Phase-4

Linear Regression Modelling

Problem Statement: -

To Find “TOP SELLING VIDEO GAMES” in NA_Sales markets by knowing characteristics of the game using Linear Regression model.

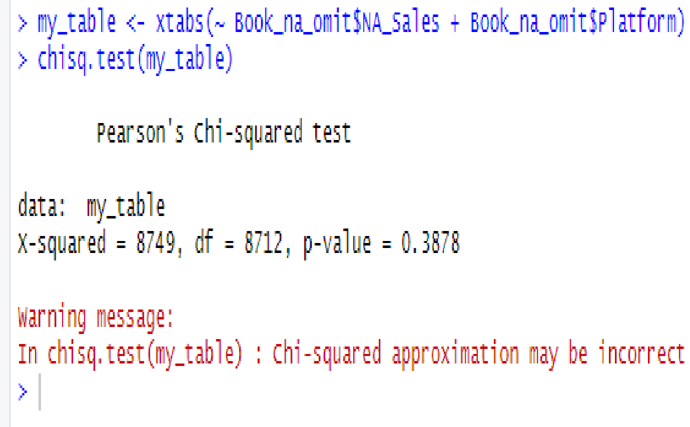
Problem solving: -

From our analysis so far we see that Platform of the videogame highly affects the sales of videogames in North America.

Calculate correlation between NA_Sales and Platform   [-1 to +1]

my_table <- xtabs(~ Book_na_omit$NA_Sales + Book_na_omit$Platform)

chisq.test(my_table)



```
> my_table <- xtabs(~ Book_na_omit$NA_Sales + Book_na_omit$Platform)
> chisq.test(my_table)

	Pearson's Chi-squared test

data:  my_table
X-squared = 8749, df = 8712, p-value = 0.3878

Warning message:
In chisq.test(my_table) : Chi-squared approximation may be incorrect
> 
```

Step 1. Create Training and Test data -

```
set.seed(100)
```

(#setting seed to reproduce results of random sampling)

```
trainingRowIndex <- sample(1:nrow(Book_na_omit),
0.8*nrow(Book_na_omit))
```

(# row indices for training data)

```
trainingData <- Book_na_omit[trainingRowIndex, ]
```

(# model training data)

```
testData  <- Book_na_omit[-trainingRowIndex, ]
```

(# test data)

Step 2.Build the model on training data -

```
lmMod <- lm(NA_Sales ~ Platform, data=trainingData)
```

(# build the model)

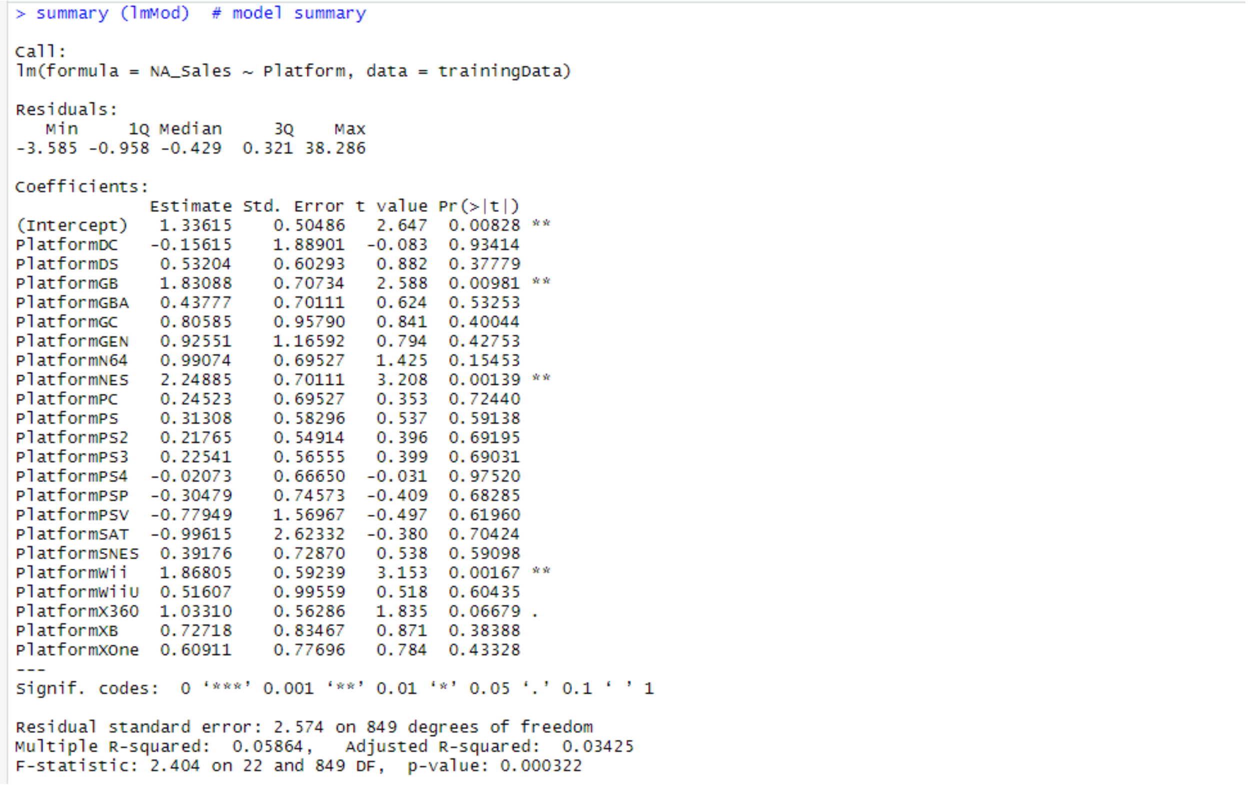
```
NA_SALESPred <- predict(lmMod, testData)
```

(# NA_salespredict distance)

Step 3: Review

summary (lmMod)

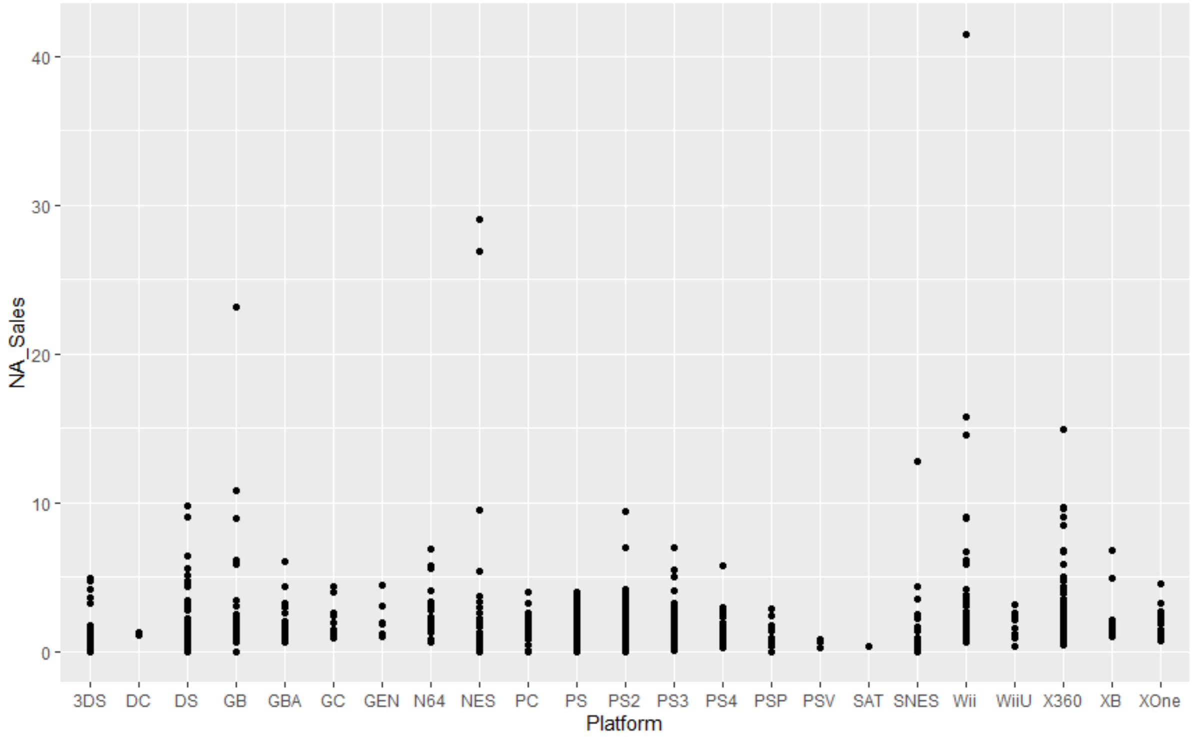
(# model summary)



```
> summary (lmMod)  # model summary

Call:
lm(formula = NA_Sales ~ Platform, data = trainingData)

Residuals:
   Min     1Q Median     3Q    Max 
-3.585 -0.958 -0.429  0.321 38.286 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)   1.33615    0.50486   2.647  0.00828 **
PlatformDC   -0.15615    1.88901  -0.083  0.93414   
PlatformDS    0.53204    0.60293   0.882  0.37779   
PlatformGB    1.83088    0.70734   2.588  0.00981 **
PlatformGBA   0.43777    0.70111   0.624  0.53253   
PlatformGC    0.80585    0.95790   0.841  0.40044   
PlatformGEN   0.92551    1.16592   0.794  0.42753   
PlatformN64   0.99074    0.69527   1.425  0.15453   
PlatformNES   2.24885    0.70111   3.208  0.00139 **
PlatformPC    0.24523    0.69527   0.353  0.72440   
PlatformPS    0.31308    0.58296   0.537  0.59138   
PlatformPS2   0.21765    0.54914   0.396  0.69195   
PlatformPS3   0.22541    0.56555   0.399  0.69031   
PlatformPS4  -0.02073    0.66650  -0.031  0.97520   
PlatformPSP  -0.30479    0.74573  -0.409  0.68285   
PlatformPSV  -0.77949    1.56967  -0.497  0.61960   
PlatformSAT  -0.99615    2.62332  -0.380  0.70424   
PlatformSNES  0.39176    0.72870   0.538  0.59098   
PlatformWii   1.86805    0.59239   3.153  0.00167 **
PlatformWiiU  0.51607    0.99559   0.518  0.60435   
PlatformX360  1.03310    0.56286   1.835  0.06679 . 
PlatformXB    0.72718    0.83467   0.871  0.38388   
PlatformXOne  0.60911    0.77696   0.784  0.43328   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 849 degrees of freedom
Multiple R-squared:  0.05864,	Adjusted R-squared:  0.03425 
F-statistic: 2.404 on 22 and 849 DF,  p-value: 0.000322
```

Step 4:- Plotting

```
library(ggplot2)

ggplot(trainingData,aes(y=NA_Sales,x=Platform))+geom_point()+geom_smooth(method="lm",col="red")
```



Conclusion: -

From the model summary, the model p value and predictor’s p value are less than the significance level, so we know we have a statistically significant model. Also, the R-Sq. and Adj R-Sq. are comparative to the original model built on full data.

Therefore we can conclude **Top Selling Video Games** in **North America** based on platform are 1.**GB,**

2.NES,

3.wii.