Case Study 3
Regression Modelling

1. TV Dataset

Jalao (2012) proposed a regression model to predict the revenue of advertising for a 30 second primetime TV show slot. Significant factors that affect the revenue of advertising were also determined. Data was obtained and compiled from multiple websites that provide information that could potentially affect the revenue of advertising. Moreover, the effect of several social media websites on the revenue of advertising was also studied.

2. Data Set Description

Table 1: Data Description and Modelling

Variable	Description	Source	Model
Revenue (Response)	Average Revenue of Advertising in a 30 second primetime advertisement slot in USD	adage.com	Continuous (Response)
Length	Either 30 minutes or 1 hour Broadcast time	Show official website site	Continuous
Viewers	Nielsen Average Number of Viewers for 2011-2012 Season	deadline.com	Continuous
18-49 Rating	Nielsen Average 18-49 Demographic Rating Share in % for 2011-2012 Season	deadline.com	Continuous
Facebook	Number of Facebook Likes from official show Facebook page	Show's official Facebook Page	Continuous
Facebook Talking About	Number of Active Social Media users talking about the show on Facebook	Show's official Facebook Page	Continuous
Twitter	Number of Tweeter Followers from official tweeter pages	Show's official Twitter Page	Continuous
Age	Number of Episodes Aired	Show official website	Continuous
Network	Network that broadcasts the show: ABC, CBS, CW, Fox or NBC. Baseline is CW since it has the lowest average revenue of advertising for all shows.	Show official website	$Network\_ABC = \begin{cases} 1 & if\ show\ is\ in\ ABC \\ 0 & o/w \end{cases}$ $Network\_CBS = \begin{cases} 1 & if\ show\ is\ in\ CBS \\ 0 & o/w \end{cases}$ $Network\_Fox = \begin{cases} 1 & if\ show\ is\ in\ Fox \\ 0 & o/w \end{cases}$

			$Network\_NBC = \begin{cases} 1 & \text{if show is in Fox} \\ 0 & o/w \end{cases}$
Day	Day of show broadcast, Sunday through Friday. No data points for Saturday. Baseline is Friday since it has the lowest average revenue of advertising for all shows.	Show official website	$Day\_Su = \begin{cases} 1 & \text{if show is on Sunday} \\ 0 & o/w \end{cases}$ $Day\_M = \begin{cases} 1 & \text{if show is on Monday} \\ 0 & o/w \end{cases}$ $Day\_T = \begin{cases} 1 & \text{if show is on Tuesday} \\ 0 & o/w \end{cases}$ $Day\_W = \begin{cases} 1 & \text{if show is on Wednesday} \\ 0 & o/w \end{cases}$ $Day\_Th = \begin{cases} 1 & \text{if show is on Thursday} \\ 0 & o/w \end{cases}$
Type	Type of Show: Drama, Sit-com, Sports or Reality TV. Baseline is Reality TV.	Show official website	$Type\_D = \begin{cases} 1 & \text{if show is a Drama} \\ 0 & o/w \end{cases}$ $Type\_C = \begin{cases} 1 & \text{if show is a sitcom} \\ 0 & o/w \end{cases}$ $Type\_S = \begin{cases} 1 & \text{if show is Sport event} \\ 0 & o/w \end{cases}$

3. Loading Data to R Studio

3.1. Initialize R: Setting Working Directory

3.1.1.Open R Studio

3.1.2.On the file explorer tab click on Files. Files Plots Packages Help Viewer

3.1.3.Click on Explore ...

3.1.4.Go to the Desktop Folder -> Module 3 Datasets -> Case 3

3.1.5.Click on More. More▾. Click on Set as Working Directory.

3.2. Load Bank Dataset into R.

3.2.1.Click on File-> New File -> R Script.

3.2.2.In the new tab script Untitled1* ×, type the following code:

```
•  options(scipen=999,digits=2)
•  tvdataset = read.csv("tvdataset.csv")
```

3.2.3.Highlight the two lines and click on Run Run. As a result, the data is loaded in the Environment

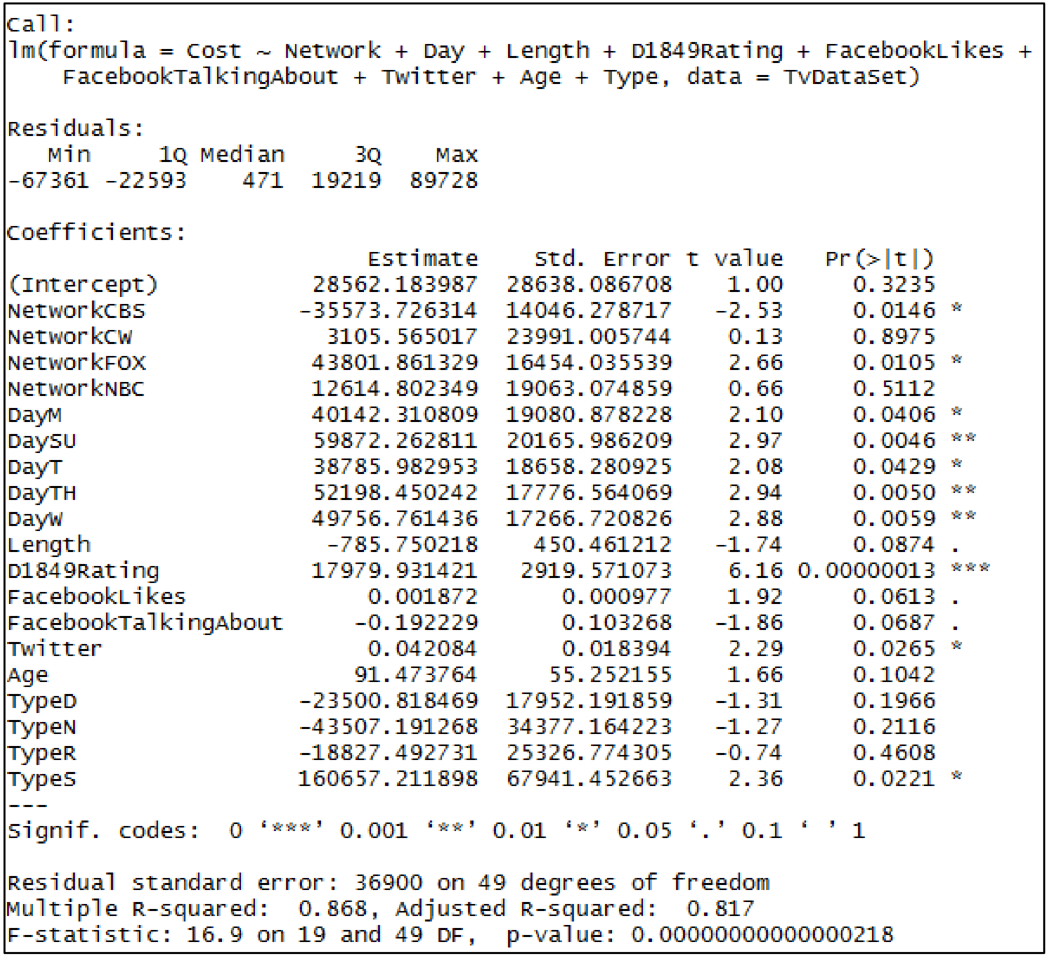
3.3. Fitting the Full Model

3.3.1.In the new tab script Untitled1* ×, type the following code:

```
•  tvdataset.fit =lm(cost~network + day + length + d1849rating +
   facebooklikes + facebooktalkingabout + twitter+ age + type, data=
   tvdataset)
•  summary(tvdataset.fit)
```

3.3.2.Highlight the two lines of code and click on Run Run.

3.3.3. The result of the linear regression fit would be as follows:



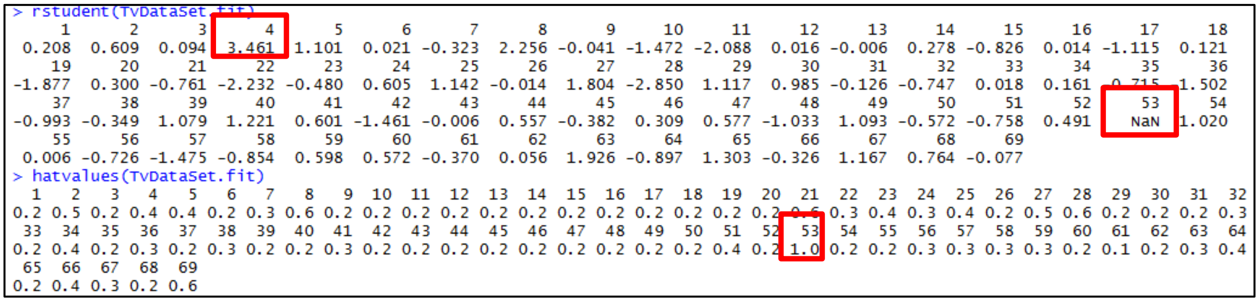
```
Call:
lm(formula = Cost ~ Network + Day + Length + D1849Rating + FacebookLikes +
    FacebookTalkingAbout + Twitter + Age + Type, data = TvDataSet)

Residuals:
   Min     1Q Median     3Q    Max 
-67361 -22593    471  19219  89728 

Coefficients:
                         Estimate    Std. Error t value   Pr(>|t|)    
(Intercept)          28562.183987  28638.086708    1.00     0.3235    
NetworkCBS          -35573.726314  14046.278717   -2.53     0.0146 *  
NetworkCW             3105.565017  23991.005744    0.13     0.8975    
NetworkFOX           43801.861329  16454.035539    2.66     0.0105 *  
NetworkNBC           12614.802349  19063.074859    0.66     0.5112    
DayM                 40142.310809  19080.878228    2.10     0.0406 *  
DaySU                59872.262811  20165.986209    2.97     0.0046 ** 
DayT                 38785.982953  18658.280925    2.08     0.0429 *  
DayTH                52198.450242  17776.564069    2.94     0.0050 ** 
DayW                 49756.761436  17266.720826    2.88     0.0059 ** 
Length                -785.750218    450.461212   -1.74     0.0874 .  
D1849Rating          17979.931421   2919.571073    6.16 0.00000013 ***
FacebookLikes            0.001872      0.000977    1.92     0.0613 .  
FacebookTalkingAbout    -0.192229      0.103268   -1.86     0.0687 .  
Twitter                  0.042084      0.018394    2.29     0.0265 *  
Age                     91.473764     55.252155    1.66     0.1042    
TypeD               -23500.818469  17952.191859   -1.31     0.1966    
TypeN               -43507.191268  34377.164223   -1.27     0.2116    
TypeR               -18827.492731  25326.774305   -0.74     0.4608    
TypeS               160657.211898  67941.452663    2.36     0.0221 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36900 on 49 degrees of freedom
Multiple R-squared:  0.868, Adjusted R-squared:  0.817 
F-statistic: 16.9 on 19 and 49 DF,  p-value: 0.0000000000000218
```

3.4. Model Adequacy Checking

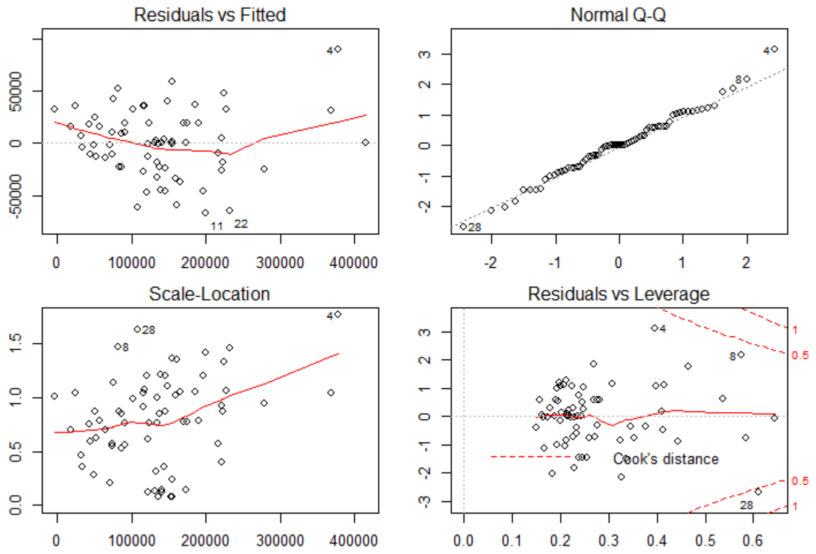
3.4.1. To check for diagnostics as well as studentized residuals and Leverage (hat values) we type the following.

```
• par(mfrow =c(2,2),mar=c(2,2,2,2))
• plot(tvdataset.fit)
• rstudent(tvdataset.fit)
• hatvalues(tvdataset.fit)
```

3.4.2. Highlight these lines of code and click on Run Run .



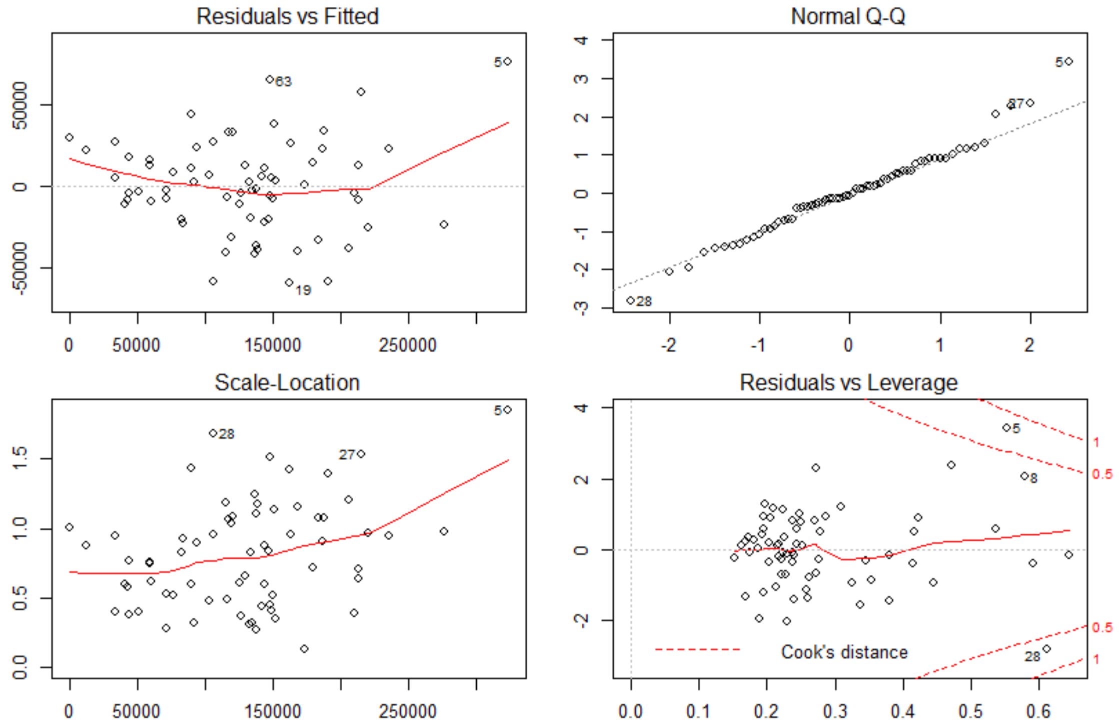
```
> rstudent(TvDataSet.fit)
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18 
 0.208  0.609  0.094  3.461  1.101  0.021 -0.323  2.256 -0.041 -1.472 -2.088  0.016 -0.006  0.278 -0.826  0.014 -1.115  0.121 
    19     20     21     22     23     24     25     26     27     28     29     30     31     32     33     34     35     36 
-1.877  0.300 -0.761 -2.232 -0.480  0.605  1.142 -0.014  1.804 -2.850  1.117  0.985 -0.126 -0.747  0.018  0.161  0.715  1.502 
    37     38     39     40     41     42     43     44     45     46     47     48     49     50     51     52     53     54 
-0.993 -0.349  1.079  1.221  0.601 -1.461 -0.006  0.557 -0.382  0.309  0.577 -1.033  1.093 -0.572 -0.758  0.491    NaN  1.020 
    55     56     57     58     59     60     61     62     63     64     65     66     67     68     69 
 0.006 -0.726 -1.475 -0.854  0.598  0.572 -0.370  0.056  1.926 -0.897  1.303 -0.326  1.167  0.764 -0.077 
> hatvalues(TvDataSet.fit)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32 
0.2 0.5 0.2 0.4 0.4 0.2 0.3 0.6 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.6 0.3 0.4 0.3 0.4 0.2 0.5 0.6 0.2 0.2 0.2 0.3 
 33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64 
0.2 0.4 0.2 0.3 0.2 0.3 0.2 0.2 0.3 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.4 0.2 1.0 0.2 0.2 0.3 0.3 0.3 0.3 0.2 0.1 0.2 0.3 0.4 
 65  66  67  68  69 
0.2 0.4 0.3 0.2 0.6 
```



3.4.3.It seems that observations 4, and 53 are outliers. We thus eliminate these rows, refit the regression model and plots as follows:

```
•  reducedtvdataset=tvdataset[-c(4, 53), ]
•  reducedtvdataset.fit =lm(cost~network + day + length +
   d1849rating + facebooklikes + facebooktalkingabout +twitter+ age
   + type, data= reducedtvdataset)
•  summary(reducedtvdataset.fit)
•  par(mfrow =c(2,2),mar=c(2,2,2,2))
•  plot(reducedtvdataset.fit)
```

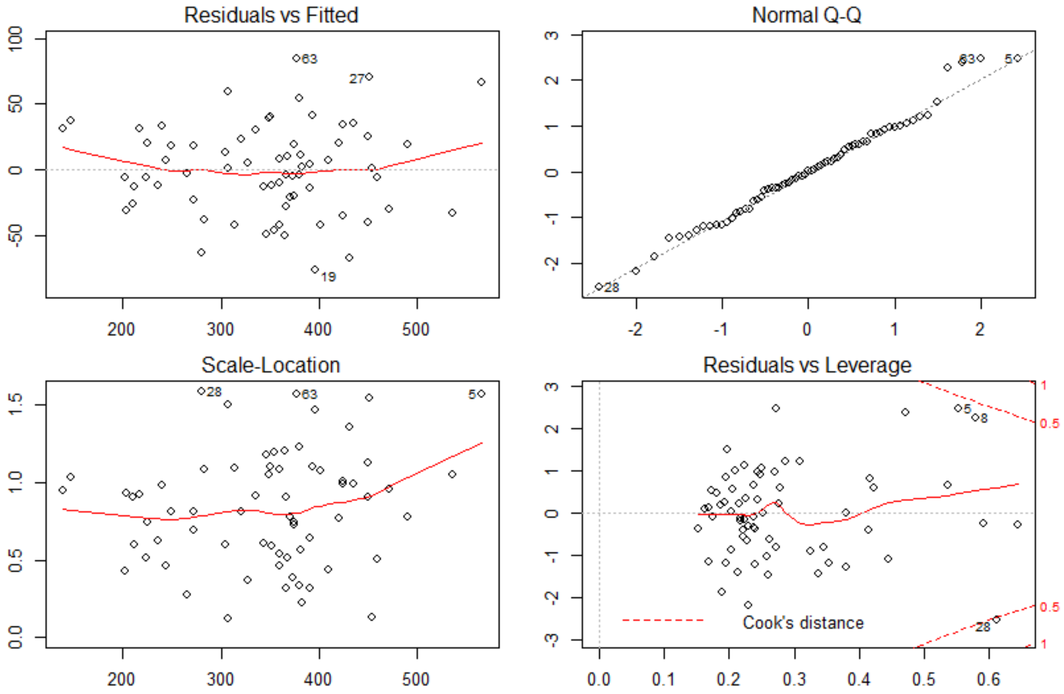
3.4.4.Highlight these lines of code and click on Run Run.



3.4.5.Based on the Residuals vs. Fitted graph, the constant variance assumption does not hold. We then transform the Cost variable as follows:

```
#Transform Data Squareroot
reducedtvdataset.fit =lm(cost^0.5~network + day + length +
d1849rating + facebooklikes + facebooktalkingabout +twitter + age
+ type, data= reducedtvdataset)
par(mfrow =c(2,2),mar=c(2,2,2,2))
plot(reducedtvdataset.fit)
```

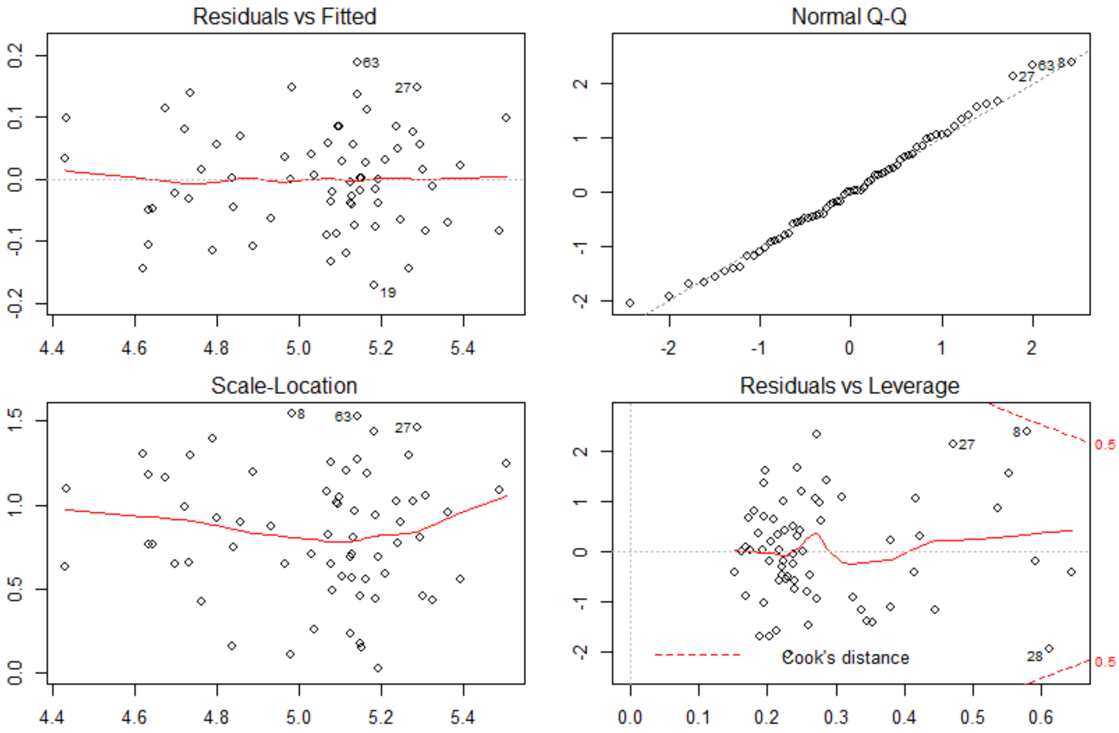
3.4.6.The Residuals vs Fits plot for Square Root transformation is as follows:



3.4.7.Based on the Residuals vs. Fitted graph, the constant variance assumption still does not hold. We further transform the Cost variable further as follows:

```
#Transform Data Log10
reducedtvdataset.fit =lm(log10(cost)~network + day + length +
d1849rating + facebooklikes + facebooktalkingabout + twitter +
age + type, data= reducedtvdataset)
par(mfrow =c(2,2),mar=c(2,2,2,2))
plot(reducedtvdataset.fit)
```

3.4.8. The plot for Log10 transformation is as follows:



3.4.9. Based on the graph, the constant variance assumption holds.

3.5. Variable Selection

3.5.1. We now choose the most relevant variables for the regression model. Type the following code and run it.

```
base.fit =lm(log10(cost)~1, data= reducedtvdataset)
forward = step(base.fit, scope = list(lower=~1,upper=~network + day +
      length + d1849rating + facebooklikes + facebooktalkingabout
      +twitter+ age + type), direction = "both", trace=1)
summary(forward)
```

3.5.2. The result of the regression model is as follows:

```
Start:  AIC=-183
log10(cost) ~ 1

                       Df  Sum of Sq  RSS  AIC
+ network               4      2.307 1.93 -228
+ d1849rating           1      1.557 2.68 -212
+ day                   5      1.357 2.88 -199
+ facebooklikes         1      0.763 3.48 -194
+ facebooktalkingabout  1      0.648 3.59 -192
+ type                  3      0.704 3.54 -189
+ twitter               1      0.355 3.89 -187
+ age                   1      0.173 4.07 -184
+ length                1      0.138 4.10 -183
<none>                              4.24 -183
```

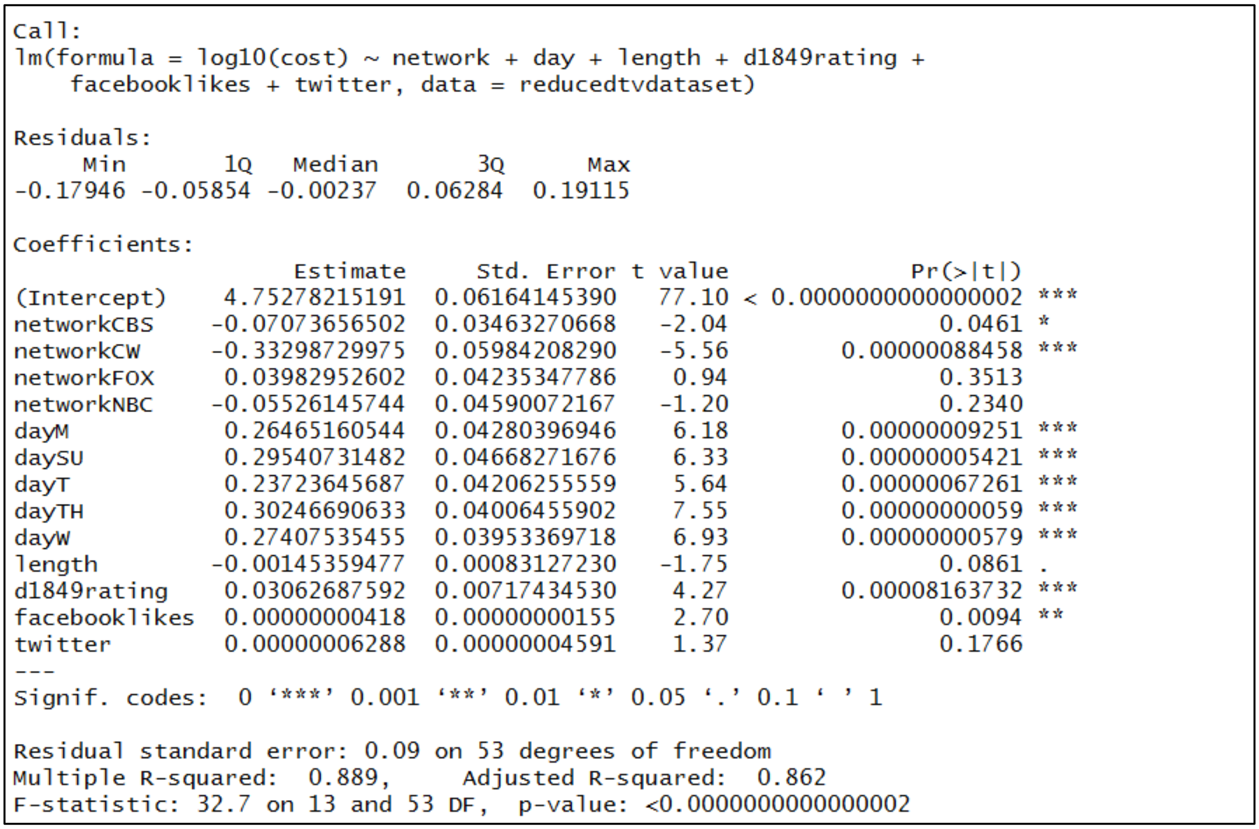
```
#Deleted Results Here…

#Final Model Results:
Step:  AIC=-304
log10(cost) ~ network + day + facebooklikes + d1849rating + length +
    twitter

                       Df Sum of Sq   RSS  AIC
<none>                              0.470 -304
- twitter               1     0.017 0.487 -304
+ age                   1     0.008 0.462 -304
- length                1     0.027 0.497 -302
+ facebooktalkingabout  1     0.000 0.470 -302
+ type                  3     0.023 0.448 -302
- facebooklikes         1     0.065 0.535 -298
- d1849rating           1     0.162 0.632 -286
- network               4     0.591 1.061 -258
- day                   5     0.671 1.141 -255
```

3.6. Fitting the Final Model:

3.6.1.Type the following code to determine the final regression model:

- `Finaltvdataset.fit =lm(log10(cost)~network + day +length+ d1849rating + facebooklikes + twitter, data= reducedtvdataset)`
- `summary(Finaltvdataset.fit)`

3.6.2.Run these lines of code and the results of the regression modelling would be as follows:



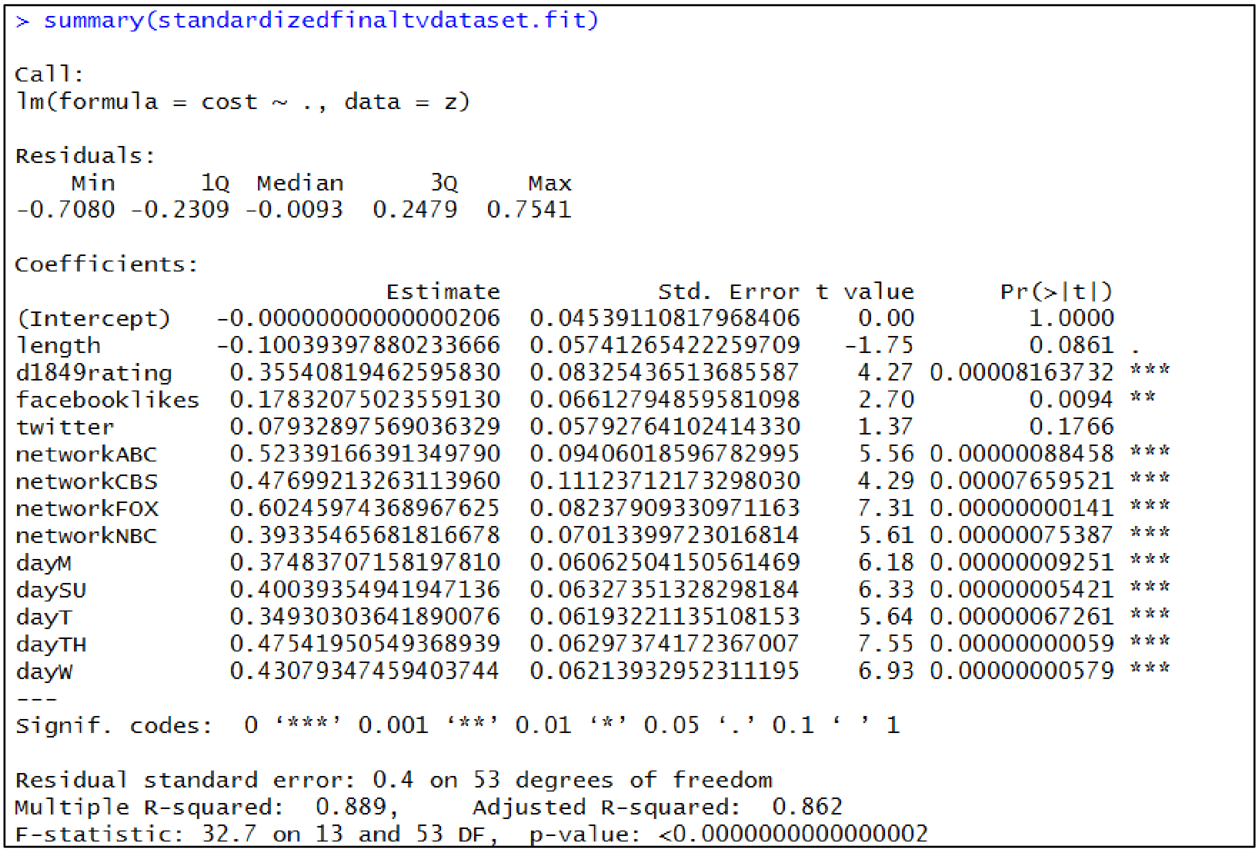
```
Call:
lm(formula = log10(cost) ~ network + day + length + d1849rating +
    facebooklikes + twitter, data = reducedtvdataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17946 -0.05854 -0.00237  0.06284  0.19115

Coefficients:
                    Estimate     Std. Error t value            Pr(>|t|)
(Intercept)    4.75278215191  0.06164145390   77.10 < 0.0000000000000002 ***
networkCBS    -0.07073656502  0.03463270668   -2.04               0.0461 *
networkCW     -0.33298729975  0.05984208290   -5.56         0.00000088458 ***
networkFOX     0.03982952602  0.04235347786    0.94               0.3513
networkNBC    -0.05526145744  0.04590072167   -1.20               0.2340
dayM           0.26465160544  0.04280396946    6.18         0.00000009251 ***
daySU          0.29540731482  0.04668271676    6.33         0.00000005421 ***
dayT           0.23723645687  0.04206255559    5.64         0.00000067261 ***
dayTH          0.30246690633  0.04006455902    7.55         0.00000000059 ***
dayW           0.27407535455  0.03953369718    6.93         0.00000000579 ***
length        -0.00145359477  0.00083127230   -1.75               0.0861 .
d1849rating    0.03062687592  0.00717434530    4.27         0.00008163732 ***
facebooklikes  0.00000000418  0.00000000155    2.70               0.0094 **
twitter        0.00000006288  0.00000004591    1.37               0.1766
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09 on 53 degrees of freedom
Multiple R-squared:  0.889,     Adjusted R-squared:  0.862
F-statistic: 32.7 on 13 and 53 DF,  p-value: <0.0000000000000002
```

3.7. Fitting the Final Model with Standardized Coefficients:

3.7.1.Type the following code to determine the final regression model:

- `#Convert To Numerical`
- `#Network`
- `networkind =model.matrix( ~ network - 1, data = reducedtvdataset)`
- `#Set CW as Baseline`
- `networkind = subset(networkind, select = -c(networkCW) )`
- `#Day`
- `dayind =model.matrix( ~ day - 1, data = reducedtvdataset)`
- `dayind = subset(dayind, select = -c(dayF) )`
- `x = cbind(subset(reducedtvdataset, select = c(3,6,8,9,11)),networkind,dayind)`
- `z = data.frame(scale(x, center = TRUE, scale = TRUE))`
- `z$cost = scale(log10(x$cost), center = TRUE, scale = TRUE)`
- `standardizedfinaltvdataset.fit =lm(cost~., data= z)`
- `summary(standardizedfinaltvdataset.fit)`

3.7.2.Run these lines of code and the results of the regression modelling would be as follows:



```
> summary(standardizedfinaltvdataset.fit)

Call:
lm(formula = cost ~ ., data = z)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7080 -0.2309 -0.0093  0.2479  0.7541 

Coefficients:
                          Estimate          Std. Error t value      Pr(>|t|)    
(Intercept)   -0.00000000000000206 0.04539110817968406    0.00        1.0000    
length        -0.10039397880233666 0.05741265422259709   -1.75        0.0861 .  
d1849rating    0.35540819462595830 0.08325436513685587    4.27 0.00008163732 ***
facebooklikes  0.17832075023559130 0.06612794859581098    2.70        0.0094 ** 
twitter        0.07932897569036329 0.05792764102414330    1.37        0.1766    
networkABC     0.52339166391349790 0.09406018596782995    5.56 0.00000088458 ***
networkCBS     0.47699213263113960 0.11123712173298030    4.29 0.00007659521 ***
networkFOX     0.60245974368967625 0.08237909330971163    7.31 0.00000000141 ***
networkNBC     0.39335465681816678 0.07013399723016814    5.61 0.00000075387 ***
dayM           0.37483707158197810 0.06062504150561469    6.18 0.00000009251 ***
daySU          0.40039354941947136 0.06327351328298184    6.33 0.00000005421 ***
dayT           0.34930303641890076 0.06193221135108153    5.64 0.00000067261 ***
dayTH          0.47541950549368939 0.06297374172367007    7.55 0.00000000059 ***
dayW           0.43079347459403744 0.06213932952311195    6.93 0.00000000579 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4 on 53 degrees of freedom
Multiple R-squared:  0.889,	Adjusted R-squared:  0.862 
F-statistic: 32.7 on 13 and 53 DF,  p-value: <0.0000000000000002
```

3.7.2.1. Which variable is the most influential in terms of predicting revenue?

____