Correlation and Linear Regression Lab

2025-09-30

Have you ever been to Boston? It's actually one of my favorite cities to explore. Today, we are going to dig into one of the annual events the city hosts called the Boston Marathon. In particular, we will be working with the top 100 fastest finishers from the 2023 marathon. Please read in the below data and load the relevant packages

```
library(ggplot2)

runners <- read.csv( 'https://vinnys-classes.github.io/data/boston_marathon_2023.csv')

head(runners)
```

```
##   age_group place_overall place_gender place_division            name  team
## 1     18-39             1            1              1   Chebet, Evans Team-
## 2     18-39             2            2              2   Geay, Gabriel Team-
## 3     18-39             3            3              3 Kipruto, Benson Team-
## 4     18-39             4            4              4   Korir, Albert Team-
## 5     18-39             5            5              5  Talbi, Zouhair Team-
## 6     18-39             6            6              6 Kipchoge, Eliud Team-
##   bib_number half_time finish_net finish_gun gender half_time_sec
## 1          1 1H 2M 20S  2H 5M 54S  2H 5M 54S      M          3740
## 2          3 1H 2M 20S   2H 6M 4S   2H 6M 4S      M          3740
## 3          5 1H 2M 19S   2H 6M 6S   2H 6M 6S      M          3739
## 4         19 1H 2M 20S   2H 8M 1S   2H 8M 1S      M          3740
## 5         31 1H 2M 20S  2H 8M 35S  2H 8M 35S      M          3740
## 6          2 1H 2M 19S  2H 9M 23S  2H 9M 23S      M          3739
##   finish_net_sec finish_gun_sec finish_net_minutes
## 1           7554           7554           125.9000
## 2           7564           7564           126.0667
## 3           7566           7566           126.1000
## 4           7681           7681           128.0167
## 5           7715           7715           128.5833
## 6           7763           7763           129.3833
```

The variables we will be most interested in is...

finish_net_sec Net time to run the race from when they crossed the starting line to when they crossed the finish line

finish_gun_sec Time to run the race from when the starter pistol was fired until the runner crossed the finish line

half_time_sec The time taken to reach the halfway point of the race.

place_overall Gives the rankings of the runners. The fastest runner is given place 1 based on their finish_net_sec value.

Scatterplots

Q1

Create a scatterplot of the variable half_time_sec and finish_net_sec. State which one you believe should be the explanatory variable and which should be the response variable.

Q2

Describe the relationship between time it takes to get to the half way point and the time it takes to run the full marathon using the plot you made. Be sure to mention outliers, form, strength, direction please.

Correlations

Q3

Is it better to use Pearson's correlation coeffient or Spearman's correlation coefficent to quantify the relationship between the two variables?

Q4

Calculate and report both correlation coeffiencts from question 3. HINT: use the cor() function and it's help page to find how to calculate both.

Q5

Calculate a correlation matrix for half_time_sec, finish_net_sec, and finish_gun_sec. HINT: there is a code along (on correlations) that does this.

Q6

Comment on which of pair of the variables is most heavily correlated. Does this make sense?

Q7

Describe what type of variable "place_overall" is.

Q8

To find the association between "place_overall" and "finish_net_sec", please state which of the two types of correlation we discussed that you'd use and why. Please find that correlation.

Linear Regression

Q9

Build a linear model using finish_net_sec as the response variable and half_time_sec as the explanatory variable. In order to do this replace parts of the following code with the correct variables and remove the hashtags

```
#my_model <- lm(RESPONSE ~ EXPLANATORY_VAR,
#                data = DATA)
```

Q10

Using the predict() and resid() functions to calculate the predictions and the residuals of your model, respectively, create a scatterplot with the predicted values on the x-axis and the residuals on the y-axis.

Q11

We need to check our assumptions. Things that the plot can help check is...

1) homogenity (constant variance)
2) normality (equal spread of points above and below the line without any notable patterns)

Q12

Check for independence. That is, ask yourself if there is there a reasonable way one race runner's time could affect another runner's time in a noticable way? Would knowing information about any given runner tell us information about another runner?

Q13

Given your answers to Q2 (form = linear relationship), Q11 (homogeniety and normality), and Q12 (Independence) please state whether you believe the assumptions for the linear model are met.

Q14

Calculate the Pearson correlation coefficient for between the residuals and the predicted values. Why does this value make sense?

Q15

Regardless of how you answered Q13, please write down the estimated linear regression equation. Be sure to say the left hand side of the regression equation is the ESTIMATED response variable or you can use the ˆ symbol. Either works.

HINT: Apply the summary() function to the model and look for the section entitled "Coefficients". The estimates will be below the word "Estimate" (and your values should be close-ish to 2,000 and 1.5).

NOTE: When R uses e+03 it means it wants the left hand number multipled by 10ˆ3... it's how R represents scientific notation. For example 5e+02 = 5*(10ˆ2) = 5*100 = 500. Similarly 2e-04 means 2 * (10ˆ(-4)).

Q16

Please interpret the slope estimate. Does this seem reasonable? Please be sure to use the name of the variables and units in the context of the problem and don't use "response variable".

Q17

Please interpret the intercept estimate. Does this value make sense in this context? Please be sure to use the name of the variables and units 1in the context of the problem and don't use "response variable".

Q19

Please make a prediction for the time it'll take a runner who reachs the halfway mark after 4,000 seconds to finish the race. Be sure to indicate that the time to finish is *estimated* or *predicted* and not the exact time it'll take.

Q20

Copy and paste the code from question Q2 here. Add the best-fit-line by adding geom_smooth(method = 'lm') to your code (make sure you use a +).

R^2

Q21

Report the value for R^2 for your linear model. In the summary() function's output near the bottom you'll see the line Multiple R-squared. That is the number we want.

NOTE: Adjusted R^2 is a distinct statistic that we are not focusing on currently.

Q22

The following is intricate and you MUST be detail orientated when you go through these steps. Calculate the following:

1) Square your residuals from the model.

HINT: To raise a number to an exponent you use ^ symbol. Eg 3^2 = 3 squared = 3 * 3 = 9.

2) Sum your answer to question 1. This value is called the Sum of the Squares of the Errors (SSE), among other names.

HINT: the function sum() will be useful for this.

3) Find the mean of finish_net_sec.

4) Subtract the mean of finish_net_sec from the actual finish_net_sec values.

HINT: The easiest way to do this for a variable MY_VAR is the code MY_VAR - mean(MY_VAR)

5) Take your answer to 4 and square the values.

6) Find the sum of your result in question 5. This value is called the Total Sum of Squares (TSS), among other names.

7) Subtract the mean of finish_net_sec from your predictions from your model.

8) Take your answer to 7 and square the values.

9) Find the sum of your answer to question 8. This value is called the Sum of the Squares of the Model (SSM), among other names.

10) Add your answers from question 9 and question 2. Note how this compares to your answer to question 6.

11) Divide your answer to question 9 by your answer to question 6. Compare this with the R^2 value your found earlier.

Q23

Interpret your R^2 value.