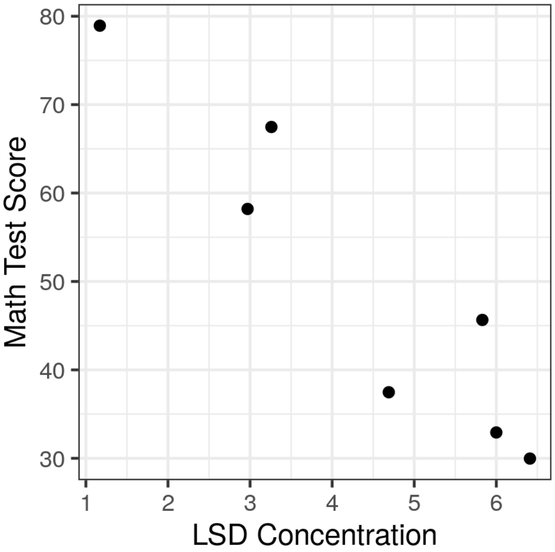
Chapter 10 Scenarios

Math on LSD

In the 1960s, there was considerable research into the effects of drugs such as LSD. In one experiment, researchers gave a group of 5 volunteers 2 micrograms of LSD intravenously, and then administered math tests to the participants at 7 timepoints (5, 15, 30, 60, 120, 240, 480 minutes) while measuring the concentration of LSD in the volunteers tissue. The average of the 5 scores, and the average concentration of LSD were recorded.



LSD conc.	Math Score
1.17	78.9
2.97	58.2
3.26	67.5
4.69	37.5
5.83	45.6
6.00	32.9
6.41	30.0

Wagner, Agahajanian, and Bing (1968). Correlation of Performance Test Scores with Tissue Concentration of Lysergic Acid Diethylamide in Human Subjects. Clinical Pharmacology and Therapeutics, Vol.9 pp635-638.

1. What is the explanatory variable? What is the response variable?

   The explanatory variable is average LSD tissue concentration across subjects. The response variable is average math test score.

2. Is this an experiment? Does it have random sampling? Random assignment?

   This is an experiment, because researchers are intervening and assigning treatments to participants. It does not have random sampling (the participants are volunteers). It also does not have random assignment - while the researchers are assigning treatments (LSD injection, observation intervals) they do not control the LSD concentration directly, so they are not assigning levels of the explanatory variable to the response variable.

3. What conclusions can you make as a result of this experimental design? What conclusions can you not make? Why?

   We cannot make causal inferences (because we lack random assignment). We can argue for generalizing this result to the population, with the argument that LSD concentration would be expected to decrease similarly in all humans with normal metabolisms, but we have to make the argument - it is not a given that our sample generalizes to our population because we do not have a random sample of the population.

4. Describe the relationship between LSD concentration and math test score (form, direction, strength,

high influence points)

There is a negative linear relationship between LSD concentration and math test score; while it is difficult to determine the strength because there are only 7 points, it does appear that most of the points lie very close to the imaginary line of best fit. There are no high influence points (there are only 7 points, so that isn't remarkable).

5. What do you know at this point about the correlation coefficient? What do you know about the value of b, the slope of the regression line?

   The direction of the relationship is negative, which means that the correlation coefficient, r, as well as the slope, are negative. Thus, the correlation coefficient will be $-1 \leq r < 0$ *and the slope will be* $b < 0$ *(but because slope can take any value, positive or negative, we can't come up with a lower bound).*

6. Enter the data in the table above into the Two Quantitative Variables/regression applet. Get the equation of the fitted regression line and the sample correlation coefficient from the applet.

   - Regression equation:

     $$\widehat{Score} = 89.12 - 9.01(LSD)$$

   - Correlation coefficient:

     $$r = -0.937$$

7. Interpret the value of the intercept in the regression equation in the context of the problem.

   The sample intercept is 89.12 and is the predicted average score on the math test when LSD tissue concentration is equal to 0.

8. Interpret the value of the slope in the regression equation in the context of the problem.

   The sample slope is -9.01, which means that as LSD concentration increases by 1, the predicted average math score will decrease by about 9 points.

9. What is the null hypothesis, in words (you do not have to specify a specific statistic or parameter)? What is the alternative hypothesis?

   H_0 : *There is no relationship between participants' average LSD tissue concentration and participants' average score on a mathematics test.*

   H_A: *There is a relationship between participants' average LSD tissue concentration and participants' average score on a mathematics test.*

10. Conduct a simulation study designed to test the hypothesis. Select the slope as your statistic of choice. Report your p-value and interpret the results. Can you reject the null hypothesis?

    $p = 2/1000 = 0.002$

    With $p = 0.002$, *I have very strong evidence against* H_0 *that there is no relationship between participants' average performance on the math test and participants' tissue LSD concentration. I reject* H_0 *and conclude that there is a relationship between participants' average performance on the math test and participants' tissue LSD concentration.*

11. Without re-shuffling, change the statistic of interest to the correlation coefficient. Report your p-value and interpret the results. Can you reject the null hypothesis? How does your answer compare to the answer on the previous question?

    $p = 2/1000 = 0.002$

    With $p = 0.002$, *I have very strong evidence against* H_0 *that there is no relationship between participants' average performance on the math test and participants' LSD tissue concentration. I reject* H_0 *and conclude that there is a relationship between participants' average performance on the math test and participants' LSD tissue concentration.*

    The hypothesis test using ρ, *the correlation coefficient, as the statistic has the same results as the hypothesis test using* β, *the slope as the statistic. Both statistics are measures of linear association, so the results of the two simulations should be the same (so long as we do not re-shuffle.)*

12. What are the validity conditions for theory-based inference on the slope? Are these conditions met?

    The validity conditions are 1) the points follow a linear trend, 2) the data is distributed symmetrically about the regression line, and 3) the variability around the regression line is approximately equal.

    These conditions are met (though the equal variance condition is hard to determine with only 7 points). We can use theory-based inference for the slope.

13. Check the box in the center-left of the applet that says "Regression Table". Using the table values, create and interpret a 2SD theory-based confidence interval for the slope. (Note: using the 95% interval provided by the applet will not get the correct answer - do the calculation by hand)

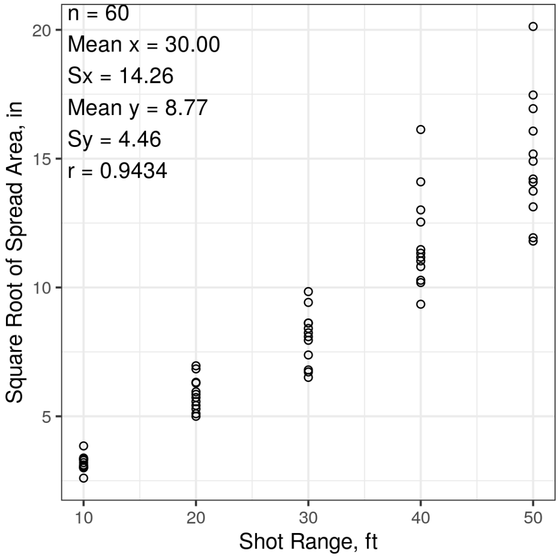
    $statistic \pm 2SE(statistic) = -9.01 \pm 2 * 1.50 = (-12.01, -6.01)$
    I am 95% confident that as participants' average LSD tissue concentration increases by 1, the predicted average score on the math test will decrease between 6 and 12 points.

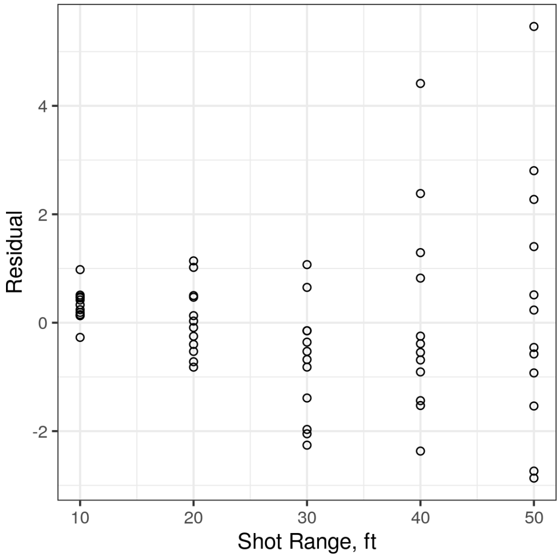
14. Calculate the coefficient of determination (R^2). Interpret it in the context of the problem.

    $R^2 = 100 \times (-0.937)^2 = 87.8\%$. *The linear relationship between LSD tissue concentration and math test score accounts for 87.8% of the variability in math test score.*

Shotgun Scatter

Forensic firearms examiners often need to examine evidence at a crime scene to determine where the perpetrator was standing when a shot was fired. In one experiment, examiners measured the square root of the area spread of shotgun pellets to range of fire, for shotgun cartridge types, to provide data for estimation of firing distance from the spread of pellets at a crime scene.





1. What is the explanatory variable? What is the response variable?

   The explanatory variable is the range of the shot, in ft. The response variable is the square root of the spread area, in inches.

2. Describe the association between shot range and the square root of spread area.

   Short range and square root of spread area are positively associated, with a strong linear relationship. There are indications that the relationship weakens as shot range increases.

3. Calculate the value of the slope of the regression line.

   $$b = r\frac{S_y}{S_x} = 0.9434(4.46)/(14.26) = 0.295$$

4. Calculate the value of the intercept of the regression line.

   $$a = \overline{y} - b\overline{x} = 8.77 - 0.295(30.00) = -0.08$$

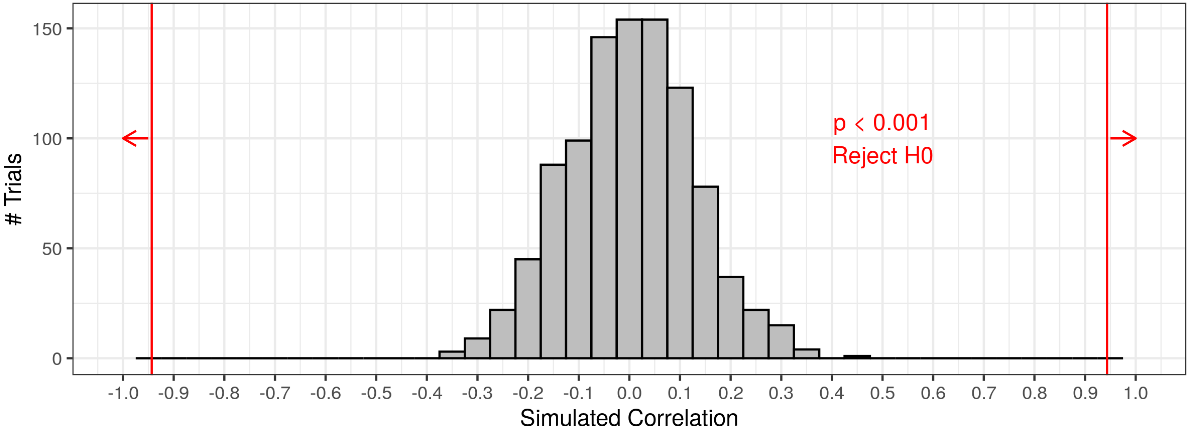
5. What does the value you got for the slope mean in the context of the problem?

   As shot range increases by 1 foot, the square root of the spread area of the pellets will increase by 0.265 inches.

6. What does the value you got for the intercept mean in the context of the problem? Does this make sense?

   At point blank range, the pellets will spread -0.08 inches. This doesn't completely make sense (you can't have a negative pellet spread), but if we think of the intercept as very close to 0, then it's reasonable - the spread of the shot at point blank range should be very close to 0.

7. To test the null hypothesis that there is no relationship between scatter area and shot range using the correlation coefficient, you conduct a simulation study using the correlation coefficient as the statistic. The simulated distribution is shown below. Draw one or more vertical lines indicating the cutoff for values considered "extreme" on the chart below. Shade in the extreme values and decide whether you think the p-value is low enough to reject the null hypothesis. (You do not have to calculate the p-value explicitly, just make your best guess from the area you shaded in.)



8. Using the p-value you estimated in the previous problem, what are your conclusions in the context of the problem?

   With $p < 0.001$ *I have very strong evidence against the null hypothesis that there is no association between shot range and the square root of the area of pellet spread. I reject* H_0 *and conclude that there is evidence of an association between shot range and the square root of the area of pellet spread.*

9. What are the validity conditions for theory-based inference? Are they met? Why or why not?

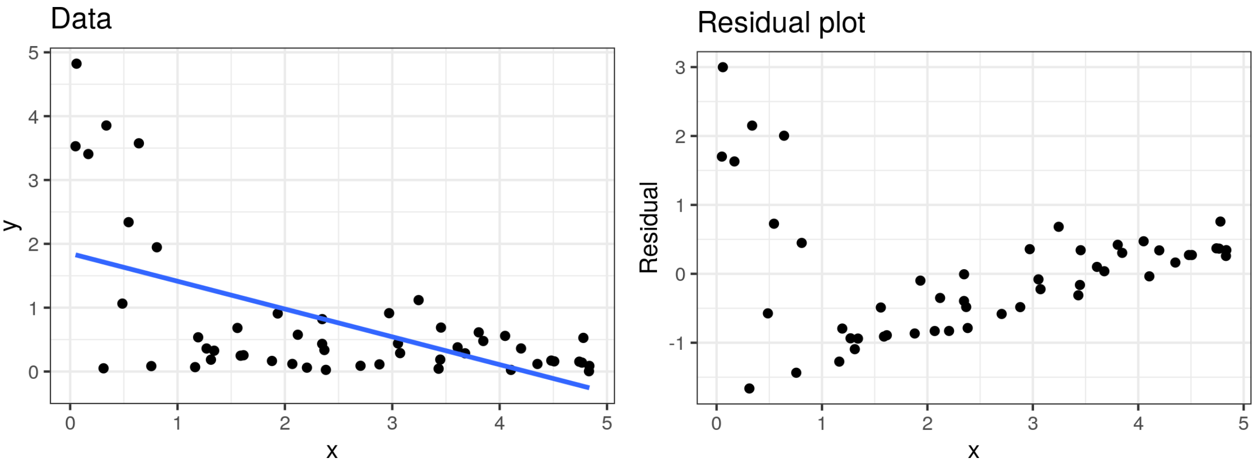
   The validity conditions are 1) the points follow a linear trend, 2) the data is distributed symmetrically about the regression line, and 3) the variability around the regression line is approximately equal.

   The points appear to follow a linear trend, but examination of the residual plot reveals a subtle quadratic curve that the linear regression does not account for. This causes asymmetry in the distribution of points around the regression line at each measured x value (violating assumption 2). Finally, it is very evident from both the data plot and the residual plot that the variability around the regression line increases with the shot range. So no, it is not appropriate to use theory-based inference on this problem, because all of the assumptions are violated.

Describing plots

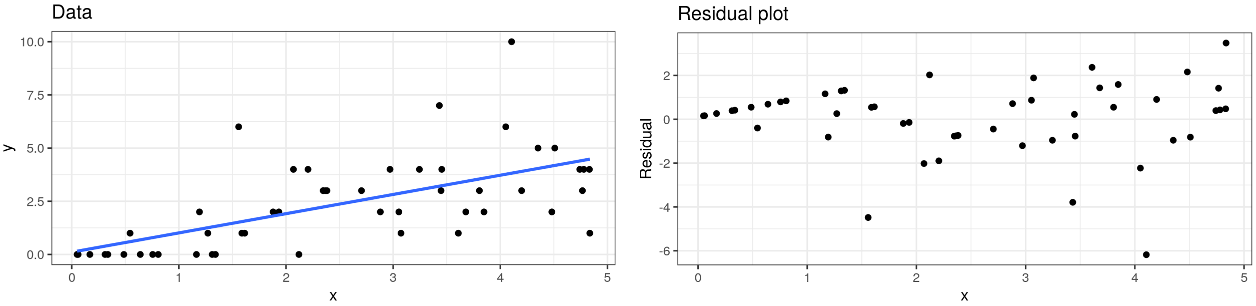
For each set of graphs below, describe the association between x and y and determine if theory based inference is appropriate using the corresponding residual plot.

1.



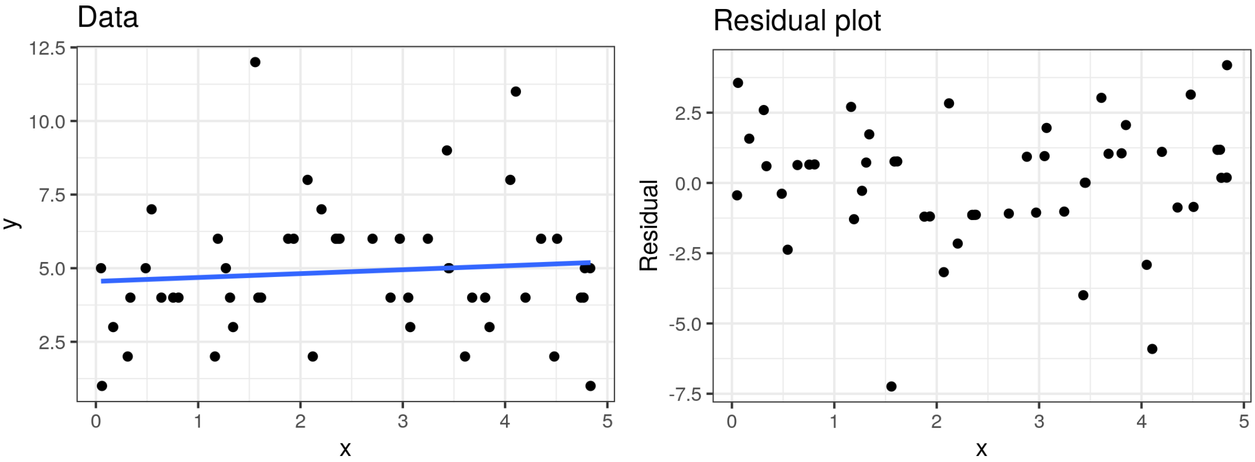
x and y are negatively associated, with a nonlinear relationship that shows a steep decrease in the value of y with an initial increase in x (e.g. similar to the form $y = 1/x$), but quickly levels off to a more homogeneous state. There are several high-influence points between 0 and 1 which will have an outsized effect on any linear regression line which is fit to these data. Theory-based inference is not appropriate for this data because there is evidence of nonlinearity, the points are not equally distributed above and below the regression line, and the variability of the points changes significantly with the value of x.

2.



x and y are positively associated and seem to have a moderately strong linear relationship. There is a high influence point at approximately $x = 4.5$, $y = 8$ which may have an outsized influence on the regression line. Theory-based inference is not appropriate for this data: even though points are symmetrically distributed around the regression line and there is no evidence of nonlinearity, there is evidence that the variability around the regression line increases with the value of x.

3.



x and y show no discernable association; it appears that the value of y does not change with respect to the value of x. There are two values which may be outliers in y at approximately $y = 12$ and $y = 11$; these points may affect the regression fit to some degree. Theory based inference is appropriate for this data as there is no evidence of nonlinearity, the points are symmetrically distributed about the regression line, and there does not appear to be any significant change in the variability in y.