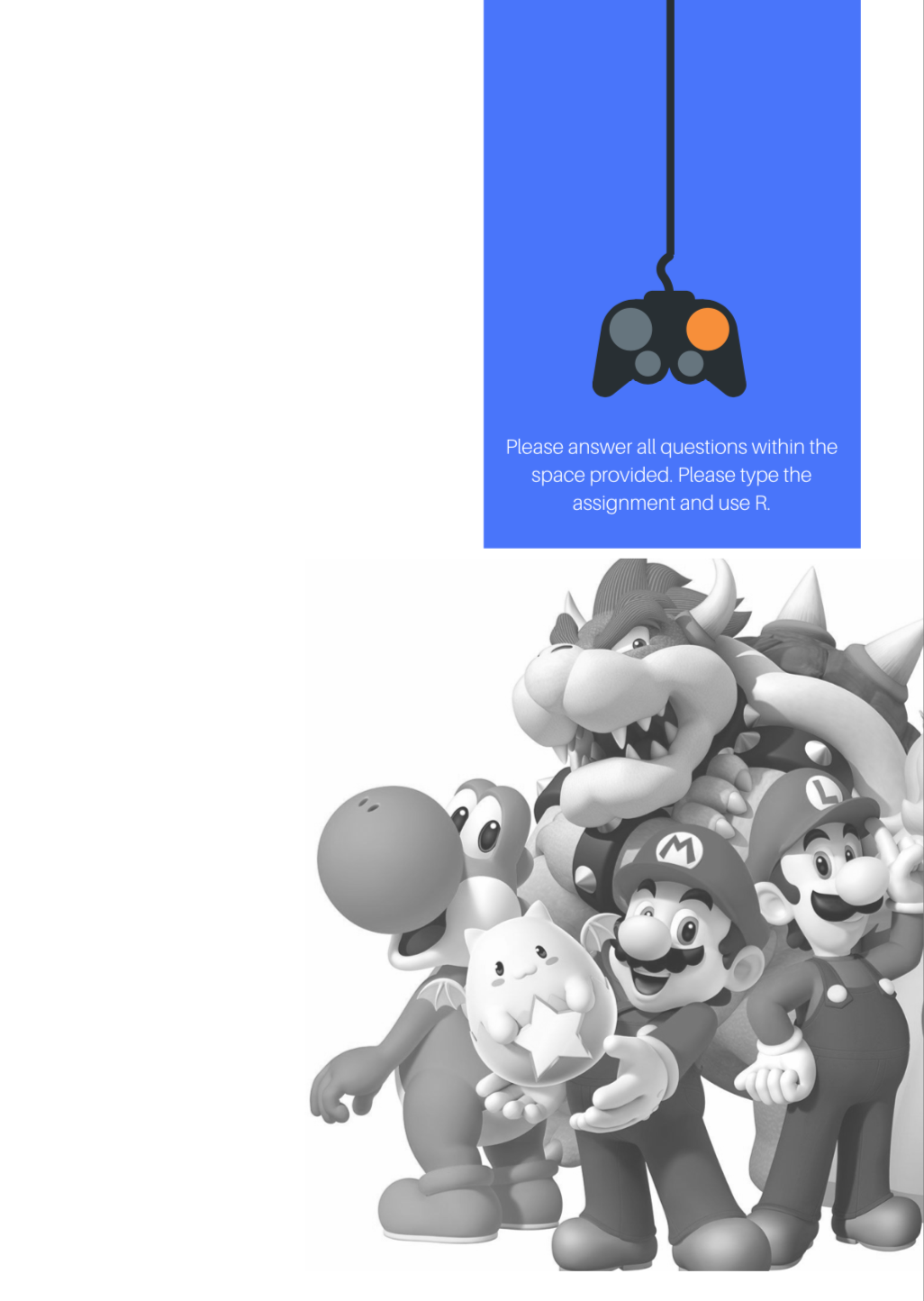
II

| LAB 1: LINEAR REGRESSION  Due date : September 18, 2023, 11:59 PM  **To access lab, go to file>make a copy.**  **The lab will be in your own google drive, and you can modify it.** |
| --- |

Your last name:

Your first name:

Your student ID:

Your last name:

Your first name:

Your student ID:

**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 |
| --- |

# 1. [Visualizing variables](https://www.youtube.com/watch?v=mnipB_8Br8U&frags=pl%2Cwn) (15 points - Lecture 2)

Let’s open the video\_games dataset.For each of the variables below, provide the information requested.

| i) Variable: score |
| --- |

1. **(0.5 points)** Summary statistics

* Min:
* Max:
* Quartile 1:
* Quartile 3:
* Median:
* Mean:

1. (1 point) Create a box plot and attach it below. The inside of the box plot should be blue.[[1]](#footnote-0)

| **Insert box plot here** |
| --- |

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

| **Insert histogram here** |
| --- |

| ii) Variable: sales\_global |
| --- |

1. (0.5 points) Summary statistics

* Min:
* Max:
* Quartile 1:
* Quartile 3:
* Median:
* Mean:

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

| **Insert box plot here** |
| --- |

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

| **Insert histogram here** |
| --- |

| iii) Variable: release\_year |
| --- |

1. (0.5 points) Summary statistics

* Min:
* Max:
* Quartile 1:
* Quartile 3:
* Median:
* Mean:

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

| **Insert box plot here** |
| --- |

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

| **Insert histogram here** |
| --- |

| iv) Variable: count\_critic |
| --- |

1. (0.5 points) Summary statistics

* Min:
* Max:
* Quartile 1:
* Quartile 3:
* Median:
* Mean:

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

| **Insert box plot here** |
| --- |

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

| **Insert histogram here** |
| --- |

1. (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.

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.

| **Insert matrix of plots here** |
| --- |

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

Run three simple linear regressions (**Y=bo+b1x**) — one for each of the three predictors. Attach the regression results.

| score = bo+ b1(sales\_global) |
| --- |

1. (2 points) **Fill in the blanks:**

* bo: \_\_\_\_\_\_\_\_\_
* b1: \_\_\_\_\_\_\_\_\_
* r-squared: \_\_\_\_\_\_\_\_\_\_
* For b1, please provide:
  + 95% Confidence interval: \_\_\_\_\_\_\_\_
  + t-test statistic: \_\_\_\_\_\_

What is the probability that there **is not** a statistically significant relationship between these two variables? \_\_\_\_\_\_\_\_\_

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

| **Insert graph here** |
| --- |

| score = bo+ b1(release\_year) |
| --- |

1. (2 points) Fill in the blanks:

* bo: \_\_\_\_\_\_\_\_
* b1: \_\_\_\_\_\_\_
* r-squared: \_\_\_\_\_\_\_\_\_\_
* For b1, please provide:
  + 95% Confidence interval: \_\_\_\_\_\_\_\_
  + t-test statistic: \_\_\_\_\_\_\_

What is the probability that there **is** **not** a statistically significant relationship between these two variables? \_\_\_\_\_\_\_\_\_

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

| **Insert graph here** |
| --- |

| score = bo+ b1(count\_critic) |
| --- |

1. (2 points) Fill in the blanks:

* bo: \_\_\_\_\_\_\_\_\_
* b1: \_\_\_\_\_\_\_\_\_
* r-squared: \_\_\_\_\_\_\_\_
* For b1, please provide:
  + 95% Confidence interval: \_\_\_\_\_\_\_\_
  + t-test statistic: \_\_\_\_\_\_\_

What is the probability that there is **not** a statistically significant relationship between these two variables? \_\_\_\_\_\_\_\_

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

| **Insert graph here** |
| --- |

1. (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:** |
| --- |

# 

# 3. Predictions (3 points - Lecture 2)

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

1. (1 point) Consider the simple regression (*score*=b0+b1*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:** |
| --- |

1. (1 point) Consider the simple regression (*score*=b0+b1*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:** |
| --- |

1. (1 point) Consider the simple regression (*score*=b0+b1*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:** |
| --- |

***Hint: Use the “Coef()” function***

# 

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

Suppose we are thinking of running the following multiple regression:

| score = bo+ b1(sales\_global) + b2(release\_year) + b3(count\_critic) |
| --- |

1. (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:** |
| --- |

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

| **Insert a screenshot of your R-output here** |
| --- |

1. (1 point) Which of the above coefficients are statistically significant at the 99% level?

| **Your answer:** |
| --- |

1. (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:** |
| --- |

1. (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:** |
| --- |

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

1. (2 points) Run the following model:

| score = bo+ b1(release\_year) + b2(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:**   * bo: * b1: * b2: |
| --- |

1. (2 points) Now, interpret the coefficient b2 to a manager who knows nothing about statistics. Avoid jargon.

| **Your answer:** |
| --- |

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).

| **Insert your plot here** |
| --- |

# 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:

| **Insert a screenshot of your R-output here** |
| --- |

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):

| **Insert a screenshot of your R-output here** |
| --- |

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:** |
| --- |

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

| **Higher score:** |
| --- |

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

| **Lower score:** |
| --- |

# 

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

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

| score = bo+ b1(Nintendo)+ b2(strategy)+b3(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:

| **Insert a screenshot of your R-output here** |
| --- |

1. (1 point) In two sentences, what does the coefficient b3 tell you?

| **Your answer:** |
| --- |

1. (2 points) Consider the following model:

| score = bo+ b1(release\_year) + b2(Nintendo) + b3(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).

| **Insert your plot here** |
| --- |

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

| **Your answer:** |
| --- |



* **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.**

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! [↑](#footnote-ref-0)
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. [↑](#footnote-ref-1)