Untitled

2024-12-01

1.) (Use R) Consider the dataset "Homework 6 data.xlsx." It consists of 5 randomly selected student's scores on Test 1 and Test 2 in my introductory statistics course. We want to answer 2 questions:

```r
setwd("~/Desktop/Personal_save/Stat_405_Module_14/Module_14_Homework")
#setwd("C:/Users/jake pc/Desktop/Personal_save/Stat_405_Module_14/Module_14_Homework")
HW_6 <- read.csv(file="Homework_6.csv",header=TRUE)
HW_6
```

```
##   Student Test.1 Test.2
## 1       1     82     90
## 2       2     74     87
## 3       3     65     68
## 4       4     62     83
## 5       5     88     92
```

a. First, we want to see if there is a difference in the two tests. Paired two-tailed t-test

```r
t.test(HW_6$Test.1, HW_6$Test.2, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  HW_6$Test.1 and HW_6$Test.2
## t = -2.9629, df = 4, p-value = 0.04143
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -18.9832784  -0.6167216
## sample estimates:
## mean difference
##            -9.8
```

The p-value is less than 0.05, reject the null hypothesis that the means of Test 1 and Test 2 are equal.

b. Second, we want to see if there was improvement over the course of the semester. H0: Test1 - Test2 < 0

```r
t.test(HW_6$Test.1, HW_6$Test.2, paired = TRUE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  HW_6$Test.1 and HW_6$Test.2
## t = -2.9629, df = 4, p-value = 0.02072
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##       -Inf -2.748774
## sample estimates:
## mean difference
##            -9.8
```

Reject the null hypothesis that the mean difference of Test 1 minus Test 2 is equal to zero. Tentatively conclude that the mean difference of test 1 minus test 2 is less than zero, and therefore that the the grades of the second test were greater (better) than the first.

2.) (Use R) The data called "plasma" from Anderson et al. (1981) consists of measurements of plasma concentrations in micromoles/liter from 10 subjects at times of 8 am, 11am, 2pm, 5 pm, and 8 pm. Analyze the data in a 1-way ANOVA model choosing time as factor.

```
plasma <- read.csv(file="plasma.csv",header=TRUE)
plasma$time <- factor(plasma$time,levels=c("8am", "11am", "2pm", "5pm", "8pm"),
                       labels = c("8am", "11am", "2pm", "5pm", "8pm"))

plasma
```

```
##    subjects time plasma
## 1         1  8am     93
## 2         2  8am    116
## 3         3  8am    125
## 4         4  8am    144
## 5         5  8am    105
## 6         6  8am    109
## 7         7  8am     89
## 8         8  8am    116
## 9         9  8am    151
## 10       10  8am    137
## 11        1 11am    121
## 12        2 11am    135
## 13        3 11am    137
## 14        4 11am    173
## 15        5 11am    119
## 16        6 11am     83
## 17        7 11am     95
## 18        8 11am    128
## 19        9 11am    149
## 20       10 11am    139
## 21        1  2pm    112
## 22        2  2pm    114
## 23        3  2pm    119
## 24        4  2pm    148
## 25        5  2pm    125
## 26        6  2pm    109
## 27        7  2pm     88
## 28        8  2pm    122
## 29        9  2pm    141
## 30       10  2pm    125
## 31        1  5pm    117
## 32        2  5pm     98
## 33        3  5pm    105
## 34        4  5pm    124
## 35        5  5pm     91
## 36        6  5pm     80
## 37        7  5pm     91
## 38        8  5pm    107
## 39        9  5pm    126
## 40       10  5pm    109
```

```
## 41        1  8pm    121
## 42        2  8pm    135
## 43        3  8pm    102
## 44        4  8pm    122
## 45        5  8pm    133
## 46        6  8pm    104
## 47        7  8pm    116
## 48        8  8pm    119
## 49        9  8pm    138
## 50       10  8pm    107
```

```
plasma_model <- lm(plasma ~ time, data = plasma)
anova(plasma_model)
```

```
## Analysis of Variance Table
##
## Response: plasma
##           Df  Sum Sq Mean Sq F value Pr(>F)
## time       4  2803.9  700.98  1.9838 0.1132
## Residuals 45 15901.2  353.36
```

The p-value from this test is greater than 0.05, we therefore fail to reject the null hypothesis that blood plasma levels are not different at different times.

3.) Two friends play a computer game and each of them repeats the same level 10 times. The scores obtained are:

```
scores <- read.table(file="scores.txt",header=TRUE)
```

```
## Warning in read.table(file = "scores.txt", header = TRUE): incomplete final
## line found by readTableHeader on 'scores.txt'
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
long <- scores %>%
            pivot_longer(cols = X1:X10, names_to = "trials", values_to = "scores") %>%
                select(-trials)
long$ID <- factor(long$ID, levels = c("Player1","Player2"), labels = c("Player1","Player2"))

write.csv(long, file="long.csv")
```

a. Player 2 insists that he is the better player and suggests to compare their mean performance. Use a t-test to test whether there is a difference between their mean performance (alpha = 0.05).

   We are testing for difference of mean on two separate individuals —> 2 sample - unpaired - two sided t-test

```r
scores <- t(as.matrix(scores))
colnames(scores) <- scores[1,]
scores <- as_tibble(scores[-1,])
library(dplyr)

scores <- scores %>%
  mutate(across(everything(), as.numeric))

scores
```

```
## # A tibble: 10 x 2
##    Player1 Player2
##      <dbl>   <dbl>
##  1      91     261
##  2     101      47
##  3     112      40
##  4      99      29
##  5     108      64
##  6      88       6
##  7      99      87
##  8     105      47
##  9     111      98
## 10     104     351
```

```r
shapiro.test(scores$Player1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  scores$Player1
## W = 0.94628, p-value = 0.6247
```

```r
shapiro.test(scores$Player2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  scores$Player2
## W = 0.75335, p-value = 0.00392
```

```r
t.test(long$scores ~ long$ID)
```

```
##
##  Welch Two Sample t-test
##
## data:  long$scores by long$ID
## t = -0.033723, df = 9.0898, p-value = 0.9738
## alternative hypothesis: true difference in means between group Player1 and group Player2 is not equal
## 95 percent confidence interval:
##  -81.57617  79.17617
## sample estimates:
## mean in group Player1 mean in group Player2
##                 101.8                 103.0
```

reject the null hypothesis that the players 2 scores are normally distributed, therefore player 1 and 2 could never have equal variances test for difference of scores –> two-tailed

test results in a p-value of 0.9738, fail to reject the null hypothesis that the difference in means is equal to zero.

b. Player 1 insists that he is the better player. He proposes to use the Wilcoxon rank-sum test for the comparison. What are the results (alpha = 0.05)?

```r
wilcox.test(long$scores ~ long$ID, alternative = "greater")
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  long$scores by long$ID
## W = 78, p-value = 0.01875
## alternative hypothesis: true location shift is greater than 0
```

The resulting p-value is less than 0.05, we therefore reject the null hypothesis that the true location shift is equal to zero and conclude that player 1 is better than player 2.

4.) (Use R)
A random sample of 90 adults is classified according to gender and the number of hours of television watched during a week:

Use a 0.01 level of significance and test the hypothesis that the time spent watching television is independent of whether the viewer is male or female.

```r
table <- matrix(data=c(15,29,27,19),nrow=2,ncol=2,byrow=TRUE,dimnames = list(c("Over 25 hours", "Under 2
table <- t(table)
table
```

```
##        Over 25 hours Under 25 hours
## Male              15             27
## Female            29             19
```

```r
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 4.5262, df = 1, p-value = 0.03338
```

The p-value obtained is 0.03338 which is greater than 0.01, we therefore fail to reject the null hypothesis that time spent watching television is independent of whether the viewer is male or female.

5.) (Use R)

The data set named "Movies" contains a random sample of 35 movies released in 2008. This sample was collected from the Internet Movie Database (IMDb). The goal of this problem is to explore if the information available soon after a movie's theatrical release can successfully predict total revenue. All dollar amounts (i.e., variables Budget, Opening, and USRevenue) are measured in millions of dollars. Consider three explanatory variables:

- The movie's budget (variable named Budget).
- Opening weekend revenue (variable named Opening).

- Number of theaters showing the movie (variable named Theaters).

```r
Movies <- read.csv(file="Movies.csv",header=TRUE)
Movies
```

```
##                                  Title Rating     Genre Budget USRevenue
## 1            Madagascar: Escape 2 Africa     PG Animation  150.0     180.0
## 2                       Sex and the City      R    Comedy   65.0     152.6
## 3                              The Ruins      R    Horror    8.0      17.4
## 4                              Stop-Loss      R     Drama   25.0      10.9
## 5    The Curious Case of Benjamin Button  PG-13     Drama  150.0     127.5
## 6                                Redbelt      R    Action    7.0       2.3
## 7                  The Secret Life of Bees  PG-13     Drama   11.0      37.8
## 8                          Kung Fu Panda     PG Animation  130.0     215.4
## 9                          The Happening      R     Drama   60.0      64.5
## 10            Zach and Miri Make a Porno      R    Comedy   24.0      31.5
## 11                        The Strangers      R    Horror   10.0      52.5
## 12                           Prom Night  PG-13    Horror   20.0      43.8
## 13                       The Dark Knight  PG-13    Action  185.0     533.3
## 14                            Baby Mama  PG-13    Comedy   30.0      60.3
## 15                               Wanted      R    Action   75.0     134.3
## 16                           Changeling      R     Drama   55.0      35.7
## 17                              Yes Man  PG-13    Comedy   70.0      97.7
## 18                          The Express     PG     Drama   40.0       9.6
## 19                                   W.  PG-13     Drama   25.1      25.5
## 20 The Mummy: Tomb of the Dragon Emporer  PG-13    Action  145.0     102.2
## 21                            Eagle Eye  PG-13    Action   80.0     101.1
## 22                   Burn After Reading      R    Comedy   37.0      60.3
## 23                                Saw V      R    Horror   10.8      56.7
## 24                  Miracle and St Anna      R    Action   45.0       7.9
## 25          The Day the Earth Stood Still  PG-13     Drama   80.0      79.4
## 26                       Be Kind Rewind  PG-13    Comedy   20.0      11.2
## 27                               Jumper  PG-13    Action   85.0      80.2
## 28                              Hancock  PG-13    Action  150.0     227.9
## 29                          Speed Racer     PG    Action  120.0      43.9
## 30                              The Eye      R     Drama   12.0      31.4
## 31                           Death Race      R    Action   45.0      36.1
## 32                              College      R    Comedy    6.5       4.7
## 33                            Blindness      R     Drama   25.0       3.1
## 34                             Iron Man  PG-13    Action  140.0     318.3
## 35                    Lakeview Terrace  PG-13     Drama   22.0      39.3
##    Opening Theaters
## 1     63.1     4056
## 2     56.8     3285
## 3      8.0     2812
## 4      4.6     1291
## 5     26.9     2988
## 6      1.1     1379
## 7     10.5     1591
## 8     60.2     4114
## 9     30.5     2986
## 10    10.1     2735
## 11    21.0     2466
## 12    20.8     2700
## 13   158.4     4366
```

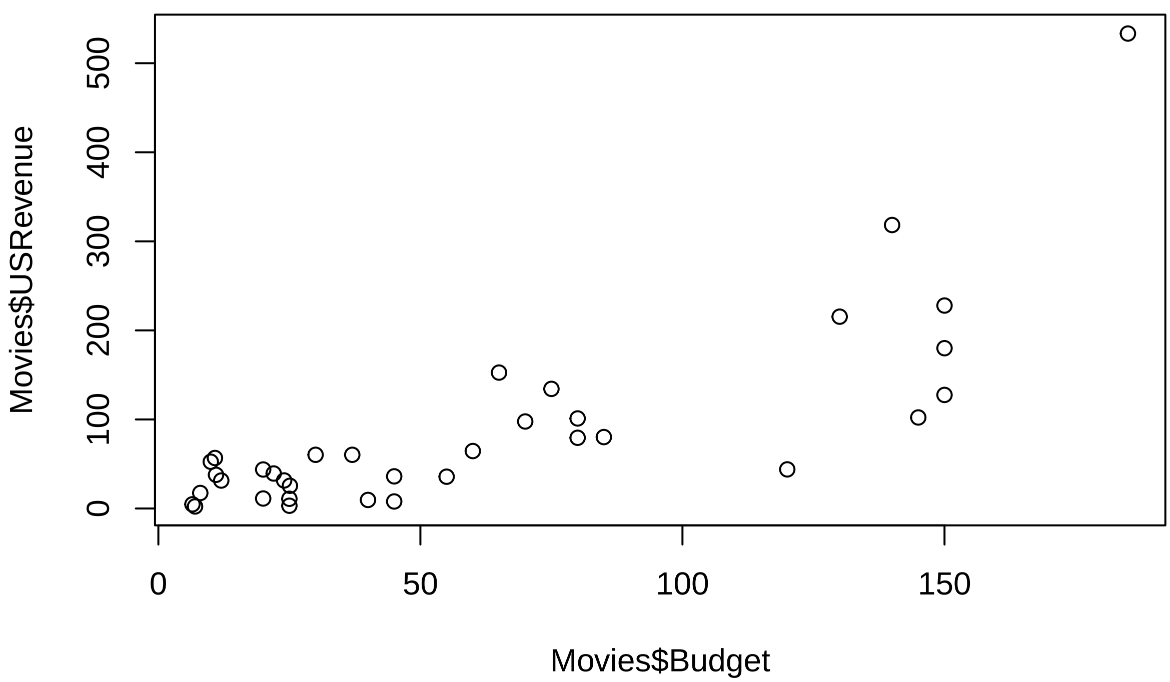
```
## 14     17.4      2543
## 15     50.9      3175
## 16     10.0      1850
## 17     18.3      3434
## 18      4.6      2808
## 19     10.5      2030
## 20     40.5      3760
## 21     29.2      3510
## 22     19.1      2651
## 23     30.1      3060
## 24      3.5      1185
## 25     30.5      3560
## 26      4.1       808
## 27     32.1      3428
## 28     62.6      3965
## 29     18.6      3606
## 30     12.4      2436
## 31     12.6      2532
## 32      2.6      2123
## 33      2.0      1690
## 34    102.1      4105
## 35     15.0      2464
```

This problem considers using each of these explanatory variables to attempt to predict a movie's total US revenue (variable named USRevenue).

a. Investigate the relationship between the explanatory variable Budget and response variable USRevenue by doing the following:

i) Make a scatterplot.

```
plot(x = Movies$Budget, y = Movies$USRevenue)
```



ii) Calculate the correlation coefficient.

```
cor(x = Movies$Budget, y = Movies$USRevenue)
```

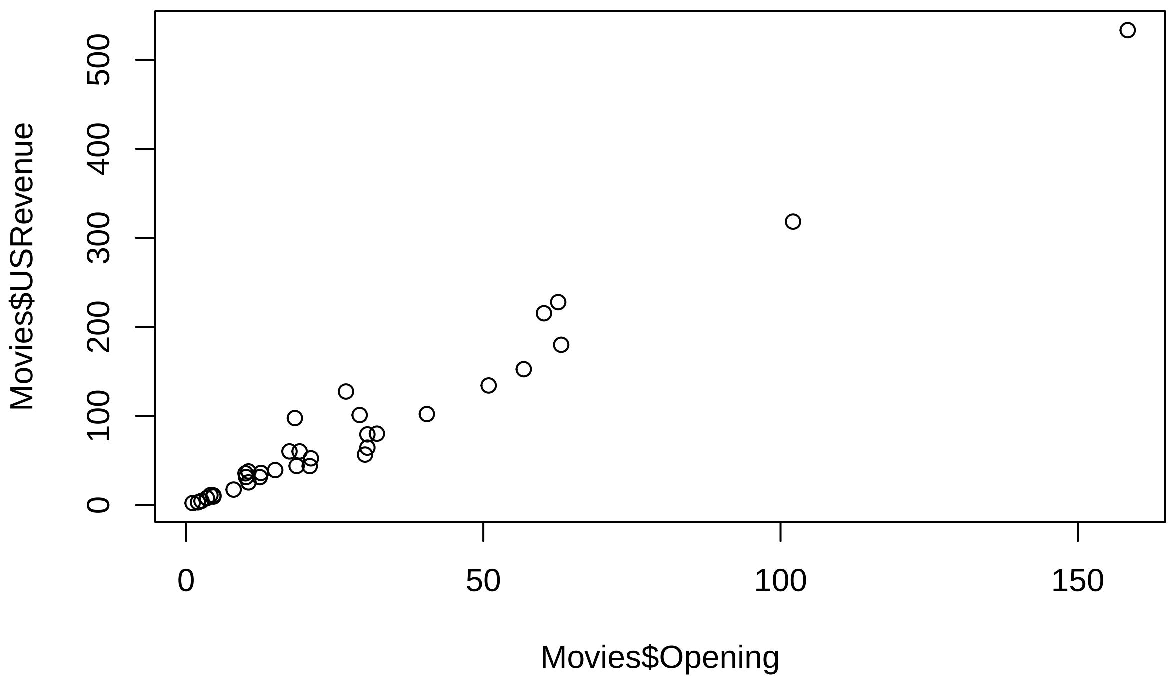
```
## [1] 0.7918636
```

iii) Interpret the scatterplot and correlation coefficient in terms of trend, strength, and shape.

The correlation coefficient is in the moderate strength range but just below the strong threshold of plus/minus 0.8, the trend is positive such that as budget increases US revenue increases. The scatter plot has a wedge shape such that as budget increases the USRevenue becomes more variable, this is a problem and will produce a "wedge shape" in the residuals plot.

b. Repeat part (a) for the explanatory variable Opening.

```
plot(x = Movies$Opening, y = Movies$USRevenue)
```



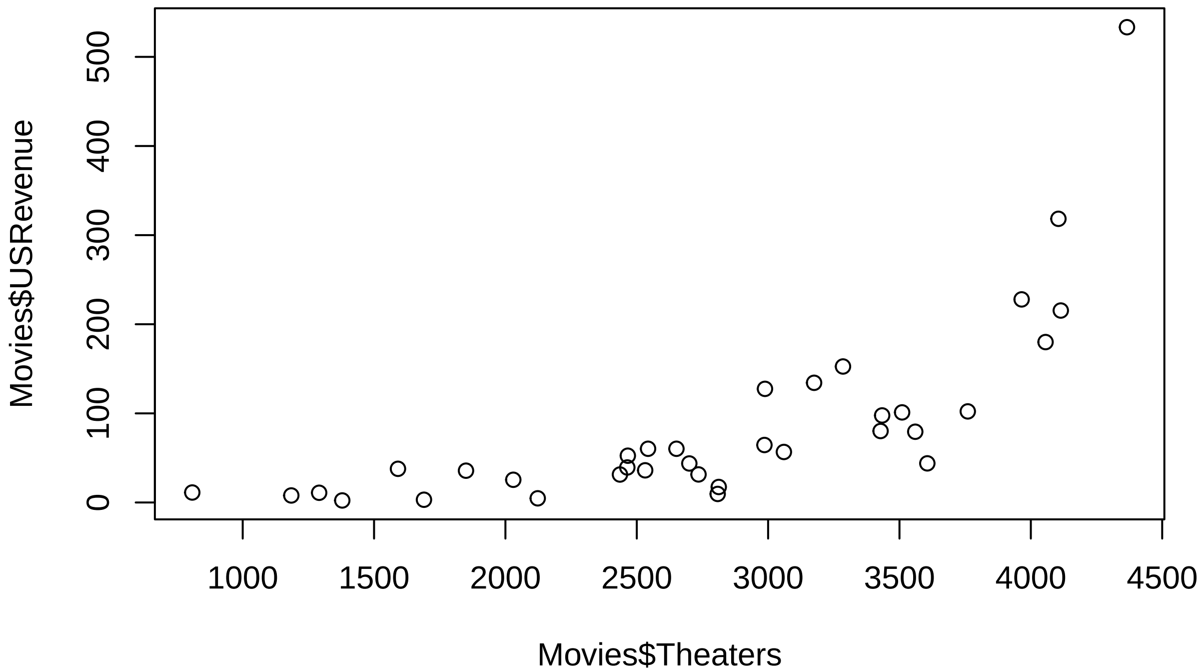

```
cor(x = Movies$Opening, y = Movies$USRevenue)
```

```
## [1] 0.9840177
```

The trend is positive such that as the variable opening increase so the variable USRevenue. The variables have a correlation coefficient of 0.9840177 which is classified as a strong correlation but it is almost perfect. The relationship between these variables has pattern other than a linear relationship between predictor and response, and has no disturbing patters of increasing variability as the predictor variable increases as the last one did.

c. Repeat part (a) for the explanatory variable Theaters.

```
plot(x = Movies$Theaters, y = Movies$USRevenue)
```




```
cor(x = Movies$Theaters, y = Movies$USRevenue)
```

```
## [1] 0.7153432
```

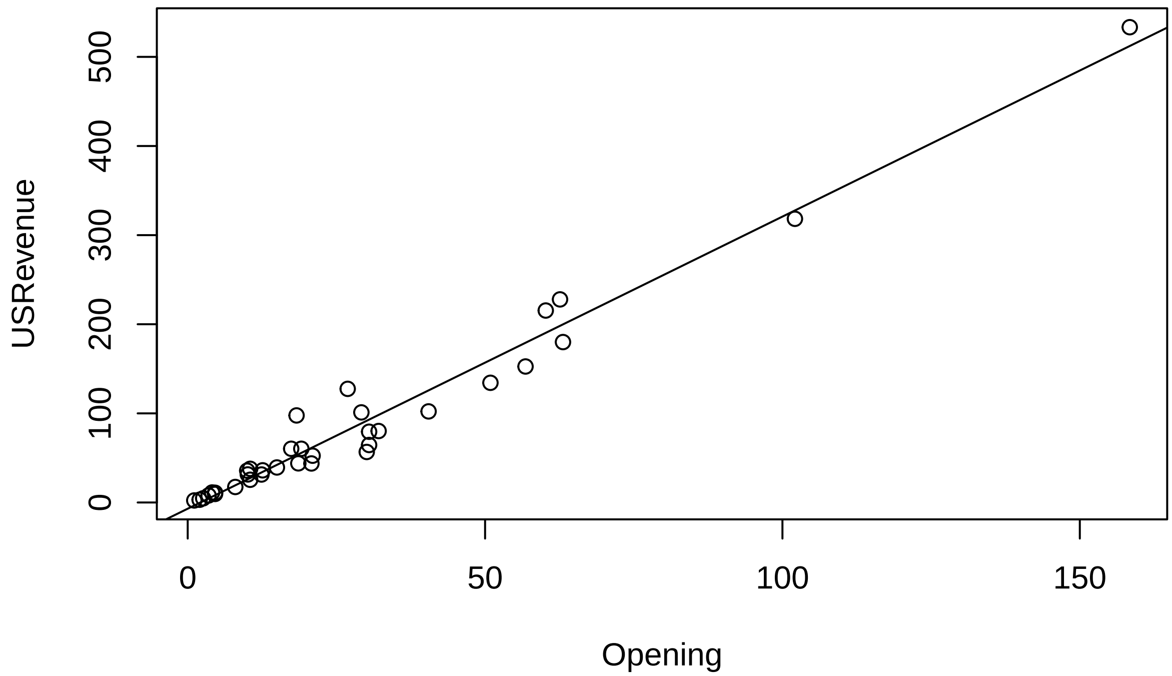
The trend is positive such that as variable theaters increases so does the variable USRevenue. The variables have a correlation coefficient of 0.7153432 which is classified as a moderate correlation. The relationship between these variables has a very obvious curvature which appears almost exponential.

d. Based on your findings in parts (a) through (c), which of the three explanatory variables would be most appropriate for predicting the response variable USRevenue? Justify your choice in a few sentences.

Opening is the best choice as predictor variable. The trend is positive such that as the variable opening increase so the variable USRevenue. The relationship between Opening and Us Revenue has a positive correlation coefficient that is almost perfect (.98) and the scatter plot indicates no pattern other than a positive linearlity. The variable Budget is completely unsuitable as the residuals of the fitted values would increase as budget increased. Theaters is unsuitable as it's relationship is non-linear.

e. For the "most appropriate" variable identified in part (d), run a Simple Linear Regression analysis.

```
Opening_Model <- lm(USRevenue ~ Opening, data=Movies)
plot(USRevenue ~ Opening, data=Movies)
intercept_slope <- coefficients(Opening_Model)
abline(a=intercept_slope[1],b=intercept_slope[2])
```




```r
summary(Opening_Model)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening, data = Movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.996 -11.855   1.763   7.771  46.293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.9619     4.3875  -1.587    0.122
## Opening       3.2777     0.1033  31.744   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.2 on 33 degrees of freedom
## Multiple R-squared:  0.9683, Adjusted R-squared:  0.9673
## F-statistic:  1008 on 1 and 33 DF,  p-value: < 2.2e-16
```

f. State the regression equation.

Predicted US Revenue = 3.277*Opening - 6.9619

g. Interpret the slope of the regression line (in context of this data set).

As Opening sales increases by 1 million, Predicted US Revenue will increase by 3.277 million.

h. Is it meaningful to interpret the y-intercept? Why or why not?

The y-intercept in this model is not meaningful as it's p-value from testing the null hypothesis that the intercept is equal to zero is greater than 0.05 and thus the null hypothesis that the true y-intercept is zero fails to be rejected.

i. State r-squared (i.e., the coefficient of determination) and explain what this value means (in context of the data set).

The R-squared of the model is .9683 which means that about 97% of the variance of the response variable US Revenue is explained by the value of the Predictor variable opening.

j. Use the regression equation from part (f) to predict the total US revenue for the movie named Get Smart. (Budget was 80 million dollars; it was shown in 3911 theaters; and its opening weekend revenue was 38.7 million dollars.) State your predicted value in a sentence that is in context of the data. Don't forget units!

```
predicted_USRevenue <- predict(Opening_Model, newdata = data.frame(Opening = 38.7))
cat("The linear model predictes that the movie Get Smart which had an opening weekend revenue of 38.7 mi
```

```
## The linear model predictes that the movie Get Smart which had an opening weekend revenue of 38.7 mil
```

The linear model predicts that the movie Get Smart which had an opening weekend revenue of 38.7 million dollars would have a total US Revenue of 119.884 million dollars.