent variable will change by the coefficient on the independent variable. For example, for a one unit increase in X2, X1 will increase by 3.66; for a one unit increase in X3, X1 will increase 7.62; and for a one unit increase in X4, X1 will increase by

.82.

6 p-value of the coefficients

```
a<-summary(model)
a$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 7.6760285  6.7602276 1.135469 0.299491477
## X2          3.6616040  1.1177514 3.275866 0.016909724
## X3          7.6210513  1.6573172 4.598426 0.003698129
## X4          0.8284681  0.5393591 1.536023 0.175439839
```

The coefficients fro X2 and X3 are significant at alpha=.05, wile X4 is only significant at alpha=.2. The coefficient for the intercept has a p-value of .29. You can interpret these p-values as the probability that a process with actual mean coefficient = 0 would produce these data.

7 What can you comment about multicollinearity?

```
cor(movie.dataset[2:4])
```

```
##           X2        X3        X4
## X2 1.0000000 0.7899575 0.4291329
## X3 0.7899575 1.0000000 0.2987612
## X4 0.4291329 0.2987612 1.0000000
```

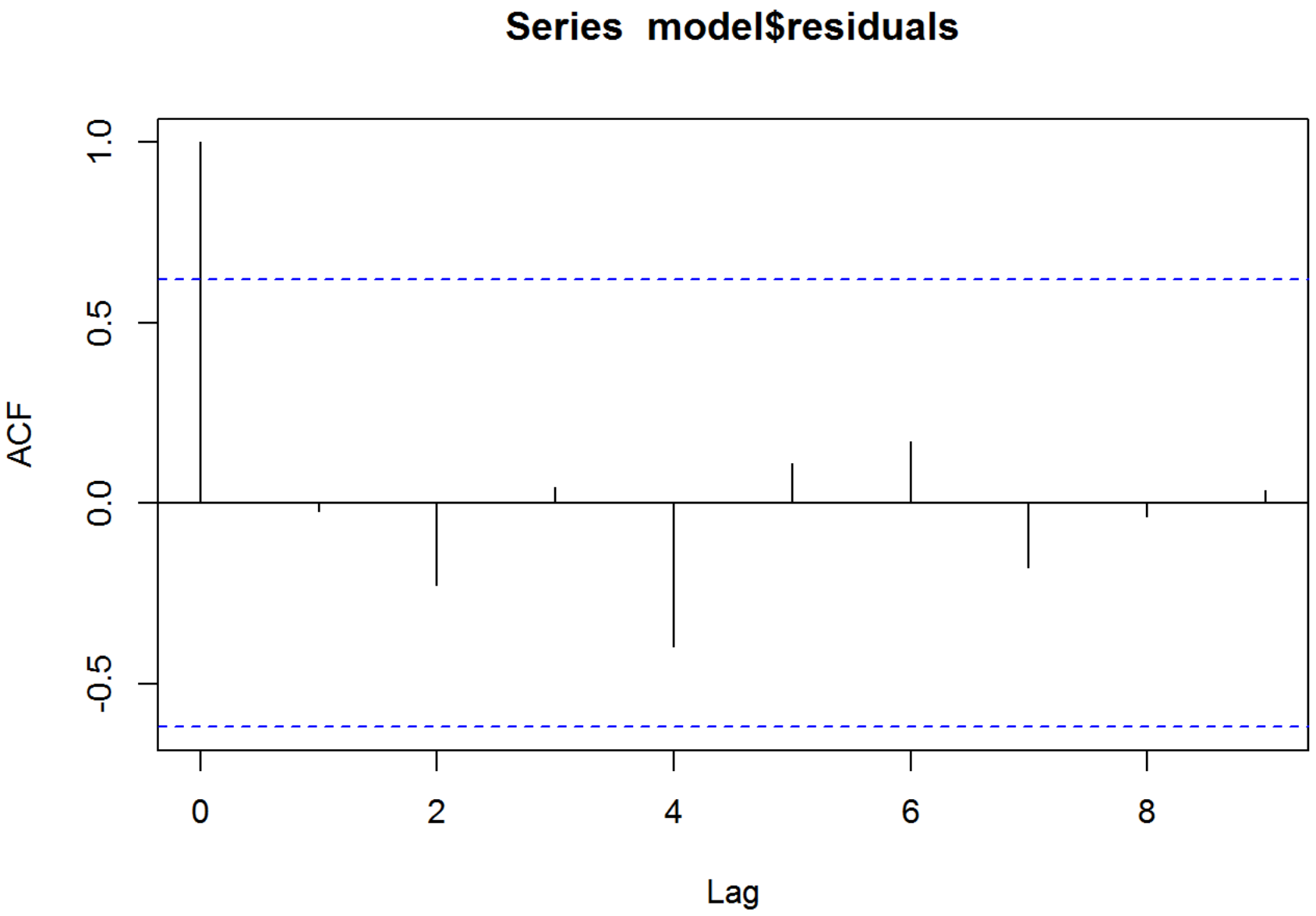
```
vif(movie.dataset[2:4])
```

```
##   Variables      VIF
## 1        X2 2.984943
## 2        X3 2.673920
## 3        X4 1.232227
```

Multi-collinearity is present because X2 and X3 are highly correlated with eachother. The vif test indicates that there is a lot of inflation in variance of the X2 and X3 variables due to multi-collinearity.

8 plot the ACF of the residuals

```
acf(model$residuals)
```



Testing the auto-correlation of residuals. There are no lags that are outside the confidence band for acf=0, so we can conclude that auto-correlation is not present.