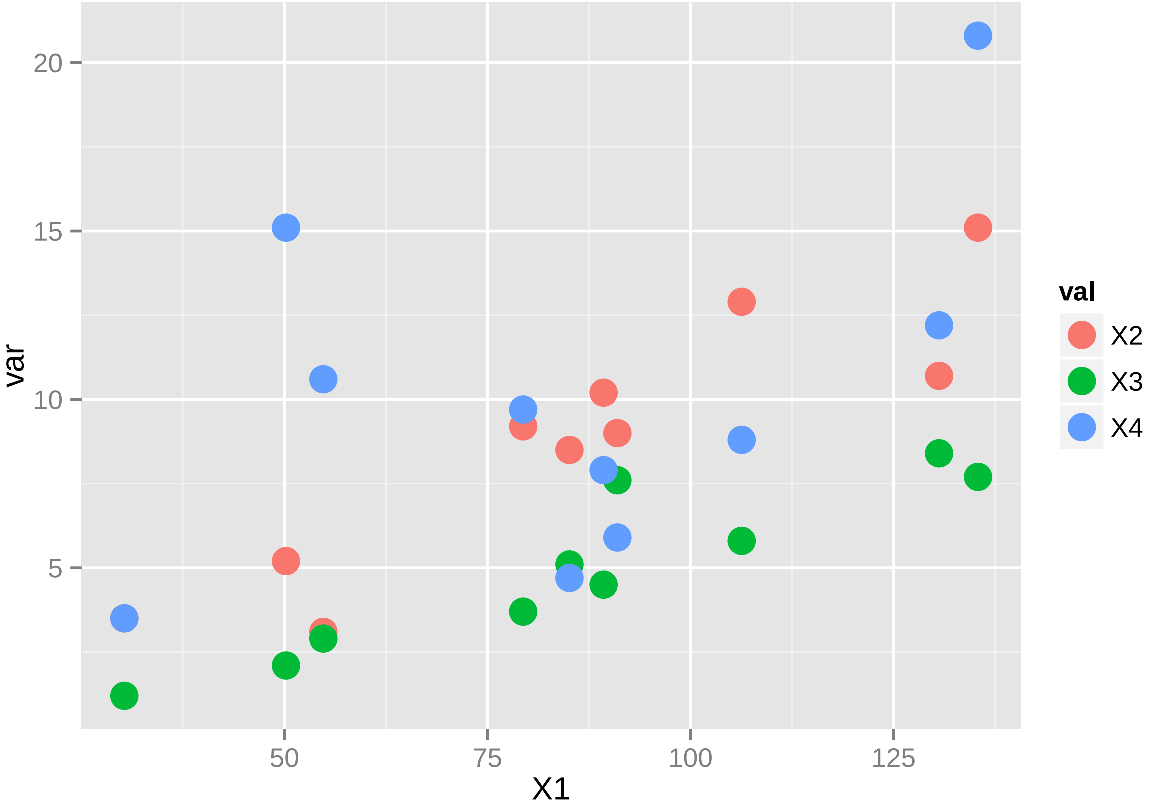
TS_hw2_markdown

Patrick Kelly

Thursday, April 30, 2015

Assignment #2 - Regression

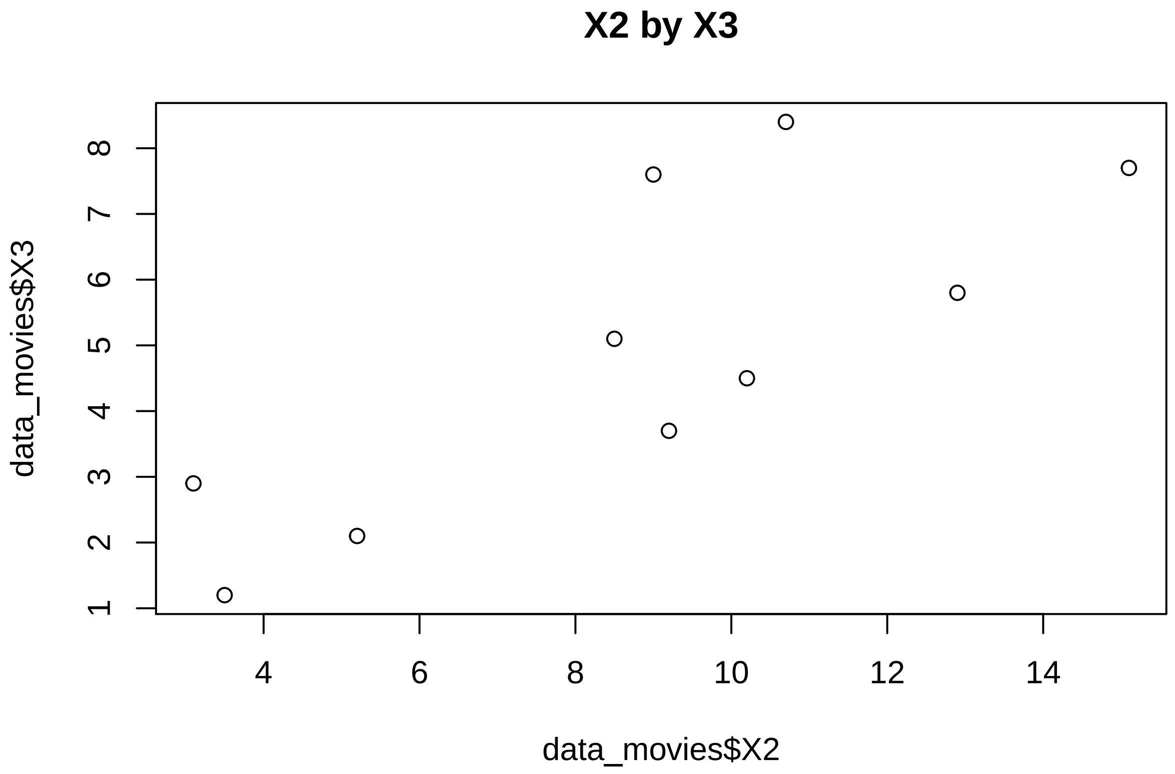
```
library(ggplot2)
library(tseries) # for adf.test() fn
library(car) # for vif() fn

#bring in dataset
data_movies <- read.csv("~/R/UChicago/Time_Series/hollywood_movies.csv")

# 1. Plot the independent variables X2, X3 X4 together in a single plot - what do you conclude in terms
d1 <- cbind(data_movies[,c(1,2)],val=rep("X2",10))
colnames(d1) <- c("X1","var","val")
d2 <- cbind(data_movies[,c(1,3)],val=rep("X3",10))
colnames(d2) <- c("X1","var","val")
d3 <- cbind(data_movies[,c(1,4)],val=rep("X4",10))
colnames(d3) <- c("X1","var","val")

data_long <- rbind(d1,d2,d3)

ggplot(data=data_long,aes(x=X1,y=var,col=val)) + geom_point(size=5)
```



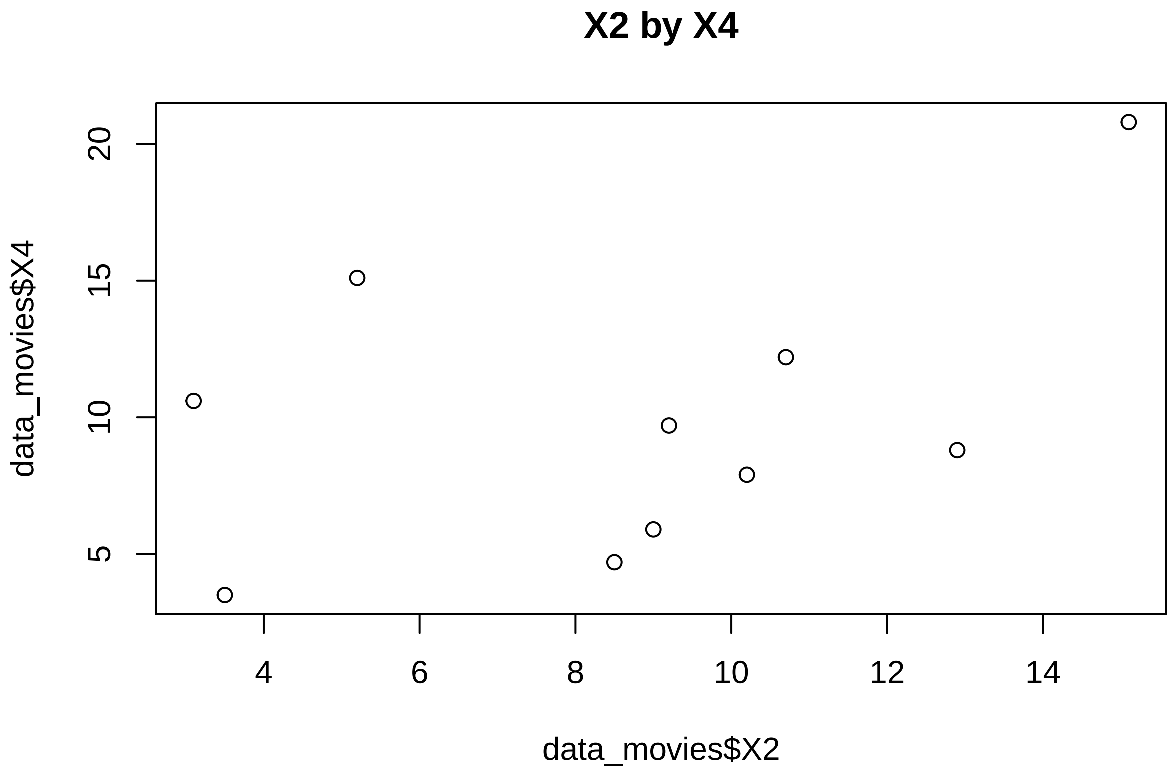
What do I see regarding the relationship between them?

The do appear to have a generally positive relationship / correlation. X2, X3 and X4 all seem to have a positive linear relationship with X1. Also X2 and X3 appear to have a strong correlation than X4 has with either of the two.

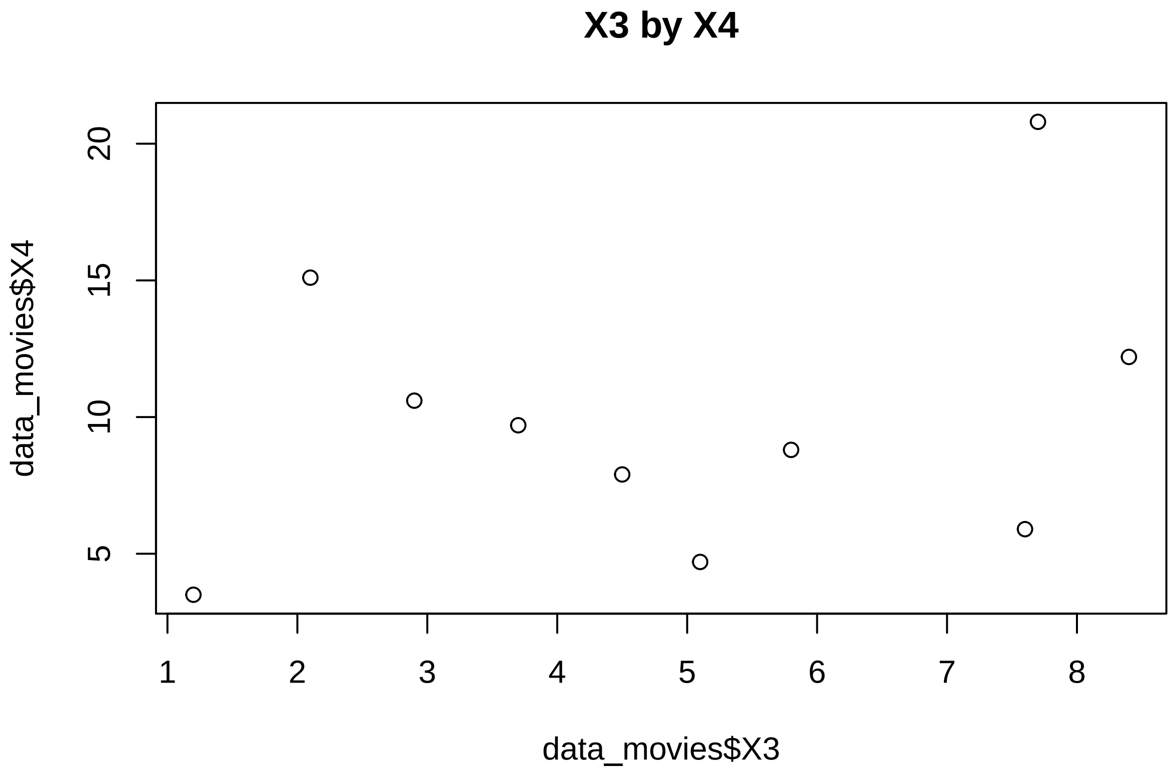
```r
# 2. Scatter plot among the variables
plot(data_movies$X2,data_movies$X3, main="X2 by X3")
```



```
plot(data_movies$X2,data_movies$X4, main="X2 by X4")
```



```
plot(data_movies$X3,data_movies$X4, main="X3 by X4")
```



The above scatterplots support my previous comments.

```
# 3. ADF of the independent variables
adf.test(data_movies$X2)
```

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

```
adf.test(data_movies$X3)
```

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

```
adf.test(data_movies$X4)
```

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

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

Of the three variables, only X4 rejected the null hypothesis of non-stationarity. So we can only conclude that X4 is stationary but cannot make conclusions about either X2 or X3.

```
# 4. Regression output - R2 and any other metric you want to mention
lm1 <- lm(X1 ~ X2 + X3 + X4, data=data_movies)
summ_lm1 <- summary(lm1)
summ_lm1
```

```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4, data = data_movies)
##
## 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 linear model above has a high (good) R2 value of 0.97. Also the overall p-value of the model is significant - as expected with the high R2.

```
# 5. Regression coefficients
summ_lm1$coefficients[,1]
```

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

```
# 6. p-value of the coefficients
summ_lm1$coefficients[,4]
```

```
## (Intercept)          X2          X3          X4
## 0.299491477 0.016909724 0.003698129 0.175439839
```

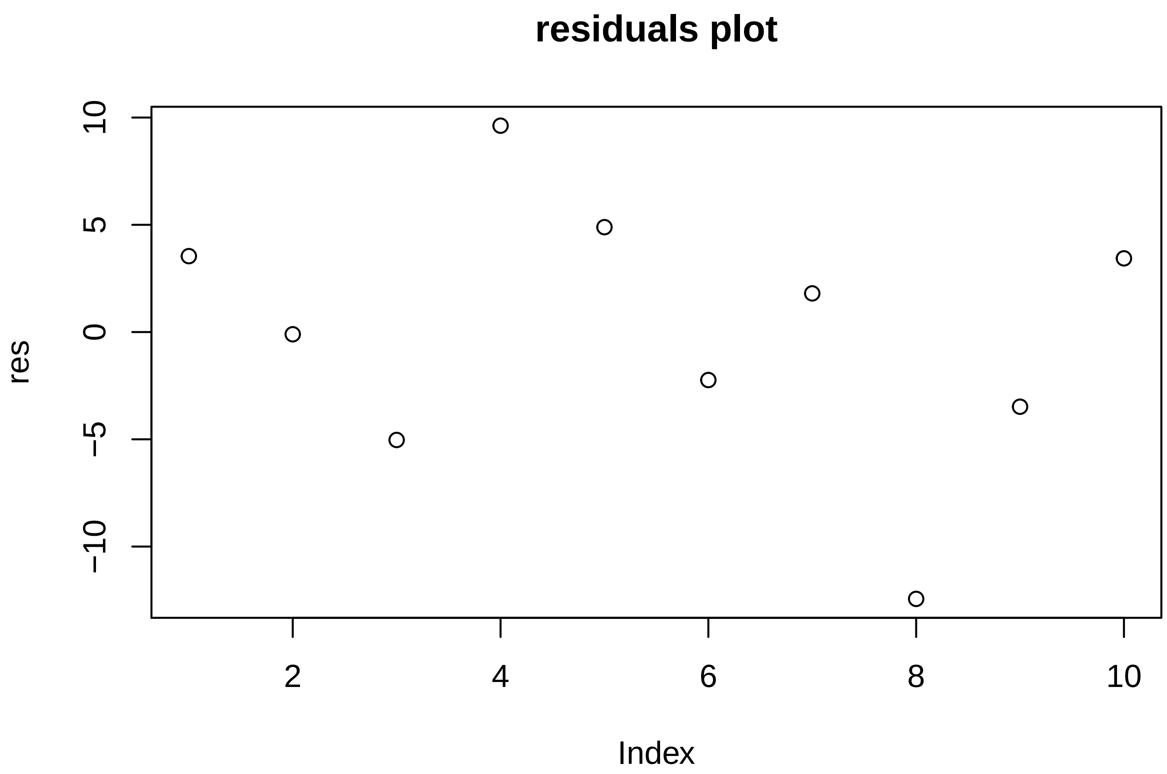
```
# 7. What can you comment about multicollinearity?
vif(lm1)
```

```
##       X2       X3       X4
## 2.984943 2.673920 1.232227
```

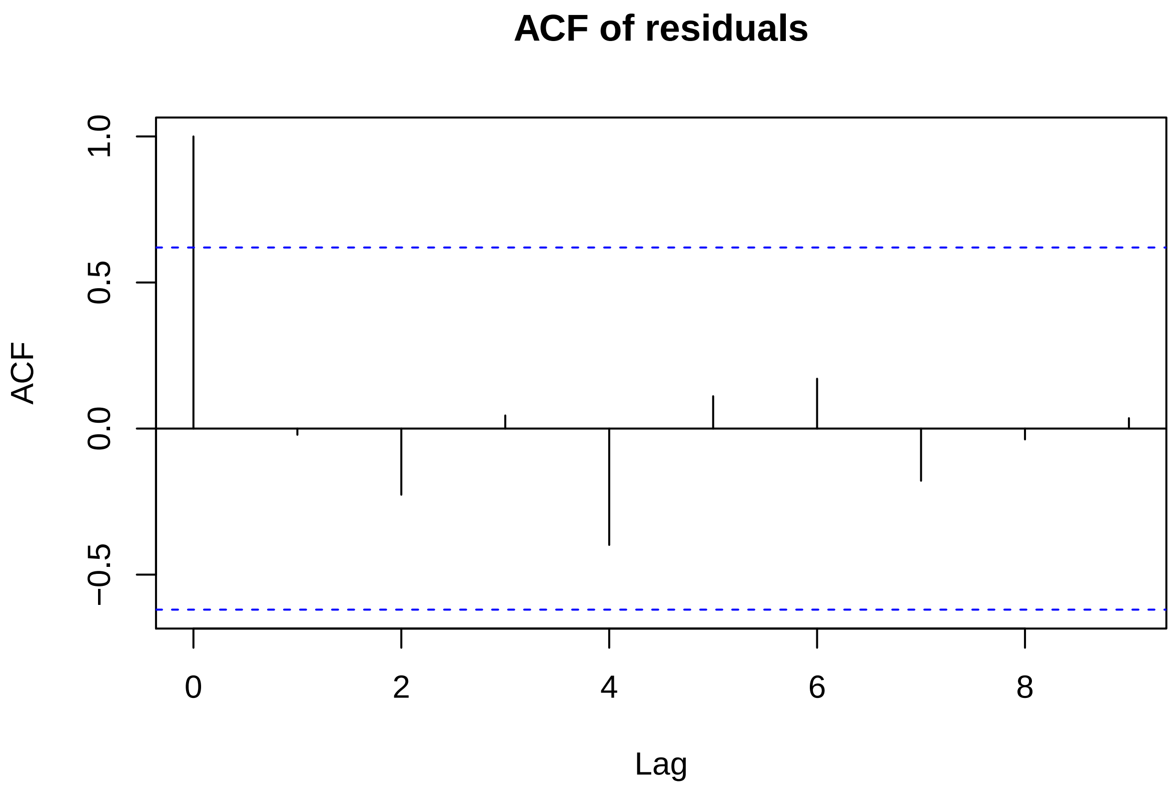
As we did see some evidence from the initial charts that multicolinearity could be a potential issue, the values outputted from the `vif()` fn - which produces the variance inflation factor - are mildly concerning.

The X4 vif value is safetly low (1.23) and does not raise concerns about multicolinearity, but the values for X2 and X3 are approaching 3. Although the X2 and X3 values are not totally alarming they may warrant further investigation, particularly if you were to collect more data.

```
# 8. plot the ACF of the residuals
res <- residuals(lm1)
plot(res, main="residuals plot")
```



```
acf(res, main="ACF of residuals")
```



The plot of the ACF of the residuals indicates that there are not any autocorrelation concerns in the residuals.