Economic Journal Subscription

Supervised By:

Prof. Douglas Jones

Submitted By:

Ankit Tyagi(RUID: 175000523)

Dhananjay Shukla(RUID: 173009014)

Harshal Thummar(RUID: 173008545)

Siddhant Naik(RUID: 175000499)

INTRODUCTION

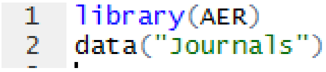
Journals are regular publications that contain articles on current research. Usually the author of a journal article will have carried out some primary research; perhaps they have carried out a case study where they have interviewed or surveyed participants, or they may have carried out a clinical trial. The purpose of the article is for the authors to outline their research and present their findings and conclusions. This in-depth information is not likely to be found in books and this is why journal articles are important in finding supporting evidence for assignments. Researchers routinely submit papers to their journals of choice and ask libraries to subscribe to those journals. But those same researchers often have little idea of the institutional subscription price of those journals, how subscribing to them may impact the ability of their institution to retain subscriptions to other titles, or the impact on maintaining an equitable distribution of library budget dollars across subject areas. Researchers who are informed about the prices and value of the journals in their subject areas and who consider subscription prices, rates of increase, and various measures of value in determining which journals they will submit papers to, review for, or be involved with editorially can positively impact the ability of their institution to maximize access to journal content across all subject areas. While keeping these things in mind about the kind of data we have we can help the publishers from an economic point of view also the consumers or to be more specific the readers it will greatly help the users to know which economic journals have higher subscription and which has good research work in it. With keeping these kind of objectives we have performed our analysis using various models. We will talk about the various models used further in the report.

PROJECT OBJECTIVE:

Our main aim of this project is to predict the number of subscription of a particular Economic Journal. For the analysis of the same we have used R Studio and we have used different models to predict the subscription of the economic journals. We also decide the number of variables on which the subscription depends. This helps greatly the readers as well as business point of view for the publishers.

DATA SOURCE:

We have used the data source provided by the R library(“AER”) package

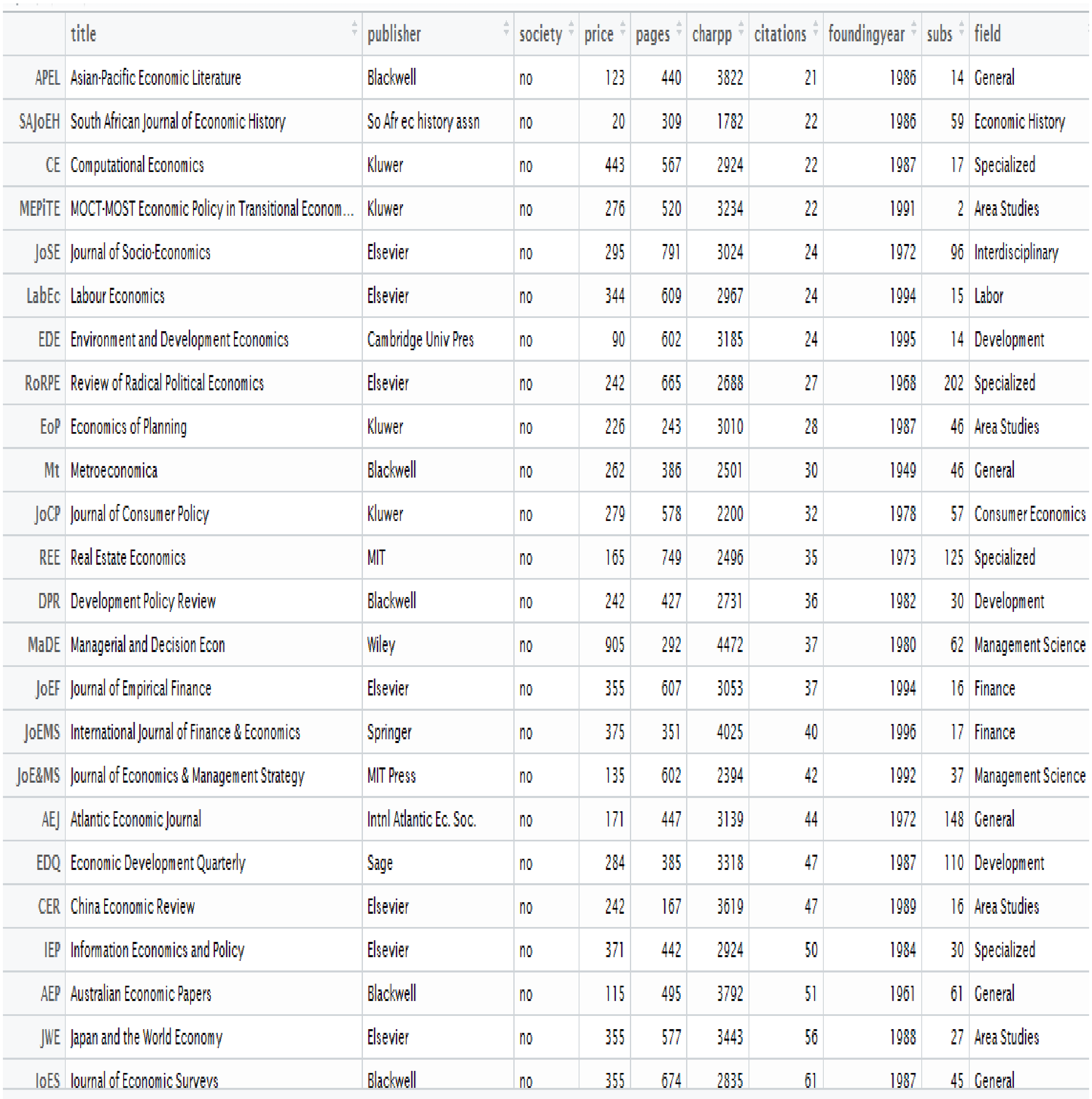


```
library(AER)
data("Journals")
```

TOOLS:

For our Project, we have used the R Studio and R

DATA SET:



	title	publisher	society	price	pages	charpp	citations	foundingyear	subs	field
APEL	Asian-Pacific Economic Literature	Blackwell	no	123	440	3822	21	1986	14	General
SAJoEH	South African Journal of Economic History	So Afr ec history assn	no	20	309	1782	22	1986	59	Economic History
CE	Computational Economics	Kluwer	no	443	567	2924	22	1987	17	Specialized
MEPiTE	MOCT-MOST Economic Policy in Transitional Econom...	Kluwer	no	276	520	3234	22	1991	2	Area Studies
JoSE	Journal of Socio-Economics	Elsevier	no	295	791	3024	24	1972	96	Interdisciplinary
LabEc	Labour Economics	Elsevier	no	344	609	2967	24	1994	15	Labor
EDE	Environment and Development Economics	Cambridge Univ Pres	no	90	602	3185	24	1995	14	Development
RoRPE	Review of Radical Political Economics	Elsevier	no	242	665	2688	27	1968	202	Specialized
EoP	Economics of Planning	Kluwer	no	226	243	3010	28	1987	46	Area Studies
Mt	Metroeconomica	Blackwell	no	262	386	2501	30	1949	46	General
JoCP	Journal of Consumer Policy	Kluwer	no	279	578	2200	32	1978	57	Consumer Economics
REE	Real Estate Economics	MIT	no	165	749	2496	35	1973	125	Specialized
DPR	Development Policy Review	Blackwell	no	242	427	2731	36	1982	30	Development
MaDE	Managerial and Decision Econ	Wiley	no	905	292	4472	37	1980	62	Management Science
JoEF	Journal of Empirical Finance	Elsevier	no	355	607	3053	37	1994	16	Finance
JoEMS	International Journal of Finance & Economics	Springer	no	375	351	4025	40	1996	17	Finance
JoE&MS	Journal of Economics & Management Strategy	MIT Press	no	135	602	2394	42	1992	37	Management Science
AEJ	Atlantic Economic Journal	Intnl Atlantic Ec. Soc.	no	171	447	3139	44	1972	148	General
EDQ	Economic Development Quarterly	Sage	no	284	385	3318	47	1987	110	Development
CER	China Economic Review	Elsevier	no	242	167	3619	47	1989	16	Area Studies
IEP	Information Economics and Policy	Elsevier	no	371	442	2924	50	1984	30	Specialized
AEP	Australian Economic Papers	Blackwell	no	115	495	3792	51	1961	61	General
JWE	Japan and the World Economy	Elsevier	no	355	577	3443	56	1988	27	Area Studies
JoES	Journal of Economic Surveys	Blackwell	no	355	674	2835	61	1987	45	General

DATA DESCRIPTION:

Variable	Description
title	Journal title
publisher	factor with publisher name.
society	factor. Is the journal published by a scholarly society?
price	Library subscription price.
pages	Number of pages.
charpp	Characters per page.
citations	Total number of citations.
foundingyear	Year journal was founded.
subs	Number of library subscriptions.
field	factor with field description.

R-CODE:

```
library(AER)

library("faraway")

library("car")
library("ggplot2")
library("gridExtra")
library("scatterplot3d")

library("rgl")

data("Journals")
library("ggplot2")
library(GGally)

# View(Journals)


data("Journals")


summary(Journals)

##     title                          publisher  society       price
##  Length:180         Elsevier             :42   no :164   Min.   :  20.0
##  Class :character   Blackwell            :26   yes: 16   1st Qu.: 134.5
##  Mode  :character   Kluwer               :16             Median : 282.0
##                     Springer             :10             Mean   : 417.7
##                     Academic Press       : 9             3rd Qu.: 540.8
##                     Univ of Chicago Press: 7             Max.   :2120.0
##                     (Other)              :70
##      pages            charpp       citations        foundingyear
##  Min.   : 167.0   Min.   :1782   Min.   :  21.00   Min.   :1844
##  1st Qu.: 548.8   1st Qu.:2715   1st Qu.:  97.75   1st Qu.:1963
##  Median : 693.0   Median :3010   Median : 262.50   Median :1973
##  Mean   : 827.7   Mean   :3233   Mean   : 647.06   Mean   :1967
##  3rd Qu.: 974.2   3rd Qu.:3477   3rd Qu.: 656.00   3rd Qu.:1982
##  Max.   :2632.0   Max.   :6859   Max.   :8999.00   Max.   :1996
##
##       subs                       field
##  Min.   :   2.0   General             :40
##  1st Qu.:  52.0   Specialized         :14
##  Median : 122.5   Public Finance      :12
##  Mean   : 196.9   Development         :11
##  3rd Qu.: 268.2   Finance             :11
##  Max.   :1098.0   Urban and Regional: 8
##                   (Other)             :84
```

MODEL CREATION:

```
g<- lm(subs~.-field-title-publisher,data = Journals)
g1 <- step(g)
```

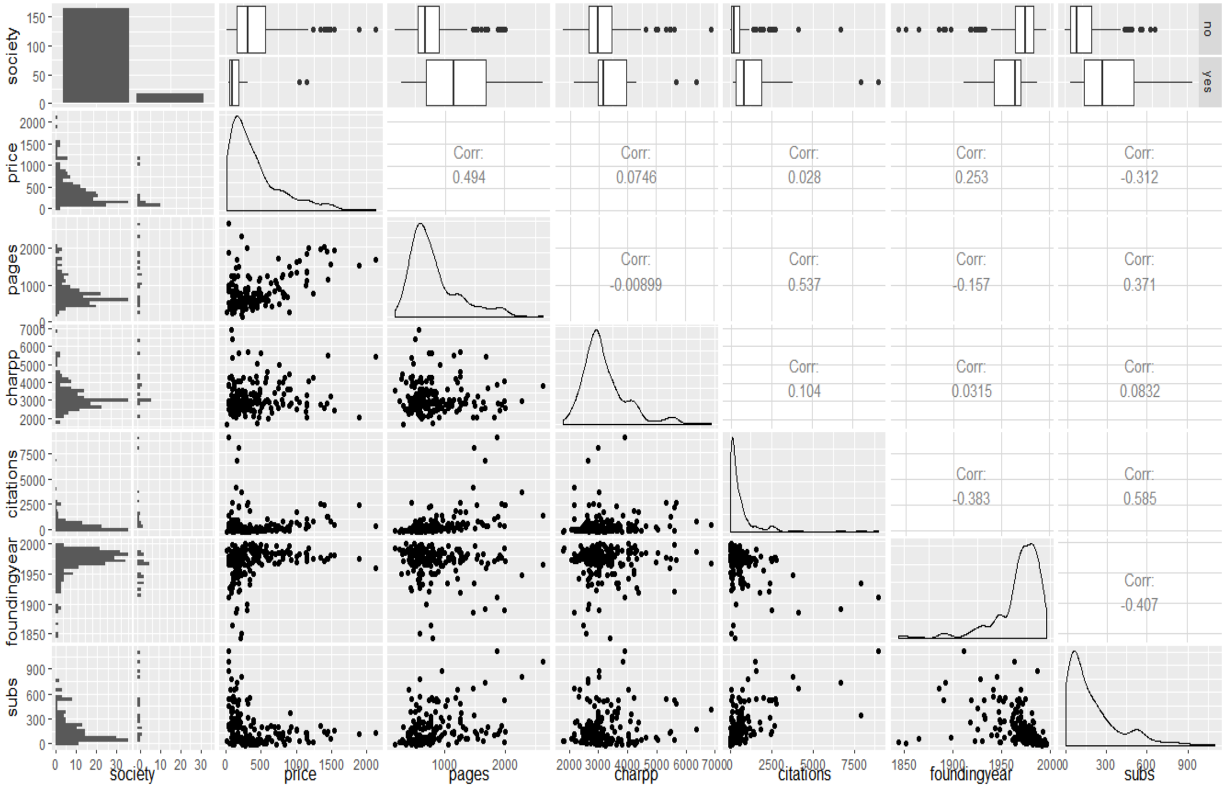
```
## Start:  AIC=1780.96
## subs ~ (title + publisher + society + price + pages + charpp +
##     citations + foundingyear + field) - field - title - publisher
##
##                Df Sum of Sq     RSS    AIC
## - society       1       230 3300057 1779.0
## <none>                      3299828 1781.0
## - foundingyear  1     41599 3341427 1781.2
## - charpp        1     63275 3363103 1782.4
## - citations     1    414434 3714262 1800.2
## - pages         1    636096 3935924 1810.7
## - price         1   1122475 4422303 1831.7
##
## Step:  AIC=1778.97
## subs ~ price + pages + charpp + citations + foundingyear
##
##                Df Sum of Sq     RSS    AIC
## <none>                      3300057 1779.0
## - foundingyear  1     41371 3341428 1779.2
## - charpp        1     67108 3367165 1780.6
## - citations     1    425311 3725369 1798.8
## - pages         1    709989 4010046 1812.0
## - price         1   1269262 4569319 1835.5
```

```
summary(g)
```

```
##
## Call:
## lm(formula = subs ~ . - field - title - publisher, data = Journals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -473.18  -81.60  -31.83   69.58  505.38
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1360.89705  904.70271   1.504   0.1343
## societyyes      4.57756   41.71326   0.110   0.9127
## price          -0.27930    0.03641  -7.671 1.19e-12 ***
## pages           0.21010    0.03638   5.775 3.50e-08 ***
## charpp          0.02383    0.01308   1.821   0.0703 .
## citations       0.05425    0.01164   4.661 6.25e-06 ***
## foundingyear   -0.67812    0.45919  -1.477   0.1415
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

## Residual standard error: 138.1 on 173 degrees of freedom


## Multiple R-squared:  0.5593, Adjusted R-squared:  0.544
## F-statistic: 36.59 on 6 and 173 DF,  p-value: < 2.2e-16
```

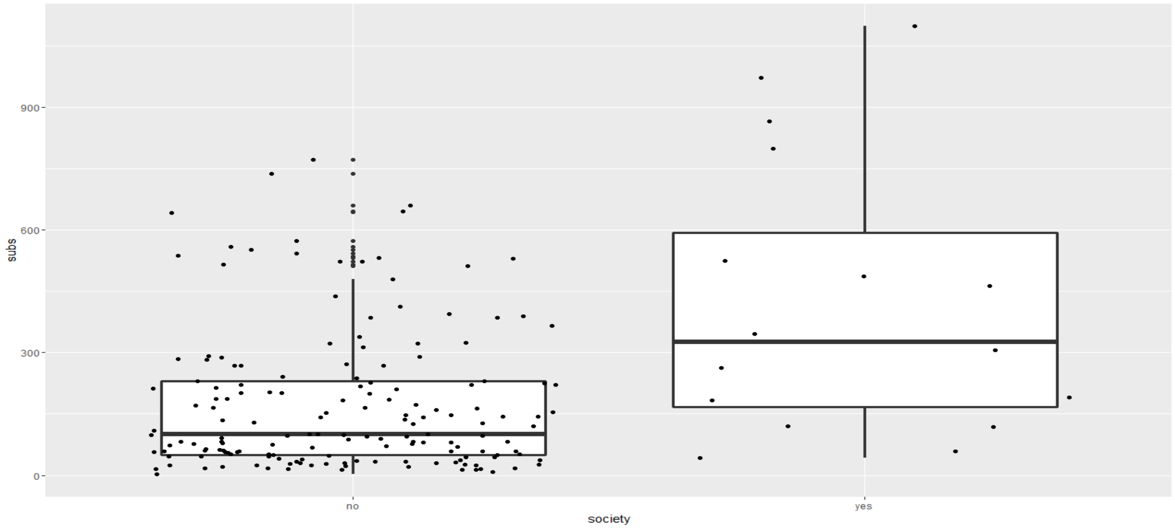
```
ggpairs(Journals,c(3:9))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From this plot, we make a conclusion that there are good predictors present and there are outliers present. We make a conclusion that citations and pages are good predictors.

```
boxplot(Journals$subs,Journals$price, main = "Box Plot")
```



The above obtained plot gives us an idea about the subscription that when there are journals form a scholarly society then the number of subscriptions are more and when there are journals from non-scholarly society then the subscriptions are less.

In this step, we have modified certain instances so that we can use them for our models that we are going to create for prediction and we can find the factors affecting subscription.

```
journals <- Journals[, c("subs", "price")]
journals$citeprice <- Journals$price/Journals$citations
journals$age <- 2000 - Journals$foundingyear
journals$chars <- Journals$charpp*Journals$pages/10^6
summary(journals)
```

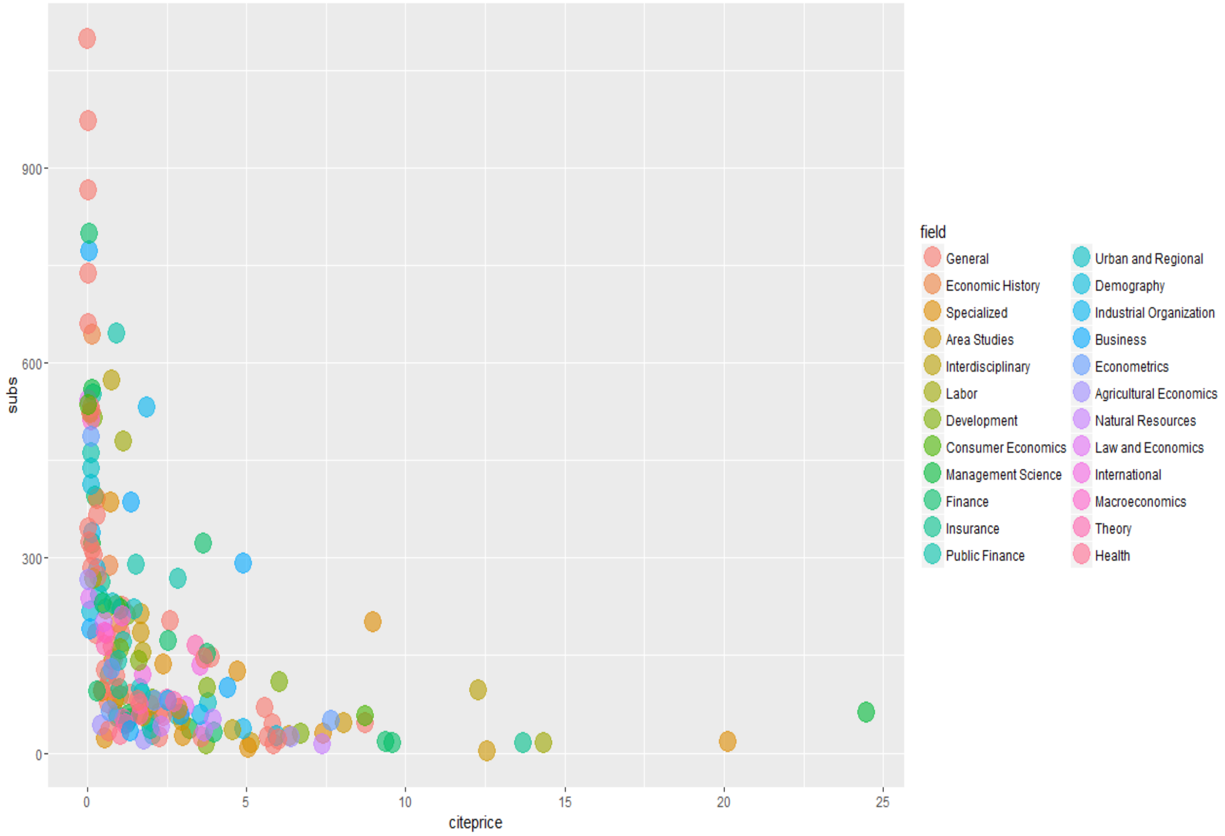
```
##       subs            price          citeprice              age
##  Min.   :   2.0   Min.   :  20.0   Min.   : 0.005223   Min.   :  4.00
##  1st Qu.:  52.0   1st Qu.: 134.5   1st Qu.: 0.464495   1st Qu.: 17.75
##  Median : 122.5   Median : 282.0   Median : 1.320513   Median : 27.00
##  Mean   : 196.9   Mean   : 417.7   Mean   : 2.548455   Mean   : 33.09
##  3rd Qu.: 268.2   3rd Qu.: 540.8   3rd Qu.: 3.440171   3rd Qu.: 37.25
##  Max.   :1098.0   Max.   :2120.0   Max.   :24.459459   Max.   :156.00
##      chars
##  Min.   : 0.5506
##  1st Qu.: 1.6577
##  Median : 2.1877
##  Mean   : 2.6727
##  3rd Qu.: 3.1829
##  Max.   :10.1279
```

After running this R Code, we can find that a new variable citeprice has been introduced which is price per citation. We have introduced this variable since the significance here is that it helps us to determine the high profit. Higher the citations lower the price and lower the citations higher price.

Also, we have introduced a new variable age, this variable determines the age of the journal we attain this variable by taking the value (2000-foundingyear). Here 2000 is the year of the Journals dataset obtained.

The other new variable introduced is chars it the character per page multiplied by the number of pages. Since this value would be very large we divide it by a million so that we get a realistic and a process able value.

```
s = ggplot(data = journals, aes(x = citeprice, y = subs, color = Journals$field ))
s + geom_point(size = I(6), alpha = I(0.6))
```

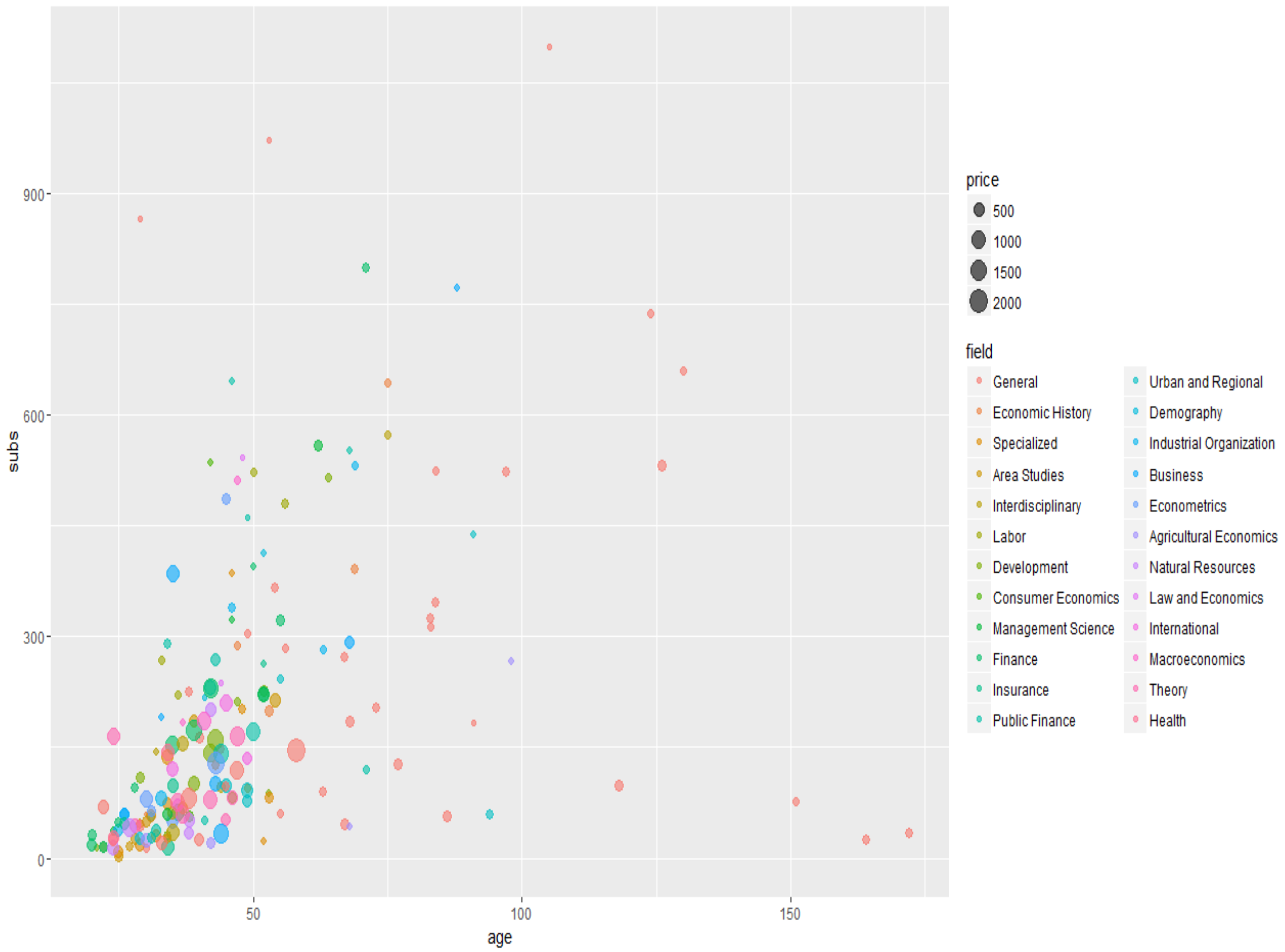


Here, we observe that as the citeprice is less the subscription is more. As the citeprice goes on increasing the number of subscription is getting lower and lower.

We have plotted the legend along the plotted graph. Various colors are used to represent various fields on the scatterplot as shown below.

Here we used the alpha function so that we can see over lapping of plotted points by making the color of plotted points transparent.

```
age1 = ggplot(data = journals, aes(x = age, y = subs, color = Journals$field, size = Journals$price))
age1 + geom_point(alpha = I(0.6))
```



Here, we can observe that till the 50-point mark of age there are lot of subscription after that the graph is scattered randomly.

Here the size of the dots indicates the price of the subscription in various fields ranging from 500 to 2000 as shown in the graph below.

The shown legend along the plot shows the various fields of subscription, age of the subscription, price of the subscription with respect to the various fields.

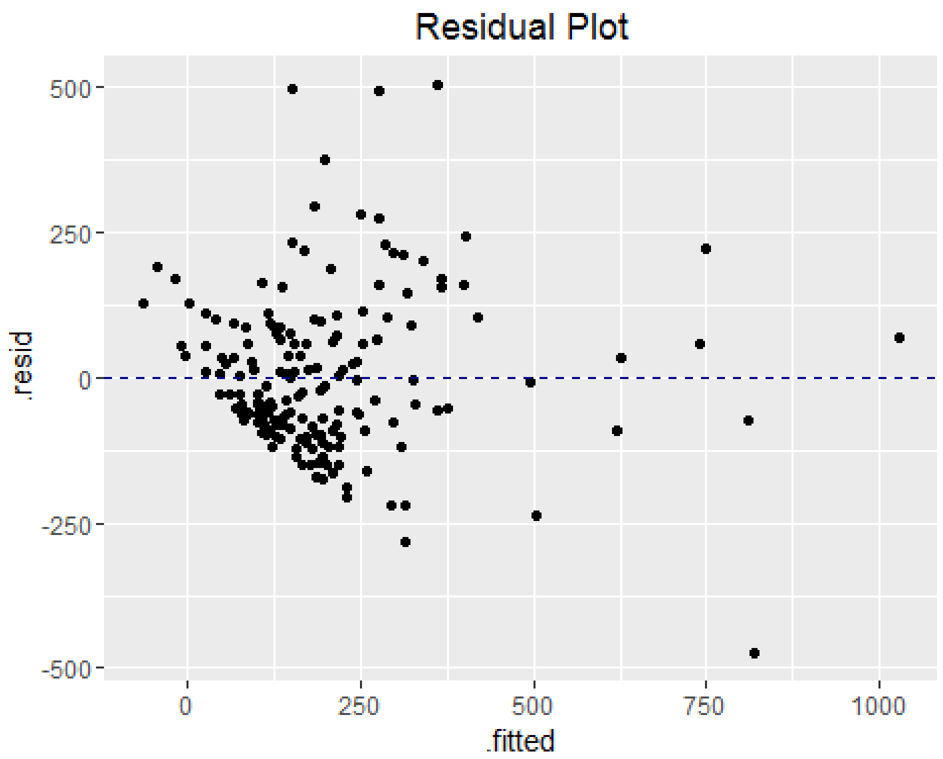
```
g <- lm(subs~.-field-title-publisher,data = Journals)
g1 <- step(g)

## Start:  AIC=1780.96
## subs ~ (title + publisher + society + price + pages + charpp +
##     citations + foundingyear + field) - field - title - publisher
```

```
##                Df Sum of Sq     RSS    AIC
## - society       1       230 3300057 1779.0
## <none>                      3299828 1781.0
## - foundingyear  1     41599 3341427 1781.2
## - charpp        1     63275 3363103 1782.4
## - citations     1    414434 3714262 1800.2
## - pages         1    636096 3935924 1810.7
## - price         1   1122475 4422303 1831.7
## 
## Step:  AIC=1778.97
## subs ~ price + pages + charpp + citations + foundingyear
## 
##                Df Sum of Sq     RSS    AIC
## <none>                      3300057 1779.0
## - foundingyear  1     41371 3341428 1779.2
## - charpp        1     67108 3367165 1780.6
## - citations     1    425311 3725369 1798.8
## - pages         1    709989 4010046 1812.0
## - price         1   1269262 4569319 1835.5
```

```
summary(g)
```

```
## 
## Call:
## lm(formula = subs ~ . - field - title - publisher, data = Journals)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -473.18  -81.60  -31.83   69.58  505.38 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1360.89705  904.70271   1.504   0.1343    
## societyyes      4.57756   41.71326   0.110   0.9127    
## price          -0.27930    0.03641  -7.671 1.19e-12 ***
## pages           0.21010    0.03638   5.775 3.50e-08 ***
## charpp          0.02383    0.01308   1.821   0.0703 .  
## citations       0.05425    0.01164   4.661 6.25e-06 ***
## foundingyear   -0.67812    0.45919  -1.477   0.1415    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 138.1 on 173 degrees of freedom
## Multiple R-squared:  0.5593, Adjusted R-squared:  0.544 
## F-statistic: 36.59 on 6 and 173 DF,  p-value: < 2.2e-16
```

```
ggplot(g, aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept=0, color="dark blue", linetype="dashed") +
  ggtitle("Residual Plot")
```

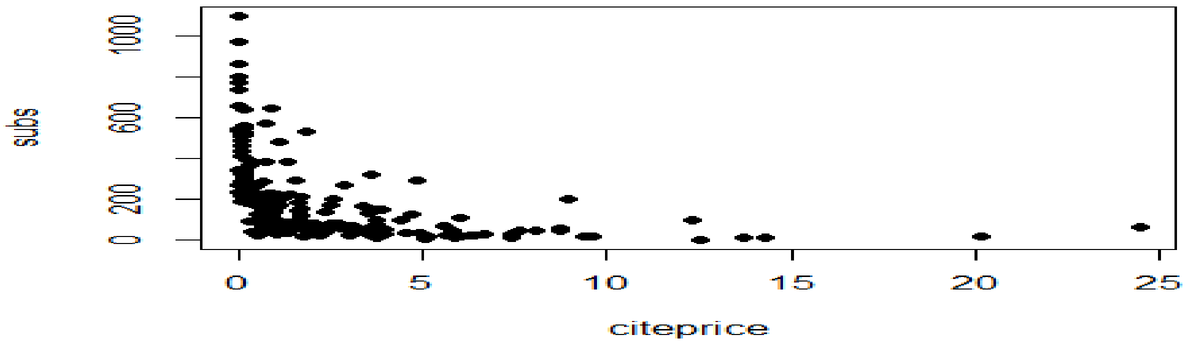


Here we can observe that the variability is not normal i.e. the variance is not constant.

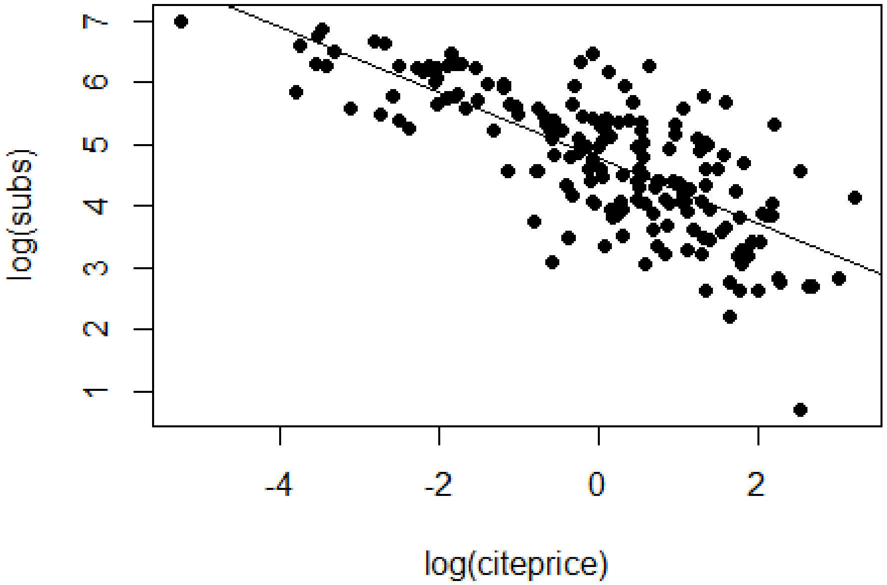
The above graph demonstrates Heteroscedasticity.

From this plot obtained we see that there is a floor effect present and there is a pattern observed in the plot so we conclude that we need to create a new model with better predictors.

```
plot(subs ~ citeprice, data = journals, pch = 19)
```



```
plot(log(subs) ~ log(citeprice), data = journals, pch = 19)
fm1 <- lm(log(subs) ~ log(citeprice), data = journals)
abline(fm1)
```



```
g2 <- lm(subs ~ citeprice + age + chars, data = log(journals))
summary(g2)

##
## Call:
## lm(formula = subs ~ citeprice + age + chars, data = log(journals))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51996 -0.41211  0.03173  0.44657  1.80588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.20665    0.31412  10.208  < 2e-16 ***
## citeprice   -0.40772    0.04196  -9.717  < 2e-16 ***
## age          0.42365    0.08972   4.722 4.75e-06 ***
## chars        0.20561    0.10747   1.913   0.0573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

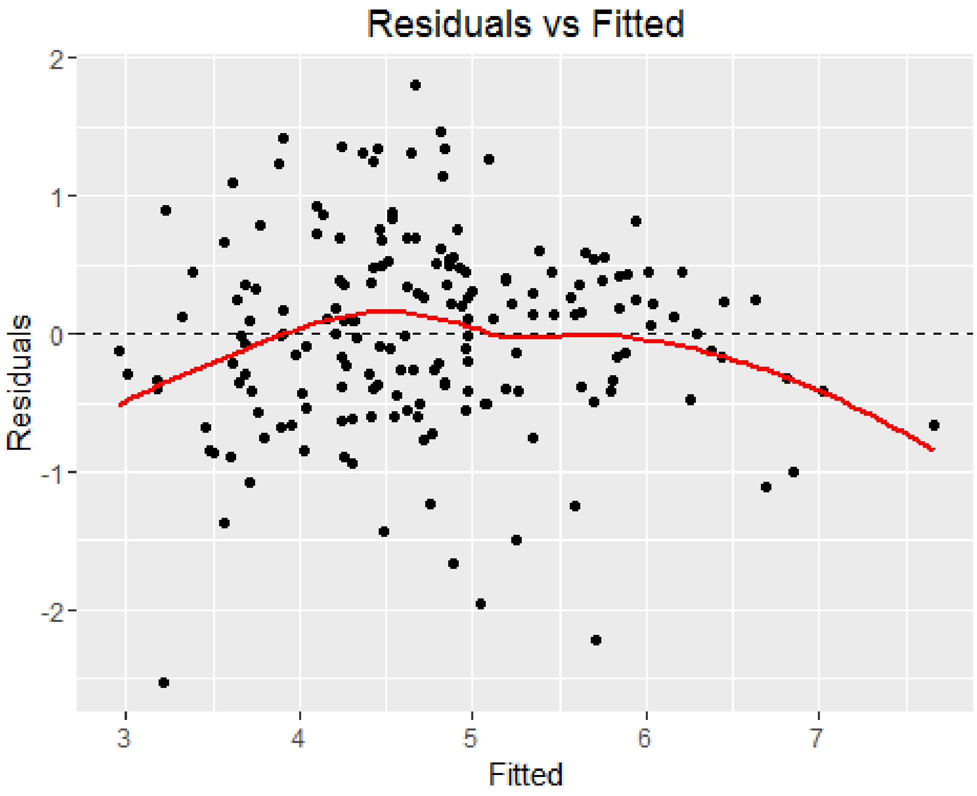
```
## Residual standard error: 0.7047 on 176 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.6066
## F-statistic: 93.01 on 3 and 176 DF,  p-value: < 2.2e-16
```

```
shapiro.test(residuals(g2))# it is not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(g2)
## W = 0.98325, p-value = 0.02958
```

```
mod2 <- fortify(g2)
p2 <- qplot(.fitted, .resid, data = mod2) + geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals") + geom_smooth(color = "red",  se = F)
p2
```

From the summary obtained above we see that the variable char has a p value 0.0573 which is greater than 5% level of significance. Also by performing the shapiro wilk normality test we observe that the p value is 0.02958 which is greater than 5% level of significance. Now we have a look at the residuals vs fitted plot.

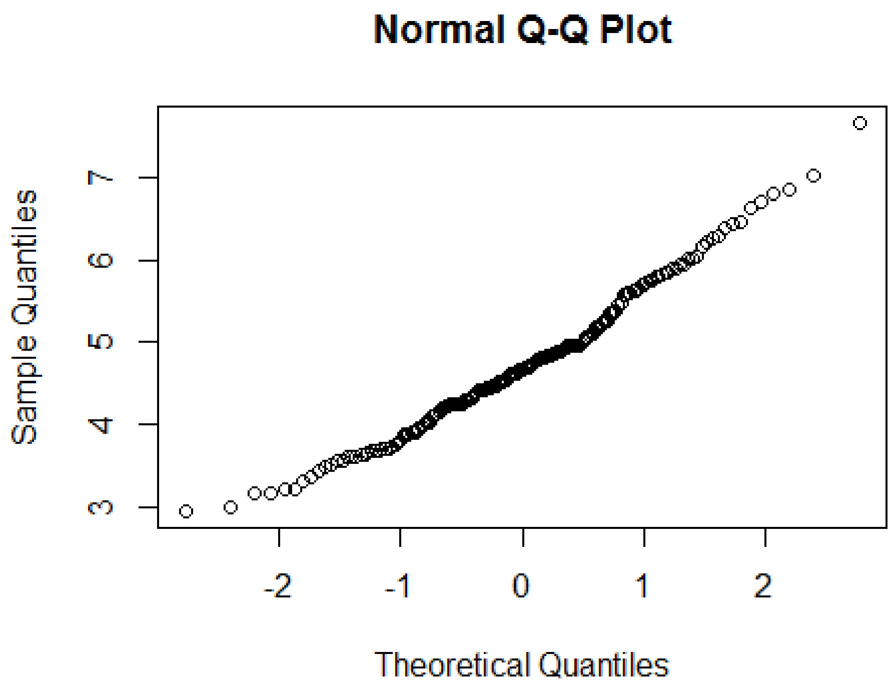


Here we can observe that, there is still not a linear regression line. The variability is still not the same.

There is no constant variance. In this plot also we can observe Heteroscedasticity. There are still many outliers as seen in the graph. But this graph is a better model than the previous model. As we can see, the adjusted R^2 is getting closer to 1. Adjusted R^2 is always between 0 to 1. The closer it is to 1 the better is the model.

Adjusted R^2 has a penalization factor which increases the Adjusted R^2 only when the predictors are significant unlike the sum of R^2 which increases even when the predictor is not significant.

```
qqnorm(fitted.values(g2))
```



Now we use a polynomial model to improve the accuracy of the model, so we have introduced new variables citeprice^2 and citeprice^3 also age*citeprice. When we have a look at the scatterplot of price vs subs we find that it is not linear and so we introduce these quadratic terms to improve the accuracy as well as the for better predictions.

```
g3 <- lm(subs ~ citeprice + I(citeprice^2) + I(citeprice^3) +
         age + I(age * citeprice) + chars, data = log(journals))
summary(g3)
```

```
##
## Call:
## lm(formula = subs ~ citeprice + I(citeprice^2) + I(citeprice^3) +
##     age + I(age * citeprice) + chars, data = log(journals))
##
```

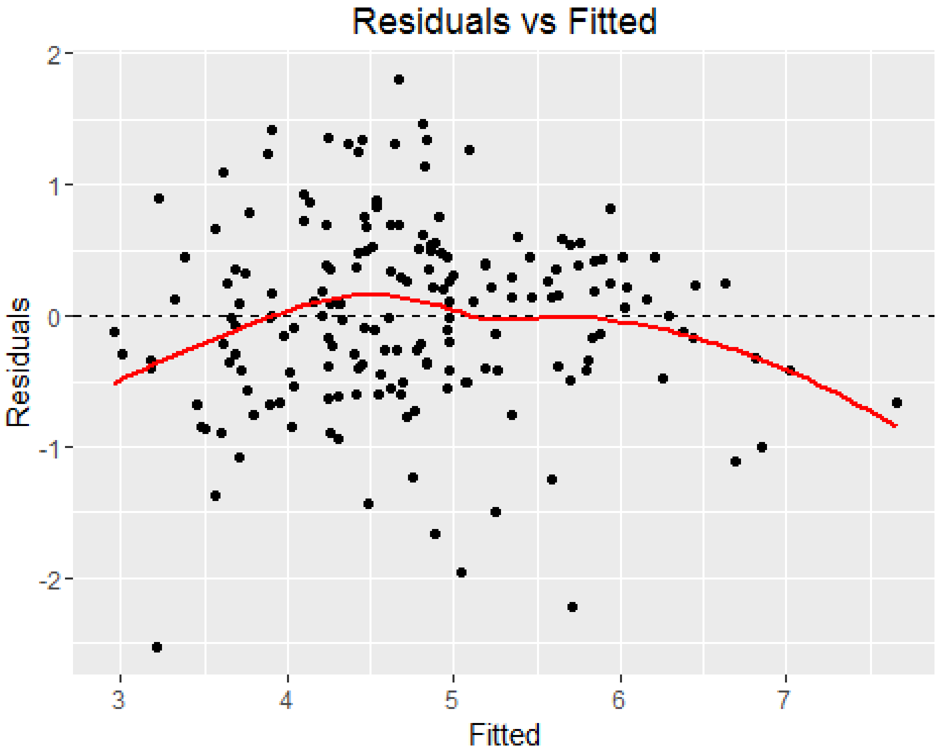
RESIDUALS:

```
##      Min       1Q   Median       3Q      Max 
## -2.25623 -0.41477  0.07605  0.40056  1.77711 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       3.407596   0.318414  10.702  < 2e-16 ***
## citeprice        -0.960936   0.189220  -5.078 9.78e-07 ***
## I(citeprice^2)    0.016510   0.024135   0.684  0.49484    
## I(citeprice^3)    0.003667   0.006862   0.534  0.59380    
## age               0.373054   0.089361   4.175 4.72e-05 ***
## I(age * citeprice)  0.155777   0.055050   2.830  0.00521 ** 
## chars             0.234618   0.106133   2.211  0.02838 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6906 on 173 degrees of freedom
## Multiple R-squared:  0.6349, Adjusted R-squared:  0.6223 
## F-statistic: 50.15 on 6 and 173 DF,  p-value: < 2.2e-16
```

```
shapiro.test(residuals(g3))# it is not normally distributed
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(g3)
## W = 0.979, p-value = 0.008116
```

```
mod3 <- fortify(g3)
p3 <- qplot(.fitted, .resid, data = mod2) + geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals") + geom_smooth(color = "red",  se = F)
p3
```



From the summary, we see that the p values of citeprice^2 and citeprice^3 are greater than 0.05 level of significance so we need to remove these variables and check for the accuracy of the model. To further improve the accuracy of the model we use the Automated Model Selection using the Step function.

```
g4 <- step(g3)
```

At the start we have the AIC value to be -126.43 so all the variables with the AIC values lower than this will have be removed first and then we get the final model.

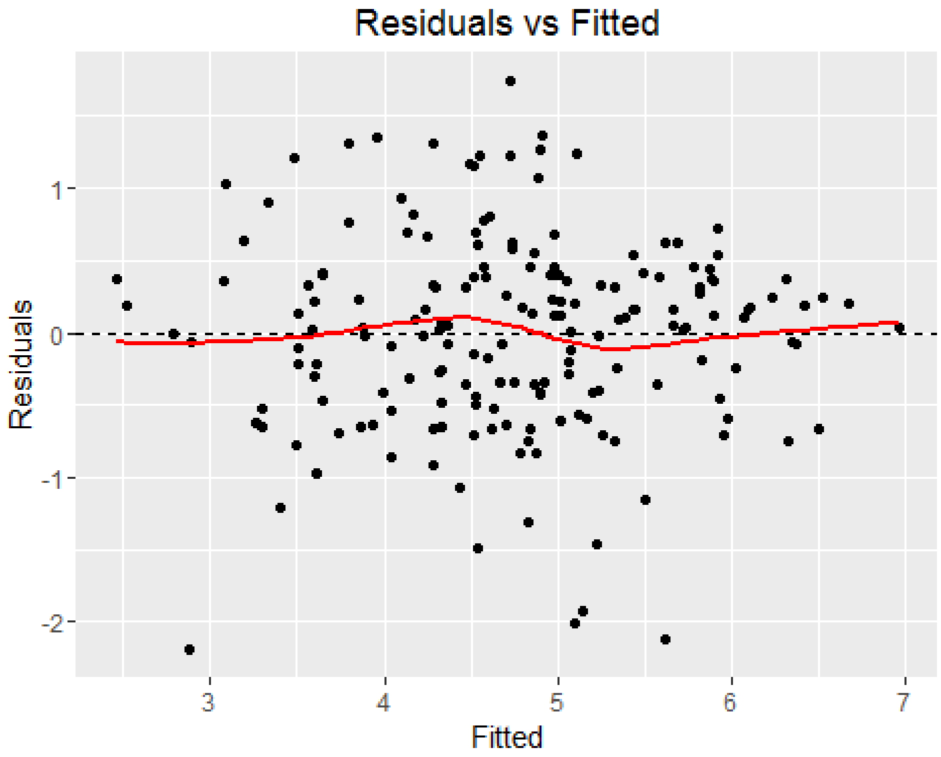
```
## Start:  AIC=-126.43
## subs ~ citeprice + I(citeprice^2) + I(citeprice^3) + age + I(age *
##     citeprice) + chars
##
##                    Df Sum of Sq    RSS     AIC
## - I(citeprice^3)      1    0.1362 82.636 -128.13
## - I(citeprice^2)      1    0.2232 82.723 -127.94
## <none>                         82.500 -126.43
## - chars               1    2.3304 84.830 -123.42
## - I(age * citeprice)  1    3.8186 86.318 -120.28
## - age                 1    8.3110 90.811 -111.15
## - citeprice           1   12.2988 94.798 -103.42
##
## Step:  AIC=-128.13
## subs ~ citeprice + I(citeprice^2) + age + I(age * citeprice) +
##     chars
##
##                     Df Sum of Sq    RSS     AIC
## - I(citeprice^2)      1    0.1017 82.738 -129.91
## <none>                            82.636 -128.13
## - chars               1    2.2406 84.876 -125.32
## - I(age * citeprice)  1    3.8094 86.445 -122.02
## - age                 1    8.3266 90.962 -112.85
## - citeprice           1   12.2911 94.927 -105.17
##
## Step:  AIC=-129.91
## subs ~ citeprice + age + I(age * citeprice) + chars
##
##                     Df Sum of Sq    RSS     AIC
## <none>                            82.738 -129.91
## - chars               1    2.2526 84.990 -127.08
## - I(age * citeprice)  1    4.6733 87.411 -122.02
## - age                 1    8.3321 91.070 -114.64
## - citeprice           1   14.6456 97.383 -102.58
```

We see that first the variable citeprice^3 with AIC value of -128.13 gets removed then the variable citeprice^2 with AIC value of -127.94 gets removed and finally we have only the values higher than the AIC value -126.43 in the model, so we finally have a model with good predictors of subs – citeprice , age , citeprice * age.

```
anova(g4,g3)

## Analysis of Variance Table
##
## Model 1: subs ~ citeprice + age + I(age * citeprice) + chars
## Model 2: subs ~ citeprice + I(citeprice^2) + I(citeprice^3) + age + I(age *
##     citeprice) + chars
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    175 82.738
## 2    173 82.500  2    0.2379 0.2494 0.7795
```

From the results of the Anova function we get that the big models used has a p value greater than the level of significance. So, we use the small model for the prediction of the subscription of journals.

```
mod4 <- fortify(g4)
  p4 <- qplot(.fitted, .resid, data = mod4) + geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals") + geom_smooth(color = "red",  se = F)
  p4
```



The results of the plot show that the error distribution is less and also there is no floor effect or ceiling effect observed, so we conclude that this model is the best model by far.

```
summary(lm(abs(residuals(g4)) ~ fitted(g4)))

##
## Call:
## lm(formula = abs(residuals(g4)) ~ fitted(g4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64263 -0.29529 -0.09586  0.16662  1.66076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83835    0.17349   4.832  2.9e-06 ***
## fitted(g4)  -0.06666    0.03597  -1.853   0.0655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4305 on 178 degrees of freedom
## Multiple R-squared:  0.01893,    Adjusted R-squared:  0.01342
## F-statistic: 3.435 on 1 and 178 DF,  p-value: 0.06549
```

```
anova(g4,g3)
```

```
## Analysis of Variance Table
##
## Model 1: subs ~ citeprice + age + I(age * citeprice) + chars
## Model 2: subs ~ citeprice + I(citeprice^2) + I(citeprice^3) + age + I(age *
##     citeprice) + chars
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    175 82.738
## 2    173 82.500  2    0.2379 0.2494 0.7795
```

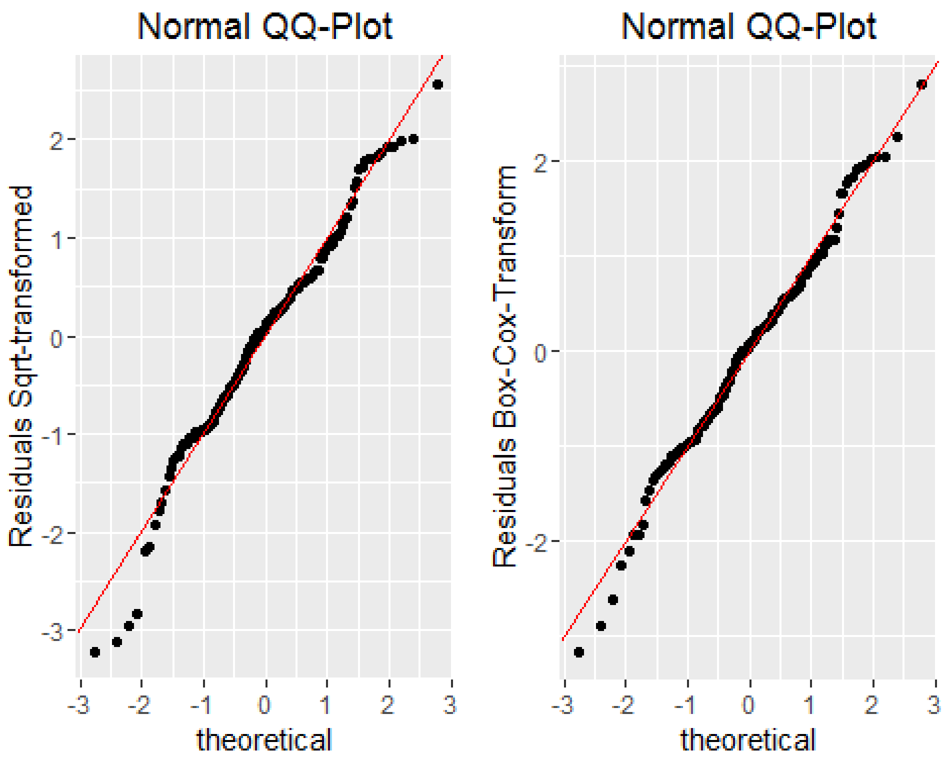
```
# Conclusion: The F-Ratio 0.2494 is small, therefore take Model g4 (the small model) since p-value 0.7795 is g
reater than 0.05
```

BOX COX TRANSFORM:

```
library(car)
  (lambda <- powerTransform(g4))
```

```
## Estimated transformation parameters
##       Y1
## 1.528275
```

```
lam <- lambda$lambda
  glam <- lm(subs^lam ~ citeprice + age + I(age * citeprice) + chars, data = log(journals))
  modlam <- fortify(glam)
  p1 <- qplot(sample = scale(.resid), data = mod4) + geom_abline(intercept = 0,
                                                                   slope = 1, color = "red") + labs(title = "Normal QQ-Plot", y = "Residu
als Sqrt-transformed")
  p2 <- qplot(sample = scale(.resid), data = modlam) + geom_abline(intercept = 0,
                                                                    slope = 1, color = "red") + labs(title = "Normal QQ-Plot", y = "Residu
als Box-Cox-Transform")
  grid.arrange(p1, p2, nrow = 1)
```



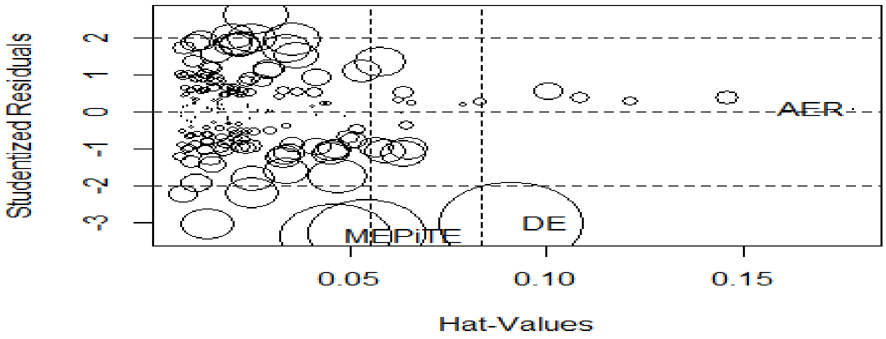
```
p4 <- qplot(.fitted, .resid, data = modlam) + geom_hline(yintercept = 0, linetype = "dashed") +
   labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals") + geom_smooth(color = "red",  se = F)
```

From the above figure, we can compare the Normal QQ plot and it is observed that the QQ plot is improved with the box cox transform and the model is normally distributed.

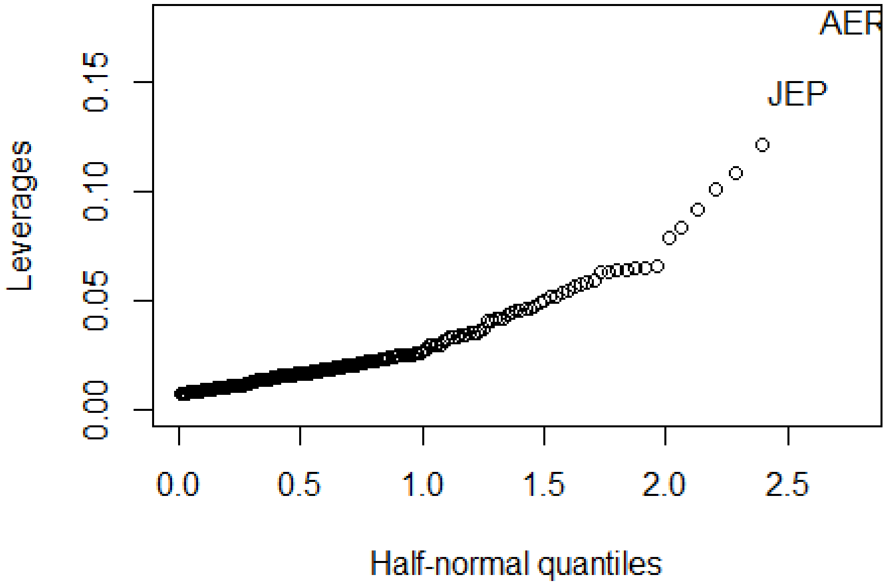
UNUSUAL OBSERVATIONS:

Unusual Observation is also known as influential observation. We can try to remove these observations and improve the fitness of model.

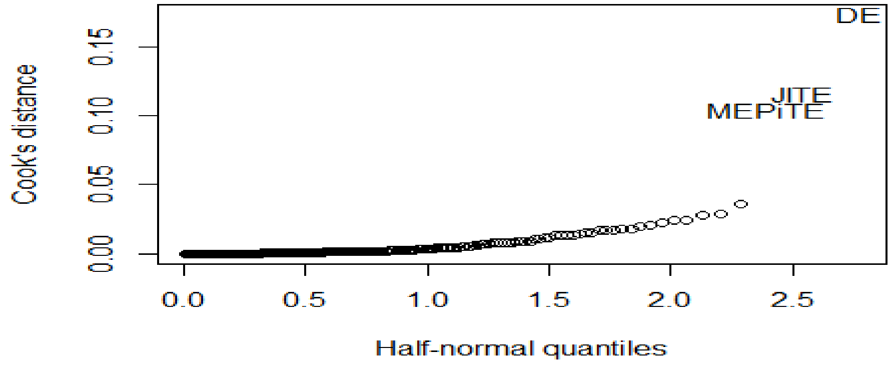
```
library("faraway")
influencePlot(g4)
```



```
##           StudRes        Hat         CookD
## MEPiTE -3.35602560 0.04643377 0.1036127486
## DE     -3.00407492 0.09115848 0.1730969150
## AER     0.05886862 0.17785226 0.0001507952

journ <- row.names(journals)
halfnorm(lm.influence(g4)$hat, labs = journ, ylab = "Leverages")
```



```
cook <- cooks.distance(g4)
halfnorm(cook, 3, labs = journ, ylab = "Cook's distance")
```



From the above plots, we can observe that the unusual observation from our dataset is Journal DE, we have removed the journal DE from our dataset and then we have fitted the model.

```
g41 <- lm(subs ~ citeprice + age + I(age * citeprice) + chars, data = log(journals),subset = (cook < max(cook)))
compareCoefs(g4, g41)
```

```
## 
## Call:
## 1: lm(formula = subs ~ citeprice + age + I(age * citeprice) + chars, 
##   data = log(journals))
## 2: lm(formula = subs ~ citeprice + age + I(age * citeprice) + chars, 
##   data = log(journals), subset = (cook < max(cook)))
##                     Est. 1    SE 1  Est. 2    SE 2
## (Intercept)         3.4335  0.3149  3.2546  0.3136
## citeprice          -0.8989  0.1615 -0.9357  0.1584
## age                 0.3735  0.0890  0.4369  0.0895
## I(age * citeprice)  0.1410  0.0448  0.1575  0.0442
## chars               0.2295  0.1051  0.2209  0.1028
```

```
summary(g41)
```

```
## 
## Call:
## lm(formula = subs ~ citeprice + age + I(age * citeprice) + chars, 
##     data = log(journals), subset = (cook < max(cook)))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.23874 -0.45256  0.07666  0.39232  1.71446 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         3.25456    0.31359  10.378  < 2e-16 ***
## citeprice          -0.93574    0.15840  -5.907 1.79e-08 ***
## age                 0.43693    0.08953   4.880 2.38e-06 ***
## I(age * citeprice)  0.15751    0.04419   3.565  0.00047 ***
## chars               0.22088    0.10284   2.148  0.03311 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6724 on 174 degrees of freedom
## Multiple R-squared:  0.6483, Adjusted R-squared:  0.6402 
## F-statistic: 80.19 on 4 and 174 DF,  p-value: < 2.2e-16
```

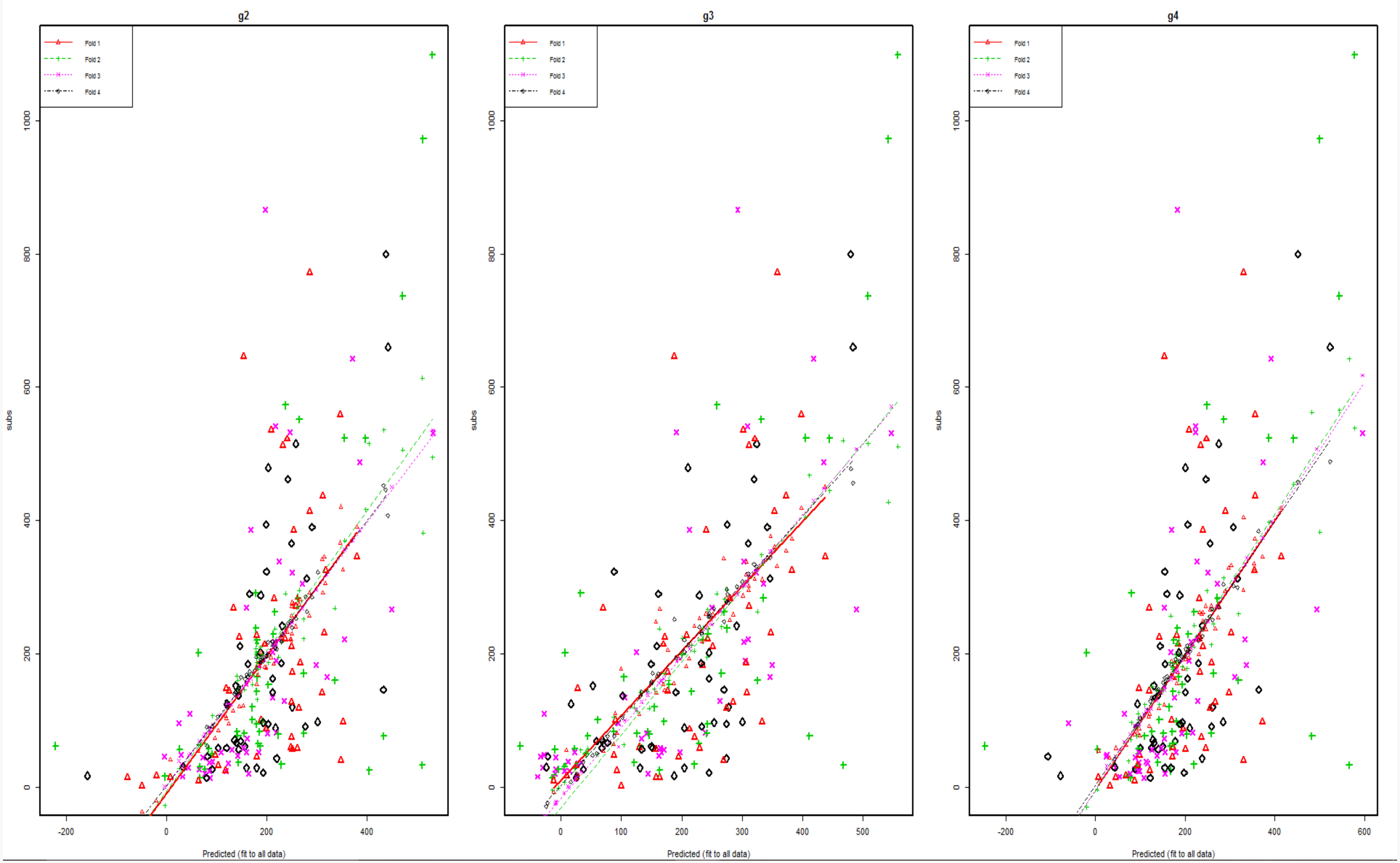
CROSS VALIDATION:

A way of measuring the predictive performance of a statistical model. Useful for overcoming the problem of over-fitting. Cross validation techniques tend to focus on not using the entire data set when building a model. “Testing Set” – data that is removed, “Training Set” – remaining data in the model

R CODE:

```
for (i in 1:10) {
seed <- round(runif(1, min=0, max=100))
oldpar <- par(mfrow=c(1,3))
mse.g2 <- CVlm(data = journals,
form.lm=g2,
m=4,
seed=seed,
printit=F,
main = "g2")
mse.g3 <- CVlm(data = journals,
form.lm=g3,
m=4,
seed=seed,
printit=F,
main = "g3")
mse.g4 <- CVlm(data = journals,
form.lm=g4,
m=4,
seed=seed,
printit=F,
main = "g4")
par(oldpar)
df.temp <- data.frame(
mse.g2=attr(mse.g2, "ms"),
mse.g3=attr(mse.g3, "ms"),
mse.g4=attr(mse.g4, "ms")
)
df <- rbind(df,df.temp)

}
df
```

OUTPUT:



CONFIDENCE INTERNALS:

A Range of values so defined that there is a specified probability that the value of a parameter lies within it.

In statistics, a confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that potentiall y includes the unobservable true parameter of interest. How frequently the observed interval contains the true pa rameter if the experiment is repeated is called the confidence level.

confint(g4)

```
##                        2.5 %      97.5 %
## (Intercept)         2.81211505  4.0549264
## citeprice          -1.21766428 -0.5801552
## age                 0.19791431  0.5491154
## I(age * citeprice)  0.05247246  0.2294457
## chars               0.02198764  0.4369445
```

confint(g41)

```
##                          2.5 %     97.5 %
## (Intercept)         2.63563114  3.8734834
## citeprice          -1.24838113 -0.6231032
## age                 0.26022963  0.6136226
## I(age * citeprice)  0.07029923  0.2447172
## chars               0.01791284  0.4238442
```

All the variables have 95% confidence intervals.

We can reject null hypothesis as all the variables does not contain zero in confidence interval.

CONCLUSION:

After doing our analysis we find that the model g4 is the best model with the predictors age , age * citeprice, chars . We performed the Box Cox transform and made the model to fit better and also we removed the unusual observation and by sub setting maximum COOK'D distance value. We were able to conclude that the model g4 fits the data better. By calculating the confidence interval for each variable the model rejects the null hypothesis for each variable.