BIOSTAT 702: Exercise 4.2

Simple Linear Regression: Visualization and Assumptions

Fall 2025

Contents

Learning Objectives **1**

How to Do This Exercise **1**

Grading Rubric **2**

Question 1 **2**

Question 2 **2**

Question 3 **2**

Question 4 **3**

Question 5 **3**

Learning Objectives

1. Practice visualizing data with 1 continuous predictor and one continuous outcome to assess validity of assumptions of SLR
2. Practice performing simulations to help develop intuition regarding the assumptions of SLR

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.

Question 1

We will be simulating data from a designed experiment, with 100 patients. Patient 1 receives X=1, patient 2 receives X=2, ..., patient 100 receives X=100.

1. Simulate a SLR model meeting all its assumptions. The slope should be 0.5, the intercept should be 2, and the errors should be normal with mean 0 and standard deviation 1. Set the seed value for the pseudorandom number generator to 1, so that everyone obtains identical results.
2. Create a scatterplot of Y vs X and add the best-fitting regression line to the plot. How good does the fit appear to be?
3. Create a Q-Q plot to check the normality of the residuals. Do the residuals appear to be normal?
4. Create a residual plot to check the homogeneity of variances assumption. Do the variances look homogeneous?
5. Fit a SLR model and calculate the R-square statistic. What is it ?

Question 2

1. Repeat the above analyses, but change the standard deviation to 15. What do you find?

   Take the scatterplot and add a LOESS function. Visually, how close does the smoothed curve look like a straight line?
2. Repeat the above analysis, but change the standard deviation to 100. What do you find? Even though we know that the signal in the data still exists, the level of noise is sufficiently high to make it impossible to find.

Question 3

1. Repeat the above analysis, with the regression parameters as before, but change the distribution of the error term to 10*K, where K is a random number derived from a t-distribution with 4 degrees of freedom. Does the scatterplot still look good? What about the Q-Q plot? Specifically, the t-distribution has heavier tails than the normal. How can this be discovered in the Q-Q plot?
2. Repeat the above analysis, but change the distribution of the errors to exp(1/5). How does the shape of the Q-Q plot differ from that of a Q-Q plot with errors that follow a t-distribution?

Question 4

Now, let's create an example where the regression function is correct but the assumption about the error term isn't. Change the distribution of the error term to $X * N(0, 1)$. Thus, the level of noise should increase as X increases. Does the scatterplot look like a fan? What does the residual plot show? Why isn't the Q-Q plot perfectly consistent with a normal distribution? In practice, the next step in response to a scatterplot that looks like this might be to research models with a non-constant error term.

Question 5

Now, let's create examples where a transformation of either Y or X is needed.

1. Return to the N(0,15) case, and set $Y^* = Y^2$. So, we know by construction that a square root transformation will be ideal, although that might or might not be apparent from the data. What do the various plots show? You should find that the variability of Y* increases as X increases, although the functional form of the relationship might not be obvious. However, this scatterplot pattern suggests trying transformations of Y* such as $\log(Y^*)$ and $\sqrt{Y^*}$, since if they work (i.e., fix the functional form and stabilize the variability) you can use a simple model with a constant error term.
2. Now, let's create an example where the error term is correct but the regression function isn't. Keep the N(0,15) error, but change the regression function to $Y = 0.1 * (X - 50)^2$. The issue should be clear from simply looking at the scatterplot, but go ahead and apply a loess fit (it should look quadratic) and append a best-fitting straight line (it should look awful). Using this information, as an analyst your first thought would probably be to fit a quadratic function.