Regression in Business Analytics

Ahmed Yahya Khaled

6/27/2020

Loading Libraries

```
library(tidyverse)
library(plot3D)
```

Problem 1

The Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 (S&P 500) indexes are used as measures of overall movement in the stock market. The DJIA is based on the price movements of 30 large companies; the S&P 500 is an index composed of 500 stocks. Some say the S&P 500 is a better measure of stock market performance because it is broader based. The closing price for the DJIA and the S&P 500 for 15 weeks, beginning with January 6, 2012, follow (Barron's Web site, April 17, 2012).

data : data1.csv

Import Data

```
# set working directory
d1 <- read.csv("data1.csv", header = T)
#View(d1)
head(d1)
```

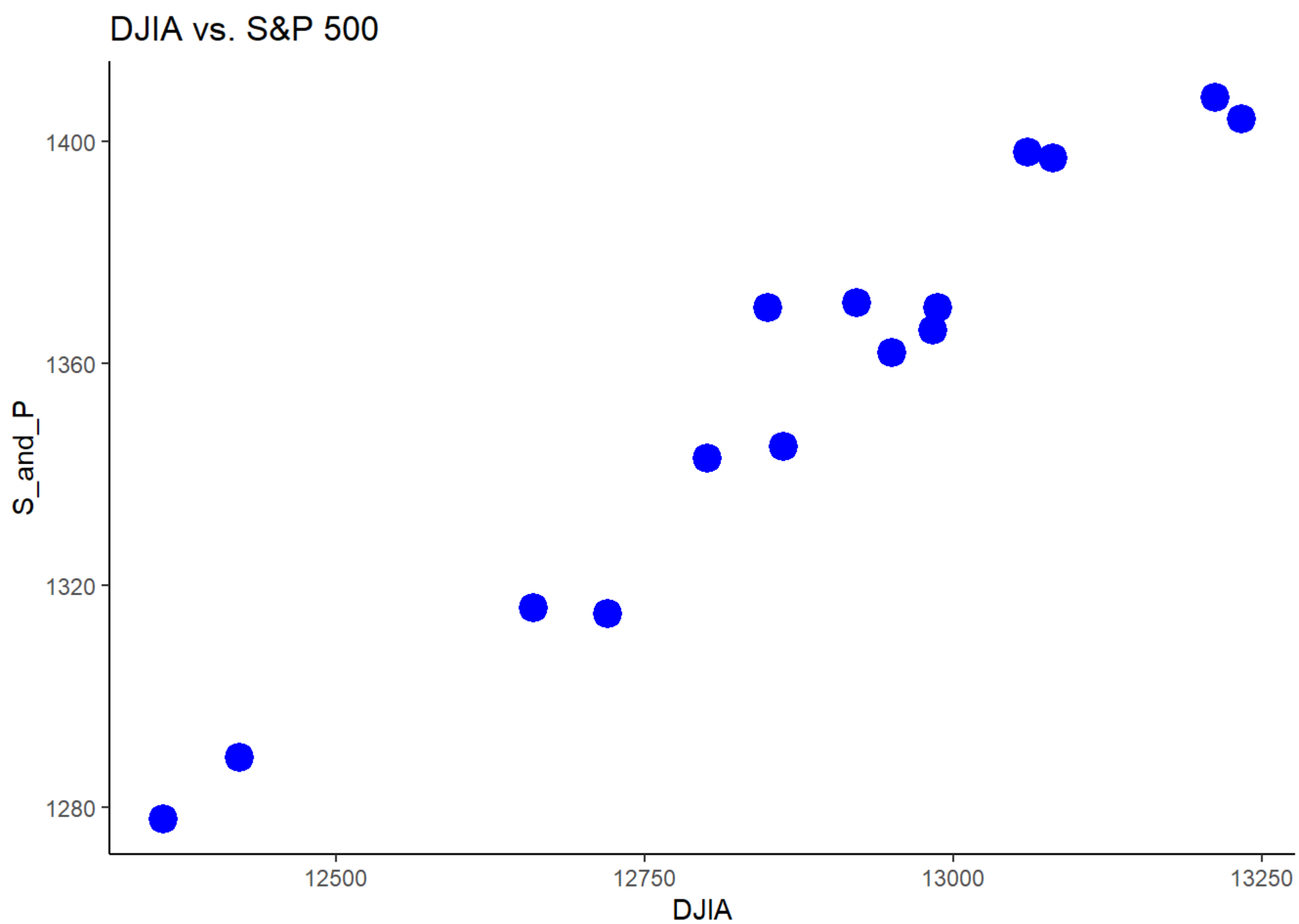
```
##          Date  DJIA S_and_P
## 1   January 6 12360    1278
## 2  January 13 12422    1289
## 3  January 20 12720    1315
## 4  January 27 12660    1316
## 5  February 3 12862    1345
## 6 February 10 12801    1343
```

```
glimpse(d1)
```

```
## Observations: 15
## Variables: 3
## $ Date    <fct> January 6, January 13, January 20, January 27, February 3, ...
## $ DJIA    <int> 12360, 12422, 12720, 12660, 12862, 12801, 12950, 12983, 129...
## $ S_and_P <int> 1278, 1289, 1315, 1316, 1345, 1343, 1362, 1366, 1370, 1371,...
```

EDA

```
ggplot(d1)+
  geom_point(aes(x = DJIA, y = S_and_P), stroke = 3, color = 'blue')+
  ggtitle("DJIA vs. S&P 500")+
  theme_classic()
```



1.a)

Develop an estimated regression equation showing how S&P 500 is related to DJIA. What is the estimated regression model?

```
linReg1 <- lm(S_and_P ~  DJIA , d1)
linReg1$coefficients
```

```
##  (Intercept)         DJIA
## -666.5546463    0.1570681
```

Estimated Regression equation

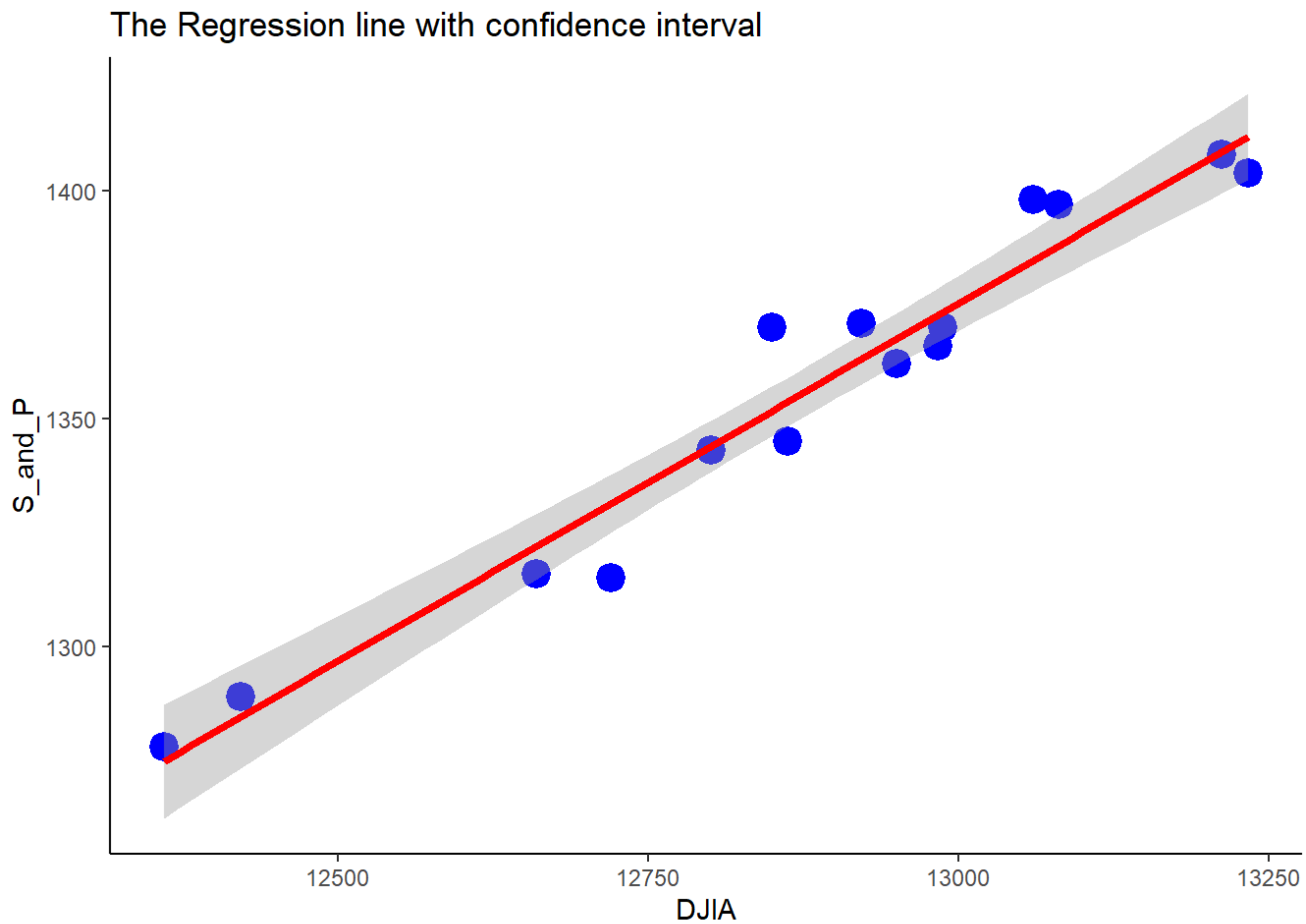
S_and_P.hat = (-666.5546463) + 0.1570681(DJIA)

Estimated regression model

```
summary(linReg1)
```

```
##
## Call:
## lm(formula = S_and_P ~ DJIA, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.352  -6.294  -1.074   6.188  18.230
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -666.55465  131.00208  -5.088 0.000208 ***
## DJIA           0.15707    0.01017  15.438 9.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.638 on 13 degrees of freedom
## Multiple R-squared:  0.9483, Adjusted R-squared:  0.9443
## F-statistic: 238.3 on 1 and 13 DF,  p-value: 9.676e-10
```

```
ggplot(d1)+
  geom_point(aes(x = DJIA, y = S_and_P), stroke = 3, color = 'blue')+
  # geom_abline(intercept = linReg1$coefficients[1], slope =   linReg1$coefficients[2],lwd = 1.
5,color = 'red')+
  geom_smooth(aes(x = DJIA, y = S_and_P), method = 'lm', lwd = 1.5, color = 'red')+
  ggtitle("The Regression line with confidence interval")+
  theme_classic()
```



1.b)

What is the 95 percent confidence interval for the regression parameter b1? Based on this interval, what conclusion can you make about the hypotheses that the regression parameter b1 is equal to zero?

```
confint(linReg1, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) -949.567444 -383.5418489
## DJIA           0.135088    0.1790482
```

The 95 percent confidence interval for the regression parameter b1

[0.135 , 0.179]

When the regression parameter b1(slope) is equal to zero, we fail to reject the **Null Hypothesis** that there is no linear relationship between S_and_P and DJIA.

1.c)

What is the 95 percent confidence interval for the regression parameter b0? Based on this interval, what conclusion can you make about the hypotheses that the regression parameter b0 is equal to zero?

```
confint(linReg1, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) -949.567444 -383.5418489
## DJIA           0.135088    0.1790482
```

The 95 percent confidence interval for the regression parameter b0

[-949.567 , -383.542]

When the regression parameter b0(intercept) is equal to zero, we can conclude that S_and_P and DJIA are same in terms of the measure for stock market performance

1.d)

Suppose that the closing price for the DJIA is 13,500. Estimate the closing price for the S&P 500.

```
S_and_P.pred <- predict(linReg1, newdata = data.frame(DJIA = 13500))
S_and_P.pred
```

```
##        1
## 1453.865
```

The estimated closing price of S&P 500 is **1453.865**

1.e)

Should we be concerned that the DJIA value of 13,500 used to predict the S&P 500 value we have just calculated is beyond the range of the DJIA used to develop the estimated regression equation?

It is one of the benefits of the regression analysis to be able to predict the dependent variable for a independent variable value which is beyond the range of the value with which the model was trained.

Problem 2

Dixie Showtime Movie Theaters, Inc., owns and operates a chain of cinemas in several markets in the southern United States. The owners would like to estimate weekly gross revenue as a function of advertising expenditures. Data for a sample of eight markets for a recent week follow.

data : data2.csv

set working directory

Import Data

```
d2 <- read.csv("data2.csv", header = T)
#View(d2)
head(d2)
```

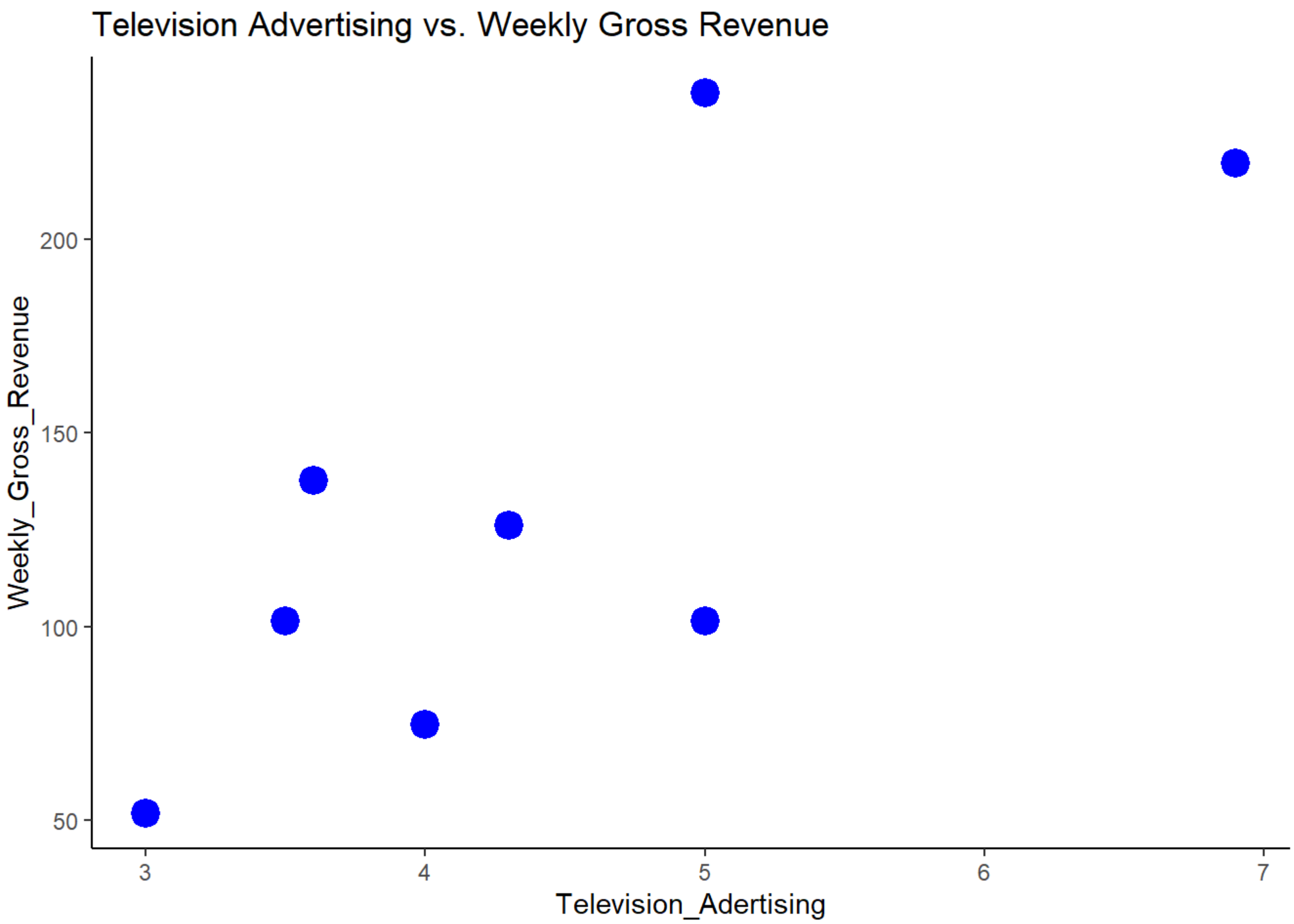
```
##        Market Weekly_Gross_Revenue Television_Adertising Newspaper_Advertising
## 1      Mobile                101.3                   5.0                   1.5
## 2  Shreveport                 51.9                   3.0                   3.0
## 3     Jackson                 74.8                   4.0                   1.5
## 4  Birmingham                126.2                   4.3                   4.3
## 5 Little Rock                137.8                   3.6                   4.0
## 6      Biloxi                101.4                   3.5                   2.3
```

```
glimpse(d2)
```

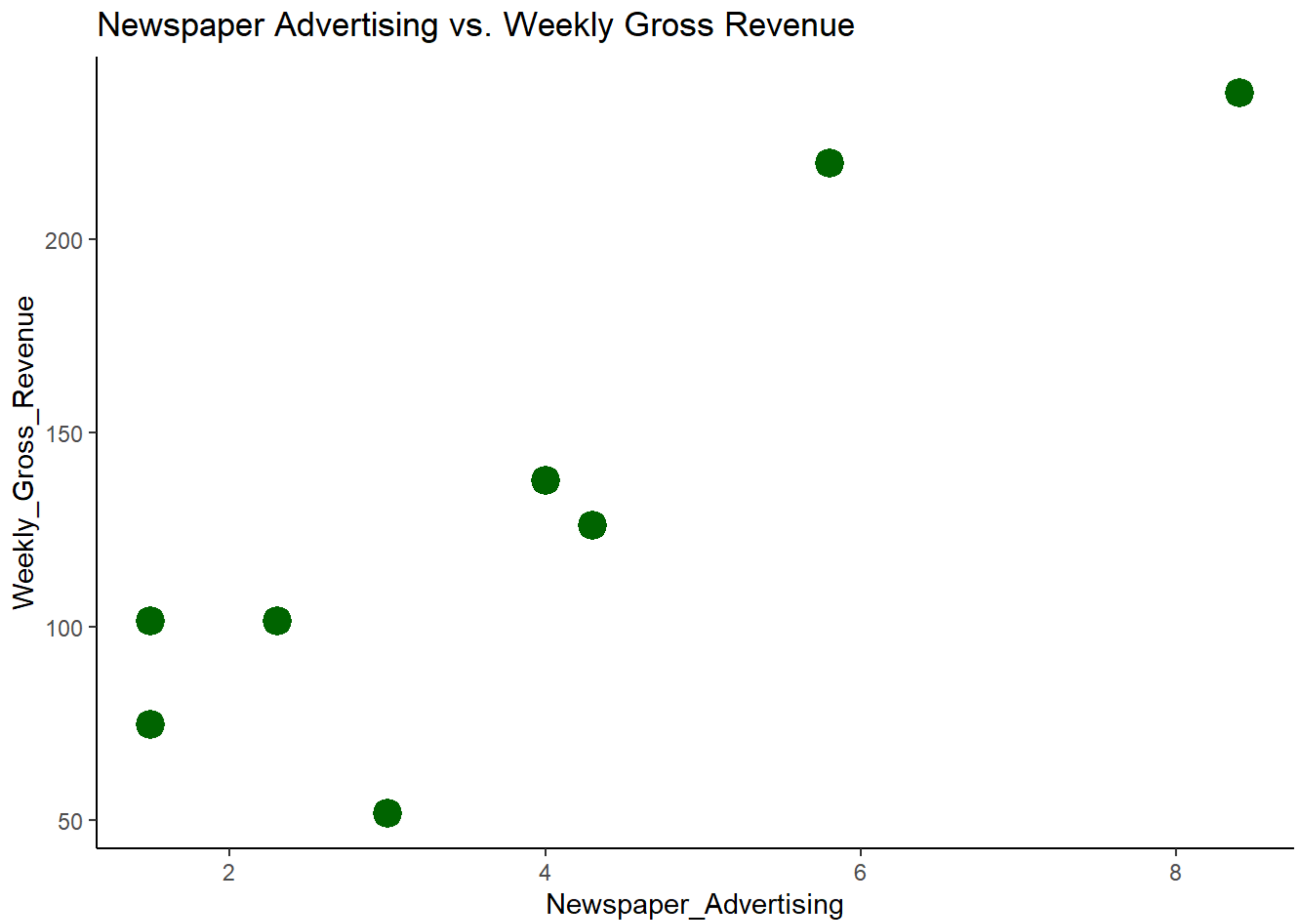
```
## Observations: 8
## Variables: 4
## $ Market                <fct> Mobile, Shreveport, Jackson, Birmingham, Litt...
## $ Weekly_Gross_Revenue  <dbl> 101.3, 51.9, 74.8, 126.2, 137.8, 101.4, 237.8...
## $ Television_Adertising <dbl> 5.0, 3.0, 4.0, 4.3, 3.6, 3.5, 5.0, 6.9
## $ Newspaper_Advertising <dbl> 1.5, 3.0, 1.5, 4.3, 4.0, 2.3, 8.4, 5.8
```

EDA

```
ggplot(d2)+
  geom_point(aes(x = Television_Adertising,
                 y = Weekly_Gross_Revenue),
             stroke = 3, color = 'blue')+
  ggtitle("Television Advertising vs. Weekly Gross Revenue") +
  theme_classic()
```




```
ggplot(d2)+
  geom_point(aes(x = Newspaper_Advertising,
                y = Weekly_Gross_Revenue),
            stroke = 3, color = 'darkgreen')+
  ggtitle("Newspaper Advertising vs. Weekly Gross Revenue") +
  theme_classic()
```



2.a)

Develop an estimated regression equation with the amount of television advertising as the independent variable. Test for a significant relationship between television advertising and weekly gross revenue at the 0.05 level of significance. What is the interpretation of this relationship?

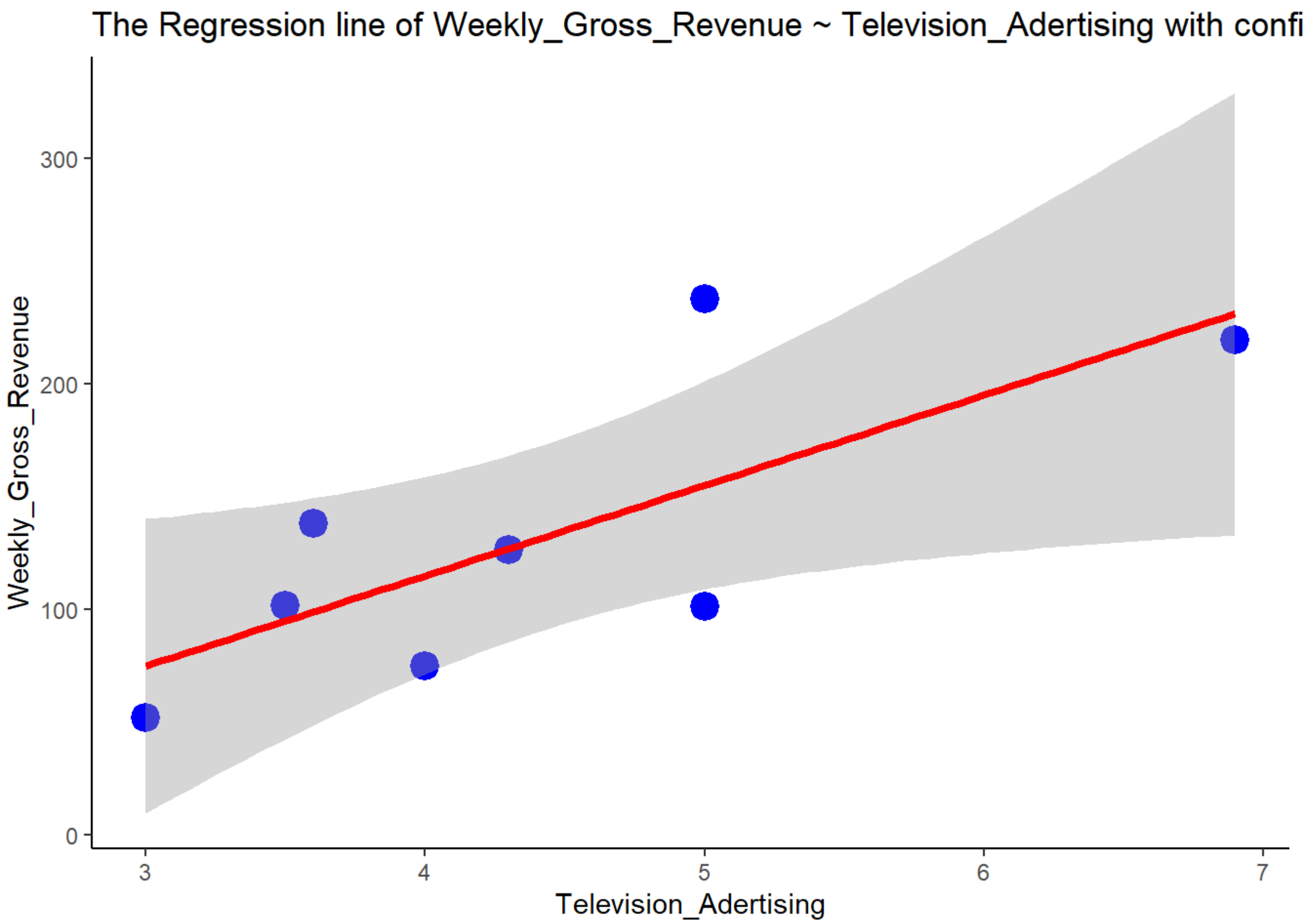
```
linReg2 <- lm(Weekly_Gross_Revenue ~  Television_Adertising , d2)
linReg2$coefficients
```

```
##          (Intercept) Television_Adertising
##            -45.43235              40.06399
```

Estimated Regression equation

Weekly_Gross_Revenue.hat = (-45.43235) + 40.06399(Television_Adertising)

```
ggplot(d2)+
  geom_point(aes(x = Television_Adertising, y = Weekly_Gross_Revenue),
             stroke = 3, color = 'blue')+
  geom_smooth(aes(x = Television_Adertising, y = Weekly_Gross_Revenue),
              method = 'lm', lwd = 1.5, color = 'red')+
  ggtitle("The Regression line of Weekly_Gross_Revenue ~ Television_Adertising with confidence i
nterval")+
  theme_classic()
```




```
summary(linReg2)
```

```
##
## Call:
## lm(formula = Weekly_Gross_Revenue ~ Television_Adertising, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.588 -27.151  -6.026  14.707  82.912
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -45.43      66.75  -0.681   0.5215
## Television_Adertising    40.06      14.64   2.737   0.0339 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.55 on 6 degrees of freedom
## Multiple R-squared:  0.5552, Adjusted R-squared:  0.481
## F-statistic: 7.489 on 1 and 6 DF,  p-value: 0.03389
```

Here, p-value = 0.03389
alpha = 0.05

Ho = p-value > alpha
Ha = p-value <= alpha

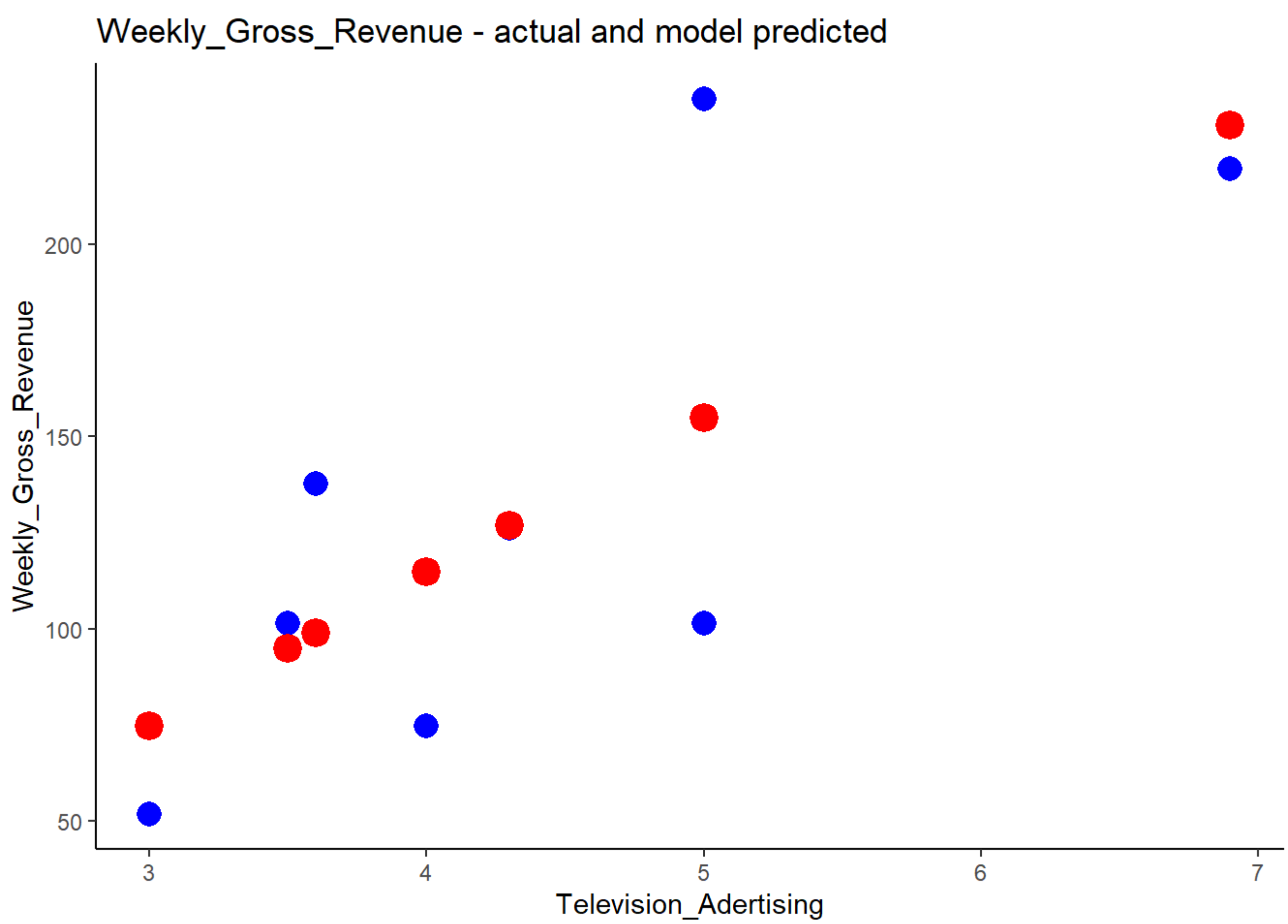
as 0.03389 < 0.05 , we can reject the null hypothesis and conlcude that **there is a significant relationship between television advertising and weekly gross revenue.**

2.b)

How much of the variation in the sample values of weekly gross revenue does the model in part a explain?

R-Squared = 0.5552, so *55.52%* of the variation in the sample values of weekly gross revenue does the model in part a explain.

```
d2$ypredtv <- predict(linReg2, newdata = d2)
ggplot(d2)+
  geom_point(aes(x = Television_Adertising, y = Weekly_Gross_Revenue),
             stroke = 2.5, color = 'blue') +
  geom_point(aes(x = Television_Adertising, y = ypredtv),
             stroke = 3, color = 'red')+
  ggtitle('Weekly_Gross_Revenue - actual and model predicted')+
  theme_classic()
```




```
d2 <- d2[, 1:4]
```

2.c)

Develop an estimated regression equation with both television advertising and newspaper advertising as the independent variables. Is the overall regression statistically significant at the 0.05 level of significance? What is the interpretation of this relationship?

```
linReg3 <- lm(Weekly_Gross_Revenue ~  Television_Adertising +
                Newspaper_Advertising , d2)
linReg3$coefficients
```

```
##           (Intercept) Television_Adertising Newspaper_Advertising
##             -42.56959              22.40224              19.49863
```

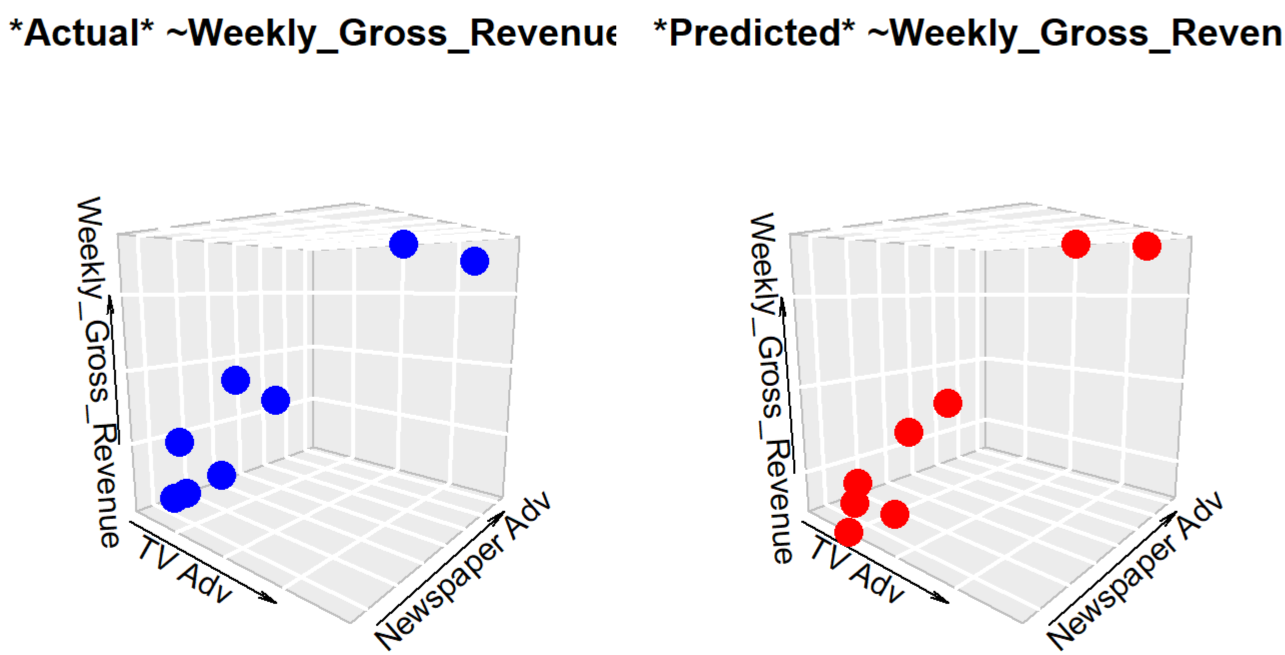
Estimated Regression equation

Weekly_Gross_Revenue.hat = (-42.56959) + 22.40224(Television_Adertising) + 19.49863(Newspaper_Advertising)

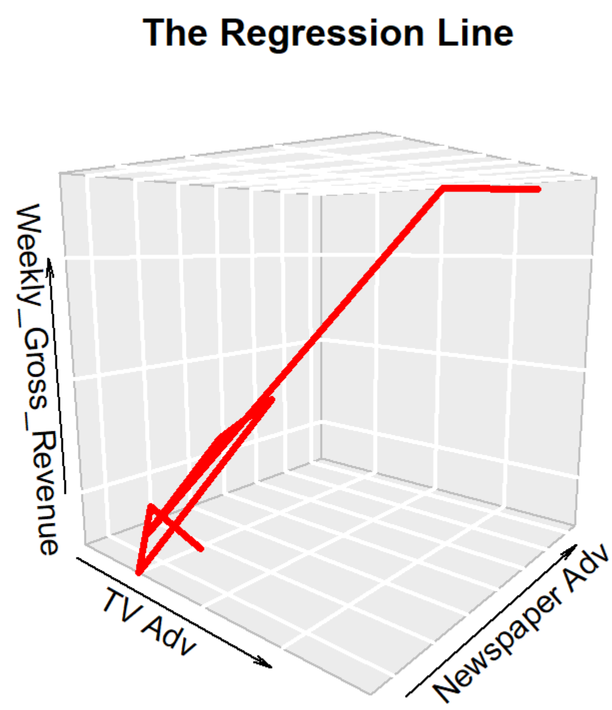
```
d2$ypred_tv_np <- predict(linReg3, newdata = d2)
x <- d2$Television_Adertising
y <- d2$Newspaper_Advertising
za <- d2$Weekly_Gross_Revenue
zb <- d2$ypred_tv_np

par(mfrow = c(1, 2))
scatter3D(x, y, za, colvar = NULL, bty = "g",
          col = "blue", pch = 16, cex = 2,
          phi = 10,
          xlab = 'TV Adv',
          ylab = 'Newspaper Adv',
          zlab = 'Weekly_Gross_Revenue',
          main = '*Actual* ~Weekly_Gross_Revenue')


scatter3D(x, y, zb, colvar = NULL, bty = "g",
          col = "red", pch = 16, cex = 2,
          phi = 10,
          xlab = 'TV Adv',
          ylab = 'Newspaper Adv',
          zlab = 'Weekly_Gross_Revenue',
          main = '*Predicted* ~Weekly_Gross_Revenue')
```




```
scatter3D(x, y, zb, colvar = NULL, bty = "g", type = "l",
          col = "red", lwd = 3.5,
          phi = 10,
          xlab = 'TV Adv',
          ylab = 'Newspaper Adv',
          zlab = 'Weekly_Gross_Revenue',
          main = 'The Regression Line')
```




```
d2 <- d2[, 1:4]
```

```
summary(linReg3)
```

```
##
## Call:
## lm(formula = Weekly_Gross_Revenue ~ Television_Adertising + Newspaper_Advertising,
##     data = d2)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##   2.610 -31.233  -1.487 -11.404  21.727  20.715   4.570  -5.498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -42.570     28.547  -1.491  0.19611
## Television_Adertising    22.402      7.099   3.156  0.02522 *
## Newspaper_Advertising    19.499      3.697   5.274  0.00326 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.33 on 5 degrees of freedom
## Multiple R-squared:  0.9322, Adjusted R-squared:  0.9051
## F-statistic: 34.39 on 2 and 5 DF,  p-value: 0.001196
```

Here, p-value = 0.001196
alpha = 0.05

Ho = p-value > alpha
Ha = p-value <= alpha

as 0.001196 < 0.05 , we can reject the null hypothesis and conlcude that **there is a significant relationship between television advertising and weekly gross revenue.**
and if we compare the p-value, adding the second independent variable results in lower p-value thus generated a better regression model.

2.d)

How much of the variation in the sample values of weekly gross revenue does the model in part c explain?

R-Squared = 0.9322, so *93.22%* of the variation in the sample values of weekly gross revenue does the model in part c explain.

2.e)

Given the results in parts a and c, what should your next step be? Explain.

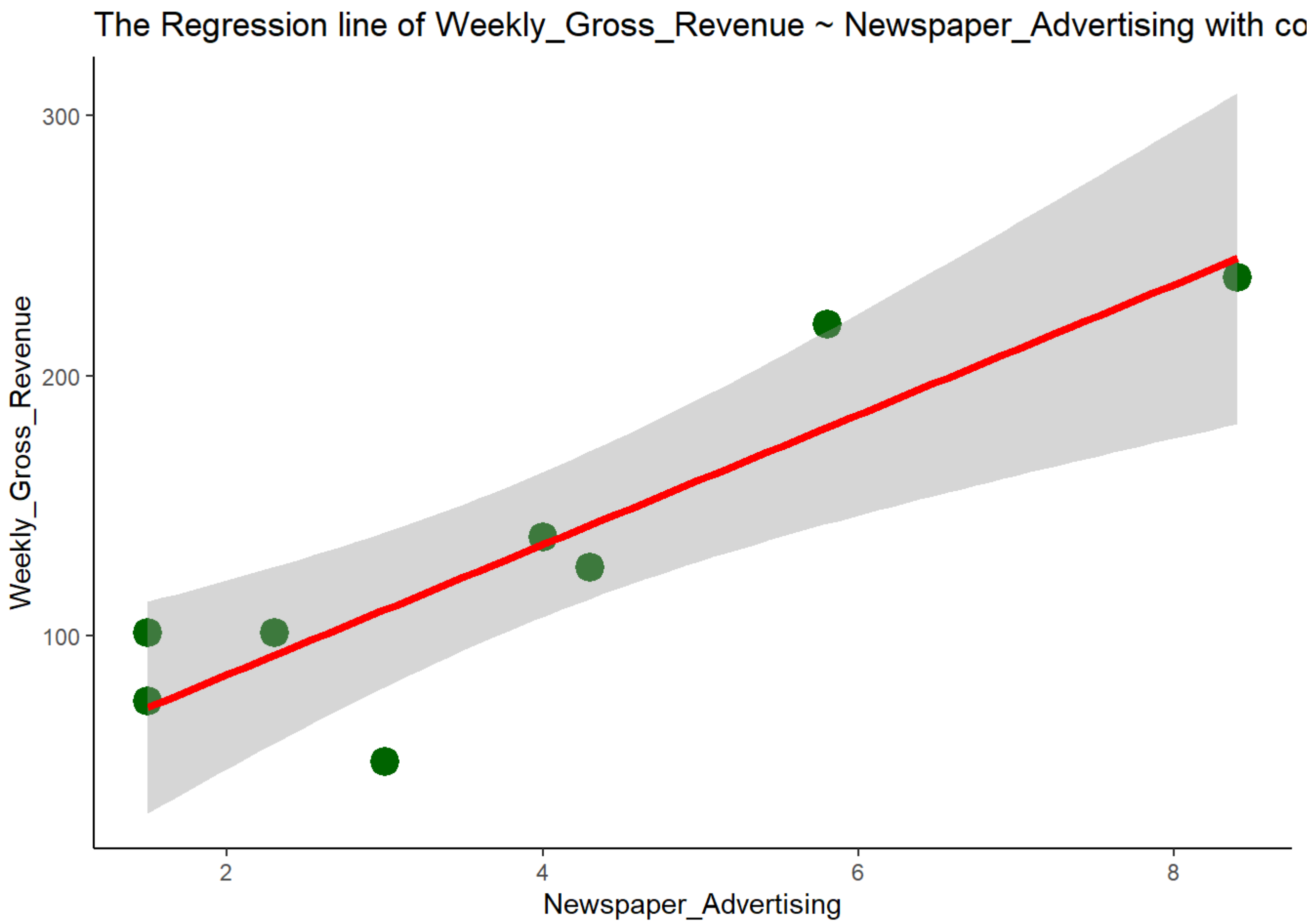
The next step can be to check the relation of weekly gross revenue with the newspaper advertising only

```
linReg4 <- lm(Weekly_Gross_Revenue ~  Newspaper_Advertising , d2)
summary(linReg4)
```

```
##
## Call:
## lm(formula = Weekly_Gross_Revenue ~ Newspaper_Advertising, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.199  -9.580   2.451  13.777  39.497
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             35.095     22.836   1.537  0.17524
## Newspaper_Advertising   25.001      5.147   4.858  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.1 on 6 degrees of freedom
## Multiple R-squared:  0.7973, Adjusted R-squared:  0.7635
## F-statistic:  23.6 on 1 and 6 DF,  p-value: 0.00283
```

R-squared = 0.7973 ; p-value = 0.00283

```
ggplot(d2)+
  geom_point(aes(x = Newspaper_Advertising, y = Weekly_Gross_Revenue),
             stroke = 3, color = 'darkgreen')+
  geom_smooth(aes(x = Newspaper_Advertising, y = Weekly_Gross_Revenue),
              method = 'lm', lwd = 1.5, color = 'red')+
  ggtitle("The Regression line of Weekly_Gross_Revenue ~ Newspaper_Advertising with confidence i
nterval")+
  theme_classic()
```



2.f)

What are the managerial implications of these results?

- *Newspaper Advertising has more impact over Weekly Gross Revenue*
- *Television and Newspaper Advertising combinely has very good impact over Revenue*
- *Weekly Gross Revenue can be predicted with any combination of advertising expense category*
- *In case of stringent advertising budget more portion can be allocated to Newspaper advertising than Television advertising*