# LAB 1: LINEAR REGRESSION

Due date : September 18, 2023, 11:59 PM

Please answer all questions within the space provided. Please type the assignment and use R.

Your last name: Suri

Your first name: Jatin

Your student ID: 261152263

Your last name: Agarwal

Your first name: Lakshya

Your student ID: 261149449

# Lab 1 Linear regression: Predicting Video Game ratings

The video game market is one of the fiercest out there. Out of thousands of video games, only a few thrive. The likelihood of making a profitable video game is even tinier. In order to boost the game's popularity and success, it is necessary to have good ratings.

To understand what drives a video game's ratings, I downloaded data for thousands of video games. These data were gathered from different video game websites. We will apply linear regression to this dataset, to build a model that helps us predict each video game's rating. In this dataset, we have the following variables:

## Data dictionary

**Output/Outcome variable:**

- **score:** (From 1 to 100) The score given to the game  by professional critics

**Input/Independent variables:**

**Identity variables:**

- **title:** The video game's name

**Numeric variables:**

- **release_year:** The year the video game was released
- **sales_na:** Sales in North America (in millions)
- **sales_eu:** Sales in Europe (in millions)
- **sales_jp:** Sales in Japan (in millions)
- **sales_others**: Sales in the rest of the world (in millions)
- **sales_global**: Sales globally (in millions)
- **count_critic:** The number of critics that have rated the game

**Categorical variables:**

- **platform:** The video game's platform (WII, PlayStation, Nintendo DS, XBox360, etc.)
- **genre:** The video game genre (sports, role-playing, puzzle, shooter, simulation, etc.)
- **publisher:** The video game's publisher company (Nintendo, Sega, Ubisoft, etc)
- **developer:** The company that developed the game
- **content_rating:** The maturity level of the game

© Juan Camilo Serpa.  This content is copyrighted. Do not upload or distribute without written permission.

# 1. Visualizing variables (15 points - Lecture 2)

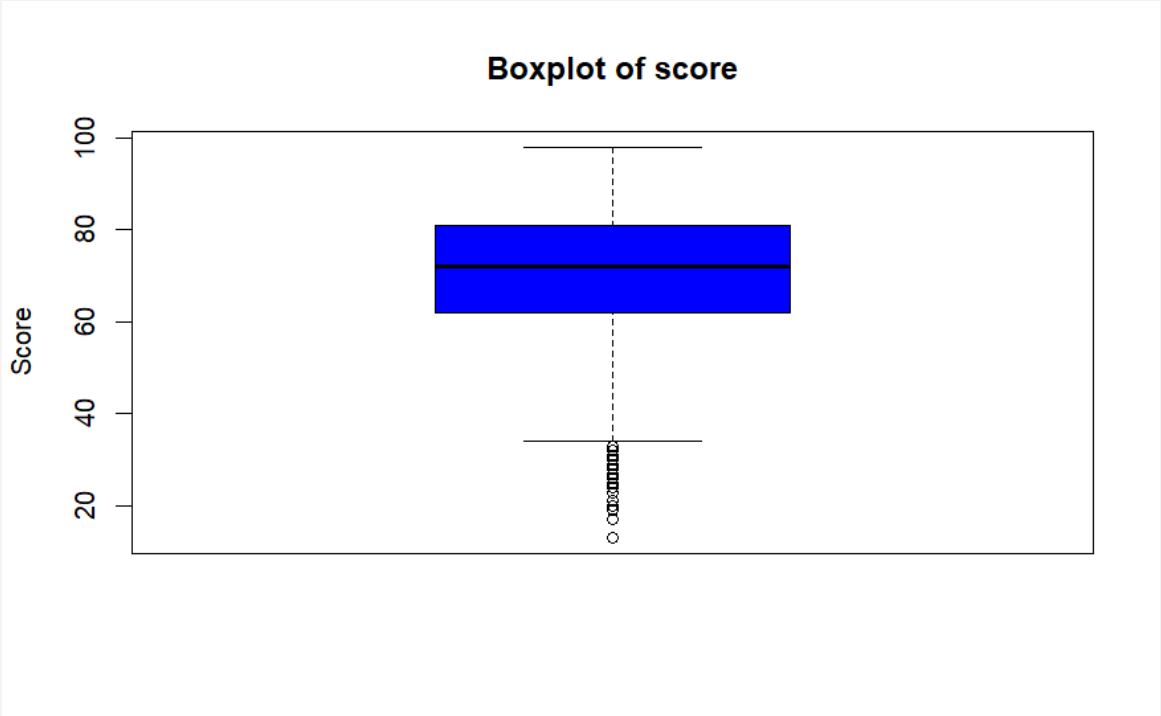
Let's open the video_games dataset. For each of the variables below, provide the information requested.

## i) Variable: score

A. (0.5 points) Summary statistics

- Min: 13.0
- Max: 98.0
- Quartile 1: 62.0
- Quartile 3: 81.0
- Median: 72.0
- Mean: 70.3

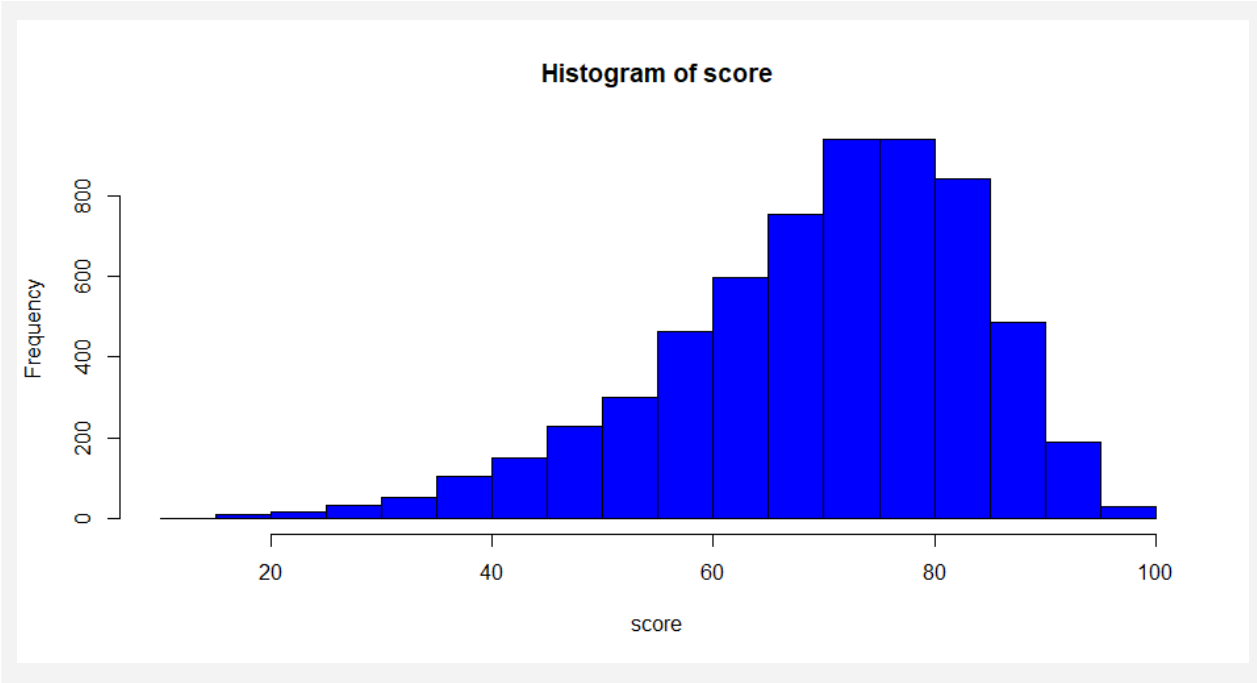
B. (1 point) Create a box plot and attach it below. The inside of the box plot should be blue.[1]



[1] To figure this one out, you might need to figure out the code. The idea is to get you used to searching for these codes on the world wide web. Throughout this course you will discover how collaborative the data analytics community is!

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

C. (1 point) Create a histogram and attach it. Make sure the histogram has 20 breaks, and that the breaks are blue.[2]



## ii) Variable: sales_global

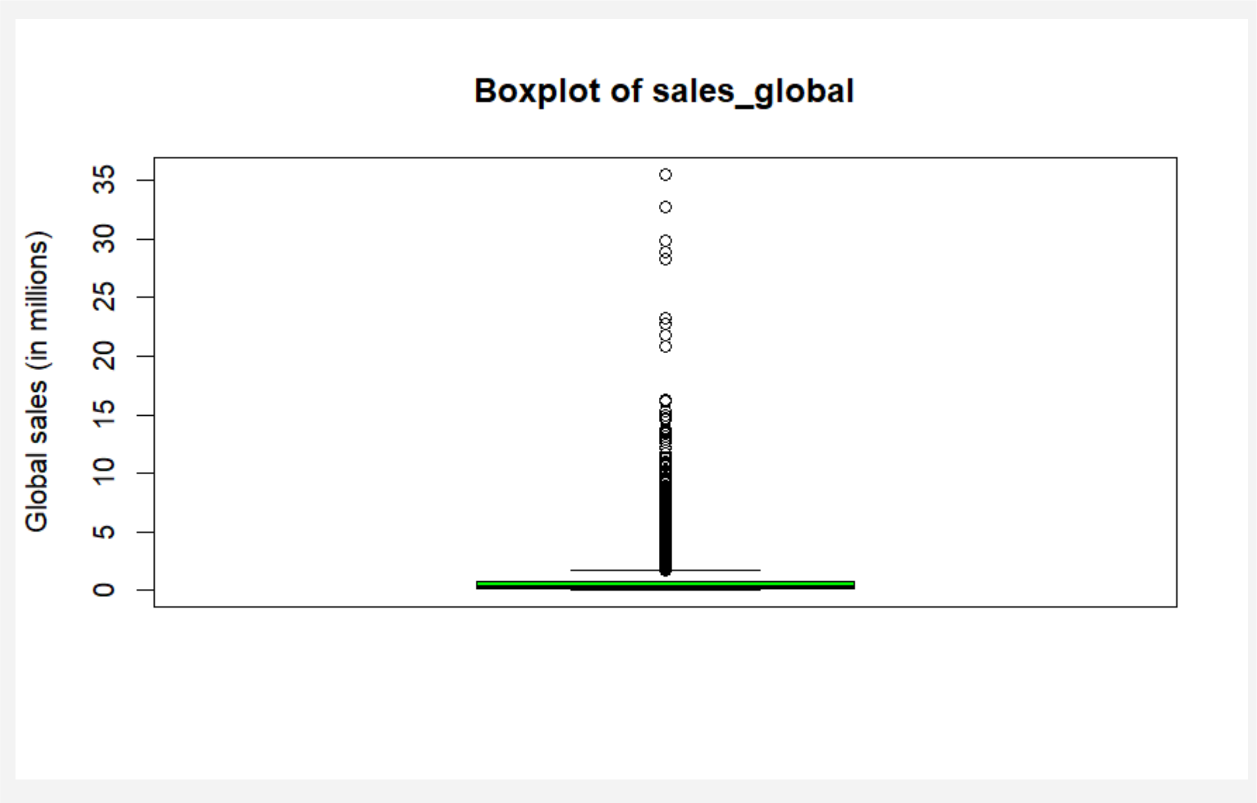
D. (0.5 points) Summary statistics

- Min: 0.0100
- Max: 35.5200
- Quartile 1: 0.1100
- Quartile 3: 0.7600
- Median: 0.3000
- Mean: 0.7639

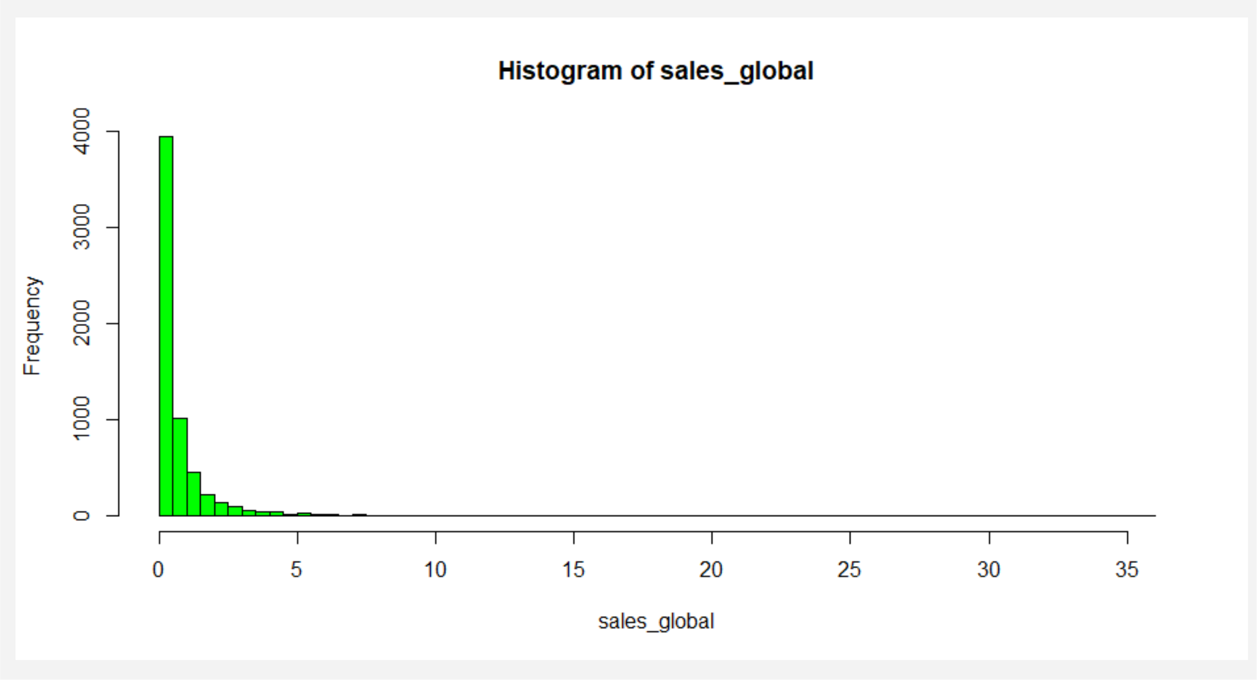
E. (1 point) Create a box plot and attach it below. The inside of the box plot should be green.

[2] You are not doing anything wrong if the number of breaks is different from the number of columns that you see in the histogram. The number of breaks means "into how many buckets are we splitting the data." If you select, say, 5 breaks, it means the data will be split into 5 buckets. But you might end up with only 3 or 4 columns in the histogram if, for example, some of those buckets contain no data points.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



F. (1 point) Create a histogram and attach it. Make sure the histogram has 100 breaks, and that the breaks are green.



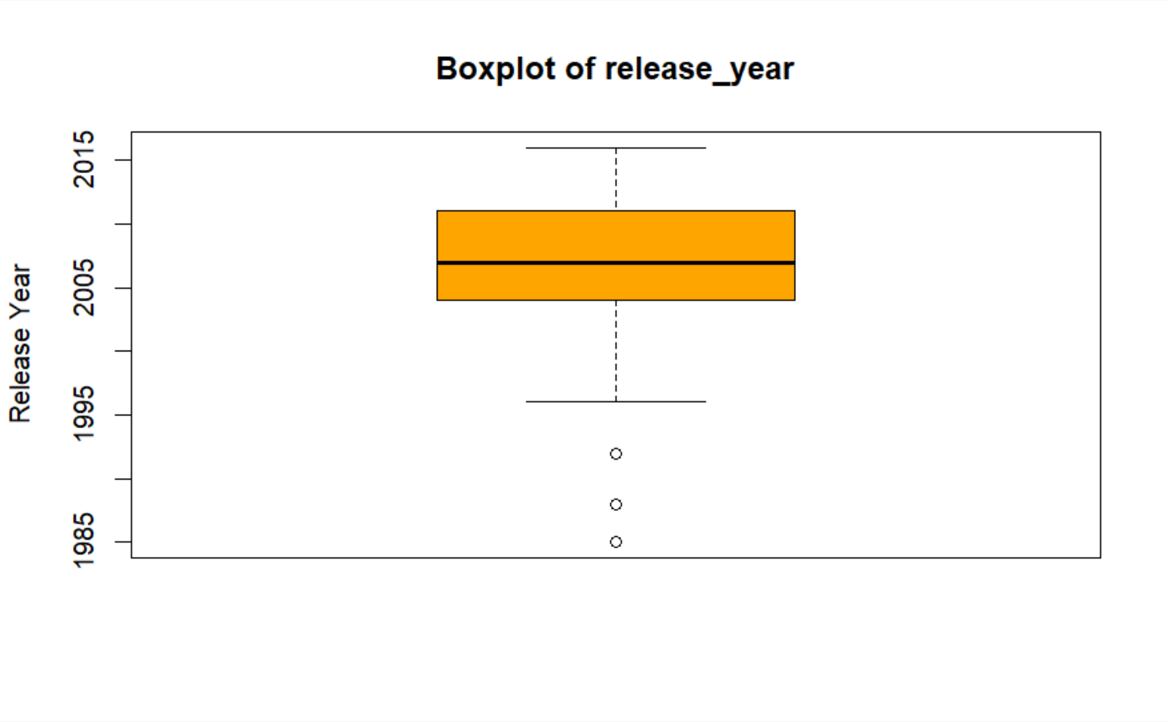
## iii) Variable: release_year

G. (0.5 points) Summary statistics

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

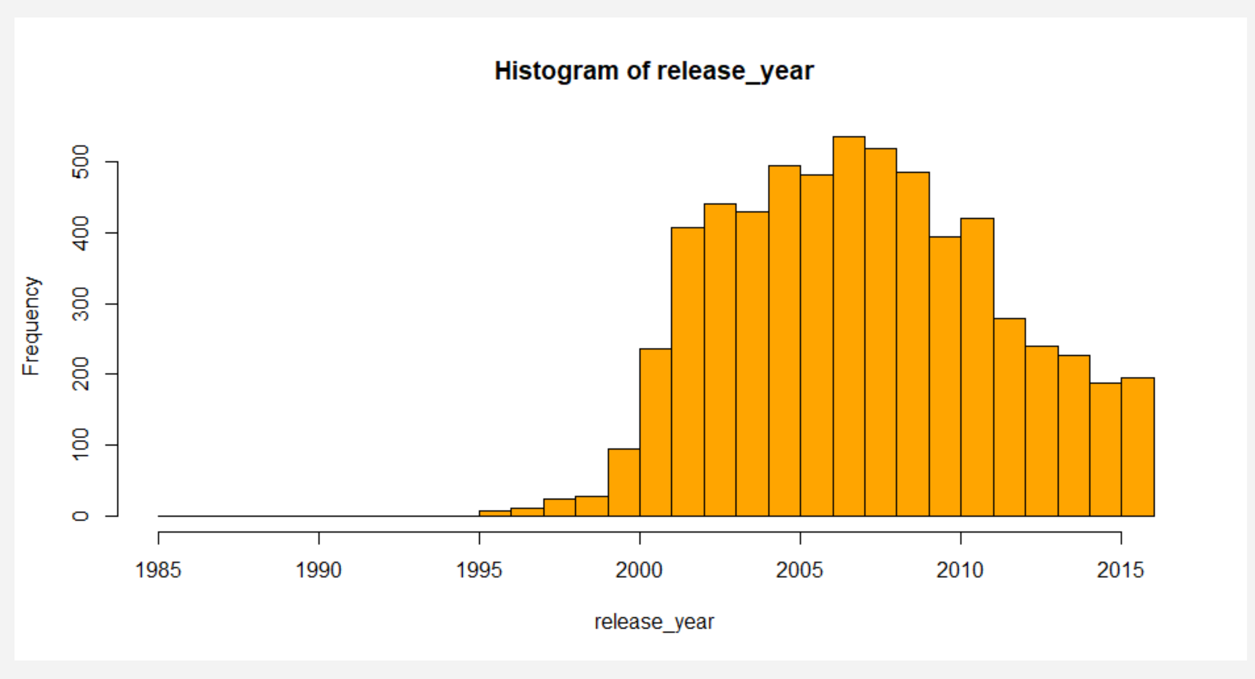
- Min: 1985
- Max: 2016
- Quartile 1: 2004
- Quartile 3: 2011
- Median: 2007
- Mean: 2007

H. (1 point) Create a box plot and attach it below. The inside of the box plot should be orange.



I. (1 point) Create a histogram and attach it below. Make sure the histogram has 25 breaks, and that the breaks are orange.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



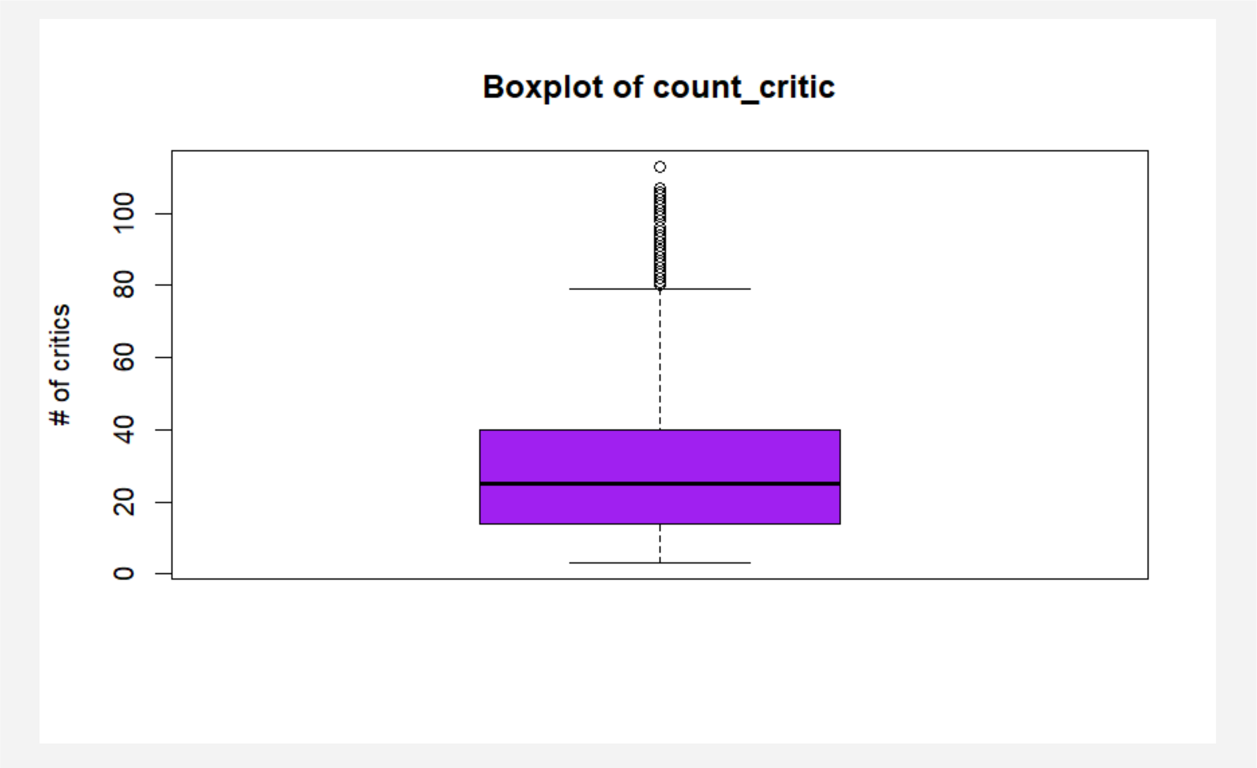
## iv) Variable: count_critic

J. (0.5 points) Summary statistics

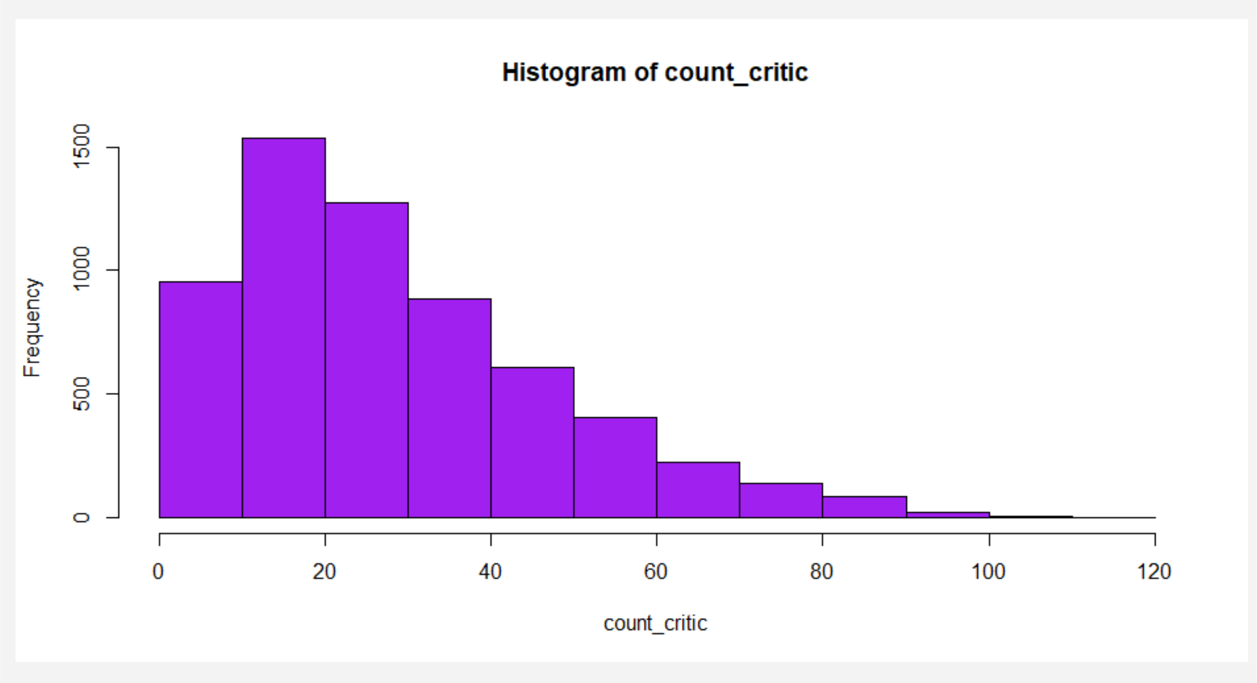
- Min: 3.00
- Max: 113.0
- Quartile 1: 14.00
- Quartile 3: 40.00
- Median: 25.00
- Mean: 29.09

K. (1 point) Create a box plot and attach it below. The inside of the box plot should be purple.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



L. (1 point) Create a histogram and attach it. Make sure the histogram has 10 breaks, and that the breaks are purple:

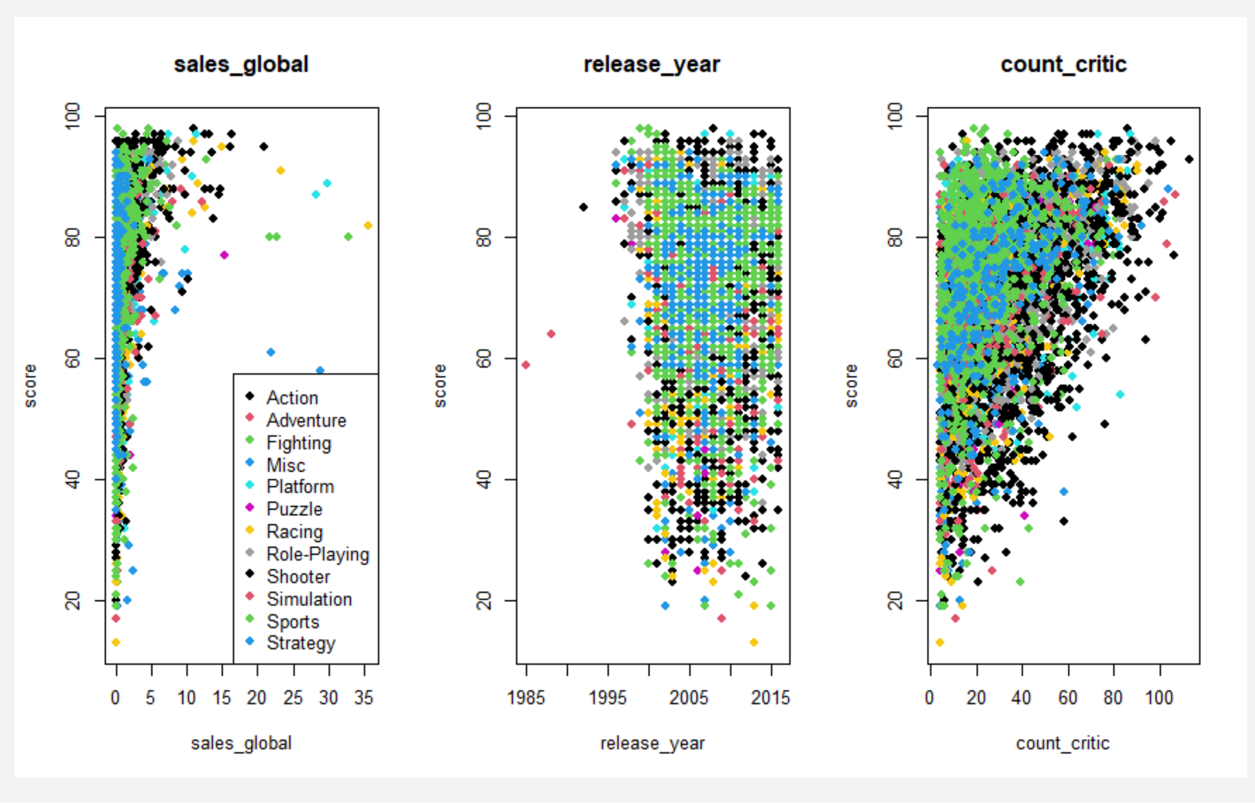


M. (5 points) Create three scatter plots. Each scatter plot should have the response variable (*score*) on the y-axis, and each respective predictor (*sales_global*, *release_year*, *count_critic*) on the x-axis.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

Using *R,* create a scatterplot matrix, with 1 row and 3 columns. In each scatterplot, the dots should be coloured based on its genre. In other words, each genre should have a distinct colour within each scatterplot. Also, make sure you resize the plot's window so that the plots are *roughly* proportional, horizontally and vertically.

Note: The objective of this question is to get you used to coding subsets of the data, which will be essential for some visualization techniques.



© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 2. Simple linear regression (8 points - Lecture 2)

Run three simple linear regressions ($Y=b_o+b_1x$) — one for each of the three predictors. Attach the regression results.

## score = $b_o$+ $b_1$(sales_global)

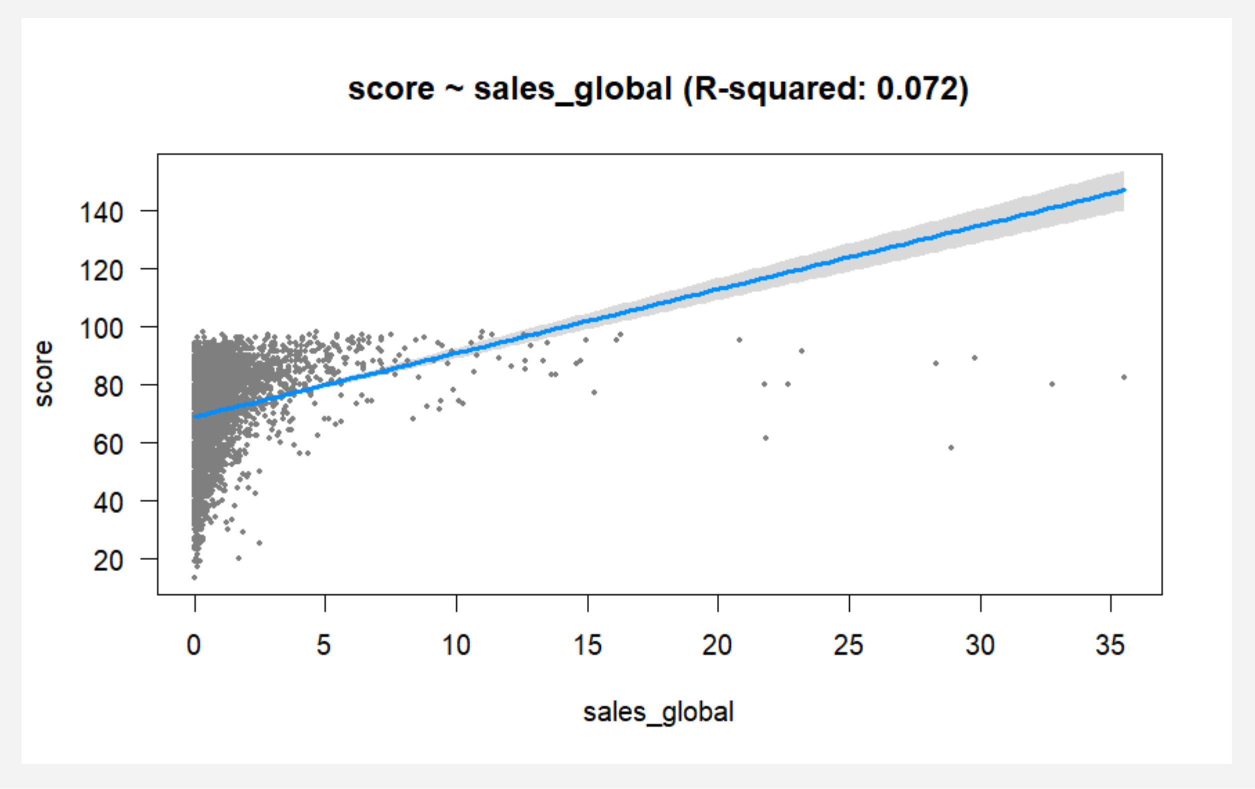
A. (2 points) Fill in the blanks:

- $b_o$: 68.6168
- $b_1$: 2.2015
- r-squared: 0.07174
- For $b_1$, please provide:
  - 95% Confidence interval: [2.003385, 2.399578]
  - t-test statistic: 21.79

What is the probability that there is not a statistically significant relationship between these two variables? ~0%

Please attach below a regression graph showing the line of best fit, the 95% confidence intervals, and the variable's scatterplot:

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



score = $b_o$+ $b_1$(release_year)

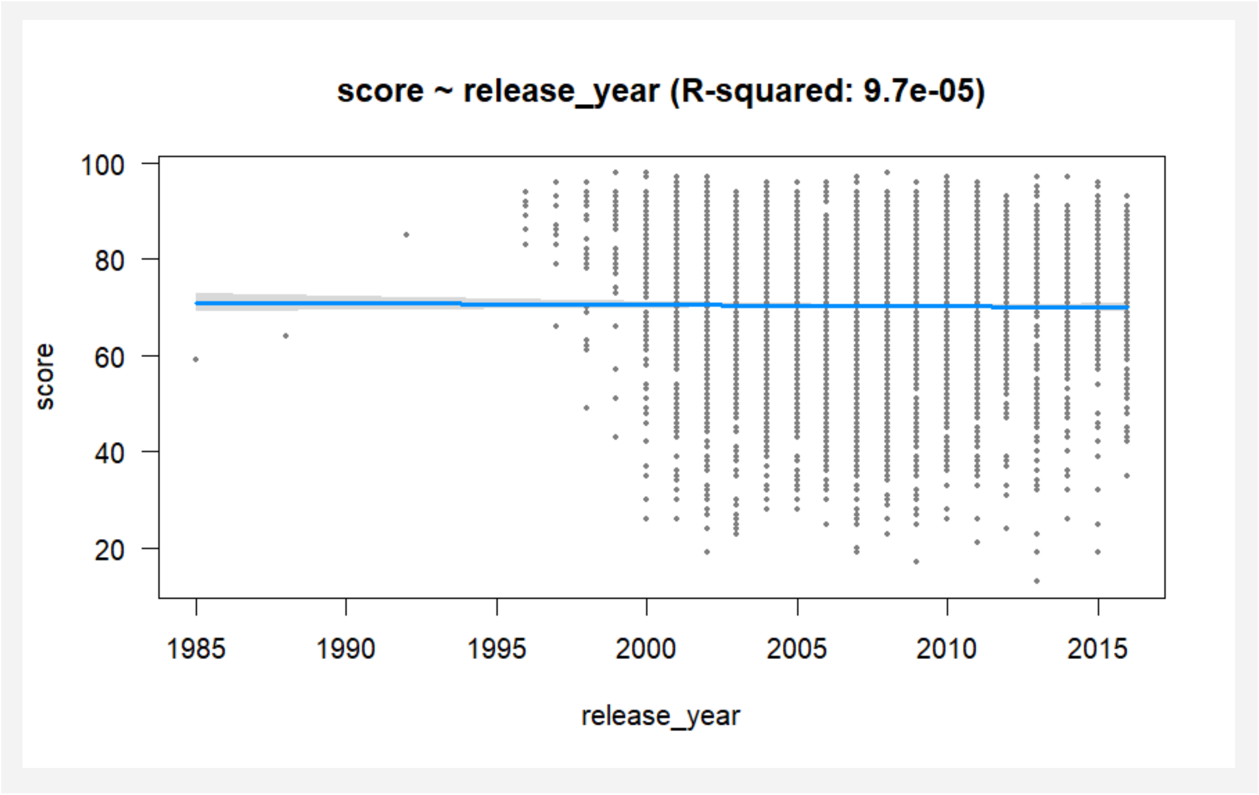
B. (2 points) Fill in the blanks:

- $b_o$: 135.45636
- $b_1$: -0.03246
- r-squared: 0.00009726
- For $b_1$, please provide:
  - 95% Confidence interval: [-0.1147883, 0.0498721]
  - t-test statistic: -0.773

What is the probability that there is not a statistically significant relationship between these two variables? ~44%

Please attach below a regression graph showing the line of best fit, the 95% confidence intervals, and the variable's scatterplot:

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



## score = $b_o$+ $b_1$(count_critic)

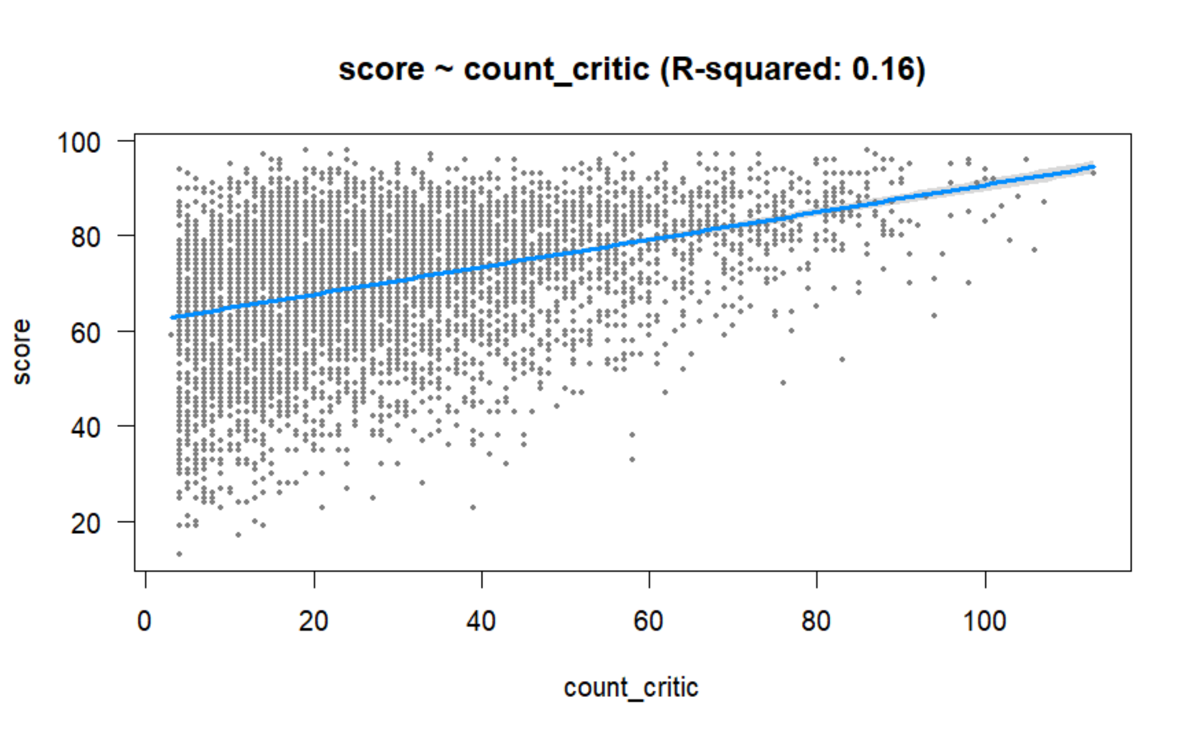
C. (2 points) Fill in the blanks:

- $b_o$: 61.954394
- $b_1$: 0.286812
- r-squared: 0.1574
- For $b_1$, please provide:
  - 95% Confidence interval: [0.2702145, 0.3034092]
  - t-test statistic: 33.88

What is the probability that there is not a statistically significant relationship between these two variables? ~0%

Please attach below a regression graph showing the line of best fit, the 95% confidence intervals, and the variable's scatterplot:

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



D. (2 points) What did you learn? Suppose I am the owner of a video game company. I want to know which factors affect the score my video game gets. What can you tell me, based on these regressions? **Note**: this is an open-ended question. Discuss what you have learned using statistical language, significance, etc. Try to interpret the regression equation. Please limit your answer to one paragraph.

**Your Response:**
Based on the simple linear regressions, the global sales amount and the number of critics who have reviewed the game are statistically significant in predicting the video game score, while the release year of a game is not statistically significant in predicting the video game score. For every $1 million increase in global sales, the video game score increases by approximately 2 points, and for every new critic that reviews the video game, the score increases by approximately 0.3 points, assuming other variables stay constant.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 3. Predictions (3 points - Lecture 2)

Based on the above results, predict the video game critic score if:

A. (1 point) Consider the simple regression (*score*=$b_0$+$b_1$*sales_global*), from the previous section. If I had a video game with 750,000 sales globally, what score would this video game have?

**Your prediction:**

70.26787

B. (1 point) Consider the simple regression (*score*=$b_0$+$b_1$*release_year*), from the previous section. If I had a video game that was released in 2009, what score would this video game have?

**Your prediction:**

70.24801

C. (1 point) Consider the simple regression (*score*=$b_0$+$b_1$*count_critic*), from the previous section. If I had a video game that was reviewed by 80 critics, what score would this video game have?

**Your prediction:**

84.89934

*Hint: Use the "Coef()" function*

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 4. Multiple Regression (7 points - Lecture 3)

Suppose we are thinking of running the following multiple regression:

$$score = b_o + b_1(sales\_global) + b_2(release\_year) + b_3(count\_critic)$$

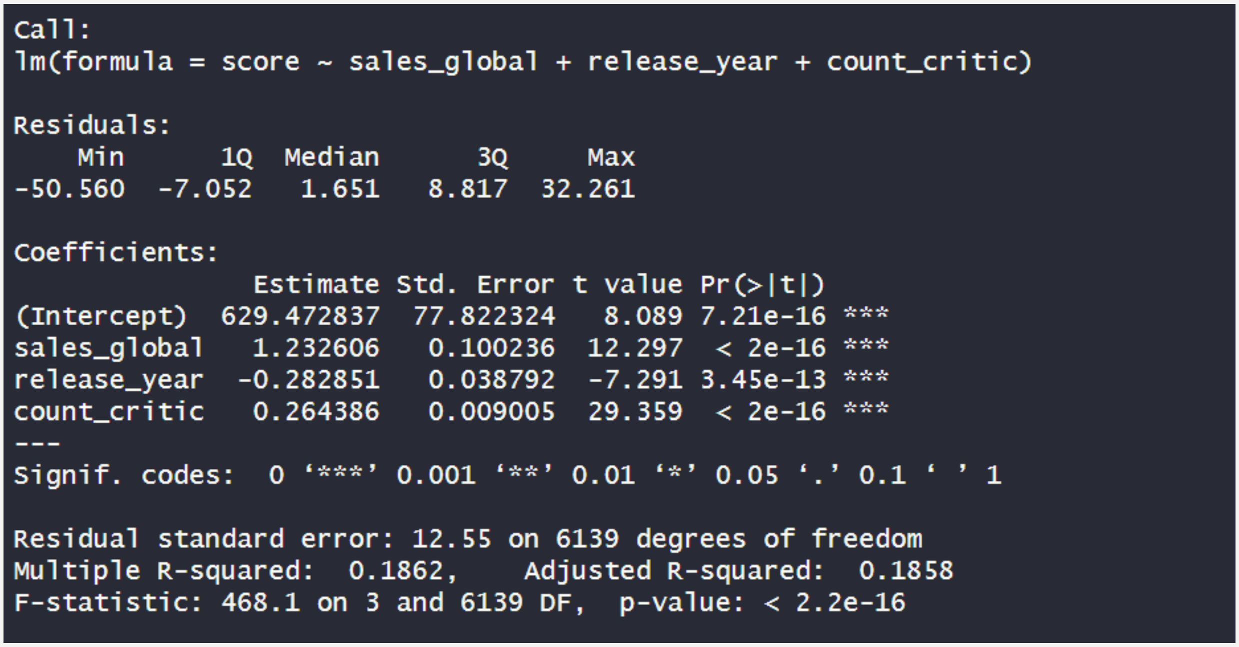
A. (1 point) Why would we want to do this, as opposed to three separate simple linear regressions (as we did above)? Answer in the space provided below giving the two reasons we discussed in class.

**Your answer:**

Reason 1: We cannot make a joint prediction about the response variable.

Reason 2: Simple regressions don’t take into account interaction between predictors.

B. (1 point) Now, run the multiple regression and provide the R-output below:



```
Call:
lm(formula = score ~ sales_global + release_year + count_critic)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.560  -7.052   1.651   8.817  32.261 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  629.472837  77.822324   8.089 7.21e-16 ***
sales_global   1.232606   0.100236  12.297  < 2e-16 ***
release_year  -0.282851   0.038792  -7.291 3.45e-13 ***
count_critic   0.264386   0.009005  29.359  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.55 on 6139 degrees of freedom
Multiple R-squared:  0.1862,	Adjusted R-squared:  0.1858 
F-statistic: 468.1 on 3 and 6139 DF,  p-value: < 2.2e-16
```

C. (1 point) Which of the above coefficients are statistically significant at the 1% level?

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

**Your answer:**
All the 3 coefficients (`sales_global`, `release_year,` `count_critic`) are statistically significant at the 1% level.

D. (2 points) Based on the above results (i.e., the multiple regression), predict the critic score of a game if: (i) it has 750,000 in global sales, (ii) was released in 2009, and (iii) was reviewed by 80 critics.

**Your prediction:**
The predicted critic score is: 83.30009

E. (2 points) In four sentences or less, interpret the results of the regression above. You need to discuss statistical significance, p-values, and r-squared. But this interpretation needs to be geared at a video game manager who has no knowledge of statistics. Thus, you will need to avoid jargon.

**Your interpretation:**
The analysis shows that the more a game sells, the higher its critic score tends to be. Specifically, for every extra million dollars in sales, the score goes up by about 1.23 points. However, older games tend to have slightly lower scores, with each passing year reducing the score by about 0.28 points. Interestingly, the more critics that review a game, the higher its score, with each additional critic raising the score by about 0.26 points. These three factors together explain about 19% of what determines a game's score.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 5. Categorical Variables (7 points - Lecture 3)

A. (2 points) Run the following model:

$$score = b_o + b_1(release\_year) + b_2(Nintendo)$$

Where *Nintendo=1* if the game was published by Nintendo (in the 'publisher' field), and Nintendo=0 if it was published by any other. Note that you will need to figure out how to create this variable (hint: use the *ifelse()* function).

**Your results:**

- $b_o$: 147.63273
- $b_1$: -0.03866
- $b_2$: 6.29066

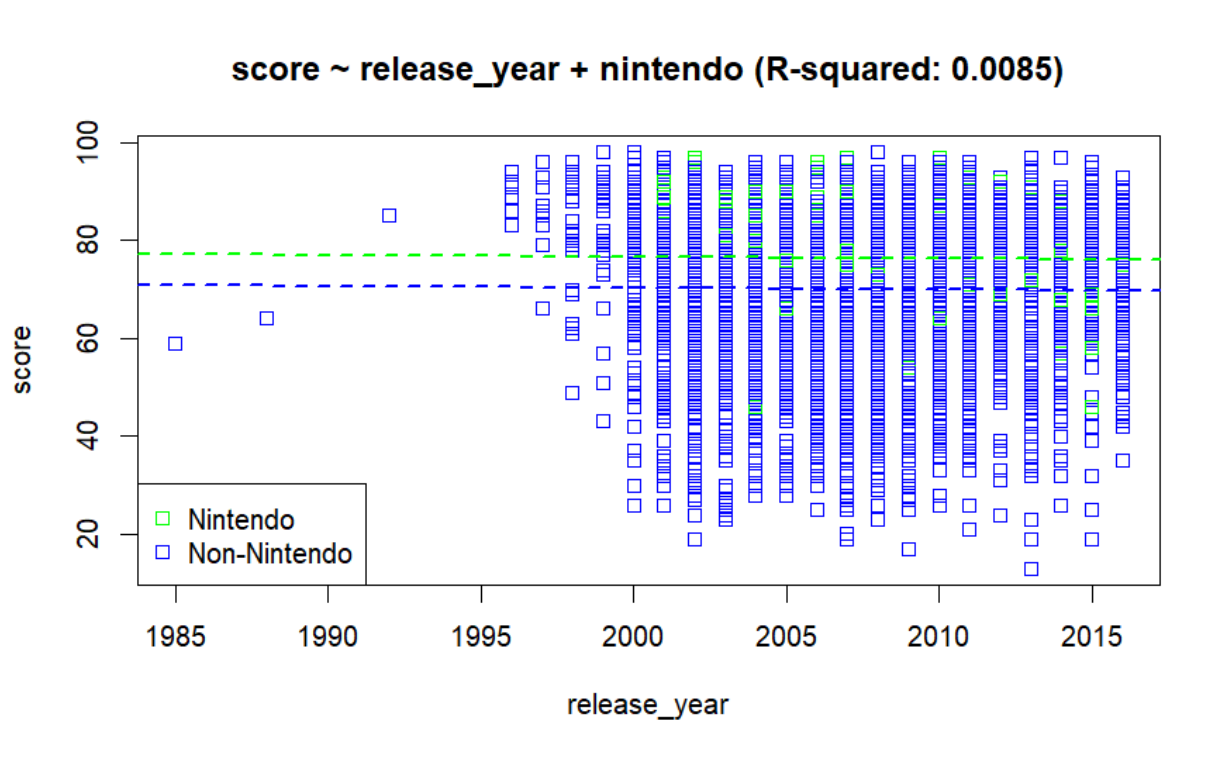
B. (2 points) Now, interpret the coefficient $b_2$ to a manager who knows nothing about statistics. Avoid jargon.

**Your answer:**
A game published by Nintendo, on average, increases the critic score by 6.29 compared to games not published by Nintendo, everything else being equal. This suggests that Nintendo's brand has a positive impact on the perceived quality of a game.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

C. (3 points) Draw the regression line from the above model (on a scatterplot) for (i) games that are published by Nintendo and (ii) games that are not published by Nintendo. Make sure you follow the instructions below:

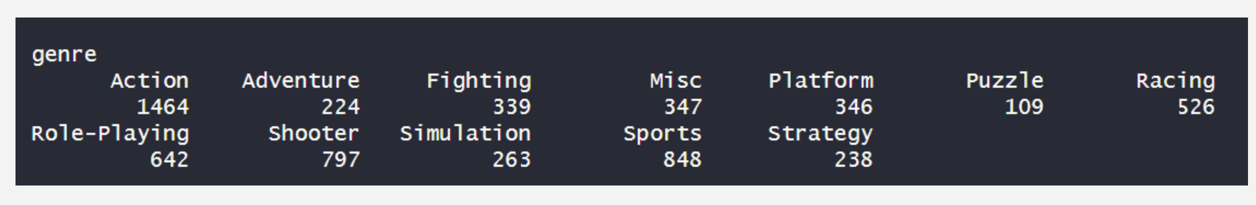
- Games that are published by Nintendo should be in green; games that aren't should be blue (both the regression lines and the scatter plot dots).
- Make sure you create a legend.
- Instead of circles, I want squares in the dots of the scatter plot (you will need to figure this on your own).
- I want the regression lines to have a width of 2 and be a dashed line (you will need to figure how to make them dashed on your own).



© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 6. Categorical Variables with Multiple Categories (5 points - Lecture 3)

A. (1 point) How many video game genres are there? Please paste the name of the categories and the number of observations per category below, using the *table()* function:

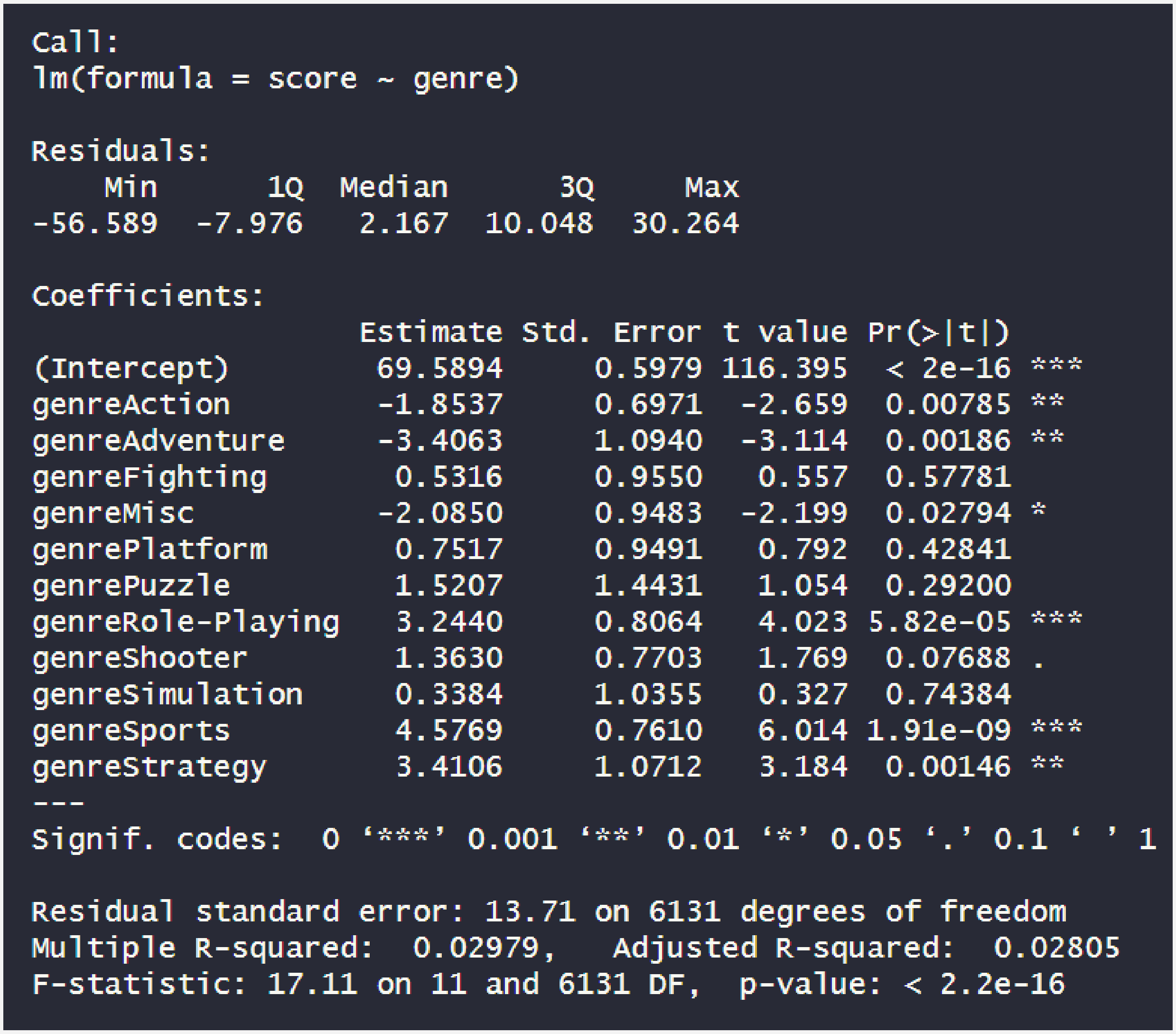


```
genre
      Action    Adventure     Fighting         Misc     Platform       Puzzle       Racing 
        1464          224          339          347          346          109          526 
Role-Playing      Shooter   Simulation       Sports     Strategy 
         642          797          263          848          238 
```

B. (1 point) I want to know which genre (Sports, Shooter, Simulation, etc.) has better ratings. To test this, run a model where (i) the dependent variable is *score*, and (ii) the predictors are the categories found in *genre* (no other predictors). Create a multiple linear regression, where the excluded dummy is *Racing*.

Paste the R-results of the regression below (as an image):

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.



```
Call:
lm(formula = score ~ genre)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.589  -7.976   2.167  10.048  30.264 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        69.5894     0.5979 116.395  < 2e-16 ***
genreAction        -1.8537     0.6971  -2.659  0.00785 ** 
genreAdventure     -3.4063     1.0940  -3.114  0.00186 ** 
genreFighting       0.5316     0.9550   0.557  0.57781    
genreMisc          -2.0850     0.9483  -2.199  0.02794 *  
genrePlatform       0.7517     0.9491   0.792  0.42841    
genrePuzzle         1.5207     1.4431   1.054  0.29200    
genreRole-Playing   3.2440     0.8064   4.023 5.82e-05 ***
genreShooter        1.3630     0.7703   1.769  0.07688 .  
genreSimulation     0.3384     1.0355   0.327  0.74384    
genreSports         4.5769     0.7610   6.014 1.91e-09 ***
genreStrategy       3.4106     1.0712   3.184  0.00146 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.71 on 6131 degrees of freedom
Multiple R-squared:  0.02979,	Adjusted R-squared:  0.02805 
F-statistic: 17.11 on 11 and 6131 DF,  p-value: < 2.2e-16
```

C. (1 point) Write the regression equation below, with the value of the coefficients you found above (two decimal points is enough, for each coefficient). (it should be in the form of y=bo+b1var1+ …). Yes, it's going to be a slightly long equation!

**Regression equation:**
y = 69.59 - 1.85(genreAction) - 3.41(genreAdventure) + 0.53(genreFighting) -2.09(genreMisc) + 0.75(genrePlatform) + 1.52(genrePuzzle) + 3.24(genreRole-Playing) + 1.36(genreShooter) + 0.34(genreSimulation) + 4.58(genreSports) + 3.41(genreStrategy)

D. (1 point) Which game genres did you find to have a statistically significant higher score than Racing games (at the 1% significance level)?

**Higher score:**
Role-playing, Sports & Strategy

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

E. (1 point) Which games have a statistically significant lower score than Racing games (at the 1% significance level)?

**Lower score:**
Action & Adventure

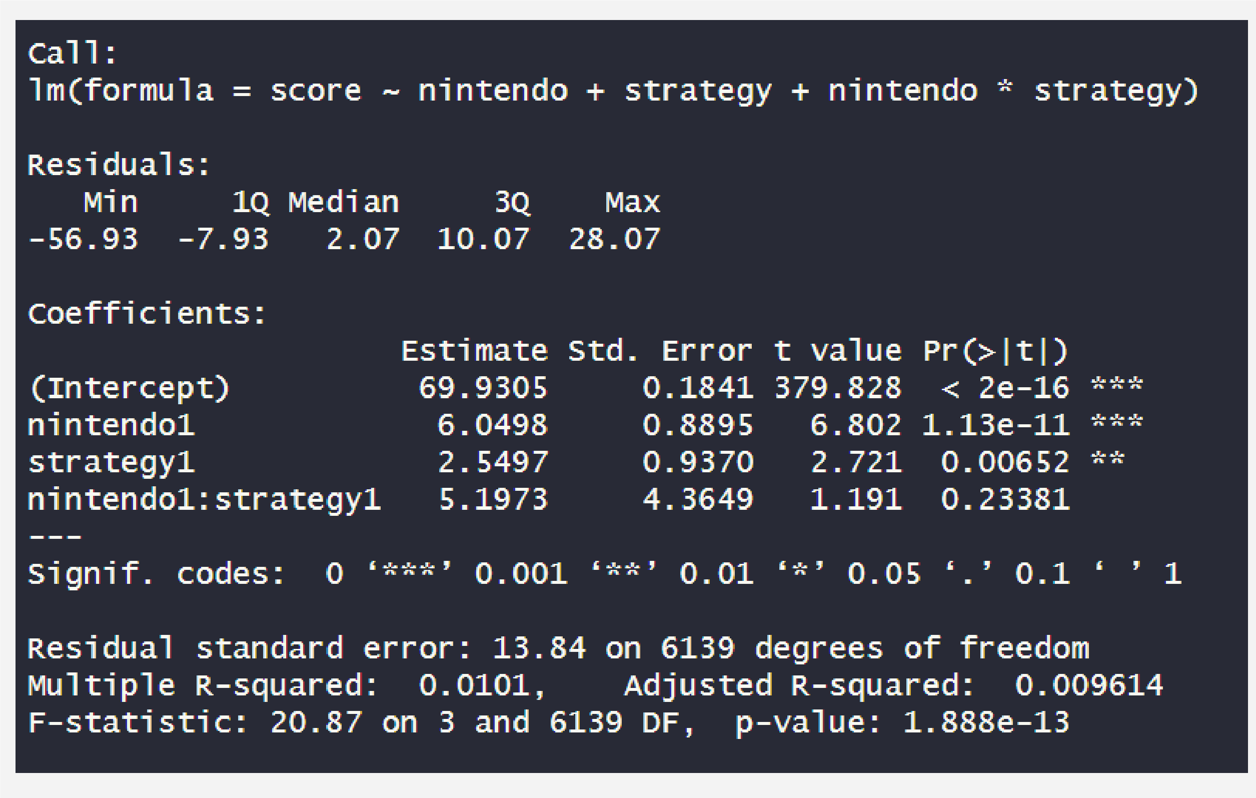
© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

# 7. Interaction terms (5 points - Lecture 3)

A. (1 point) Run the following interaction model:

score = $b_o$+ $b_1$(Nintendo)+ $b_2$(strategy)+$b_3$(strategy*Nintendo)

*Nintendo* is the variable you created in the previous question. S*trategy* is a variable that you need to create in this question, where *strategy*=1 if the genre of the game is *strategy* and *strategy*=0 if the game has another genre. Paste the regression output below:



```
Call:
lm(formula = score ~ nintendo + strategy + nintendo * strategy)

Residuals:
   Min     1Q Median     3Q    Max 
-56.93  -7.93   2.07  10.07  28.07 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         69.9305     0.1841 379.828  < 2e-16 ***
nintendo1            6.0498     0.8895   6.802 1.13e-11 ***
strategy1            2.5497     0.9370   2.721  0.00652 ** 
nintendo1:strategy1  5.1973     4.3649   1.191  0.23381    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.84 on 6139 degrees of freedom
Multiple R-squared:  0.0101,	Adjusted R-squared:  0.009614 
F-statistic: 20.87 on 3 and 6139 DF,  p-value: 1.888e-13
```

B. (1 point) In two sentences, what does the coefficient $b_3$ tell you?

**Your answer:**
The coefficient $b_3$ tells the increase in critic score for a game published by Nintendo with a genre of “Strategy”.

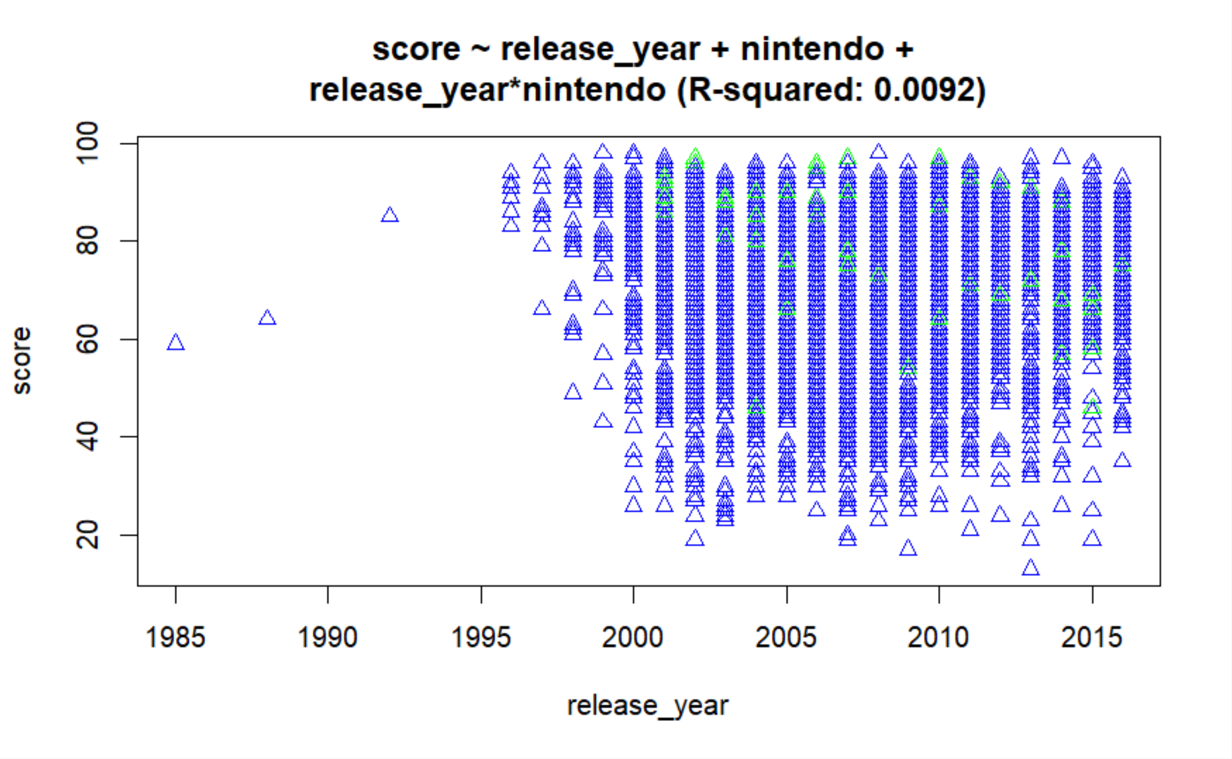
© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

C. (2 points) Consider the following model:

score = $b_o$+ $b_1$(release_year) + $b_2$(Nintendo) + $b_3$(release_year*Nintendo)

I want you to run the regression for the above model. Then draw the regression lines for (i) Nintendo games and (ii) non-Nintendo games.

- Games that are published by Nintendo should be in green; games that aren't should be blue (both the regression lines and the scatter plot dots).
- Make sure you create a legend.
- Instead of circles, I want triangles in the dots of the scatter plot (you will need to figure this on your own).
- I want the regression lines to have a width of 2 and be a dashed line (you will need to figure how to make them dashed on your own).



D. (1 point) Based on the above regression, what can you say about the quality of Nintendo games throughout the years?

**Your answer:**
Based on the above regression, the quality of Nintendo games has gone down throughout the years, as evidenced by a declining line.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.

- Submission: Please save in colour as a PDF and submit via MyCourses. If you don't submit a color PDF, there will be a 2-point penalty.

- Code: Submit code in a separate file.

© Juan Camilo Serpa. This content is copyrighted. Do not upload or distribute without written permission.