Exploration 10.4-5: How do Invasive Plants Suppress the Growth of Native Trees?

Inference for Regression Slope

LEARNING GOALS

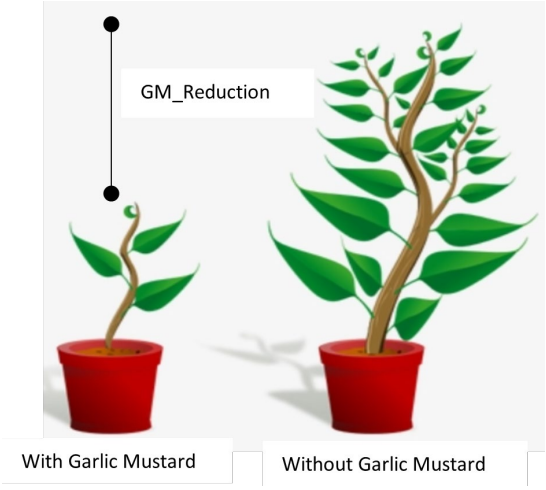
- Apply the 3S strategy when evaluating the hypothesis of association using the slope as the statistic.
- Articulate how to conduct a tactile simulation to implement the 3S strategy for testing a slope.
- Define the p-value in the context of the 3S strategy using simulated slopes under the null hypothesis of no association.
- Know that a test of association based on slope is equivalent to a test of association based on a correlation coefficient.
- Realize that both simulation-based approaches to testing correlation coefficients and slopes can, under certain conditions, be well predicted by the theory-based approach known as a *t*-test.
- Evaluate a scatterplot for the two validity conditions for a theory-based test of correlation coefficients/slopes: symmetry and consistent variability around the regression line.
- State hypotheses in terms of population slopes and correlations.
- Interpret a confidence interval for the population slope.

STEP 1. Ask a research question.

Arbuscular mycorrhiza fungi (AMF) are microorganisms in the soil that form a symbiotic relationship with some native plants and can help them acquire nutrients needed for growth. Alternatively, invasive plants, like garlic mustard, can change or displace native plants through things as simple as competing for water, nutrients, or light. This can result in a reduction of growth from the native plant. The dependency on AMF is known to vary for different types of native plants, and it is believed that garlic mustard disrupts the AMF, and therefore also disrupts the growth of plants that are highly dependent on AMF. Researchers Stinson et al. (2006) conjectured that there is a positive association between a plant's dependency on AMF and its reduction in growth due to the presence of garlic mustard.

STEP 2. Design a study and collect data.

Sixteen species of native plants were used to study the relationship between a plant's AMF dependency and its reduction in growth due to the presence of the invasive garlic mustard. First, to determine a plant's dependency on AMF (i.e., **AMF_Score**), the researchers added the fungi to the soil in which the plant was potted, and then performed a chemical analysis of the roots 6 weeks later to measure how much AMF was present. This was translated to an AMF score, with higher values indicating more dependency on AMF.



Once AMF dependency was understood, researchers then tried to determine the effect garlic mustard has on the respective plant species. For each species of plant studied, the researchers planted two individuals, one with garlic mustard and one without garlic mustard. After 3 months, the difference in growth (measured in cm) was calculated by comparing the individuals. A larger reduction in growth (i.e., difference in growth) is represented by higher values of the variable **GM_Reduction**.

Again, the goal is to see whether larger values of plants' dependency on AMF are associated with larger reduction in plant growth due to garlic mustard. Additionally, we want to see if reduction in plant growth due to garlic mustard can be predicted by plants' dependency on AMF.

1. Identify the observational (or experimental) units and sample size in this study.

2. Describe the explanatory variable and the response variable and classify each as quantitative or categorical.

3. Is this an observational study or a randomized experiment? Justify your answer.

4. Write the null and alternative hypothesis for this study in words using the term association.

STEP 3. Explore the data.

5. Paste the data (**GarlicMustard**) into the **Corr/Regression** applet. Make sure to delete the last empty row if there is one. Examine a scatterplot and correlation coefficient where reduction in plant growth due to garlic mustard is the response variable (vertical axis) and AMF dependency is the explanatory variable (horizontal axis). Describe the direction, form, and strength of association between the variables as revealed in the scatterplot. Are there any unusual observations?

6. Check the **Show Regression Line** box in the applet to determine the least squares regression line for predicting reduction in plant growth due to garlic mustard based on AMF dependency. What is the value of the slope of the regression line? What does this number imply in terms of reduction in plant growth due to garlic mustard and AMF dependency?

STEP 4. Draw inferences.

You should have found the expected positive association between reduction in plant growth due to garlic mustard and ATM dependency in the sample. The question, however, is if there were no association between reduction in plant growth due to garlic mustard and AMF dependency in the population, how likely is it that we would get a slope as large or larger as we did in a sample of 16 plants? The 3S process can be used to answer this question, using the slope as the sample statistic.

7. Let's model the null hypothesis by assuming there is no association between the two observations in each data pair. That is, any response value is just as likely to be paired with any explanatory variable value. Check the **Show Shuffle Options** box, select the **Slope** from the **Choose statistic** pull-down menu to specify the slope as the statistic, and press **Shuffle Y-values** to shuffle the response variable values (reduction in plant growth due to garlic mustard), reassigning them at random to the explanatory variable values (AMF dependency).

    a. How does the regression line fit to the re-randomized data (blue) observed regression line (red)? What is the value of the slope for the shuffled data?

    b. Use the applet to repeat the random re-shuffling five times and write down the five simulated slopes you get. Are any of them as large as or larger than the value of the actual sample slope?

    c. Now do at least 1,000 shuffles and describe the behavior of the 1,000 regression lines across the different shuffles. (How would you describe the graph of the lines to someone who can't see it?)

    d. Use the applet to count how many of the shuffled slopes are at least as extreme as the observed slope. Report your estimated p-value.

    e. Can you conclude that there is strong evidence of a genuine positive association between AMF dependency and reduction in plant growth due to garlic mustard?

8. Use the pull-down menu to change the **Statistic** to the **correlation** coefficient and find the p-value corresponding to the observed correlation coefficient. How does this p-value compare with the p-value corresponding to the slope?

> **Key Idea**
> For a given data set, the significance test for slope is analogous to the test for correlation coefficient.

THEORY-BASED APPROACH

When using the theory-based approach for regression you can write your hypotheses in terms of population parameters with the assumption that the underlying relationship between the variables is linear. The relevant population parameters are the population slope (indicated by the Greek letter β) and the population correlation coefficient (indicated by the Greek letter ρ).

9. Rewrite the above null and alternative hypotheses in terms of the population slope.

10. Change the **Statistic** in the applet back to **slope**. What is the shape of the null distribution of shuffled slopes in the applet?

    By now, you've seen many times that when the null distribution of statistics takes a familiar, mound-shaped curve, we can often use theory-based methods to predict the null distribution of related standardized statistics—as long as certain validity conditions are met. This is no different for regression (and correlation). In regression, the standardized statistic (*t*-statistic) is computed using one of the following equations:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{b}{SE(b)}.$$

    Notice that the *t*-statistic can be computed based on either the correlation coefficient or the slope, yielding the same value, and in both cases is modeled by a *t*-distribution with *n* - 2 degrees of freedom.

11. Use the correlation coefficient to find the *t*-statistic using the first part of the equation shown above. Include a one-sentence interpretation of this standardized statistic.

12. In the applet, use the **Statistic** pull-down menu to select the ***t*-statistic**, and use the *t*-statistic you calculated to approximate the p-value for this test. How does it compare to the simulation-based p-value you found in Question #7d?

    There are three validity conditions for regression which are needed in order to use the theory-based approach to yield a p-value.

Validity conditions
for a theory-based test for regression

1. The general pattern of the points in the scatterplot should follow a linear trend; the pattern should not show curved or other nonlinear patterns.
2. There should be approximately the same distribution of points above the regression line as below the regression line (symmetry about the regression line).
3. The variability of the points around the regression line should be similar regardless of the value of the explanatory variable; the variability (spread) of the points around the regression line should not differ as you slide along the *x*-axis (equal variance/standard deviation).

13. Based on the scatterplot from the applet:

    a. Is Validity Condition 1 met? Namely, is the general pattern of the scatterplot linear?

    b. Is Validity Condition 2 met? Namely, is the distribution of points above the regression line the same as below the line?

    c. Is Validity Condition 3 met? Namely, is the variability of the points around the line similar regardless of the value of the explanatory variable?

14. To find a theory-based p-value, check the **Overlay *t*-distribution** box.

    a. Does the *t*-distribution appear to do a good job of predicting the distribution of the simulated *t*-statistics?

    b. What is the theory-based p-value?

    c. How does the theory-based p-value compare to the simulation-based p-value?

    Most statistical packages will directly report the (two-sided) theory-based p-value.

15. In the applet, check the box for **Regression Table** (lower in left panel). This table provides the observed *t*-statistic for the slope (coefficient of AMF dependency) as well as the corresponding two-sided, theory-based test p-value. Explain why this p-value is double the theory-based p-value you just found.

    STEP 5. Formulate conclusions.

16. Based on any of the p-values you found (they should all be similar) write out a complete conclusion to this test.

17. Check the box in the applet to reveal the **95% Confidence interval for slope** (it's below the regression table). Record and interpret this interval. (*Hint*: Make sure that your interpretation refers to how to interpret the slope coefficient in the population.)

18. Remember that the sample used here was not randomly selected, but seedings selected from 16 different species.

    a. Describe a population in which you would be comfortable drawing inferences. Explain your reasoning.

    b. Can we conclude that an increase in AMF dependency causes the reduction in plant growth from garlic mustard?

    STEP 6. Look back and ahead.

    In the study you considered, it was found that plants that are efficient at absorbing nutrients provided by arbuscular mycorrhiza fungi (AMF) are also the same plants that are highly affected by growth reduction due to garlic mustard presence. It may lead one to ask the question, "If we provide those plants more AMF, does this help their immunity to garlic mustard?"

    Besides the results from the data that you explored, the researchers conducted other experiments that showed that garlic mustard specifically caused the AMF decline in soils thus disrupting the association many plants need with AMF for healthy growth. They also showed that woody plants (they specifically tested some species of ash, maple, and cherry) are both more likely to have AMF dependency and an increase in reduction of growth due to garlic mustard compared to herbaceous plants.

It is still unclear exactly what phytochemicals produced by garlic mustard cause the disruption in AMF. It is also not known how plants in Europe (where garlic mustard is native) buffers the effect of garlic mustard's antifungal properties. Further research is still needed to explore these things.

Reference

Stinson KA, Campbell SA, Powell JR, Wolfe BE, Callaway RM, Thelen GC, et al. (2006) Invasive Plant Suppresses the Growth of Native Tree Seedlings by Disrupting Belowground Mutualisms. PLoS Biol 4(5): e140. https://doi.org/10.1371/journal.pbio.0040140

Additional references

Information about garlic mustard from The Nature Conservancy: https://www.nature.org/en-us/about-us/where-we-work/united-states/indiana/stories-in-indiana/garlic-mustard/