BIOSTAT 702: Exercise 4.1

Simple Linear Regression: Calculations

Fall 2025

Contents

Learning Objectives

1. To practice some of the details around the various SLM calculations, including the derivation of the predictions and the ANOVA table. Our premise is that by programming these yourself, one step at a time, you'll attain a deeper understanding of the content.
2. To practice two types of R programming: (1) use of standard R functions to perform statistical analysis; and (2) use of matrices to perform calculations and create/extend statistical procedures.

How to Do This Exercise

We recommend that you read this entire document prior to answering any of the questions. If anything is unclear please ask for help from the instructors or TAs before getting started. You are also allowed to ask for help from the instructors or TAs while you are working on the assignment. You may collaborate with your classmates on this assignment—in fact, we encourage this–and use any technology resources available to you, including Internet searches, generative AI tools, etc. However, if you collaborate with others on this

assignment please be aware that *you must submit answers to the questions written in your own words. This means that you should not quote phrases from other sources, including AI tools, even with proper attribution.* Although quoting with proper attribution is good scholarly practice, it will be considered failure to follow the instructions for this assignment and you will be asked to revise and resubmit your answer. In this eventuality, points may be deducted in accordance with the grading rubric for this assignment as described below. Finally, you do not need to cite sources that you used to answer the questions for this assignment.

Grading Rubric

The assignment is worth 20 points (4 points per question). The points for each question are awarded as follows: 3 points for answering all parts of the question and following directions, and 1 point for a correct answer. Partial credit may be awarded at the instructor's discretion.

Resources

We will be using the original ultrarunning dataset and associated paper for this exercise.

Question 1

Remember from Exercise 1.1 that we are interested in quantifying the effect of emotional intelligence (teique_sf) on ultramarathon times (pb100k_dec). To do so, we will first apply the matrix version of the SLR model.

1. Prepare your data. Select only the variables you need and remove any observations with missing data for the purposes of this analysis. We typically add a column of 1's (when doing the calculations "by hand" like this) to represent the intercept. Append a column of 1's to your data.
2. We already looked at univariate distributions of these variables in previous exercises. Now, make a scatterplot of personal best 100k times in hours vs emotional intelligence score. Append a best fit line.
3. Transform your data into a matrix X (which should be of dimension $2 \times n$) and a vector Y.
4. Multiply $(X'X)^{-1}$ by $X'Y$. Name the resulting matrix Beta. Print Beta, which contains the parameter estimates from the SLR model. What are they?

Question 2

1. Now run a SLR on your data using the `lm()` function to create an output object named lm_obj (note you no longer need the intercept column). Then run the `summary()` function on lm_obj. Among others, a table of parameter estimates is printed. Do these parameter estimates match what you obtained using matrix operations?
2. Although we will go into the composition of lm_obj in more detail elsewhere, run the `names()` function to find the names of its components. How many components do you see?
3. Print `lm_obj$coefficients`. What does the output represent?
4. In R code, how would you refer to the element of `lm_obj$coefficients` that contains β_1?

5. Using the information from `names()`, assign the fitted values to a new object named Fitted and print the first 5 fitted values

6. Verify that the components of Fitted are identical to the result of applying the `predict()` function to lm_obj.

7. Check the calculation of the first fitted value – it should be $11.03 + (0.71 * 5.73)$, since their emotional intelligence score is 5.73. Does the calculation match the first fitted value?

Question 3

Now we will generate the inputs to the ANOVA table.

1. The first task is to collect Y, Y_m, and Y_p, all as vectors. You already have Y and Y_p. Calculate the mean of Y and copy the results into a vector using the `rep()` function.

2. Create a vector that (1) calculates $Y - Y_m$; and then (2) squares the results. Sum the elements of this vector, and name the result SST.

3. Create a vector that (1) calculates $Y - Y_p$; and then (2) squares the results. Sum the elements of this vector, and name the result SSE.

4. Create a vector that (1) calculates $Y_p - Y_m$; and then (2) squares the results. Sum the elements of this vector, and name the result SSR.

5. You have now calculated the 3 sums of squares in the ANOVA table. What are their values?

6. Generate an ANOVA table directly from lm_obj by applying the `anova()` function. Do you obtain the same sums of squares? As a default, R doesn't print SST. From a statistical perspective, why not?

7. How could you use the output object from the `anova()` function to calculate SST? Verify that your code works.

Question 4

The variance-covariance matrix contains the variance of β_0, the variance of β_1, and the covariance between β_0 and β_1. We'll focus on the variance of β_1.

1. The `vcov()` function calculates the variance-covariance matrix directly. Apply this function to lm_obj. What is the value of the variance of β_1? Take the square root of the variance of β_1. Does it equal the standard error of the slope coefficient from summary(lm_obj)?

2. Now consider the algebraic formula for $Var(\beta_1) : \{\sum_i (Y_i - Y_p)^2/(n-2)\}/\sum_i (X_i - X_m)^2$. The standard error is the square root of this. You have already created most of the elements of this formula. Verify that this formula yields an identical estimate of se(β_1).

3. The formula contains 3 terms. Here, $(n-2)$, the degrees of freedom associated with β_1, serves the purpose of appropriately accounting for the sample size. Considering $\sum_i (Y_i - Y_p)^2$, under what circumstances will this be small? In particular, when the model fits well will this term be large or small?

4. Considering $\sum_i (X_i - X_m)^2$, under what circumstances will this be large? When you are designing an experiment, and thus have the ability to assign the values of X, should they have a little variation or a lot of variation? Remember: se(β_1) will be small when the numerator is small and the denominator is large.

Question 5

Now, consider the null hypothesis $\beta_1 = 0$. We know that we can test this hypothesis using a t-test or an F-test.

1. Run the t-test "by hand", using the formula. Do the t-statistic and p-value match what you found from the lm() function?
2. Run the F-Test "by hand", using the formula. Do the F-statistic and p-value match what you found from the anova() function?
3. Verify that the F-statistic is the square of the t-statistic
4. Is there statistical evidence to conclude that emotional intelligence has an effect on ultramarathon times? (Assume a critical value $\alpha = 0.05$). Interpret the slope estimate in the context of the problem as well.
5. Do you think the effect is clinically significant?