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)
```

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
```

In this paired two-sided t-test, the p-value obtained is 0.04143 which is less than 0.05, we therefore reject the null hypothesis that the test means are equal and tentatively conclude that the test means were different.

b. Second, we want to see if there was improvement over the course of the semester. H0: Test 2 - Test 1 > 0

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

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

Since the mean difference (Test2 – Test1) is positive, there appears to be an increase in test scores from Test 1 to Test 2. From the p-value obtained in this test of 0.02072 which is less than alpha=0.05, we reject the null hypothesis that this difference is zero and conclude that there was improvement over the semester.

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_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
```

```
summary(plasma_model)
```

```
##
## Call:
## lm(formula = plasma ~ time, data = plasma)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.90 -10.85   0.15  11.93  45.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.500      5.944  19.935   <2e-16 ***
## time11am       9.400      8.407   1.118    0.269
## time2pm        1.800      8.407   0.214    0.831
## time5pm      -13.700      8.407  -1.630    0.110
## time8pm        1.200      8.407   0.143    0.887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.8 on 45 degrees of freedom
## Multiple R-squared:  0.1499, Adjusted R-squared:  0.07434
## F-statistic: 1.984 on 4 and 45 DF,  p-value: 0.1132
```

The Pr(>F) p-value obtained in testing the null hypothesis that there is no difference in the mean blood plasma levels of subjects at different times of the day is 0.1132, which is greater than 0.05. We therefore fail to reject the null hypothesis that there is no difference in the mean blood plasma levels of subjects at different times of the day.

The estimated mean blood plasma level of the 8am group from the test results was 118.50 micromoles/liter

Furthermore, the p-value obtained from testing the null hypothesis that the blood plasma level mean of the 8am group is equal to zero was less than 0.05, we therefore reject the null hypothesis the blood plasma level mean of the 8am group is equal to zero.

The t-tests for the remaining group are testing the null hypothesis that the 8am group mean - the group of the row the test corresponds too = 0, as illustrated in the following table. All of the t-tests testing the H0: 8am group mean - their group mean = 0 produced p-values greater than 0.05, therefore for all groups we fail to reject the null hypothesis that each groups mean is not different from the 8am groups mean and therefore conclude that the different groups means are not different from each other,

Group	H0:	P-value
time11am	H0: 8am group mean - 11 am group mean = 0	0.269 > .05, fail to reject H0
time2pm	H0: 8am group mean - 2 pm group mean = 0	0.831 > .05, fail to reject H0
time5pm	H0: 8am group mean - 5 pm group mean = 0	0.110 > .05, fail to reject H0
time8pm	H0: 8am group mean - 8 pm group mean = 0	0.887 > .05, fail to reject H0

finally, therefore we conclude there is not a difference in blood plasma levels at different times of the day

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

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

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

```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, var.equal=FALSE)
```

```
##
##  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
```

fail to reject the null hypothesis that the player 1 scores are normally distributed.

reject the null hypothesis that the player 2 scores are normally distributed, therefore player 1 and 2 could never have equal variances

two-sample independent t-test for difference of mean results in a p-value of 0.9738, fail to reject the null hypothesis that the difference in means is equal to zero. This test fails to produce sufficient evidence that the difference of players mean scores is not 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 Wilcox rank sum test having an HA that implies that Player 1 is better than Player 2 resulted in a p-value of 0.01875 which is less than 0.05, we therefore reject the null hypothesis that true location shift is equal to 0 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.

```
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
```

```
chisq.test(table, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 5.4702, df = 1, p-value = 0.01934
```

The p-value obtained is 0.01934 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).

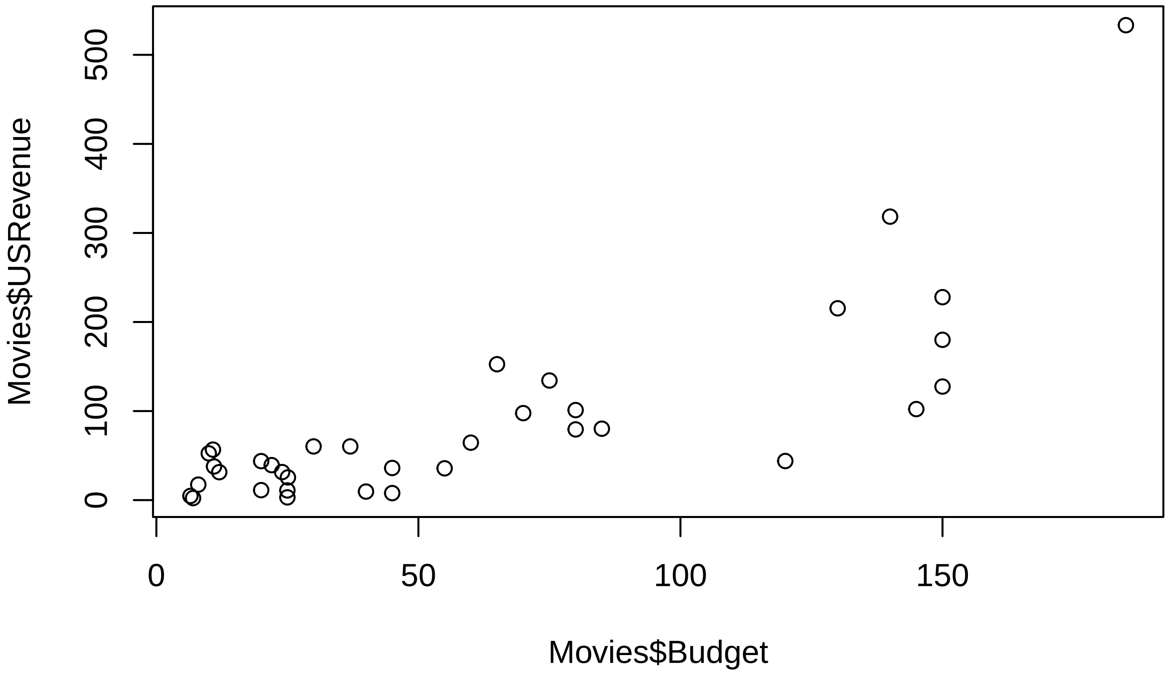
```
Movies <- read.csv(file="Movies.csv",header=TRUE)
```

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)
```

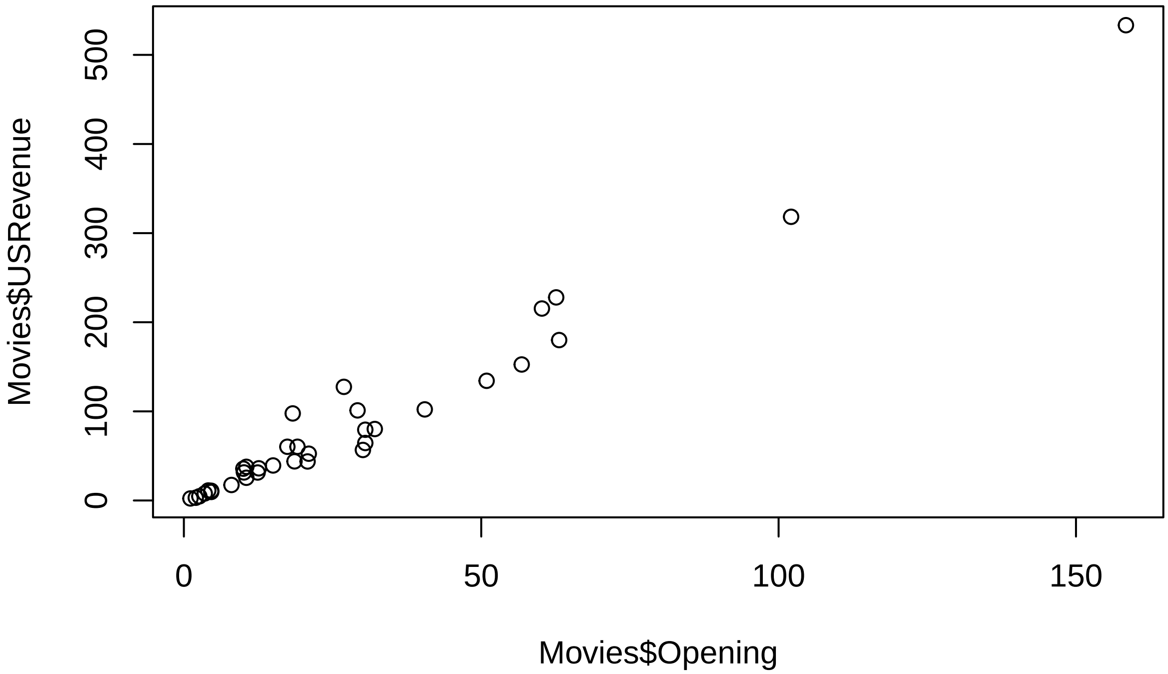
```
## [1] 0.7918636
```

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)
```



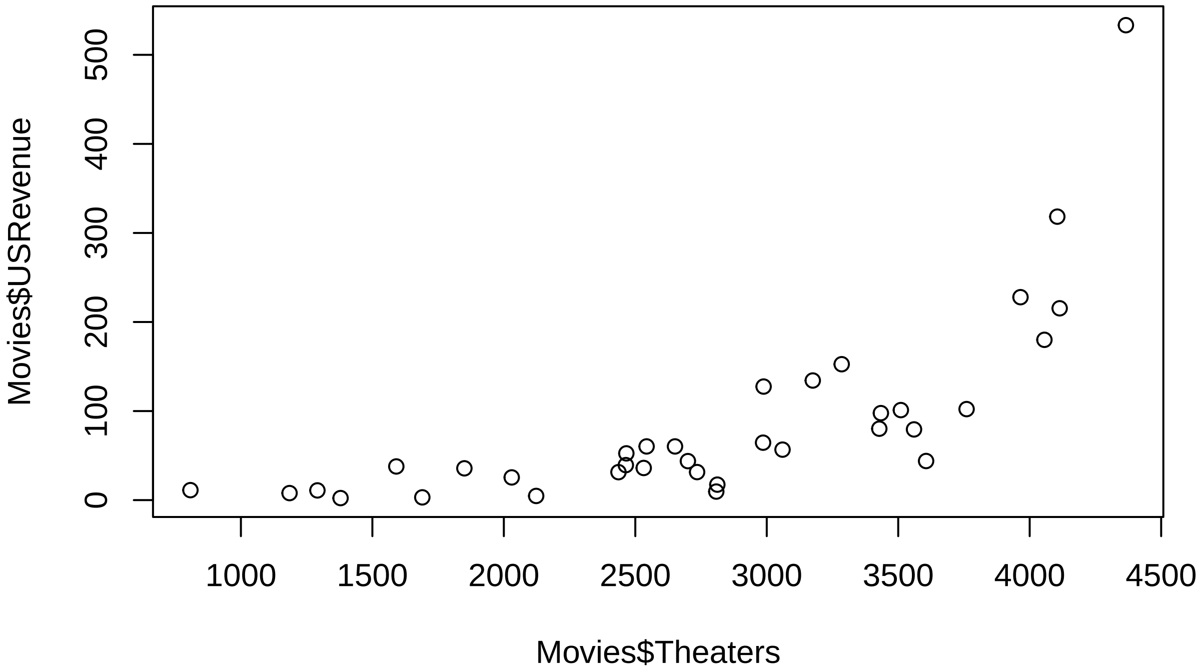
```r
cor(x = Movies$Opening, y = Movies$USRevenue)
```

```
## [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.

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



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

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

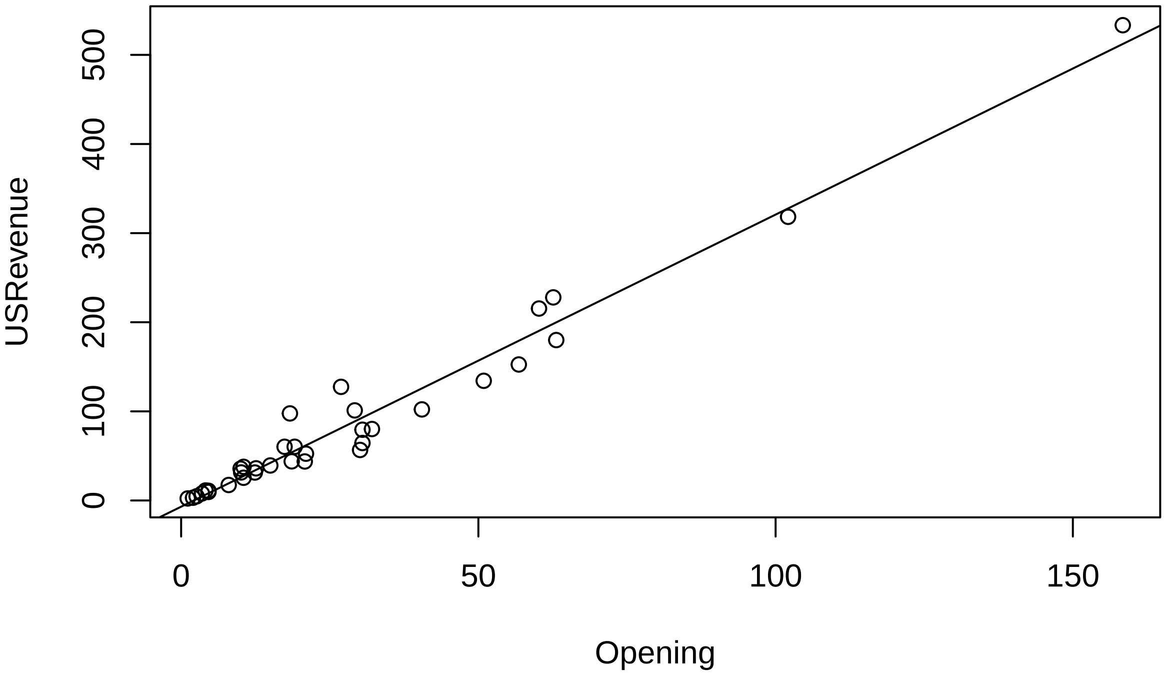
The trend is positive such that as variable theaters increases so does the variable USRevenue. The variables have a correlation coefficient of 0.7153432 which is classified as a moderate correlation. The relationship between these variables has a very obvious curvature which appears almost exponential.

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])
```



```
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: -6.9619 would imply negative revenue when Opening is 0, which is nonsensical in this context of producing a model for US Revenue.

Also,

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. This means that opening weekend revenue is a very very strong indicator of US Revenue.

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!

```r
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.